arXiv:2301.12613v1 [cs.CV] 30 Jan 2023

AudioEar：用于个性化空间音频的单视图耳朵重建

Xiaoyang Huang[1], Yanjun Wang[1],Yang Liu, Bingbing Ni[1*], Wenjun Zhang[1],Jinxian Liu[1],Teng Li[2*]

[1]上海交通大学，中国上海 200240， [2]安徽大学{huangxiaoyang, nibingbing}@sjtu.edu.cn

摘要

空间音频侧重于沉浸式三维声音渲染，在声学行业中得到广泛应用。当前空间音频渲染方法面临的关键问题之一，是缺乏基于个体不同解剖结构的个性化处理，而这对于准确定位声源至关重要。本研究从跨学科视角着手解决这一问题。空间音频渲染与人体的三维形状密切相关，尤其是耳朵。为此，我们提出通过单视图图像重建三维人耳，实现个性化空间音频。首先，为耳部重建任务建立基准，我们推出高质量三维耳朵数据集 AudioEar3D，其中包含 112 组带有 RGB 图像的点云耳朵扫描数据。为通过自监督方式训练重建模型，我们还收集了一个由 2,000 张图像组成的二维耳部数据集，每张图像均经过人工标注，标出了遮挡区域和 55 个关键点，并将其命名为 AudioEar2D。据我们所知，这两个数据集在各自类别的公开数据集中规模最大、质量最佳。此外，我们提出 AudioEarM，这是一种由在合成数据上训练的深度估计网络引导的重建方法，并针对耳部数据设计了两个损失函数。最后，为弥合视觉与声学领域之间的差距，我们开发了一套流程，将重建得到的耳朵网格与现成的三维人体模型整合，并模拟个性化的头相关传递函数（HRTF），这是空间音频渲染的核心。代码和数据已公开提供[1]。

引言

空间音频广泛应用于虚拟现实、游戏和电影制作中（Begault和Trejo），用于辨别声源位置并营造沉浸式三维声音。没有空间音频，人们就会失去声音的空间感。空间音频的渲染取决于头相关传递函数（HRTF）（Elliott、Jung和Cheer；Blauert、Allen和Press）。中高端音频设备（如立体声音响、耳机、高保真音响等）都采用HRTF。由于HRTF取决于人体的三维结构，包括耳朵、头部和躯干，因此因人而异。个性化HRTF可以通过声学仿真来近似（Pierce），但前提是拥有完整的三维人体结构；使用三维扫描仪对人体建模则十分费力。单视图重建是另一种解决方案。利用单视图图像重建个性化的人体头部和躯干已得到充分研究。然而，若不对耳部结构进行建模——耳部是人类听觉系统的核心器官——仿真的HRTF仍会存在偏差。在这项工作中，我们提出通过重建单视图图像中的人耳三维网格（Huang等人），来获取个性化HRTF。[2]

为耳部重建任务建立基准，我们使用先进的结构光三维扫描仪采集了高质量的三维耳朵数据集，命名为AudioEar3D。与采用二维耳部标志点重投影误差作为指标的既有评估方案相比，三维基准更为精确。AudioEar3D包含来自56名个体的112个耳部点云扫描及RGB图像，每个扫描包含100，000至250，000个点，可捕捉耳朵的高分辨率形状特征。据我们所知，正如第 章节所分析，这是目前公开可用的规模最大、精度最高的三维耳朵数据集。我们认为，除空间音频外，该数据集也可助力其他应用，包括数字人（Kappel等人）、形态学（Krishan、Kanchan和Thakur）以及医学手术研究（Varman等人）。

此前针对人体其他部位的重建方法，使用大规模二维数据集进行自监督训练（Zhang等人；Chen等人）。然而，现有的二维耳部数据集要么缺少语义标注，要么规模较小。这种情况极大地限制了人耳计算机视觉研究的拓展。为此，我们构建了一个大规模二维耳部图像数据集，其中包含 2,000 张高分辨率耳部图像。这些图像选自 FFHQ人脸数据集，每张图像都附有遮挡情况的人工标注和 55 个耳部标志点。标志点标注极大地丰富了耳部图像的语义信息，使其能够用于广泛的应用。

基于自监督重建流程，我们提出了 AudioEarM，这是一种专为耳朵设计的深度引导重建方法。由于缺乏公开的耳部纹-

*通讯作者：TengLi、Bingbing Ni 版权所有 © 2023年 人工智能促进协会 (www.aaai.org)。保留所有权利。[1] https://github.com/seanywang0408/AudioEar

[2]严格来说，耳部结构包括外耳、中耳和内耳。我们的工作仅关注外耳中可见的结构（耳廓）。不过，遵循先前研究（Jin等人）的做法，我们仍在论文中使用更易理解的名称 耳朵 。

理，而这阻碍了自监督训练，我们首先基于 DECA（ Feng 等人）为 UHM Ear（Ploumpis 等人）构建了耳部纹理空间；UHM Ear 是我们工作中使用的耳部形状模型。我们通过 K 近邻搜索和基于距离的加权，挖掘 UHM Ear 顶点与 DECA 纹理之间的 UV 坐标对应关系。此外，受耳部数据特征的启发，我们设计了两个损失函数：（i）轮廓损失，相较于关键点损失，它能更好地适应标注误差；（ii）相似性损失，鼓励网络为不同样本预测出具有区别的形状。此外，我们引入合成数据以提升重建质量。我们没有将合成数据直接与真实数据结合，并使用三维监督训练模型，而是利用合成数据训练单目深度模型，从耳部图像中提取深度特征。随后，我们将深度信息注入主网络，以引导整个重建过程。

最后，为了弥合耳部重建与声学应用之间的差距，我们开发了一套流程，将重建出的耳朵集成到三维网格人体躯干中，用于声学仿真。我们提出在三维顶点上进行近似德劳内三角剖分，以确保集成后人体网格的仿真有效性。我们通过仿真获得包含重建耳朵的个性化头相关传递函数，并进一步通过比较头相关传递函数，展示重建模型相较于基线模型的有效性。

相关工作

空间音频

空间音频，又称空间听觉，是声学领域长期研究的课题（Culling 和 Akeroyd；Vorl¨ander）。它关注人类如何定位声源的位置，以及如何通过音响系统重现空间感。人类通过双耳 和 单耳线索来定位声源。双耳线索包括双耳时间差（ITD）和双耳声级差（ILD）：声音信号在到达耳朵时会受到人体（包括耳廓、头部和躯干）的衍射和反射影响，从而产生到达时间和强度上的差异。单耳线索是声音的频谱失真。此外，人体的三维形状因个体而异，因此不同的人会具有不同的双耳和单耳线索。为了对不同个体的空间听觉特征进行建模，研究者提出了头相关传递函数（HRTF）（Li 和 Peissig；Xie、Zhong 和 He），并将其广泛用于空间声音渲染。HRTF 描述特定个体在各个方向上接收到的声源声压级（SPL，dB），用于将任意声音转换为仿佛源自特定位置的声音。然而，获取准确的个性化头相关传递函数十分耗时费力，因为测量需要昂贵的设备和专门的声学实验室。目前的实现通常采用平均头相关传递函数，或从头相关传递函数数据库中选取一个（ Guo 等人）。除物理测量外，还可以基于边界元法（ BEM）对头相关传递函数进行数值模拟（Conrad；Jensen）。Meshram 等人提出利用多视图立体视觉（ MVS）重建三维人体，以模拟个性化头相关传递函数

表1：AudioEar3D及其同类方法。质量根据采集设备的精度进行评估。

三维 耳朵 数据集	尺度	图像质量		可访问性
UND-J2	1，800	✓	⋆	✓
York3DEar	500	✗	⋆	✓
SYMARE-1	20	✗	⋆⋆⋆	✓
SYMARE-2	102	✗	⋆⋆⋆	✗
Ploumpis 等人	234	✗	⋆⋆⋆	✗
AudioEar3D	112	✓	⋆⋆⋆⋆	✓

表2：AudioEar2D 与其同类数据集的比较。地标 表示地标标注。占用 表示关于耳朵是否被头发或耳环部分遮挡的注释。

二维耳朵数据集	规模	来源	注释 地标。	占用。	用法
UND-E	464	有限	✗	✗	生物特征识别
AMI	700	有限	✗	✗	生物特征识别
IIT Delhi Ear	754	有限	✗	✗	生物特征识别
WPUTEDB	3,348	有限	✗	✗	生物特征识别
UBEAR	4,410	有限	✗	✗	生物特征识别
IBug-B	2,058	野外环境	✗	✗	生物特征识别
AWE	9，500	野外环境	✗	✗	生物特征识别
EarVN	28，412	野外环境	✗	✗	生物特征识别
IBug-A	605	野外环境	✓	✗	重建
AudioEar2D	2，000	野外环境	✓	✓	重建

然而，由于重建方法存在缺陷，重建出的人体严重模糊，导致头相关传递函数误差较大。

三维和二维耳部数据集

以往关于人耳的计算机视觉研究大多局限于生物特征识别应用（O’Sullivan 和 Zafeiriou），其发展并不像面部、手部和骨架等人体其他部位那样普遍。这导致缺乏丰富的三维和二维耳朵数据集，如表1、2所示。

现有三维耳部数据集存在不可访问、规模较小或质量较低等问题。Yan 和 Bowyer 使用深度传感器，以 640×480 的分辨率采集了 1，800 张耳部深度图。由于这些深度图是单视角的，因此无法呈现完整的耳朵形状。Dai、Pears 和 Smith 发布了 York3DEar，该数据集由 500 个变形的三维耳朵网格组成。这些网格并非通过仪器采集，而是基于二维耳部数据集，使用数据增强技术估计得到的，因此会引入不可预见的误差。Jin 等人提出了 SYMARE耳部数据库，其中包含通过磁共振成像（MRI）采集的 61 个个体的实测头部相关脉冲响应（HRIR）以及上躯干、头部和耳部网格模型。然而，目前只有 10 个个体（SYMARE-1）的数据可访问，其余数据（SYMARE-2）均为私有数据。Ploumpis 等人通过 CT扫描从 64 名成人中采集了 121 个耳朵模型，并从儿童中采集了 133 个耳朵模型。然而，这些数据并未公开。此外，上述大多数三维耳部数据集都缺少对应的耳部图像（ UND-J2 除外），使得在这些数据集上执行单视角重建任务变得不可行。而且，这些数据中的大多数是通过

MRI采集的，其测量误差通常约为 1 毫米（Nowogrodzki）。对于这种结构精细的器官而言，这一测量误差不可忽略，因此会在HRTF的仿真中引入额外误差。

现有的二维耳部数据集包括 UND-E（Chang等人）、AMI（Gonzalez、Alvarez和Mazorra）、IIT Delhi Ear（Kumar和Wu）、WPUTEDB（Frejlichowski和Tyszkiewicz）、UBEAR（Raposo等人）、IBug(-A/B)（Zhou 和 Zaferiou）、AWE（Emeršić等人）和 EarVN（Hoang），其中大多数用于人类身份识别。这些数据集的汇总见表2。

三维可变形模型重建

三维可变形模型（3DMM）是一种参数化模型，可将三维网格的形状和纹理编码到潜在空间中。三维可变形模型重建研究最多的对象是人脸（Ploumpis 等人；Ploumpis 等人）、手部（Chen等人；Wang等人；Zhang等人）和人体（Choi、Moon和Lee；Corona等人）。为了获得合理的性能，三维可变形模型重建算法需要大量二维图像进行自监督训练，通常还需配合语义标注，例如关键点或姿态。它们使用特征提取器和多层感知机（MLP）来回归形状和纹理的潜在编码。随后，将这些潜在编码输入可微三维可变形模型，以生成带颜色的三维网格。接着，使用可微渲染器将网格投影到图像上（Kato、Ushiku和Harada；Huang等人）。通过最小化图像上的光度损失和特征点重投影损失来训练网络。常用的面部重建算法 DECA（Feng 等人）提出通过回归 UV位移图来实现稳健的细节重建。Sun、Pears和Dai提出使用 YEM模型（Dai、Pears和Smith）重建三维耳朵网格。他们从图像中推导出颜色模型，并通过最小化光度误差和关键点误差来回归形状潜在编码。

方法

数据收集

我们为个性化空间音频收集了两个耳部数据集。其中一个是 AudioEar3D，即三维耳部扫描数据集，用于对耳朵重建任务进行基准测试。另一个是 AudioEar2D，用于训练耳朵重建模型。

AudioEar3D 在各种三维扫描设备中，我们选择结构光三维扫描仪来采集三维耳部数据，原因如下：(i) 与以往的三维耳部扫描方法（如磁共振成像和计算机断层扫描）相比，结构光扫描仪更加便捷、易于获取（可集成到小型便携设备中），因此能够轻松覆盖更多人群；(ii) 操作和后处理更加简单；(iii) 在价格相同的情况下，结构光扫描仪的分辨率高于其他设备。具体而言，我们使用 MantisVision® F6-SR [3]，这是一款便携式高分辨率三维扫描仪，专为扫描精细的医学器官模型而设计应用。其平面分辨率和深度分辨率分别为 0.1 毫米和 0.4 毫米（相比之下，大多数磁共振成像扫描仪的分辨率为 1 毫米），帧率为 8 FPS。

为缓解复杂耳部结构带来的自遮挡问题，并获取清晰完整的三维耳部扫描，我们使用上述设备从不同方向对每位个体扫描约 90 秒。我们移除身体的其他部位，仅保留耳朵部分。这样，每次耳部扫描可获得约 100，000 至 250，000 个点，能够捕捉耳朵的高分辨率形状特征。此外，我们还在与扫描仪距离相近的位置拍摄耳朵正面照片。目前，我们已收集 56 名受试者的 112 只耳朵的数据。我们计划进一步扩大数据集规模，以覆盖更多人群。图 1 展示了若干样本。出于隐私考虑，最终数据集已完全匿名化。

AudioEar2D 我们的二维耳部数据集取自一个高质量的公开人脸数据集 Flickr-Faces-HQ 数据集（FFHQ）（Karras、Laine 和 Aila），该数据集包含 70，000 张分辨率为 1，024×1，024 的面部图像，在姿态、年龄和种族方面各不相同。为了找到包含清晰耳朵的图像，我们首先使用预训练的 Yolov4（Bochkovskiy、Wang 和 Liao）在 IBug Collection-B 数据集（Zhou 和 Zafe-riou）上训练耳部检测模型；该数据集包含耳朵的边界框。随后，我们将训练好的检测模型应用于 FFHQ 数据集，根据检测置信度和边界框面积粗略筛选出高质量耳部图像。此外，我们利用预训练的头部姿态估计模型 WHENet（Zhou 和 Gregson），进一步剔除视角不佳的图像。我们根据边界框将剩余图像裁剪成正方形。为避免插值缩放造成的信息损失，我们保留裁剪后耳部图像的原始分辨率。得益于边界框面积筛选，图像分辨率大多高于 300×300，如图 1 下方 (a) 所示。接着，我们对剩余图像进行人工筛选，并按照 IBug Ear（Zhou 和 Zafe-riou）中的地标点标注规范，标注 55 个关键点。此外，我们还标注耳朵是干净的，还是被头发或耳环遮挡；在某些应用中，这些因素可能会成为噪声。遮挡情况的分布如图 1 下方 (b) 所示。最终，我们获得了 2，000 张高分辨率耳部图像，每张图像均标注了 55 个关键点。

AudioEarM：深度引导的耳部重建

我们基于常见的参数化形状重建流程开发了本方法。该流程通过自监督训练，回归三维形状、纹理、相机和光照的潜在编码。我们首先针对耳朵的特定特征，对这一流程进行适配，并做出几项有效改进。此外，为了进一步提升性能，我们提出利用合成耳朵图像训练深度估计模型，并通过多尺度特征对齐将深度信息融合到重建模型中。我们将本方法命名为 AudioEarM。

纹理获取 我们使用 UHM Ear（Ploumpis 等人），这是质量最高的耳部三维形变模型，可将形状潜在编码转换为三维网格。UHM Ear 将耳朵网格的顶点数 N

[3]产品介绍：https://mantis-vision.com/handheld-3d-scanners/

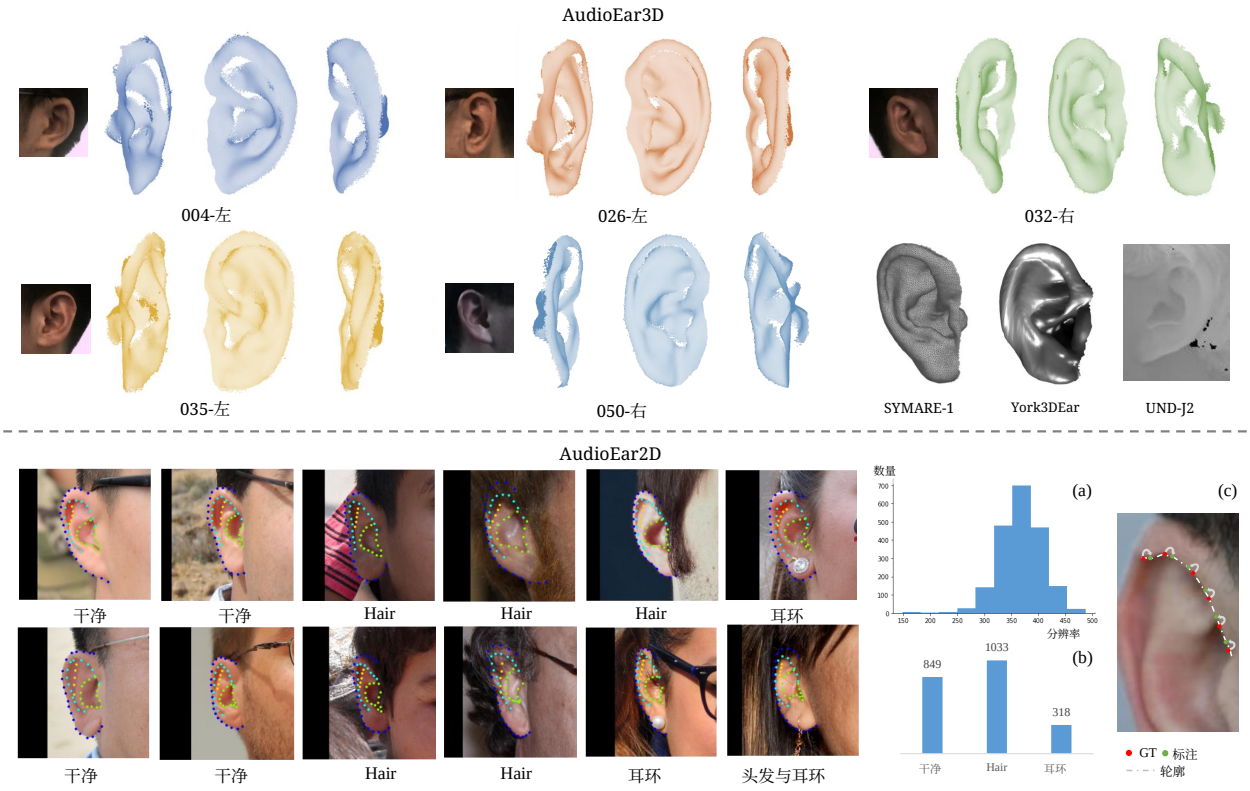


图 1：两个采集数据集的可视化结果。**AudioEar3D** （上）：展示了五个样本及其三视图渲染点云和 RGB 图像。与现有的三维耳朵数据集相比，AudioEar3D 质量更高、规模更大，并且也公开可用。**AudioEar2D** （下）：展示了若干带有遮挡标注和 55 个关键点的样本。关键点的四种颜色表示耳朵的四条主要轮廓。为便于观察，我们将左耳翻转为右耳。

（2800）通过主成分分析降至较低维度（236）。令 $\mathbf{S} \in \mathbb{R}^{3N}$ 表示平均形状，$\vec{v} \in \mathbb{R}^{236}$ 和 $\mathcal{U} \in \mathbb{R}^{3N\times 236}$ 分别表示三维形变模型的特征值和特征向量。形状变化通过特征值加权线性混合形状建模：$B_{\mathcal{U}}(\vec{\beta};\mathcal{U}) = \sum_{n=1}^{236} v_n \beta_n \mathcal{U}_n$，其中 $\vec{\beta} \in \mathbb{R}^{236}$ 是预测的形状潜在编码。最终耳朵网格通过以下方式获得：

$$M(\vec{\beta}) = \sum_{n=1}^{236} v_n \beta_n \mathcal{U}_n + \overline{\mathbf{S}} \tag{1}$$

UHM Ear 不提供纹理，而纹理对于在自监督重建流程中最小化光度损失至关重要。我们从 DECA（Feng 等人）中提取耳部纹理空间，以实现彩色渲染。在 DECA 中，给定纹理潜在编码 $\vec{\theta} \in \mathbb{R}^{|\vec{\theta}|}$，可以计算出纹理贴图$T(\boldsymbol{\theta}) \in \mathbb{R}^{h\times w\times 3}$（$h \times w$ 为分辨率）。DECA 网格中的每个顶点都对应纹理贴图中的一个 UV坐标 p_{deca}。然而，UHM 耳朵的顶点与 DECA 中的顶点并不对应。为此，我们首先将 UHM 耳朵的均值形状与 DECA 中的耳朵尽可能精确地进行视觉对齐。随后，对于 UHM 模型中的每个顶点$v_{uhm} \in \mathbb{R}^3$，我们在 DECA 的顶点中搜索其 K 近邻 $v^i_{deca} \in \mathbb{R}^3$，$i = 1, 2 \ldots, k$（在我们的实现中 k 为 3）。我们根据 KNN 距离分配加权 UV 坐标——给 UHM 顶点，权重依据 v_{uhm} 与 v^i_{deca}之间的欧几里得距离计算。具体而言，v_{uhm} 的 UV坐标计算如下：

$$p_{uhm} = \sum_{i=1}^{k} \frac{D(v_{uhm}, v^i_{deca})}{\sum_{j=1}^{k} D(v_{uhm}, v^j_{deca})} p^i_{deca} \tag{2}$$

轮廓损失。轮廓损失针对由二维特征点构成的四条轮廓。（图 1 中的 AudioEar2D 用四种不同颜色的关键点标出了这四条轮廓。）如图 1 底部 (c) 所示，沿着轮廓的地标标注存在偏差，但这些标注能够很好地表示轮廓。原因在于，轮廓对人眼而言更加醒目。因此，在训练中衡量轮廓预测的误差，优于衡量关键点误差。实际操作中，我们依次连接关键点，形成四条折线，并在这些折线上均匀地重新采样出密集点。我们将真实值与预测值之间重新采样点的倒角距离 $\mathcal{P}$作为轮廓损失：

$$L_{contour} = \frac{1}{4} \sum_{i=1}^{4} ChamferDistance(\mathcal{P}^i_{gt}, \mathcal{P}^i_{pred}) \tag{3}$$

其中，上标 i 是四个轮廓的索引。

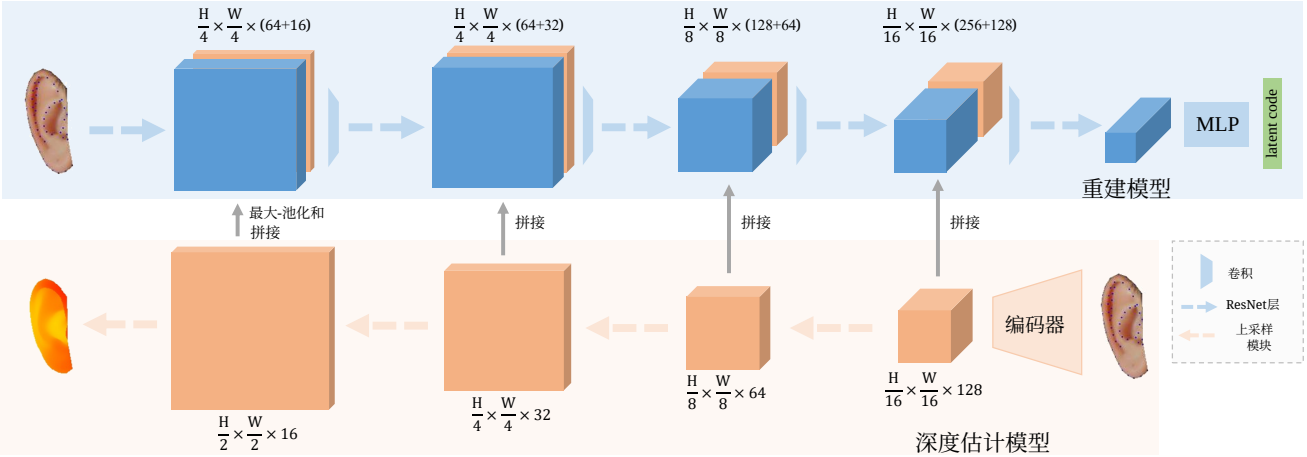


图 2：我们的深度引导重建模型 AudioEarM 的架构。深度估计模型在合成数据集上训练，并通过多尺度特征对齐引导主重建网络。

相似性损失。 在早期实验中，我们发现，预测的形状潜在编码在所有实例之间过于相似。这种相似性并不理想，因为不同个体的形状应当各不相同。我们通过测量一个批次中形状潜在编码之间的平均余弦相似度，并对较高的相似度进行惩罚，鼓励网络预测出不同的形状：

$$L_{sim} = \frac{1}{bs \times (bs-1)} \sum_{i,j=0,i\neq j}^{bs-1} \frac{\beta_i \beta_j}{|\beta_i||\beta_j|} \tag{4}$$

其中， bs 表示批次大小， β_i， β_j 分别表示一个批次中的第 i 个和第 j 个样本。

此外，我们通过 $L1$惩罚来防止相机中的旋转角度和尺度参数过大，因为图像中耳朵的角度和尺度通常处于一定范围内。我们将这三个损失函数与常用的损失函数相结合：标志点损失、光度损失、网格平滑损失和潜在编码正则化。总体损失是上述损失的加权和。所有损失函数的完整公式请参见附录。

深度引导重建 Zhang等人证明，将带有3D监督的合成数据引入训练集可以提高单视图手部重建的质量。然而，我们通过实验证实，直接将这一方法迁移到我们的任务中会产生负面影响（见消融研究）。作为替代方案，我们利用预训练的优势（Yang等人），即使用合成数据预训练单目深度估计（MDE）模型。随后，我们利用MDE模型提取的深度特征来促进整个重建过程，如图2所示。

为了生成具有 3D监督的合成耳朵数据集，我们从高斯分布中随机采样形状和纹理潜在编码，并渲染对应的 RGB 图像和深度图。我们以这种方式生成了 10,000 个样本。随后，我们在该合成数据集上训练 MDE模型，并使用深度图作为直接监督。MDE模型（Laina 等人）由编码器和解码器组成（图 2）。

ResNet-34（He 等人）的编码器从图像中提取高级特征，而解码器则根据这些特征预测深度图。解码器逐层对提取的特征进行上采样，直至恢复至原始分辨率。我们旨在利用解码器获得的深度信息来指导重建过程。为此，我们将多尺度上采样特征与主网络中每层之前的特征进行拼接。每次拼接后，我们都会额外插入一个卷积层，将增加后的通道数转换为原有通道数，以匹配下一层的通道数。由于 MDE 模型最后一层的分辨率是主网络第一层分辨率的两倍，因此我们在拼接前对 MDE模型最后一层的特征进行最大池化。通过这种方式，我们将深度信息显式融合到主网络中，以指导重建过程。消融实验表明，该方法带来了显著提升。

声学仿真

空间音频渲染取决于头相关传递函数，其定义为在耳膜处（实际测量时为耳道入口处）测得的声压级（SPL）。我们使用 COMSOL™ 仿真个性化头相关传递函数。COMSOL 是一款采用边界元法进行仿真的多物理场软件。然而，仿真需要完整的人体上半身，而不仅仅是耳朵。为此，我们开发了一套流程，将重建的耳朵整合到现成的人体三维模型中（Braren和Fels；Braren和Fels）。该流程如图3所示。

我们手动移除给定三维人体的原始耳朵，并将重建后的耳朵放置到相应的孔洞中。为了将耳朵与人体网格组合在一起，我们提出了一种名为近似 *Delaunay* 三角剖分的方法。Delaunay 三角剖分是一种针对二维点集的三角剖分方法，在三角剖分过程中使所有三角形角度中的最小角最大化，从而避免产生狭长三角形。狭长三角形不适用于边界元法仿真。由于耳朵边缘和人体孔洞边缘都不在同一二维平面上（但接近平面），我们将其投影

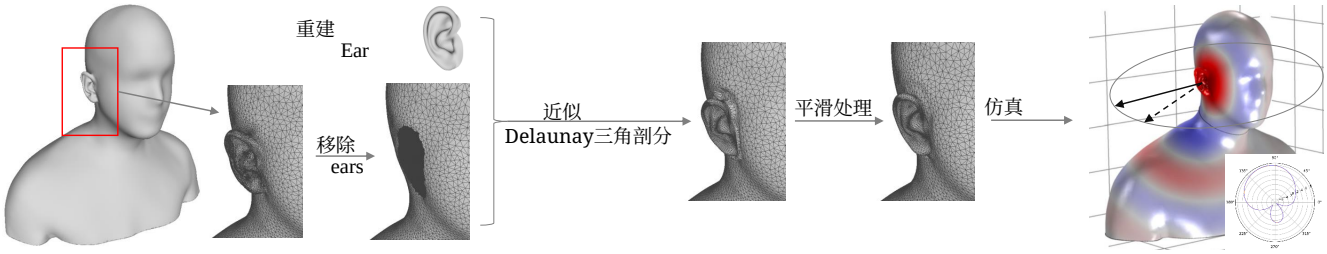


图3：所提出的集成与仿真流程。我们采用近似 *Delaunay*三角剖分 将重建的耳朵与身体结合起来，并在水平面上的所有方位角（361 个角度）下仿真个性化头相关传递函数。

到一个平面上，该平面由边缘周围网格面片的平均法线确定。这样，我们便可以对两条边缘上的顶点进行近似 *Delaunay* 三角剖分，并修补耳朵与头部之间的裂缝。由于投影不会显著改变边缘的几何形状，因此仍可尽可能避免狭长三角形。我们使用 Detri2 程序 [4] 进行 Delaunay 三角剖分。最后，我们对生成的网格进行平滑处理，并将其用于仿真。我们按照通用流程，在水平面上对全部方位角（共 361 个角度）以及三个频率（f = 1033.6 Hz、2067.5 Hz 和 3962.1 Hz）进行头相关传递函数仿真。请注意，仿真流程需要人工操作，不属于耳部重建的评估流程，而是用于实际应用。我们在 AudioEar3D 上评估重建性能。

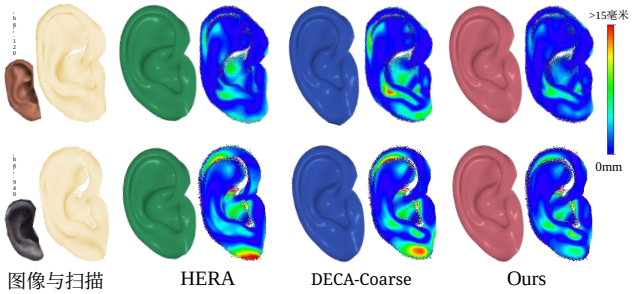


图4：我们展示了我们的方法与两种基线方法在测试集（AudioEar3D）上的**S2M误差图**。与基线方法相比，我们的方法误差更低。

实验

耳朵重建基准

评估 我们利用完整的 AudioEar3D 数据集评估耳朵重建模型的性能。我们计算真实耳朵扫描中的每个点到预测耳朵网格的距离。（点到网格的距离定义为该点到网格中最近三角形面的距离。）我们对扫描中所有点的距离取平均，作为评估指标，称为扫描到网格（S2M）距离。然而，由于扫描与预测网格的尺度和位置未对齐，我们首先采用基于梯度的方法，将网格与扫描进行配准。我们通过迭代优化尺度和位置参数来完成配准。为缩短评估时间并避免收敛到局部最小值，我们设计了三阶段配准方案。首先，最小化四个手动选定的关键点之间的距离，以实现粗配准。接着，在预测网格表面随机采样点，并最小化采样点与扫描之间的倒角距离。最后，直接最小化扫描与网格之间的 S2M 距离。由于 AudioEar3D 中密集的点云会导致大量计算，我们将扫描随机降采样至 1,000 个点用于计算 S2M，并仅在完整扫描上计算 S2M距离，

经过大量测试，我们通过经验发现，这种设置可以使大多数配准收敛到全局最小值。我们对 AudioEar3D 中的全部 112 个样本进行配准，这大约需要半小时，并将所有样本的 S2M 平均值作为最终评估指标。通过在矢状面上反射，将左耳扫描转换为右耳扫描。

设置为了进行比较，我们将 UHM 的平均耳朵输入配准流程，作为不进行个性化处理的基线。我们还将我们的方法与一种名为 HERA（Sun、Pears和Dai）的耳朵重建先前方法进行比较。由于 HERA 未公开其代码和耳朵模型，我们使用自己的耳朵模型实现了该方法。此外，我们还希望将 AudioEarM 与最先进的人脸重建算法 DECA（Feng 等人）进行比较。由于 DECA 没有提供适用于耳朵的详细 UV 位移图，我们仅实现了 DECA 中的粗略分支，并将其称为 DECA-coarse。训练配置详见附录。

结果 如表3 #6-9所示，我们的方法优于平均耳基线、HERA 和 DECA-coarse，最终 S2M距离为 1.28 mm。平均耳的 S2M距离为 1.78 mm，而 HERA 和 DECA-coarse 的 S2M距离分别为 1.70 和 1.46 mm。我们展示了测试集（AudioEar3D）上我们的方法和两种基线方法的**S2M误差图**。与基线方法相比，我们的方法误差更低。

最后一次迭代。我们使用 Adam（Kingma 和 Ba），并将初始学习率设为 0.45 进行优化。每个阶段持续 166 次迭代。进入下一阶段时，学习率乘以 0.1。

[4]http://www.wias-berlin.de/people/si/detri2.html

表3：重建方法的比较。 {#1 +tex- ture}表示在方法#1的基础上添加了纹理。其他 +、–标记的含义类似。

#	模型	S2M ↓
1	朴素方法	1.73毫米
2	#1 +纹理	1.65mm
3	#2 +相似性损失	1.63mm
4	#2 +轮廓损失	1.52mm
5	#2 +轮廓与相似性	1.47毫米
6	**我们的方法(#5 +深度引导)**	**1.28毫米**
7	平均耳朵	1.78毫米
8	HERA	1.70毫米
9	DECA-coarse	1.46 毫米

重建模型的比较

我们考察了方法中各个部分的影响（表3 #1-6）。所有模型均按照章节 中所述的相同设置进行训练。我们首先实现了一个基线模型（方法#1），将常见的自监督重建流程直接用于我们的数据，未作任何修改。我们为纹理指定单一肤色，而非使用我们的纹理空间。方法#1的 S2M距离为 1.73 mm，相较于平均耳仅有小幅提升。随后，我们用纹理空间替换单一颜色（方法#2），将 S2M误差降至 1.65 mm。在此基础上，我们将相似性损失和轮廓损失加入训练过程（方法#3-5）。同时加入这两种损失后，结果为 1.47 mm。轮廓损失的作用比相似性损失更为显著。最后，我们使用深度估计模型进行深度引导重建，得到 1.28 mm 的 S2M距离，这是我们的最佳结果。其优势的原因在于，预测规则的深度图比预测形状潜在编码更不适定。此外，深度信息能够提供物体几何形状的信息。类似的动机也启发了用于三维物体检测的伪激光雷达方法（ Qian 等人）。

HRTF 仿真

评估。 在本部分中，我们使用预测耳朵进行头相关传递函数仿真。我们将预测耳朵的头相关传递函数与真实耳朵的头相关传递函数进行比较。真实耳朵的头相关传递函数是使用 AudioEar3D 中从原始耳朵扫描数据配准得到的耳朵网格进行仿真的。具体而言，我们将平均 UHM Ear 输入到第 章节 所述的同一配准流程中，不同之处在于，第二和第三阶段的优化参数除了原有参数外，还包括 UHM 的形状潜在编码。配准后，所有扫描数据的平均 S2M 为 0.11 mm，这一数值相当低。较低的 S2M 误差表明，优化后的网格与真实耳朵形状十分接近。我们测量所有角度下真实耳朵与预测耳朵之间 SPL 误差绝对值的平均值。作为比较，我们将预测耳朵替换为平均 UHM Ear（在表 4 中记为平均值 ）。平均 UHM Ear 的 SPL 误差代表未进行个性化的空间音频效果。误差越大，空间音频效果越差。由于仿真

表 4：以声压级误差衡量的 HRTF 仿真结果。每个样本的第一行和第二行分别表示平均值和预测值结果。可以看出，重建结果越好，HRTF 就越逼真。前 5 个样本中，预测耳朵的 S2M 和声压级误差均低于平均值；而最后 2 个样本则相反。

样本	S2M ↓ (mm)	声压级误差 ↓(dB× 10) f=1kHz	f=2kHz	f=4kHz
033-右耳	1.49	1.12±0.75	3.59±5.97	4.13±5.44
	0.94	0.33±0.19	0.59±0.89	2.32±2.84
021-左	1.66	0.58±0.27	1.20±2.19	4.25±3.43
	0.88	0.30±0.12	1.30±1.47	2.39±1.85
036-left	2.23	1.23±1.17	3.52±9.16	6.97±11.97
	1.85	1.13±0.89	2.69±4.58	4.75±5.89
008-左	1.23	1.69±1.31	4.50±6.71	5.29±5.40
	0.74	1.35±0.69	3.532±5.97	6.75±6.49
048-左	1.35	0.40±0.30	1.27±1.64	5.00±4.76
	0.94	0.23±0.14	1.18±1.23	3.31±2.88
040-右侧	1.27	1.00±0.66	2.83±5.65	5.54±10.23
	1.58	1.17±0.81	3.57±6.54	5.99±10.43
029-左侧	1.33	2.11±0.98	5.95±9.76	8.06±9.47
	1.46	1.12±0.76	3.44±5.98	10.09±10.74

实验耗费人力且耗时较长，我们仅展示若干样本的结果作为示例。

结果。 表 4 比较了每个样本的平均值 与预测值 结果。每个样本中的第一行和第二行分别表示平均值 和预测值 结果。我们可以看到，更好的重建结果（更低的 S2M）能够带来更真实的头相关传递函数（更低的声压级误差）。对于那些重建后的 S2M距离低于平均 UHM Ear 的样本（前 5 个样本），声压级误差的平均值和变化幅度也都更低，这意味着与平均 UHM Ear 相比，它们获得了更接近真实值的头相关传递函数。对于最后两个样本，预测耳朵的 S2M 更高，声压级误差也相对更大，这表明所获得的头相关传递函数与重建结果高度相关。为了在即将到来的虚拟现实趋势中实现更真实的空间音效，有必要开发先进的耳部重建算法。为了更好地理解声压级误差，我们在附录中使用极坐标对其进行了说明。

结论

本研究针对从单视图图像进行三维耳部重建，以实现个性化空间音频。为此，我们收集了一个用于基准评测的三维耳朵数据集，以及一个用于训练耳朵重建模型的二维数据集。此外，我们提出了一种由深度估计网络引导的重建方法，该网络在合成数据上进行训练，并设计了两个针对耳部数据的损失函数。最后，我们开发了一套流程，将重建得到的耳朵网格与人体进行整合，以开展声学仿真并获得个性化空间音频。

致谢

本研究得到中国国家自然科学基金（U20B2072、61976137）的资助。同时，本研究还得到上海交通大学医学工程交叉研究项目 Grant YG2021ZD18 的部分资助。

参考文献

Begault, D. R.; and Trejo, L. J. 2000. 用于虚拟现实和多媒体的三维声音。编号：美国国家航空航天局/*TM-2000-209606*。

Blauert, J.; Allen, J.; and Press, M. 1997. 空间听觉：人类声音定位的心理物理学。麻省理工学院出版社。ISBN 9780262024136。

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. YOLOv4：目标检测的最佳速度和精度。*arXiv* 预印本 *arXiv:2004.10934*。

Braren, H.; and Fels, J. 2020a. 用于头相关传递函数仿真和验证的高分辨率个体三维成人头部和躯干模型：三维数据。医学声学组，技术声学研究所，亚琛工业大学。

Braren，H.；Fels，J. 2020年b。一种用于头相关传递函数模拟和验证的高分辨率个体三维成人头部和躯干模型：头相关传递函数测量。医学声学组，技术声学研究所，亚琛工业大学。

Chang，K.；Bowyer，K. W.；Sarkar，S.；以及 Victor，B. 2003年。基于外观的生物特征识别中耳朵和面部图像的比较与结合。*IEEE*《模式分析与机器智能汇刊》，25(9)：1160–1165。

Chen，W.；Ling，H.；Gao，J.；Smith，E.；Lehtinen，J.；Jacobson，A.；以及 Fidler，S. 2019年。利用基于插值的可微渲染器学习预测三维物体。神经信息处理系统进展，32：9609–9619。

Chen，Y.；Tu，Z.；Kang，D.；Bao，L.；Zhang，Y.；Zhe，X.；Chen，R.；以及 Yuan，J. 2021年。通过自监督学习进行基于模型的三维手部重建。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，10451–10460。

Choi, H.; Moon, G.; and Lee, K. M. 2020. Pose2Mesh：用于从二维人体姿态恢复三维人体姿态和网格的图卷积网络。载于欧洲计算机视觉会议，769–787。Springer。

Conrad, Z. 2018. 向边界元法致敬。*IEEE Spectrum* 多物理场仿真，30。

Corona, E.; Pumarola, A.; Alenya, G.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. SMPLicit：面向穿衣人体的拓扑感知生成模型。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，11875–11885。

Culling, J.; and Akeroyd, M. 2012. 空间听觉。牛津听觉科学手册：听觉。

Dai, H.；Pears, N.；以及 Smith, W. 2018年。耳朵的数据增强三维可变形模型。载于*2018*年第*13*届*IEEE*国际自动人脸与手势识别会议（*FG 2018*），404–408。IEEE。

Desbrun, M.；Meyer, M.；Schr¨oder, P.；以及 Barr, A. H. 1999年。使用扩散和曲率流对不规则网格进行隐式平滑。载于第*26*届计算机图形学与交互技术年度会议论文集，317–324。Elliott, S. J.；Jung, W.；以及 Cheer, J. 2018年。头部追踪将宽带声音的局部主动控制扩展至更高频率。《科学报告》，8(1)：1–7。Emeršicˇ, ˇZ.；ˇStepec, D.；ˇStruc, V.；Peer, P.；George, A.；Ahmad, A.；Omar, E.；Boult, T. E.；Safdaii, R.；Zhou, Y.；等。2017年。无约束耳朵识别挑战赛。载于*2017*年*IEEE*国际联合生物特征识别会议（*IJCB*），715–724。IEEE。Feng, Y.；Feng, H.；Black, M. J.；以及 Bolkart, T. 2021年。从野外图像中学习可动画化的精细三维人脸模型。《*ACM*图形学汇刊》（*TOG*），40(4)：1–13。Frejlichowski, D.；以及 Tyszkiewicz, N. 2010年。西波美拉尼亚理工大学耳朵数据库——用于测试生物特征识别算法的工具。载于国际图像分析与识别会议，227–234。Springer。Gonzalez, E.；Alvarez, L.；以及 Mazorra, L. ???? AMI耳朵数据库。 https://ctim.ulpgc.es/research works/ami ear database/。Guo, X.；Bai, Y.；Liu, Q.；Xiong, D.；Wang, Y.；以及 Du, J. 2020年。中国飞行员头相关传递函数的三维头部人体测量。载于人机环境系统工程国际会议，171–178。Springer。He, K.；Zhang, X.；Ren, S.；以及 Sun, J. 2016年。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，770–778。Hoang, V. T. 2019年。EarVN1. 0：一个大规模野外耳朵图像数据集。《数据简报》，27。Huang, X.；Yang, J.；Wang, Y.；Chen, Z.；Li, L.；Li, T.；Ni, B.；以及 Zhang, W. 2022a年。表示无关形状场。载于国际学习表征会议。Huang, X.；Zhang, Y.；Ni, B.；Li, T.；Chen, K.；以及 Zhang, W. 2022b年。通过辐射度映射增强点云渲染。*arXiv*预印本 *arXiv:2210.15107*。Jensen, M. H. 2009年。使用声学仿真提升助听器性能。*COMSOL*会议。Jin, C. T.；Guillon, P.；Epain, N.；Zolfaghari, R.；Van Schaik, A.；Tew, A. I.；Hetherington, C.；以及 Thorpe, J. 2013年。创建悉尼约克耳朵形态与声学录音数据库。《*IEEE*多媒体汇刊》，16(1)：37–46。

Kappel, M.; Golyanik, V.; Elgharib, M.; Henningson, J.-O.; Seidel, H.-P.; Castillo, S.; Theobalt, C.; and Magnor, M. 2021. 单目视频的高保真人体神经运动迁移。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，1541–1550。

Karras, T.; Laine, S.; and Aila, T. 2019年. 一种用于生成对抗网络的基于风格的生成器架构. 收录于

*IEEE/CVF*计算机视觉与模式识别会议论文集, 4401–4410. Kato, H.; Ushiku, Y.; and Harada, T. 2018. 神经三维网格渲染器. 收录于 *IEEE*计算机视觉与模式识别会议论文集, 3907–3916. Kingma, D. P.; and Ba, J. 2014. Adam：一种随机优化方法. 收录于*ICLR*. Krishan, K.; Kanchan, T.; and Thakur, S. 2019年. 人耳形态变异及其在个人身份识别中的应用研究. 《埃及法医学杂志》, 9(1): 1–11. Kumar, A.; and Wu, C. 2012. 利用耳部成像自动识别人体身份. 模式识别, 45(3): 956–968. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; and Navab, N. 2016. 利用全卷积残差网络进行更深层次的深度预测. 收录于*2016*年第四届国际三维视觉会议（*3DV*）, 239–248. IEEE. Li, S.; and Peissig, J. 2020. 头相关传递函数的测量：综述. 《应用科学》, 10(14): 5014. Meshram, A.; Mehra, R.; Yang, H.; Dunn, E.; Franm, J.-M.; and Manocha, D. 2014. 个性化头相关传递函数：用于高保真空间声音的高效个性化头相关传递函数计算. 收录于*2014*年*IEEE*混合与增强现实国际研讨会（*ISMAR*）, 53–61. IEEE. Nealen, A.; Igarashi, T.; Sorkine, O.; and Alexa, M. 2006. 拉普拉斯网格优化. 收录于澳大拉西亚与东南亚计算机图形学及交互技术第四届国际会议论文集, 381–389. Nowogrodzki, A. 2018. 世界上最强大的磁共振成像设备正将人体成像推向新的极限. 《自然》新闻特稿, 563(7729): 24–26. O'Sullivan, E.; and Zafeiriou, S. 2020. 用于人耳点云的三维关键点定位. 收录于*2020*年第*15*届*IEEE*自动人脸与手势识别国际会议（*FG 2020*）, 402–406. IEEE. Pierce, A. D. 2019. 声学：物理原理及其应用导论. Springer. Ploumpis, S.; Ververas, E.; O'Sullivan, E.; Moschoglou, S.; Wang, H.; Pears, N.; Smith, W.; Gecer, B.; and Zafeiriou, S. P. 2020. 朝着完整的人头三维可变形模型迈进. *IEEE*《模式分析与机器智能汇刊》. Ploumpis, S.; Wang, H.; Pears, N.; Smith, W. A.; and Zafeiriou, S. 2019. 结合三维可变形模型：大规模人脸与头部模型. 收录于*IEEE*计算机视觉与模式识别会议论文集, 10934–10943. Qian, R.; Garg, D.; Wang, Y.; You, Y.; Belongie, S.; Hariharan, B.; Campbell, M.; Weinberger, K. Q.; and Chao, W.-L. 2020. 端到端伪激光雷达：基于图像的三维物体检测. 收录于*IEEE/CVF*计算机视觉与模式识别会议论文集, 5881–5890. Raposo, R.; Hoyle, E.; Peixinho, A.; and Proenc¸a, H. 2011. UBEAR：在不受控条件下移动拍摄的耳部图像数据集。载于*2011*年*IEEE*生物特征识别与身份管理计算智能研讨会（*CIBIM*）, 84–90。IEEE。Sun, H.; Pears, N.; and Dai, H. 2021. 人耳重建自编码器。载于*2021*年第*16*届计算机视觉、成像与计算机图形学理论及应用国际联合会议（*VISIGRAPP 2021*）, 136–145。Varman, R. M.; Schrader, K.; Fort, C.; Daniel, H.; Al Mekdash, M. H.; and Demke, J. 2021. 基于医学培训年限和专业判断外耳畸形与正常外耳解剖结构的能力。《美国美容外科杂志》, 0748806821993681。Vorl¨ander, M. 2020. 可听化。Springer。Wang, J.; Mueller, F.; Bernard, F.; Sorli, S.; Sotnychenko, O.; Qian, N.; Otaduy, M. A.; Casas, D.; and Theobalt, C. 2020. Rgb2hands：基于单目RGB视频实时跟踪三维手部交互。《*ACM*图形学汇刊》（*TOG*）, 39(6)：1–16。Xie, B.; Zhong, X.; and He, N. 2015. 中国受试者头相关传递函数的典型数据与聚类分析。《应用声学》, 94：1–13。Yan, P.; and Bowyer, K. W. 2007. 利用三维耳形进行生物特征识别。*IEEE*《模式分析与机器智能汇刊》, 29(8)：1297–1308。Yang, J.; Huang, X.; He, Y.; Xu, J.; Yang, C.; Xu, G.; and Ni, B. 2021. 为三维图像重新构思二维卷积。《*IEEE*生物医学与健康信息学杂志》, 25(8)：3009–3018。Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.-L.; and Grundmann, M. 2020a. Mediapipe hands：设备端实时手部跟踪。第四届*AR/VR*计算机视觉研讨会。Zhang, Y.; Chen, W.; Ling, H.; Gao, J.; Zhang, Y.; Torralba, A.; and Fidler, S. 2020b. 图像生成对抗网络与可微渲染相结合，实现逆向图形学与可解释三维神经渲染。载于学习表征国际会议。Zhou, Y.; and Gregson, J. 2020. WHENet：适用于广泛头部姿态的实时精细估计。Zhou, Y.; and Zaferiou, S. 2017. 野外环境中的耳部可变形模型：用于对齐和识别。载于*2017*年第*12*届*IEEE*国际自动人脸与手势识别会议（*FG 2017*）, 626–633。IEEE。

附录 A：损失函数详解

轮廓损失 我们根据55个关键点在耳朵上的分布，将其分为4组： $K_i \in \mathbb{R}^2$， i =1，2 .....，55。图 2（ AudioEar2D）中用四种不同颜色的关键点标示这四组。随后，将每组中的关键点依次连接，得到4条折线。然后，在每条折线上均匀采样 N 个点，分别记为$\mathcal{P}^1$， $\mathcal{P}^2$， $\mathcal{P}^3$， $\mathcal{P}^4 \in \mathbb{R}^{N\times 2}$。我们分别对标注关键点和预测网格投影得到的关键点执行这些步骤，得到四对采样点（$\mathcal{P}^1_{proj}$， $\mathcal{P}^1_{gt}$），（$\mathcal{P}^2_{proj}$， $\mathcal{P}^2_{gt}$），（$\mathcal{P}^3_{proj}$， $\mathcal{P}^3_{gt}$），（$\mathcal{P}^4_{proj}$， $\mathcal{P}^4_{gt}$）。轮廓损失是每对中$\mathcal{P}_{proj}$ 和 $\mathcal{P}_{gt}$ 之间倒角距离的平均值：

$$L_{contour} = \frac{1}{4}\sum_{i=1}^{4} ChamferDistance(\mathcal{P}^i_{proj}, \mathcal{P}^i_{gt}) \quad (5)$$

其中，上标 i 表示四个轮廓的索引。

相机损失 我们首先定义范围损失：当变量超出指定范围 $[x_{min}, x_{max}]$ 时，通过 $L2$ 正则化对其进行惩罚：

$$L_{range}(x, x_{min}, x_{max}) = \begin{cases} (x_{min} - x)^2 & \text{if } x < x_{min} \\ (x - x_{max})^2 & \text{if } x > x_{max} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

基于范围损失，我们对尺度 s 和绕 y 轴的旋转角 r_y 进行约束。相机损失公式如下：

$$L_{cam} = L_{range}(s, 0.5, 4) + L_{range}(r_y, -1.5, -0.5)$$

关键点损失 我们采用 Zhou 和 Zaferiou 提出的归一化关键点损失。将预测的三维耳朵网格 $K^i_{pred} \in \mathbb{R}^3, i = 1, 2 \ldots, 55$ 上的关键点通过正交投影函数 $\mathbf{C}$ 投影到图像平面上。在图像平面上，计算投影后的关键点与真实二维关键点 $K^i_{gt} \in \mathbb{R}^2, i = 1, 2 \ldots, 55$ 之间的关键点损失。关键点损失公式如下：

$$L_{lmk} = \frac{\sum_{i=1}^{55} \left\| K^i_{gt} - \mathbf{C}(K^i_{pred}) \right\|_2}{55 \times D_{gt}} \quad (7)$$

其中， D_{gt} 是真实值关键点边界框的对角线长度。

光度损失 光度损失用于计算输入图像 I 与渲染图像 I_r之间的像素误差。为忽略输入图像中的背景，我们使用二维特征点构建耳朵掩膜 M_I。具体而言，将外部特征点连接起来形成一个闭合多边形，并将多边形内部填充为 1，外部填充为 0。损失计算如下：

$$L_p = \|M_I \cdot I - I_r\|_2 \quad (8)$$

网格平滑损失 我们在预测耳朵网格上应用拉普拉斯平滑（Desbrun 等人；Nealen 等人），以避免出现尖锐变形。为简化起见，我们采用均匀拉普拉斯。对于每个顶点 v_i，将其邻接顶点记为 $\mathcal{N}_{v_i}$，邻接顶点的数量记为 $N = |\mathcal{N}_{v_i}|$。v_i 的拉普拉斯定义为：

$$\mathbf{Laplacian}(v_i) = \sum_{j=1}^{N} w_{ij} \left(\mathcal{N}^j_{v_i} - v_i\right) \quad (9)$$

对于均匀拉普拉斯， $w_{ij} = 1/|\mathcal{N}_{v_i}| = 1/N$。因此，公式 9 可简化为：

$$\mathbf{Laplacian}(v_i) = \frac{1}{N}\sum_{j=1}^{N} \mathcal{N}^j_{v_i} - v_i \quad (10)$$

项 $\frac{1}{N}\sum_{j=1}^{N} \mathcal{N}^j_{v_i}$ 表示邻接顶点的质心。最终的网格平滑损失是所有顶点的拉普拉斯算子的平均值。

正则化 我们对形状潜在编码 $L1$和纹理潜在编码 $\beta \in \mathbb{R}^{236}$ $\theta \in \mathbb{R}^{50}$ 施加正则化，以避免它们过大。正则化形式如下：

$$L_{reg} = \frac{1}{236}\sum_{i=1}^{236} |\beta_i| + \frac{1}{50}\sum_{i=1}^{50} |\theta_i| \quad (11)$$

总损失是上述各项损失的加权和：

$$L_{total} = \lambda_{contour} L_{contour} + \lambda_{sim} L_{sim} + \lambda_{cam} L_{cam} + \lambda_{lmk} L_{lmk} + \lambda_{photo} L_{photo} + \lambda_{smooth} L_{smooth} + \lambda_{reg} L_{reg}. \quad (12)$$

其中，在我们的实验中， $\lambda_{contour} = 100$、 $\lambda_{sim} = 1$、 $\lambda_{cam} = 100$、 $\lambda_{lmk} = 10$、$\lambda_{photo} = 100$、 $\lambda_{smooth} = 10$ 和 $\lambda_{reg} = 0.005$ 。正则化的权重较小，因为较大的正则化会导致模型欠拟合。关键点损失的权重设为 10 而非 100，因为它不如轮廓损失精确。网格平滑也可视为对网格的正则化，因此其权重也低于其他损失。

附录 B：训练配置

对于深度估计模型，我们使用 8，000 个合成数据进行训练，使用 2，000 个进行验证。我们使用 Adam 优化器训练模型 100 个训练轮次，初始学习率为 0。001，并在第 50 个训练轮次将其降低至原来的 0。2。对于重建模型，我们将 AudioEar2D 随机按 9:1 划分为训练集和验证集。按照先前的研究，我们将所有左耳翻转为右耳，使模型只需处理单侧耳朵。我们分两个阶段训练 AudioEarM，共 100 个训练轮次。在前 20 个训练轮次中，我们固定形状和纹理潜在编码，使模型输出平均形状和纹理，并且只最小化关键点损失和正则化损失，以训练相机和光照。我们通过实验发现，这有助于获得合理的形状重建。在剩余的 80 个训练轮次中，我们通过最小化所有损失的加权和来优化所有参数。我们使用 Adam 优化器，初始学习率为 0。001，并在第 50、75 和 90 个训练轮次将学习率乘以 0。2。

附录 C：训练集的影响

表 5：训练集的影响。#2–3 分析 AudioEar2D 的遮挡情况。#4–5 比较 AudioEar2D 和 IBug-A。#6 使用合成数据进行训练，并采用显式的潜在编码监督来训练重建模型，而不是以深度引导的方式使用潜在编码监督。

#	训练集	尺度	S2M
1	AudioEar2D	2,000	**1.28 毫米**
2	#1 −耳环	1,682	1.36mm
3	#1 −头发与耳环	849	1.44mm
4	IBug-A	605	1.48毫米
5	#1 +IBug-A	2,605	1.45毫米
6	#1 +合成	12,000	1.63毫米

我们通过在不同训练集上训练 AudioEarM，考察训练集的影响。首先，我们从训练集中移除 AudioEar2D 中部分被遮挡的耳朵，以考察遮挡的影响（表 5 #2-3）。我们观察到，移除的遮挡图像越多，性能越差。与较小的训练集相比，部分遮挡对性能的影响似乎没那么大。接着，我们将 AudioEar2D 与 IBug-A（Zhou 和 Zaferiou）进行比较，后者是一个已有的带有关键点的耳朵数据集。IBug-A 包含 605 张图像，而 AudioEar2D 包含 2，000 张（表 5 #4-5）。仅使用 IBug-A 训练 AudioEarM，结果为 1。48 毫米（使用 AudioEar2D 时为 1。28 毫米）。这表明，训练集规模较小确实会导致性能下降，证明了 AudioEar2D 的价值。我们还将 AudioEar2D 与 IBug-A 合并进行跨域训练，发现尽管训练集规模更大，性能却不如仅使用 AudioEar2D 训练。我们认为这是由两个数据集之间的领域差异造成的。具体而言，IBug-A 与我们的 AudioEar2D 之间的**领域差异** 可能体现在两个方面。一方面，IBug-A 包含灰度图像，而 AudioEar2D 完全由 RGB 图像构成。另一方面，两者的地标点标注标准可能不同。由于部分关键点缺乏明确的定义，我们的标注标准可能与他们的不同，从而造成领域差异。我们认为，性能下降主要是由第一个因素导致的，因为我们的轮廓损失可以缓解第二个因素的影响。最后，我们没有使用合成数据训练深度估计模型，而是将合成数据与 AudioEar2D 合并作为训练集，并将合成数据已知的潜在编码作为显式监督，用于训练重建模型。这种做法与 MediaPipe Hands（Zhang等人）类似。结果为 1。63 毫米，表现相当差。从这一点来看，我们方法中的深度估计模型可能发挥了连接合成数据与真实数据之间差异的作用

数据。

附录D：声压级误差

为更好地理解声压级误差，我们在图5中用极坐标进行说明。在水平面上的所有方位角（共361个角度）处模拟头相关传递函数。随后，在极坐标系中绘制声压级误差，其中方位角作为角坐标，各角度处的误差作为径向坐标。最终的声压级误差取所有角度的平均值。

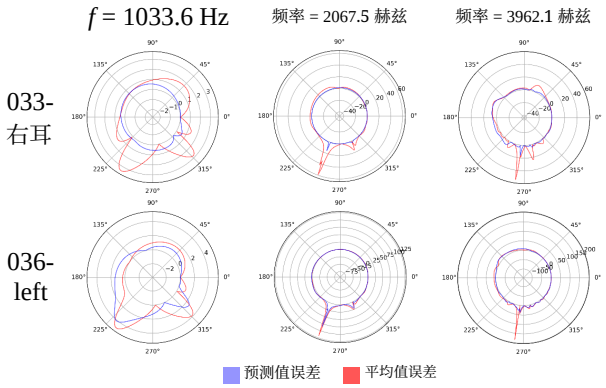


图5：三种仿真频率下极坐标中声压级误差的示意图。极角表示仿真的方位角，半径表示误差的大小。平均SPL误差是在所有方位角之间计算得出的。

附录E：局限性

尽管本文基于单视图图像的重建与平均耳相比已取得相对较好的结果，但为了获得更好的空间音频效果，仍有进一步改进的空间。在未来的工作中，我们将考虑多视图重建设置。更多视角能够展示耳朵更多的三维结构，并带来更好的重建效果。

头相关传递函数不仅取决于耳部结构，也取决于头部和身体结构。此外，重建的耳朵相对于头部的位置和方向并未自动估计。因此，模拟得到的头相关传递函数并非该个体真实的头相关传递函数。未来，我们将同时重建人体并估计耳朵的相对位置，以获得更逼真的头相关传递函数。