Software Engineering and Applications 软件工程与应用, 2024, 13(3), 302-311
Published Online June 2024 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2024.133030

Hans汉斯

基于自注意力机制的动态全息声场生成方法

杨　柳，游福成

北京印刷学院信息学院，北京

收稿日期：2024年3月5日；录用日期：2024年6月18日；发布日期：2024年6月27日

摘　要

声场控制对于扬声器设计、超声成像和声学粒子操纵等多种应用至关重要。对微米和纳米尺度物体进行精确操纵的需求导致了非接触式操纵方法的发展。然而，关于给定全息声场的反向操纵的研究很少。在本文中，我们提出了一种在相控阵技术(PAT)背景下基于自注意力机制Transformer模型(VS3D-Transformer)的方法，以实现快速准确地全息声场生成。我们的方法解决了传统CNN仅考虑局部感受野且训练精度低的缺点。此外，我们降低了传统物理方法的迭代复杂性。为了模拟声场的产生，我们采用基于活塞模型的模拟方法来产生全息声场。在仿真研究中，与传统的IB迭代算法和深度学习Acousnet算法相比，我们的模型表现出更快的训练速度和更高的精度。我们提出的模型在各种条件下(即声场相位优化准确率、损失率和训练速度)的结果表明我们的模型可以作为一种高效的替代方案。

关键词

全息声场，超声相控阵，深度学习，自注意力机制

Dynamic Holographic Acoustic Field Generation Method Based on Self-Attention Mechanism

Liu Yang, Fucheng You

College of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: Mar. 5th, 2024; accepted: Jun. 18th, 2024; published: Jun. 27th, 2024

Abstract

Acoustic field control is critical in applications as diverse as loudspeaker design, ultrasonic imaging, and acoustic particle manipulation. The need for precise manipulation of objects at the micron and nanoscale has led to the development of contactless manipulation methods. However, there

文章引用: 杨柳, 游福成. 基于自注意力机制的动态全息声场生成方法[J]. 软件工程与应用, 2024, 13(3): 302-311.
DOI: 10.12677/sea.2024.133030

are few studies on the reverse manipulation of a given holographic acoustic field. In this paper, we propose a method based on the attention mechanism transformer model (VS3D-Transformer) within the context of phased array technology (PAT) to achieve fast and accurate holographic acoustic field generation. Our method solves the shortcomings of traditional CNNs which only consider the local receptive field and possesses low training accuracy. Moreover, we reduce the iterative complexity of traditional physical methods. To simulate acoustic field generation, we use the simulation method based on the piston model to generate the holographic acoustic field. In the simulation study, our model demonstrates faster training speed and higher accuracy compared to both the traditional IB iterative algorithm and the deep learning Acousnet algorithm. The results of our proposed model under various conditions (*i.e.*, overall field generation, loss rate, and training speed) indicate that our model could serve as a highly effective alternative.

Keywords

Acoustic Holography, Phased Array Transducer, Deep Learning, Self-Attention Mechanism

1. 引言

随着我们进入科学发展的新时代，微米级和纳米级物质的操纵和控制(即全息声学)已成为对生物医学、数字喷墨打印技术等许多领域至关重要的研究领域[1]、纳米加工、材料科学与工程、可持续能源与环境保护。对用于分离、混合、旋转和聚焦微米和纳米物体的非触觉精密操纵技术的需求不断增长，增加了这项研究的相关性[2]-[6]。现有的触觉操纵技术虽然很熟练，但在同时处理多个目标时常常会对微小的细胞结构造成损害。

为了引入非接触式操纵，已经开发了光学、磁力和声学操纵技术。如光镊、磁镊和声镊，这些镊子利用光波、磁场和声波的能量与微米和纳米实体相互作用达到非接触式操纵的目的。对于声学等领域来说，利用高频超声波形成全息声场的相控阵技术(PAT)是一次重大飞跃[7]-[9]，其巨大的潜力使其成为各种应用的理想工具，包括高强度聚焦超声(HIFU)和工业无损检测形式的医学诊断[10]-[14]。然而，准确生成复杂的全息声场仍然是一个艰巨的挑战。现代解决方案，例如 Gerchberg-Saxton (GS)算法[15]、单相检索(SPR)、双相检索(DPR)和多相检索(MPR)算法，以及改进的基于 GS 算法的算法已被证明在不同程度上是有效的[16]-[19]。随着人工智能的热门趋势，科学界将注意力转向机器学习和深度学习[8]，这是一门快速发展的学科，用于解决众多领域的复杂问题。钟承熙等人在 2021 年的开创性工作中[20]。利用卷积神经网络(CNN)和特定的声全息模型来实现全息声场重建方法。然而，这种方法受到 CNN 局部感受野视野的限制，因此无法充分发挥深度学习在优化全息声场方面的全部潜力。解决这些缺点是声学领域的首要任务，这促使我们探索和应用最先进的人工智能技术来增强全息声场的生成。

在本文中，我们建议使用以其在不同领域的卓越性能而闻名的 Transformer 模型，并结合自注意力机制来克服 CNN 的只考虑局部感受野的缺点。这种新颖的方法旨在为现有的复杂性提供更全面、更有效的解决方案，从而有助于弥合当前全息声场生成方面的知识差距。本研究系统地进行了介绍，首先概述了相控阵传感器(PAT)如何影响声场的理论基础，然后介绍了我们独特的深度学习框架及其专注于注意力机制的训练策略。我们还深入研究了实验方法的细节，提供了数据集准备、预处理的全面视图，以及我们研究结果的广泛概述。

2. 相关工作

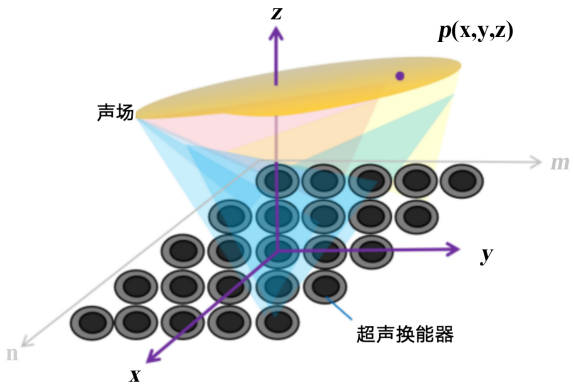

Figure 1. Schematic of the holographic sound field generated by a phased array
图 1. 相控阵产生的全息声场示意图

如图 1 所示，在超声相控阵(Phased Array Transducer, PAT)上建立 2D 坐标(m,n)和笛卡尔坐标((x,y,z)，其中换能器平面与 *mon* 平面和 *xoy* 重叠。换能器在换能器平面上按 $a\times a$ 阵列排列。由(m,n)索引的每个换能器的特征在于相位 $\varphi_{m,n}$，位于$[0,2\pi)$的范围内。PAT 发射的声波，产生的全息声场，被建模为 $H\in R^{m\times m\times m}$，产生的全息声场中的复数声压，与振幅 $A_{x,y,z}$ 和相位 $\varphi_{x,y,z}$ 之间的关系为：

$$p(x,y,z)=\sum_{i=1}^{M}A_i(x,y,z)\mathrm{e}^{j\varphi_i(x,y,z)} \tag{1}$$

假设换能器阵列有 M 个元素，每个元素的位置可以用(x_i,y_i,z_i)表示，其中，A_i 表示为第 i 个换能器的声压幅值，$\varphi_i(x,y,z)$是第 i 个换能器的声场相位函数。exp 为自然指数函数，j 为单位虚数单位。

具体的声场相位函数$\varphi_i(x,y,z)$可以根据超声换能器阵列换能器的布局、电信号激励和换能器的特性来确定。常见的超声换能器阵列包括线性阵列、矩阵阵列、环形阵列等，每一种阵列都有特定的声场相位函数，通过累积的方法将每个换能器看作是一个点声源或一个平方/圆形声源。

所研究的基于对压力振幅和相位的综合考量，所以将涉及的声场被归类为全息声场。声波从换能器阵列向前传播以产生全息声场如下：

$$f:H(x,y,z)\to T(m,n);\quad m,n\le M,\ \text{且}\ |x|,|y|\le S_1,\quad 0<z\le S_2 \tag{2}$$

其中 M 定义为相控阵数值，S_1 和 S_2 定义声场数值。然而，为了在实践中产生一个高度可控的人工声场，我们需要解决映射问题，即

$$f^*:H(x,y,z)\to T(m,n);\quad m,n\le M,\ \text{且}\ |x|,|y|\le S_1,\quad 0<z\le S_2 \tag{3}$$

在简单而紧凑的前向运动学建模中，由于高度的非线性，公式(3)的反向运动学在数学上实际上是无法解决的。这项工作的目的是提出一种基于深度学习的方法，学习从给定的全息声场 F 到阵列相位 T 的逆向映射，用于声场控制。以下是关于数据集准备和神经网络结构的细节。

3. 方法

本研究设计了一个注意力机制专门来学习逆映射如公式(3)所定义，该方法的整体逻辑如下：我们将全息声场中采样点的信息输入神经网络，并对 PAT 中的换能器相位进行预测。随后，我们计算真实相位均值与预测相位均值之间的差异，并利用该差异计算损失函数。接下来，我们使用梯度下降算法来优化

预测相位均值，从而获得满足误差要求的换能器相位。本节将详细描述我们提出的神经网络架构和损失函数的设计。

3.1. 模型架构

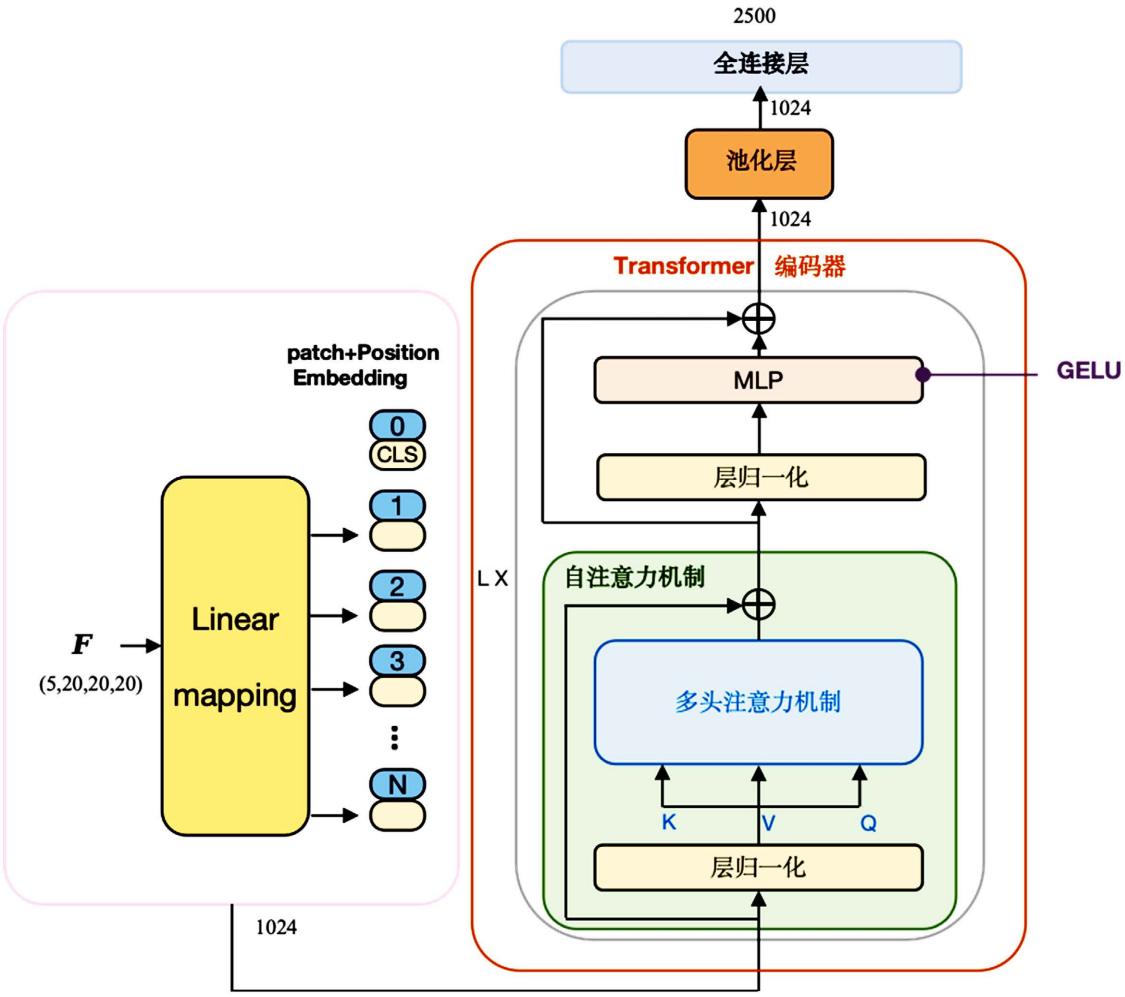

Figure 2. The input size of the VS3D-Transformer architecture is 5 × 20 × 20 × 20, and the output size is 128 × 2500
图 2. VS3D-Transformer 架构的输入尺寸为 5 × 20 × 20 × 20，输出尺寸为 128 × 2500

VS3D-Transformer 是一种用于处理全息声场数据并进行多分类任务的深度学习模型。如图 2 所示，它采用嵌入编码层将输入的全息声场数据转换为二维序列，并通过位置编码层为序列中的每个位置添加位置信息，以区分不同位置的数据。随后，模型使用多层 Transformer 编码器对序列进行编码。每个编码器由一个自注意力机制(self-Attention)和一个前馈神经网络(Feedforward)组成。通过堆叠多个编码器，模型能够从不同层次抽取序列的表示。编码器的输出经过池化层进行池化操作，得到固定维度的整个全息声场序列表示。最后，通过线性层(mlp head)将池化后的表示映射到类别数目(num classes)的向量空间，用于多分类任务。

具体而言，该网络的输入是一个 5 × 20 × 20 × 20 的全息声场数据，其中每个切片是一个 5 × 5 × 5 大小的立方体。网络的深度(depth)为 5 层，每层包含 8 个头(head)，每个头的维度为 128。最后的全连接层具有 1024 维，而用于分类的全连接层维度为 2500，对应于 50 * 50 大小的图像。

与传统的卷积神经网络(CNN)相比，VS3D-Transformer 通过自注意力机制实现了声场全局上下文的建模，从而能够考虑更大范围的感受野。自注意力机制允许每个位置的输入与整个序列中的其他位置进行交互和关联，从而在编码序列时捕捉全局上下文信息并获取更大范围的感受野。自注意力机制的优势在于它不受固定卷积核大小或池化窗口的限制，能够处理任意距离范围内的依赖关系，并利用这些信息提取顶层特征表示，以更好地适应全息声场数据的多分类任务。

3.2. 损失函数的设计

当考虑到声波的周期性时，我们使用损失函数对预测的相位和实际值之间的差异进行惩罚。具体而言，我们计算预测的阵列相位和真实值之间的余弦值，并通过最小化估计值和目标值之间的相位差异来实现声场重建的准确性和相位一致性。当输入数据的批大小为 N 时，损失函数可以表示为公式(4)：

$$L=\frac{1}{N^2}\sum_{(m,n)=(1,1)}^{(N,N)}\left(1-\cos\left(2\pi\left(\varphi_{m,n}\right)_{pred}-\left(\varphi_{m,n}\right)_{truth}\right)\right) \tag{4}$$

4. 数据集与实验结果

4.1. 数据集的准备

首先，我们需要准备一个数据集，其中包含了一系列的全息声场 H 以及对应的阵列相位 T。为了收集这些数据，采用模拟方法来准备训练数据集。已经存在几种众所周知的声源模型，例如面源模型和点源模型和活塞源模型，而活塞源模型在实践中更为常用。

如图 3 所示，矩形换能器嵌于无限大刚性板中，刚板表面振幅为零，并且假设声源上各质点振动幅值相同，声波在各向同性、无衰减的介质中传播。式(2)中的正向运动学映射 f 可以表示为以下形式：

$$p(x,y,z)=\sum_{m,n}p_0\frac{D(\theta,\beta,\omega)}{d}\mathrm{e}^{j\left(\varphi_{m,n}+kd\right)} \tag{5}$$

式 4 中，p_0 为声换能器的声压振幅，且相控阵中所有换能器的声压振幅保持不变，$D(\theta,\beta,\omega)$ 是指向性函数，它取决于声学换能器到声场中物理点 (x,y,z) 的极角 θ 和方位角 φ，相对于换能器坐标，D 是换能器到物理点 (x,y,z) 的自由空间传播距离，$\varphi_{m,n}$ 是传感器指数的初始发射相位，由 (m,n) 表示，如第二节所示，k 是从 2π 除以波长得到的波数。

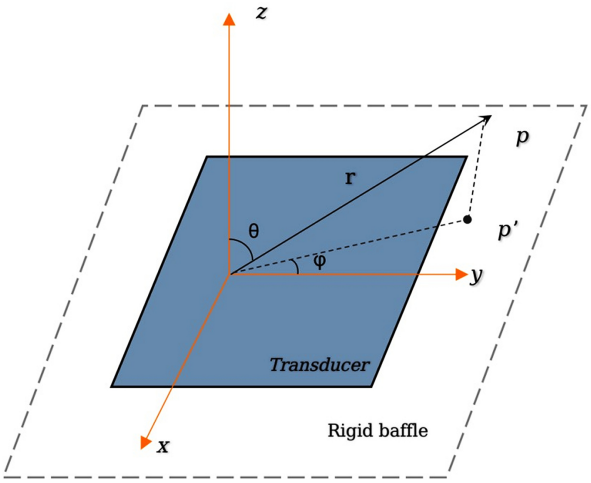

Figure 3. Piston acoustic field computation model
图 3. 活塞声场计算模型

与含有白噪声的图像相似。为此，也研究了如何生成有意义的 T。受 IB 算法[14]的启发，提出了一种生成图案化 T 的方法。整个数据准备过程见算法 1 中的代码。

算法 1 :数据集准备&处理

数据: *换能器的位置，复声压，模式相位，幅值：* $(x_m,\ y_n)$, $p(x_m,\ y_n)$, $\phi_{m,n}$, *和* $A_{m,n}$;
声场的位置，复声压，相位，幅值 $(x_l,\ y_w,\ z_d)$, $p(x_l\ , y_w,\ z_d)$, $\phi_{l,w,d}$, $A_{x,y,z}$;
an identity matrix $T_{l,w,d}$, *directivity function* $D(\theta,\beta,\omega)$, *wave number* k
结果: *归一化处理后的相位* $\phi_{m,n}$,*复声压* $p(x_l, y_w,\ z_d)$, *位置* $(x_l,\ y_w,\ z_d)$, *对应点的幅值和相位* $A_{m,n}$, $\phi_{l,w,d}$;

1. Initialize $P_0 \leftarrow T_{l,w,d}$;$d \leftarrow [(x_l - x_m)^2 + (y_w - y_n)^2 + z_d{}^2]^{0.5}$,$D \leftarrow D(\theta,\alpha,\omega)$;$\phi \leftarrow \phi_{m,n}$
2.iteration $\leftarrow 0$, *ntrue* $\leftarrow 0$
3. ***while*** *iteration* ≤ 200 *and ntrue* ≤ 1500 *do:*
i. 计算 $p(x_l, y_w,\ z_d) \leftarrow \sum P_0(D/d)\ exp(j(\phi_{m,n} + kd)$;
ii.计算 $A_{x,y,z} \leftarrow [p(x_l\ , y_w,\ z_d).\mathrm{real}]^2 + [p(x_l\ , y_w,\ z_d).\mathrm{img}]^2$;
iii.计算 $\phi_{l,w,d} \leftarrow (p(x_l\ , y_w,\ z_d).\mathrm{real}, p(x_l\ , y_w,\ z_d).\mathrm{img})$;
iv.更新 $p(x_m,\ y_n) \leftarrow \sum A_{x,y,z}\ \ exp(j\ (\phi_{l,w,d}\ + kd))$;
v. 更新 $A_{m,n} \leftarrow [p(x_m,\ y_n).\mathrm{real}]^2 + [p(x_m,\ y_n).\mathrm{img}]^2$;
vi. 更新 $\phi_{m,n} \leftarrow (p(x_m\ , y_n).\mathrm{real}, p(x_m\ , y_n).\mathrm{img})$;
vii. ***if*** $abs(\phi_{m,n}\ -\ \phi_{l,w,d}) \le \pi/100$ ***then*** *ntrue* $\leftarrow$ *ntrue + 1 end*
viii.iteration $\leftarrow$ *iteration+1*
4.归一化处理换能器相位: $\phi_{m,n} \leftarrow \phi_{m,n}/2\pi$.
5.归一化处理(l, w, d)的相位: $\phi_{l,w,d} \leftarrow \phi_{l,w,d}/2\pi$
6.极坐标:$p \leftarrow \mathrm{sqrt}[x_l^2 + y_w^2 + z_d^2]$, $\theta \leftarrow acos(x_l^2/\mathrm{sqrt}[x_l^2 + y_w^2])$,$\varphi \leftarrow [atan2(\mathrm{sqrt}[x_l^2 + y_w^2], z_d^2) + \pi]/2\pi$;

7. ***For*** *each cross-section i of the acoustic field:*
i. $\mathrm{dis}_i \leftarrow \max(A_{x,y,z})_i\ -\ \min(A_{x,y,z})_i$,
ii. $\max_i \leftarrow \max(A_{x,y,z})_i\ -\ \mathrm{dis}_i \times e^{-1.84})$,$\min_i \leftarrow min(A_{x,y,z})\ +\ \mathrm{dis}_i \times e^{-1.84}$
iii. ***if*** $A_{x,y,z} \ge max_i$ ***then*** $A_{x,y,z} \leftarrow max_i$,***elif*** $A_{x,y,z} \le min$ ***then*** $A_{x,y,z} \leftarrow min$;
end

4.2. 实验细节

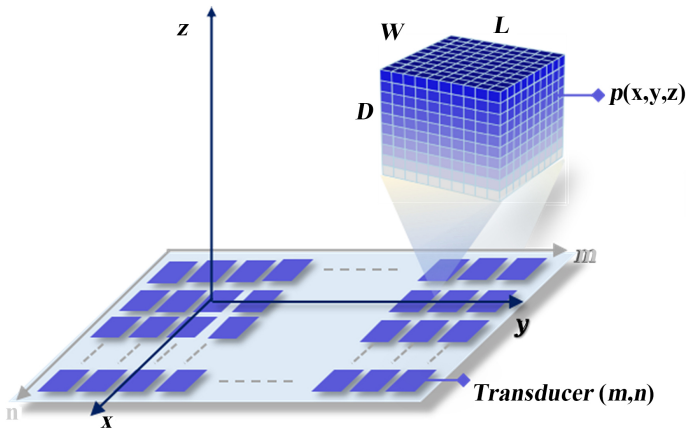

Figure 4. Illustration of data preprocessing
图 4. 数据预处理示意图

本文描述了一种模拟方法，用于生成一个包含 20,000 个数据对的训练数据集。每个数据对以{H，T}的格式存档，并用于 VS3D-Transformer 网络的训练。根据图 4，我们可以清晰地观察到换能器阵列的相位和声场分布，这些特征类似于灰度图像。为了提高网络的泛化能力，我们按照{0.8，0.1，0.1}的比例随机划分数据集为训练集、验证集和测试集，并在训练过程中使用验证集进行监控和调整。每个数据对

包含阵列相位和相应的全息声场分布，它们采用不同的 COI 封装，具有不同的尺寸和空间位置，以获得最佳的通用性。

在这里，我们面临一个重要问题，即如何有效地量化和重新组合 H，使其能作为 DNN 的输入进行处理。我们建议遵循一个原则，即 H 的采样点数量应为传感器数量的三到四倍，以在计算复杂性和数学要求之间取得平衡。上方的图 4 演示了我们提出的方法。

在处理声场 H 的过程中，我们采用了一个利用 3D 体素进行数据结构描述的方法，以实现高准确性的声场构建和控制。首先，我们选取声场 F 中的一个特定区域，将其标记为“兴趣区域”or“COI”(Cube of Interest)，并定义其尺寸为 L、W 和 D。接着，我们进一步将 COI 离散化为更小的子立方体，数量等于 H 中的样本数量。每个子立方体可以视作是一个体素，其物理尺寸不应超过换能器阵列的孔径，从而确保有效的声功率传输。在每个子立方体(即体素)中，我们随机选取一个点 $\left(x_l, y_w, z_d\right)$，通过公式(3)~(4)计算得到这一点的复声压 $p\left(x_l, y_w, z_d\right)$，从而在整个 COI 中总共得到了 $L \times W \times D$ 的样本。这种方法的适用性不受工作空间大小或位置的限制，因为无论如何，始终可以生成 $L \times W \times D$ 的样本。此外，保持深度神经网络输入结构的一致性，对于训练通用应用模型也具有重要的作用。声场的输入数据数学表达式为公式(6)：

$$H=\begin{bmatrix} x_1 & x_2 & \cdots & x_{L\times W\times H-1} & x_{L\times W\times H} \\ y_1 & y_2 & \cdots & y_{L\times W\times H-1} & y_{L\times W\times H} \\ z_1 & z_2 & \cdots & Z_{L\times W\times H-1} & z_{L\times W\times H} \\ A_1 & A_2 & \cdots & A_{L\times W\times H-1} & A_{L\times W\times H} \\ \phi_1 & \phi_2 & \cdots & \phi_{L\times W\times H-1} & \phi_{L\times W\times H} \end{bmatrix} \tag{6}$$

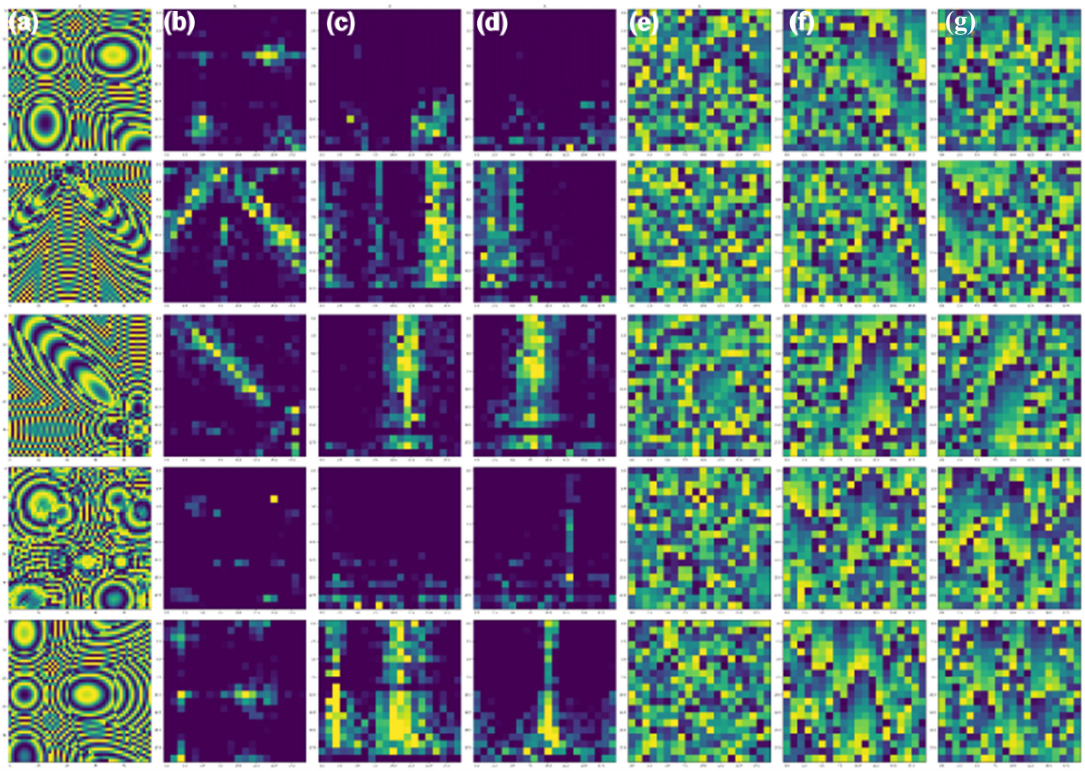

Figure 5. Examples of partial 5 groups of training data pairs
图 5. 部分 5 组训练数据对

在本实验中，我们对全息声场的实验结果图进行了详细的分析。根据图 5 的展示，我们可以清楚地观察到换能器阵列的相位和声场分布，它们可以被视为灰度图像。首先，我们关注标准阵列相位(图 5(a))，该图展示了在给定条件下阵列的相位分布。通过将标准阵列相位与其他结果图进行对比，我们可以评估全息声场的还原效果。接下来，我们观察全息声场的振幅分布。图 5(b)~(d)分别展示了在 xyz 三个横截面上的复声压振幅。通过这三幅图像，我们可以获取声场的空间振幅信息，从而了解声场的传播和分布特性。此外，我们还需要关注全息声场的相位分布。图 5(e)~(g)展示了在 xyz 横截面上的声场相位，它们显示了声场在平面上的相位变化情况。相位信息对于理解声场的波动和相位差异非常重要，可以提供有关波前形状和相干特性等方面的有价值信息。

4.2.1. 可视化实验结果

VS3D-Transformer 方法是基于 PyTorch 实现的，并在具有以下配置的环境中进行了训练：Persistence-M (GPU 服务器)、CUDA 版本 11.0、Python 3.8。为了避免梯度消失并提高模型的泛化能力，我们采用了带动量的 SGD 优化器，其中动量值设置为 0.9。初始学习率设置为 0.01，并且当学习过程停滞时，学习率会以 0.1 倍的倍率进行更新。

本研究通过比较预测的阵列相位和真实值来评估 VS3D-Transformer 的性能。还分析了期望全息声场和生成全息声场之间的差异。图 6 展示了沿着训练的每个惩罚项的渐进优化，用于性能收敛的感知和视觉评估。如图所示，训练的总损失稳定地最小化，其对应的学习率从初始的 0.01 下降至 1e−3。作为对学习表现在第 600 个 epoch 后的诠释 $\mathscr{L}_{\cos}$ 最终降低到 0.02559。

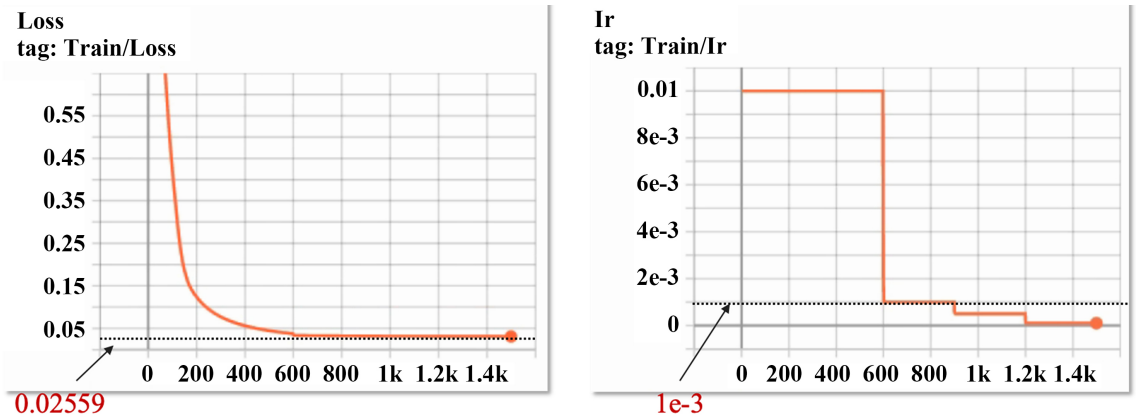

Figure 6. Training loss curve and its corresponding adaptive learning rate curve.
图 6. 训练损失曲线和其对应的自适应学习率曲线

4.2.2. 实验性能对比

图 7 以图形方式展示了对测试数据集中的 10 个样本进行全息声场预测实验的性能，使用了 VS3D-Transformer 模型。在图 7(a)列中，展示了真实阵列相位的图像；而在图 7(b)列中，显示了 VS3D-Transformer 对阵列相位的预测结果。图 7(c)列展示了预测相位和真实相位之间的直接差异。由于相位差在 θ 和 2π−θ 之间是相同的，这反映了声波振荡在设计损失函数时的周期性。图 7(c)中高对比度的区域确实突出了学习到的声网络在全息声场预测方面的出色性能。

除了全息声场的重建性能外，实时性能也是本研究的另一个关注点。根据作者的了解，现有的反射问题方法本质上依赖于迭代优化设计。由于 IB 算法作为一种 PTA 声场控制的物理方法是最先进的方法之一，因此我们将其与我们的方法进行了计算复杂度方面的比较。根据表 I 的结果，可以看出 IB 算法比

我们的方法需要更多的计算时间，这与实时映射相去甚远。无论网络的复杂性如何，我们可以得出结论，所提出的基于深度学习的方法在实时性能方面明显优于迭代反向传播 IB 方法，从而获得更实用的结果。此外，一旦网络经过良好的训练，预测时间就不会受到换能器阵列大小的影响。比较实验结果如表 1 所示。

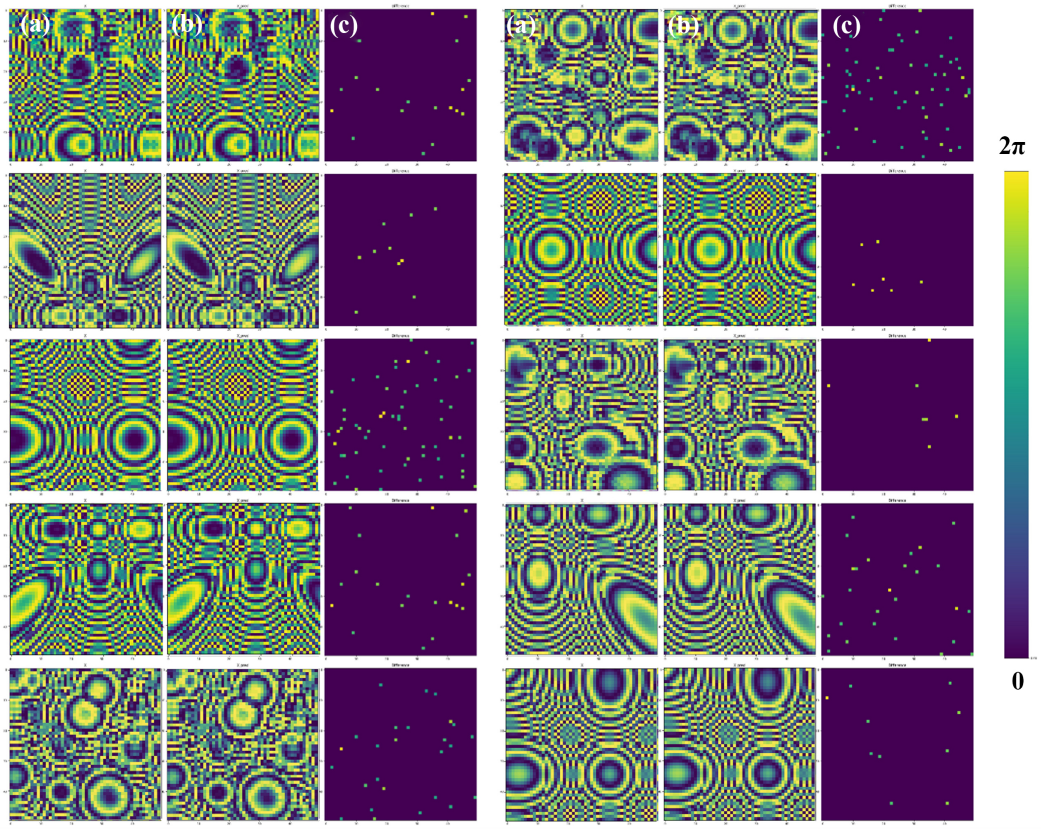

Figure 7. Examples of partial 10 groups of predicted results and their comparisons
图 7. 部分 10 组预测结果及其对比

Table 1. Experimental results of real-time performance comparison.
表 1. 实时性能对比实验结果

方法/类型	VS3D-Transformer	UIB Method [19]	Acousnet [20]
单焦点	212 ms	15.3 min	218 ms
多焦点	214 ms	16.35 min	218 ms
漩涡(2π 相位跨度)	214 ms	12 个控制点 30.72 分钟；24 个控制点未收敛	217 ms

5. 总结

本文利用 Vision transformer 模型和自注意力机制探索全息声场的深度学习，显示出提高非接触操纵技术的准确性和控制的前景。这种新方法专门解决了传统 CNN 在使用过程中仅考虑声场的局部感受野的局限性。这项工作是同类研究中的第一个，基于自注意力机制在克服与复杂全息声场生成相关的复杂性方面的潜力，增强了当前的研究成果，为未来在多个领域的全息声学应用的重大改进提供了支点，展示了跨学科研究的积极影响。

基金项目

该研究得到北京市教委和北京市自然科学基金委员会联合资助项目(KZ201710015010)和 BIGC 项目(Ec202007)的支持。

参考文献

[1] Memoli, G., Caleap, M., Asakawa, M., *et al*. (2017) Metamaterial Bricks and Quantization of Meta-Surfaces. *Nature Communications*, **8**, Article No. 14608. https://doi.org/10.1038/ncomms14608

[2] Li, B., Lu, M., Liu, C., *et al*. (2022) Acoustic Hologram Reconstruction with Unsupervised Neural Network. *Frontiers in Materials*, **9**, Article ID: 916527. https://doi.org/10.3389/fmats.2022.916527

[3] Friend, J. and Yeo, L. (2011) Microscale Acoustofluidics: Microfluidics Driven via Acoustics and Ultrasonics. *Reviews of Modern Physics*, **83**, 647-704. https://doi.org/10.1103/RevModPhys.83.647

[4] Wiklund, M., *et al*. (2006) Ultrasonic Standing Wave Manipulation Technology Integrated into a Dielectrophoretic Chip. *Lab on a Chip*, **6**, 1537-1544. https://doi.org/10.1039/B612064B

[5] Shi, J., *et al*. (2009) Continuous Particle Separation in a Microfluidic Channel via Standing Surface Acoustic Waves (SSAW). *Lab on a Chip*, **9**, 3354-3359. https://doi.org/10.1039/b915113c

[6] Frommelt, T., *et al*. (2008) Microfluidic Mixing via Acoustically Driven Chaotic Advection. *Physical Review Letters*, **100**, Article ID: 034502. https://doi.org/10.1103/PhysRevLett.100.034502

[7] Gao, Y., Yang, B.Q., Shi, S.G., *et al*. (2023) Extension of Sound Field Reconstruction Based on Element Radiation Superposition Method in a Sparsity Framework. *Chinese Physics B*, **32**, Article ID: 044302. https://doi.org/10.1088/1674-1056/ac8e55

[8] Dong, W., Chen, M. and Xiong, L. (2022) Research on Sound Transmission Performance of an Infinite Solid Plate Excited by a Vibrating Piston. *The* 32*nd International Ocean and Polar Engineering Conference*, Shanghai, June 2022. ISOPE-I-22-455.

[9] Lahoud, J., Cao, J., Khan, F.S., *et al*. (2022) 3D Vision with Transformers: A Survey.

[10] He, C., Li, R., Li, S., *et al*. (2022) Voxel Set Transformer: A Set-to-Set Approach to 3d Object Detection from Point Clouds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 8417-8427. https://doi.org/10.1109/CVPR52688.2022.00823

[11] Zhao, S., You, F. and Liu, Z.Y. (2020) Leveraging Pre-Trained Language Model for Summary Generation on Short Text. *IEEE Access*, **8**, 228798-228803. https://doi.org/10.1109/ACCESS.2020.3045748

[12] Tang, Z., Cho, J., Nie, Y., *et al*. (2022) TVLT: Textless Vision-Language Transformer. *Advances in Neural Information Processing Systems*, Vol. 35, 9617-9632.

[13] Cranston, D. (2015) A Review of High Intensity Focused Ultrasound in Relation to the Treatment of Renal Tumours and Other Malignancies. *Ultrasonics Sonochemistry*, **27**, 654-658. https://doi.org/10.1016/j.ultsonch.2015.05.035

[14] Geng, J. (2013) Three-Dimensional Display Technologies. *Advances in Optics and Photonics*, **5**, 456-535. https://doi.org/10.1364/AOP.5.000456

[15] Zhao, T. and Chi, Y. (2020) Modified Gerchberg-Saxton (GS) Algorithm and Its Application. *Entropy*, **22**, Article No. 1354. https://doi.org/10.3390/e22121354

[16] Fushimi, T., Yamamoto, K. and Ochiai, Y. (2021) Acoustic Hologram Optimisation Using Automatic Differentiation. *Scientific Reports*, **11**, Article No. 12678. https://doi.org/10.1038/s41598-021-91880-2

[17] Plasencia, D.M., Hirayama, R., Montano-Murillo, R., *et al*. (2020) GS-PAT: High-Speed Multi-Point Sound-Fields for Phased Arrays of Transducers. *ACM Transactions on Graphics* (*TOG*), **39**, Article No. 138. https://doi.org/10.1145/3386569.3392492

[18] Long, B., Seah, S.A., Carter, T., *et al*. (2014) Rendering Volumetric Haptic Shapes in Mid-Air Using Ultrasound. *ACM Transactions on Graphics* (*TOG*), **33**, Article No. 181. https://doi.org/10.1145/2661229.2661257

[19] Marzo, Y.A. and Drinkwater, B.W. (2019) Holographic Acoustic Tweezers. *Proceedings of the National Academy of Sciences*, **116**, 84-89. https://doi.org/10.1073/pnas.1813047115

[20] Zhong, C., Jia, Y., Jeong, D.C., *et al*. (2021) A Deep Learning Based Approach to Dynamic 3d Holographic Acoustic Field Generation from Phased Transducer Array. *IEEE Robotics and Automation Letters*, **7**, 666-673. https://doi.org/10.1109/LRA.2021.3130368