



本科生毕业论文（设计）

题目： 基于深度学习中 Transformer

在光声成像中的探讨

姓 名 黄梓航

学 号 19337047

院 系 数学学院（珠海）

专 业 数学与应用数学

指导教师 时聪

2023 年 5 月 9 日

基于深度学习中 Transformer 在光声成像中的探讨

Photoacoustic Imaging Reconstruction based on Transformer

姓 名 黄梓航

学 号 19337047

院 系 数学学院（珠海）

专 业 数学与应用数学

指导教师 时聪

2023 年 5 月 9 日

学术诚信声明

本人郑重声明：所呈交的毕业论文（设计），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文（设计）不包含任何其他个人或集体已经发表或撰写过的作品成果。对本论文（设计）的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本论文（设计）的知识产权归属于培养单位。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日 期： 年 月 日

【摘 要】

本项目研究如何提高光声成像重构结果。研究过程中发现在现有光声成像重建算法下，重建图像质量与采集数据的传感器数量成正比。而较多的传感器数量往往带来高昂的仪器成本，从而阻碍光声成像技术的普及。但目前光声成像局限于研究领域的现状，使得控制光声成像成本相关的研究比较匮乏。为此，本研究采用深度学习探究在低成本下得到高质量光声成像重建图像的方法。

本研究创新性地将 Transformer 模型引入光声成像重建领域，使用 k-Wave 进行光声成像仿真并生成了有关皮肤癌的医学图像数据集。基于该数据集搭建与训练了 Swin-Unet 神经网络，并运用其实现了将低质量光声重建图像优化为高质量医学图像的模型，为降低光声成像成本提供了一种解决办法。

关键词： 光声成像， Transformer， 图像重构， k-Wave

[ABSTRACT]

This project investigates how to improve the quality of photoacoustic imaging reconstruction images. During the study we find that under the existing photoacoustic imaging reconstruction algorithms, the quality of the reconstructed image is directly proportional to the number of sensors collecting data. A large number of sensors often brings high instrument costs, which hinders the popularization of photoacoustic imaging technology. However, at present, photoacoustic imaging is limited to research field, which makes the research related to controlling the cost of photoacoustic imaging relatively lacking. Therefore, this study uses deep learning to explore the method of obtaining high-quality photoacoustic imaging reconstruction images at low cost.

In this study, we innovatively introduce the Transformer model into the field of photoacoustic imaging reconstruction, use k-Wave for photoacoustic imaging simulation, and build a medical image dataset on skin cancer. Based on this dataset, a Swin-Unet neural network is built to optimize low-quality photoacoustic reconstruction images into high-quality medical images, which provides a solution for reducing the cost of photoacoustic imaging.

Keywords: Photoacoustic Imaging, Transformer, Image Reconstruction, k-Wave

目录

1 绪论	1
1.1 课题背景	1
1.2 研究问题的发现及研究课题的提出	2
1.3 基于深度学习的光声成像重建的研究现状	5
1.4 论文章节安排	6
2 光声成像的理论基础	8
2.1 光声成像的波动方程	8
2.2 光声成像的重建算法	9
3 Transformer 的理论基础	11
3.1 transformer 的简介	11
3.2 自注意力机制	11
3.3 Transformer 在 CV 领域的应用与改进	16
3.4 SwinTransformer 的优势	19
4 训练数据的生成与预处理	20
4.1 数据集的介绍	20
4.2 k-wave 的介绍与 k 空间伪谱法	20
4.3 训练数据的生成与预处理	20
5 Transformer 模型的搭建与训练	24
5.1 神经网络的结构	24
5.2 Encoder 和 Decoder 结构与 Long Skip Connection	24
5.3 Patch Embeding、Patch Merging 层、Patch Expanding 层和 Patch Expandx4 层	25
5.4 Swin Transformer Block	28
5.5 模型训练结果	29

6 模型的分析与评估	31
6.1 MSE	31
6.2 PSNR	32
6.3 SSIM	34
6.4 余弦相似度	37
6.5 哈希相似度	38
6.6 直方图相似度	39
6.7 归一化互信息	41
6.8 模型评估总结	42
6.9 模型不足与发展方向	43
参考文献	44
致谢	45

插图目录

1.1	光声成像与其他成像技术在分辨率及深度上的对比 ^[1]	2
1.2	不同传感器数量下的重建图像	3
1.3	不同传感器数量下重建图像与原图像间的 MSE、PSNR、SSIM	4
1.4	可学习的非迭代重建方案	5
1.5	可学习的迭代重建方案	6
3.1	”Attention Is All You Need” 中的 Transformer 模型 ^[2]	11
3.2	翻译任务 1	12
3.3	翻译任务 2	12
3.4	自然语言编码过程	12
3.5	注意力机制的原理	13
3.6	自然语言解码过程	13
3.7	运用注意力机制的机器翻译过程	13
3.8	将 X 经过三个线性变换后得到 Q、K、V	14
3.9	Q 与 K 的转置相乘	14
3.10	将得到的注意力权重矩阵与 V 相乘	14
3.11	自注意力机制的实现	15
3.12	二头注意力机制的实现	16
3.13	Axial-Attention 模型的结构 ^[3]	17
3.14	ViT 模型的结构 ^[4]	18
3.15	Swin Transformer 中的 SW-MSA 操作 ^[5]	18
4.1	Skin Cancer MNIST 数据集示例	20
4.2	预处理操作流程	21
4.3	k-Wave 仿真的设置参数 ^[6]	22
4.4	光声传感器接收到的压力数据示例	22
4.5	重建图像示例	23

5.1	神经网络的结构	24
5.2	tensor 在神经网络内的形状变化	25
5.3	Patch Merging 的实现原理	27
5.4	Swin Transformer Block 的结构	29
5.5	loss 随 epoch 的变化曲线	30
5.6	重建图像优化效果示例	30
6.1	测试集图像及其预测图象与原图像的 MSE	31
6.2	测试集图像及其预测图象与原图像的 MSE 的均值	32
6.3	测试集图像及其预测图象与原图像的 PSNR	33
6.4	测试集图像及其预测图象与原图像的 PSNR 的均值	33
6.5	测试集图像及其预测图象与原图像的 SSIM	36
6.6	测试集图像及其预测图象与原图像的 SSIM 的均值	36
6.7	测试集图像及其预测图象与原图像的余弦相似度的均值	37
6.8	测试集图像及其预测图象与原图像的三种哈希相似度的均值	39
6.9	皮肤癌图片的直方图	40
6.10	测试集图像及其预测图象与原图像的直方图相似度的均值	40
6.11	测试集图像及其预测图象与原图像的 NMI 的均值	42

表格目录

5.1 Patch Merging 的形状变化	26
5.2 Patch Expanding 的形状变化	26
6.1 验证集的各项评价指标的均值	42

1 绪论

1.1 课题背景

1.1.1 光声成像简介

光声成像的物理基础是光声效应。光声效应是最早由 A.G.Bell 于 19 世纪 20 年代所发现的物理现象。所谓光声效应，即当物体接受激光脉冲等光源照射后，因吸收光源的能量发生热弹性膨胀，产生压力波并通过物体传播的现象。如果利用激光束照射生物体，使其产生光声效应后，使用完全或部分围绕生物体表面 S （常被称为采集表面）上的声学传感器记录随时间 t 所发生的压力变化 $p(x, t)$ 。并根据声学测量数据 $p(x, t)$ 重建初始压力 $f(x)$ ，从而得到生物体成像的方法，就称为光声成像。

虽然光声效应被发现已久，但光声成像技术在近年来才得以被重视，并由于其高分辨率、高对比度的成像与无损害、无辐射的优势开始迅速发展。虽然如今的光声成像技术大多局限于研究领域，但随着光声成像技术的成熟，光声成像技术必将拥有更广阔前景，在未来必将广泛地应用于临床医学检测领域。

1.1.2 光声成像的优势

现今我们已经掌握了多项成熟且完善的医学成像方法，包括广泛运用于医疗场所的利用 X 射线的 CT 成像技术，超声波扫描技术，核磁共振成像技术等，还有不太为公众所知的 PET 等技术。抛开现有的成熟技术，转而开发新的光声层析成像方法 (PAT) 是否是必要的呢？一个主要的答案是光声层析成像具有其他成像方法所不具备的优点。

1.1.2.1 光声层析成像作为一种“混合”成像方法，能有效避免单一成像方法的弊端

现有的成像方法大都依赖于某一种物理方法，而单一的物理方法所能得到的测量信息是有局限的。例如，由于许多肿瘤细胞比健康细胞在某些特定的能带上能吸收更多的电磁能量，因此使用电磁波的成像方法可以提供非常高的对比度，但由于电磁波的特性，这类成像方法只能提供非常低的分辨率。

如何规避单种物理方法所带来的局限性呢？一个显而易见的改进方法是同时使用多种物理过程成像，这种成像模式被称为“混合”成像方法（Hybrid imaging methods），而光声层析成像正是一种新兴的“混合”成像方法。光声层析成像同时使用了光波与声波两种物理波，即使用电磁波照射病灶，而后基于声学测量重建图像。事实上，由于癌细胞对激光能量的高吸收率，因此光声成像能产生高对比度的图像（此时单独使用超声波无法产生良好的对比度），并且通过超声波测量还获得了良好的分辨率（而激光波对于高分辨率图像来说太长了）。因此光声成像通过使用两种类型的波，在结合了它们优点的同时，消除了它们各自的缺点。如图1.1，光声成像的两种实现模态（PACT、PAM）与其他的成像模态相比，具有更高的深度分辨率比。

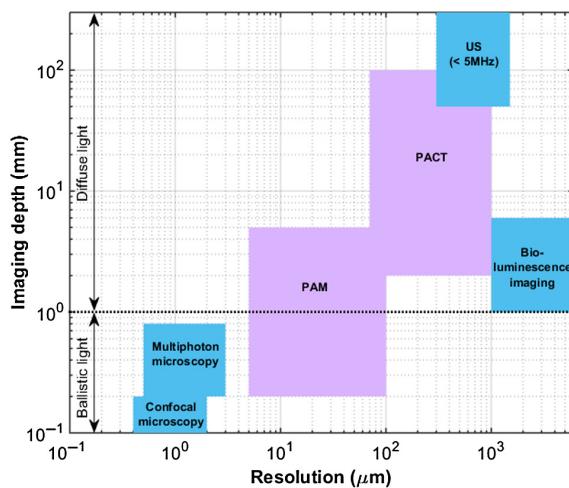


图 1.1 光声成像与其他成像技术在分辨率及深度上的对比^[1]

1.1.2.2 光声成像技术对人体更安全

现在广泛运用的成像方法都或多或少地对人体产生不好的影响。例如，X 射线及 CT 具有较强的电离辐射，频繁使用会增加人体细胞的癌变风险等。但是由于光声成像产生的激光功率密度与超声场强度都远低于生物组织损伤阈值，所以光声成像是一种较安全、对人体无损的成像技术，能更好地服务于病人和医生。

1.2 研究问题的发现及研究课题的提出

光声成像作为一项新兴的医学成像技术，在目前主要局限于研究领域，此时光声成像仪器的成本因素并不是主要考虑因素。但随着光声成像技术的日渐成熟，要想将光声成像技术广泛地应用于医疗卫生领域，其成本因素在实际推广的过程中往往不可忽视，有时甚至处于主导地位。虽然光声成像在实际的应用过程中具

有其他现有成像方法所不具备的许多优势，但若在其推广的过程中无法控制应用成本，则必将受到层层阻碍。因此，如何降低光声成像的成本（包括设备成本、人力成本、耗材成本等），使得这项利民的医学技术能更快地使病人与医疗人员受益，是一项十分重要且有意义的工作。

因此，本文拟探究在低成本下得到高质量光声成像图像的方法，以降低光声成像技术在实际推广过程中的应用成本。

1.2.1 光声成像重建算法问题的发现

在前期的研究过程中发现，在现有的光声成像重建方法中，点状声波探测器的个数往往对重建效果起到关键性的影响。对点状声波探测器个数选用了 50 个、100 个、200 个、400 个、800 个的条件下（传感器阵列均采用围绕病灶排列且等距分布的中心圆），使用 MatLab 中的 k-Wave 对 Skin Cancer MNIST 数据集中选取的 100 张图片做光声成像的仿真以及重建，对得到的重建图像与原图像使用如 MSE、SSIM、PSNR 等指标进行评价，比较其在不同传感器个数下的成像质量。

我们挑选出其中一张皮肤癌图片，在不同传感器个数下的重建结果如图1.2。

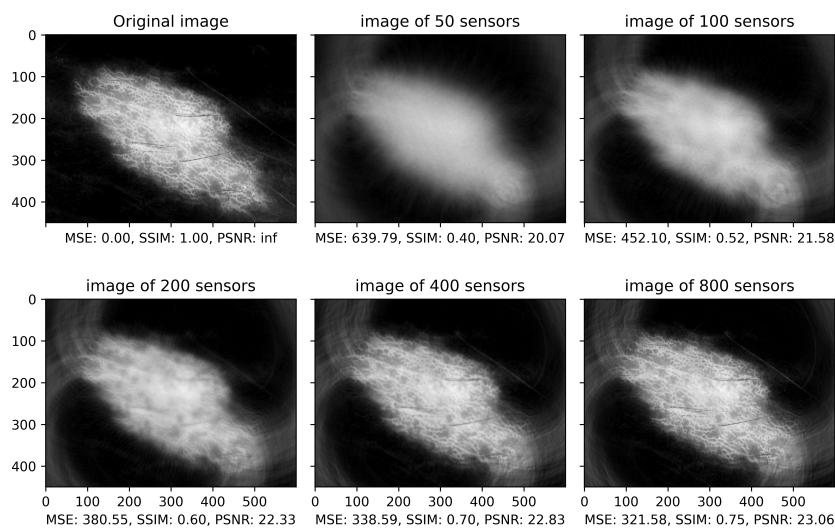


图 1.2 不同传感器数量下的重建图像

可以发现，随着传感器个数的减少，重建图片与原图片的 MSE 逐渐增加，SSIM、PSNR 值逐渐降低，这说明该图片的成像质量不断降低。

为了排除单一案例造成的误差，我们计算出在不同传感器个数下，这 100 张重建图片与原图片的 MSE、SSIM、PSNR 的平均值。最终得到如图1.3的比较结果。

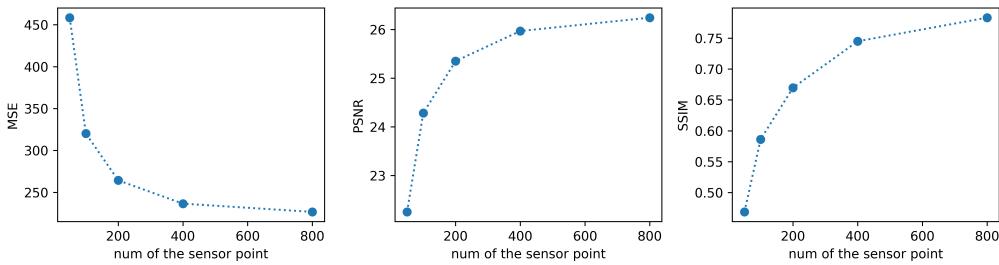


图 1.3 不同传感器数量下重建图像与原图像间的 MSE、PSNR、SSIM

从图1.3的比较结果可知，当声波探测器的个数增加时，光声重建图像的 MSE 的均值呈现下降趋势，PSNR 与 SSIM 的均值呈现上升趋势。这表明，传统光声成像重建算法的重建图像的精度和质量与仪器的声学传感器个数成正相关关系。并且可以得知，在声学传感器个数 ≥ 400 个时，现有的算法重建的图像才具备较好的成像精度。

值得注意的是，上述结果是在没有为传感器数据添加噪声以及对图像进行任何降采样的情况下得出的。在实际情况中，由于传感器所接收到的数据信号往往受到噪声的影响，重建效果往往更差。

1.2.2 本项目研究课题的提出

从上述观察可以看出，以现有的光声成像重建方法，重建的图像质量与光声成像仪器的声波探测器数量成正比。如果需要得到符合诊断标准的医学图像，就需要装备较多的声波探测器，从而进一步增加光声成像仪器的设备成本。

如何利用较少传感器数量产生的较低精度光声重建图像，并将其优化为符合医学标准的光声重建图像就成为了本项目的出发点。对此，我们提出在图像处理领域，神经网络是一种成本较低的图像优化方案（神经网络的主要成本在于数据集的搭建及神经网络的训练）。并且由于神经网络的高延展性，在少量数据集上进行训练后的神经网络能有效应用到其他的数据集。因此，该方法能有效控制光声成像的应用成本。

现今虽然已存在很多已训练的效果显著的图像优化网络。但由于医学图像的特殊性，若将这些神经网络直接运用于医学图像的优化，并使用优化后的医学图像进行疾病的诊断与治疗，不仅难以保证优化后的图像符合诊断标准，而且还有可能产生极大的医疗风险。因此，为光声成像单独搭建一组数据集并据此训练一套图像优化模型就显得十分必要了。

综上所述，本项目的研究课题拟在搭建并训练一套神经网络模型，以实现将在低传感器数量下产生的低精度光声重建图像还原成高精度光声重建图像的目的。

1.3 基于深度学习的光声成像重建的研究现状

如今基于深度学习的光声成像图像重建主要有两大类研究方向：第一类是“可学习的非迭代重建”；第二类是“可学习的迭代重建”。

1.3.1 可学习的非迭代重建

对于光声成像的整个过程（如图1.4所示），可学习的非迭代重建主要有三种：

- 1) 增强测量信号：在重建的过程中，我们接受的信号可能存在畸变，如带宽的削减、噪声、通道数的跌落。于是我们可以通过深度学习的方法来将测量信号进行增强，如增强带宽、增强通道数。然后再对增强后的信号使用传统的重建方法重建。（图1.4橙色过程）
- 2) 学习逆过程：训练一个神经网络模型，来实现将接收到的信号传入，并直接输出重建图像的功能。（图1.4蓝色过程）
- 3) 增强重建图像：用神经网络对获取的重建图像进行增强，输出一个更高质量的图像。（图1.4绿色过程）

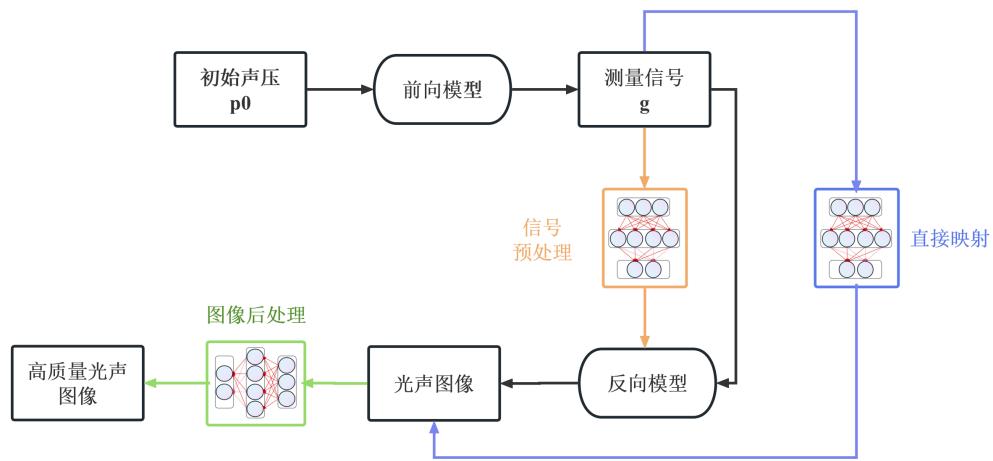


图 1.4 可学习的非迭代重建方案

1.3.2 可学习的迭代重建

可学习的迭代重建主要有两种：

- 1) 利用深度学习学习重建过程中的优化方法（图1.5蓝色过程）
- 2) 利用深度学习学习重建过程中的正则项（图1.5绿色过程）

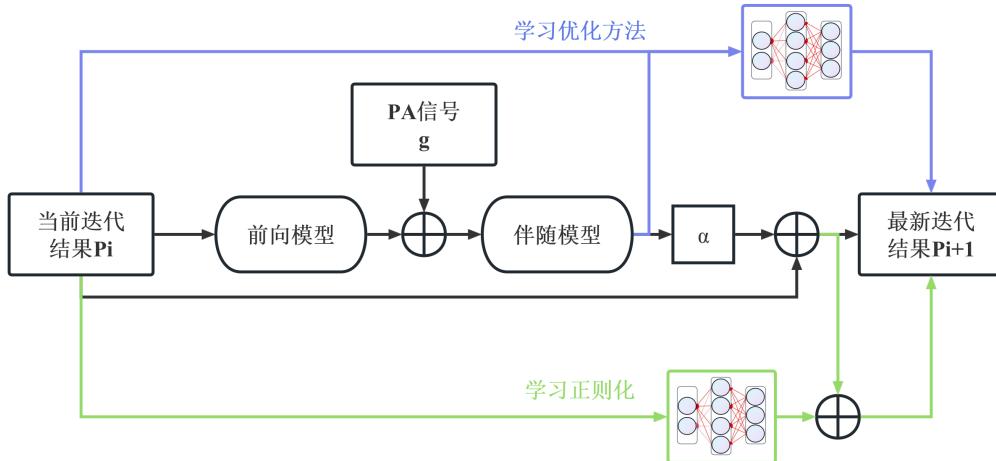


图 1.5 可学习的迭代重建方案

1.3.3 本项目的研究现状

在上述基于深度学习的光声成像重建方向中，本项目的研究课题是对非迭代重建方法中的后处理过程进行研究，以达到增强重建图像的研究目的。

关于利用深度学习进行光声成像的重建，现在也有不少的相关工作，比如用 CNN 来进行光声成像重建。而 Transformer 作为一个新兴的、热门的神经网络，在处理图像上超越了很多传统的神经网络模型。Transformer 将自注意力机制引入 CV 领域，解决了以往神经网络诸如 CNN 无法很好地学习全局信息等缺点。Axial-Attention（轴向注意力）、ViT、Swin Transformer 等不同的 Transformer 模型，都在计算机视觉领域取得了很大的成功，Swin Transformer 更是其中的佼佼者。

因此，本项目拟将现今在 CV 领域效果卓越的 Swin Transformer 运用在光声成像重建后的图像优化，尝试并解决现今传统重建算法的各种问题。

1.4 论文章节安排

本文共分为六章，各章节内容安排如下：

第一章绪论。简单说明了本文章的选题背景与意义。

第二章阐述与光声成像有关的的理论基础，包括介绍光声成像的波动方程模型以及常见的几种光声成像重建算法等。

第三章主要介绍了 Transformer 模型的相关理论基础，包括 Transformer 模型中的注意力机制的理解与实现，及综述了现今将 Transformer 模型运用于 CV 领域的若干神经网络架构的实现原理。

第四章主要介绍该项目中利用 MatLab 及 k-Wave 搭建光声成像仿真与重建数

据集的若干细节。

第五章主要介绍使用 PyTorch 实现 Transformer 模型并利用自建数据集进行训练的过程。

第六章的主要内容是对该 Transformer 模型的实验效果进行分析与评估。

2 光声成像的理论基础

2.1 光声成像的波动方程

假设 $p(x, t)$ 为位于整个采集表面 S 上位置 x 的点在时刻 $t(t \geq 0)$ 的压力值。并且记位于表面 S 上位置 y 的点状声波探测器在观测时刻 t 获得的声波压力数据为函数 $g(y, t)$, 即 $g(y, t) := p(y, t)$ for $y \in S, t \geq 0$ 。

于是我们得到一个波动方程:

$$\begin{cases} p_{tt} = c^2(x) \Delta_x p, & t \geq 0, x \in \mathbb{R}^3 \\ p(x, 0) = f(x), p_t(x, 0) = 0 \\ p|_S = g(y, t), & (y, t) \in S \times \mathbb{R}^+. \end{cases} \quad (2.1)$$

其中, $f(x)$ 是声压的初始值。因此, 光声成像的目标是使用传感器的测量数据 $g(y, t)$, 反推出上述波动方程 $p(x, t)$ 在 $t=0$ 处的初始值 $f(x)$ 。

我们做如下记号:

定义 1: 我们使用 \mathcal{W} 表示正算子, 即 $\mathcal{W}: f(x) \rightarrow g(y, t)$, 其中 f 与 g 的定义同上述波动方程。

如果成像介质是均匀的, 即 $c(x)$ 等于一个常数, 我们假设该常数在适当的单位下等于 1。此时, 波动方程为:

$$\begin{cases} p_{tt} = \Delta_x p & t \geq 0, x \in \mathbb{R}^3 \\ p(x, 0) = f(x), p_t(x, 0) = 0 \\ p|_S = g(y, t), & (y, t) \in S \times \mathbb{R}^+. \end{cases} \quad (2.2)$$

此时, 根据 Poisson Kirchhoff 公式, 我们能得到上述波动方程的解为:

$$p(x, t) = a \frac{\partial}{\partial t} (t(Rf)(x, t)). \quad (2.3)$$

其中 $(Rf)(x, r) := \frac{1}{4\pi} \int_{|y|=1} f(x + ry) dA(y)$ 是作用于函数 $f(x)$ 的球面平均算子, dA 是 \mathbb{R}^3 单位球面上的标准面积元, a 为常数。

从上述公式可以得知, $p(x, t)$ 由函数 f 的球面平均值 $(Rf)(x, t)$ 唯一决定。我

们将这个球面平均算子作用在 f 上的映射 $R : f \rightarrow Rf$ 记为 \mathcal{M} , 即:

$$\mathcal{M}f(x, t) := \frac{1}{4\pi} \int_{|y|=1} f(x + ry) dA(y), \quad x \in S, t \geq 0. \quad (2.4)$$

因此, 在成像介质是均匀的情况下, 我们可以选择使用 \mathcal{M} 来代替 $p(x, t)$ 进行研究。

2.2 光声成像的重建算法

对于成像介质是均匀介质的情况 (此时 $c(x)$ 为常数), 有上面的讨论可得, 光声成像的图像重建等效于求解球面均值变换 \mathcal{M} 的逆。下面介绍几种常见的光声成像重建方法:

2.2.1 幂级数解法

将 f 和 g 进行傅里叶分解后, 即

$$f(x) = \sum_{-\infty}^{+\infty} f_k(\rho) e^{ik\varphi}, \quad x = (\rho \cos(\varphi), \rho \sin(\varphi)). \quad (2.5)$$

$$g(y(\theta), r) = \sum_{-\infty}^{+\infty} g_k(r) e^{ik\theta}, \quad y = (R \cos(\theta), R \sin(\theta)). \quad (2.6)$$

将其代入到公式 (2.3) 中, 由等式两边系数相等可得:

$$f_k(\rho) = \mathcal{H}_m \left(\frac{1}{J_k(\lambda|R|)} \mathcal{H}_0 \left[\frac{g_k(r)}{2\pi r} \right] \right). \quad (2.7)$$

其中 $(\mathcal{H}_m u)(s) = 2\pi \int_0^\infty u(t) J_m(st) dt$ 为 Hankel 变换, $J_m(t)$ 为贝塞尔函数。

应该注意的是, 幂级数解法依赖于在球面几何中成立的变量分离, 因此这种方法仅在球面上成立。

2.2.2 特征函数展开法

设 λ_m 和 $u_m(x)$ 为封闭曲面 S 内部 Ω 的狄利克雷-拉普拉斯算子 $-\Delta$ 的特征值和特征函数的正交基, 满足:

$$\begin{cases} \Delta u_m(x) + \lambda_m^2 u_m(x) = 0, & x \in \Omega, \Omega \subseteq \mathbb{R}^n \\ u_m(x) = 0, & x \in S \\ \|u_m\|_2^2 \equiv \int_{\Omega} |u_m(x)|^2 dx = 1. \end{cases} \quad (2.8)$$

可解得：

$$u_m(x) = \int_S \Phi_{\lambda_m}(|x - y|) \frac{\partial}{\partial n} u_m(y) ds(y), \quad x \in \Omega. \quad (2.9)$$

其中 $\Phi_{\lambda_m}(|x - y|)$ 是亥姆霍兹方程的自由空间格林函数， n 是 S 的外法向量。

函数 $f(x)$ 可以展开成级数：

$$f(x) = \sum_{m=0}^{\infty} \alpha_m u_m(x), \quad \text{where } \alpha_m = \int_{\Omega} u_m(x) f(x) dx. \quad (2.10)$$

如果将表示形式 (2.9) 替换为表示形式 (2.10) 并交换积分顺序，则可以得到

α_m 。

$$\alpha_m = \int_{\Omega} u_m(x) f(x) dx = \int_S I(y, \lambda_m) \frac{\partial}{\partial n} u_m(y) dA(x),$$

where

$$(2.11)$$

$$I(y, \lambda) = \int_{\Omega} \Phi_{\lambda}(|x - y|) f(x) dx = \int_0^{diam \Omega} g(y, r) \Phi_{\lambda}(r) dr.$$

将 α_m 代入级数 (2.10) 就能得到重建公式 $f(x)$ 。

3 Transformer 的理论基础

3.1 transformer 的简介

Transformer 是最早在自然语言处理领域提出并逐渐被广泛运用于其他领域的神经网络模型。Transformer 的出现在性能上超越了很多传统的自然语言处理模型如 RNN、LSTM 等。并且 Transformer 还具有突破了并行计算的限制、更具可解释性等优点。

在最开始的使用中，transformer 包括 encoding（编码器）和 decoding（解码器）两个部分。其神经网络的结构如图3.1所示：

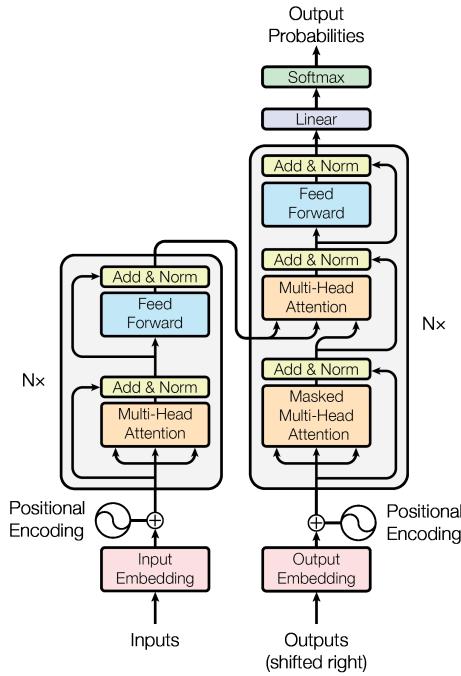


图 3.1 ”Attention Is All You Need” 中的 Transformer 模型^[2]

3.2 自注意力机制

3.2.1 注意力机制的原理

注意力机制是 Transformer 的核心机制。自注意力机制的提出是对人类获取外界信息的机制的一种抽象。当我们观察外界信息，并不是对外界的所有信息“全盘吸收”，而是会无意识地忽略某些“不重要”的信息，从而能提高对外界信息的吸收效率。我们用一个机器翻译的任务来阐述注意力机制的原理。

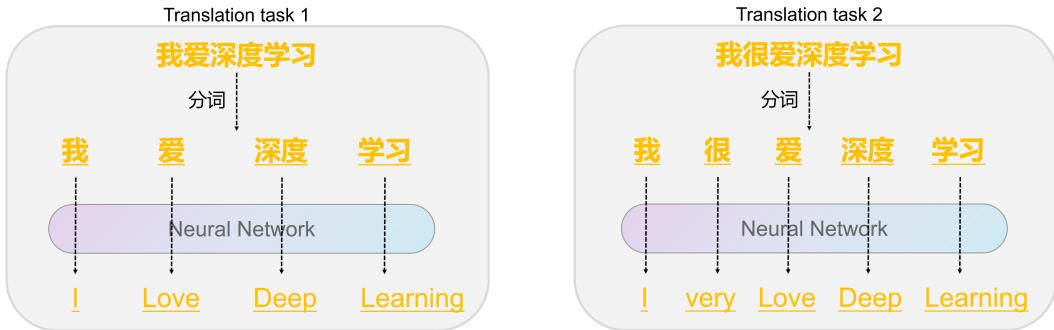


图 3.2 翻译任务 1

图 3.3 翻译任务 2

假设我们要使用神经网络来完成中译英的机器翻译任务。如果完全不考虑整个句子的上下文信息，那么机器翻译其实就是训练神经网络，使其能将特定的中文字符映射到特定的英文字符，这时只需采用最简单的 Mlp 神经网络就能实现这种功能。这种做法在某些机器翻译任务中是可行的，如图3.2。但在大多数的机器翻译任务中，这种直接映射的做法会导致语法上的错误或语义上的歧义，如图3.3的机器翻译任务。

这时，在翻译单个词时，通过结合上下文信息再进行翻译就十分重要了。而 Transformer 中的注意力机制就能很好地实现上下文信息的结合。首先在进行机器翻译的任务之前，都会对翻译的单词进行“编码”使其能被计算机运算与处理。比如将四字句子“我”“很”“爱”“深度”“学习”编码成五个向量 x_1, x_2, x_3, x_4, x_5 。并记 $X = [x_1, x_2, x_3, x_4, x_5]$ 代表整个句子，即如图3.4所示。

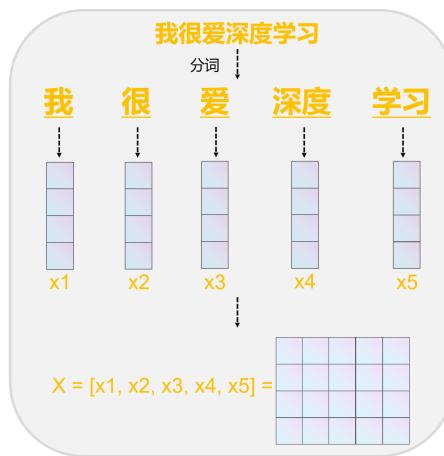


图 3.4 自然语言编码过程

在翻译词“很”的时候，我们想知道所在句子中的其他词所提供的信息量的程度。我们需要找到一个值域为 $[0, 1]$ 的注意力打分函数 F ，来衡量这种提供信息的权重。即若我们输入 $F(\text{“很”}, \text{“我”}) = F(x_2, x_1) = \omega_{21}$ ，所得到的 ω_{21} 代表在翻译“很”字时，“我”字所提供信息的权重。以此类推，我们可以得到各个字在翻译“很”字时提供的信息权重 $\omega_{21}, \omega_{22}, \omega_{23}, \omega_{24}, \omega_{25}$ ，即：

$$\left\{ \begin{array}{l} F(\vec{x}_2, \vec{x}_1) = \omega_{21} \\ F(\vec{x}_2, \vec{x}_2) = \omega_{22} \\ \vdots \\ F(\vec{x}_2, \vec{x}_5) = \omega_{25} \end{array} \right. \quad (3.1)$$

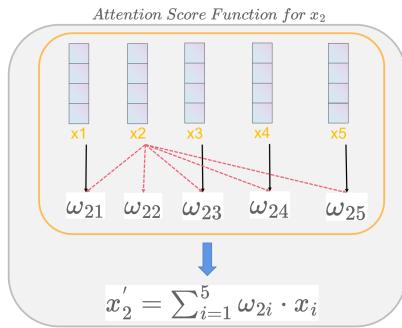


图 3.5 注意力机制的原理

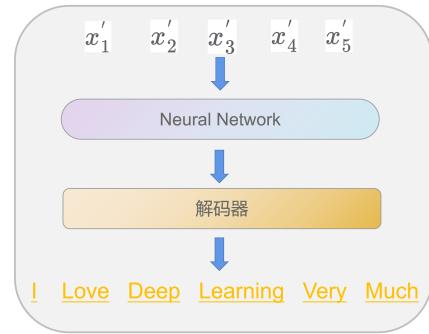


图 3.6 自然语言解码过程

这时根据注意力权重进行加权求和得到的 $x_2' = \omega_{21}x_1 + \omega_{22}x_2 + \omega_{23}x_3 + \omega_{24}x_4 + \omega_{25}x_5$ 就能代表结合了上下文信息后的“很”字，如图3.5。同样的步骤，我们能得到结合了全局信息的 x_1', \dots, x_5' ，将其作为神经网络的输入，就能得到考虑了上下文信息的翻译，如图3.6。

机器翻译的 Transformer 模型就是上述的三个步骤的相互组合，如图3.7。

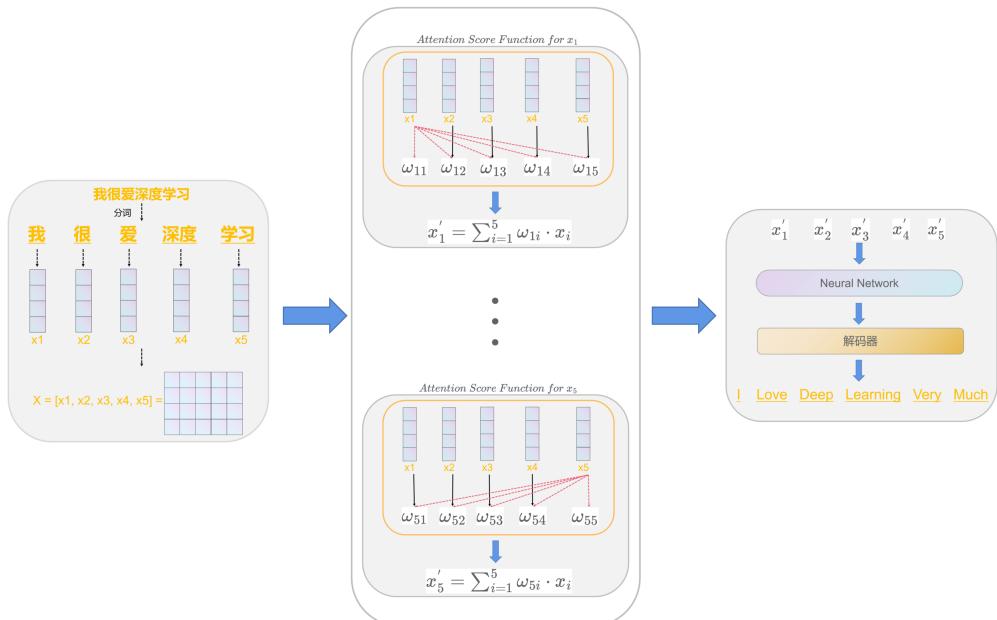


图 3.7 运用注意力机制的机器翻译过程

3.2.2 自注意力机制的实现

自注意力机制的运算实现主要分成四个部分：

3.2.2.1 将 X 经过三个线性变换后得到 Q, K, V

将多个研究对象进行编码后得到的向量 x_1, \dots, x_n (在如图3.3所示的任务中, x_1, \dots, x_5 分别代表“我”“很”“爱”“深度”“学习”五个单词), 将其按行排列而成的矩阵记为 X , 如图3.8左边。 X 矩阵与定义三个线性变换矩阵 W_q, W_k, W_v 相乘得到三个矩阵 Q, K, V 。得到的 Q, K, V 的各行元素各代表 X 矩阵中的一个向量。

其中 W_q, W_k, W_v 为可学习的参数, 这一步的目的是为了使注意力打分函数成为可学习的函数。

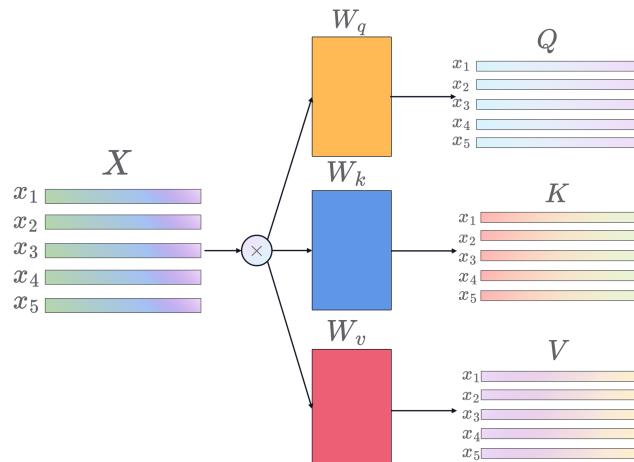


图 3.8 将 X 经过三个线性变换后得到 Q, K, V

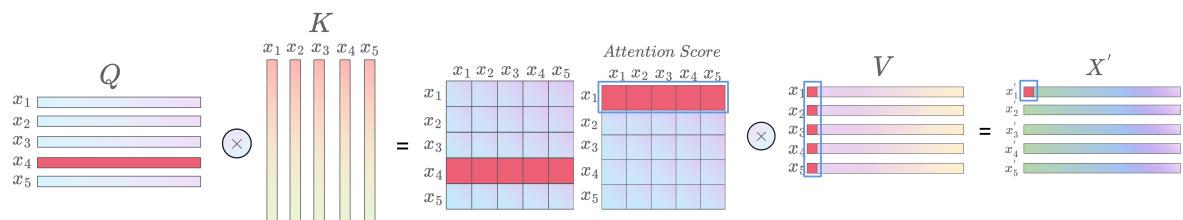


图 3.9 Q 与 K 的转置相乘

图 3.10 将得到的注意力权重矩阵与 V 相乘

3.2.2.2 将 Q 与 K 输入注意力打分函数得到注意力权重

其中注意力打分函数的实现如下:

$$F(X) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right) \quad (3.2)$$

注意力打分函数的运算由两部分组成：

- 1) 第一部分是将 Q 与 K 的转置相乘，得到一个矩阵。该矩阵的第 i 行 j 列的元素为 x_i 与 x_j 的转置相乘得到的，代表“将 x_i 和 x_j 代入到注意力打分函数”这一操作。如图3.9所示，右边矩阵红色方块所在行为 Q 中代表 x_4 的行与 K 转置中代表 x_1, x_2, \dots, x_5 所在列相乘得到的。
- 2) 第二部分将所得矩阵各元素进行标准化，而后再进行一次 softmax 运算。这一步的作用是确保注意力打分函数得到的矩阵各元素都为位于 $[0,1]$ 之间的权重。运算后得到的矩阵即为注意力权重 (Attention Score) 矩阵。

在图3.3的机器翻译任务中，第4行1列元素的值代表在翻译“深度”字时，“我”字所提供的信息权重。

3.2.2.3 将得到的注意力权重矩阵与 V 相乘

如图3.10所示，注意力权重矩阵第一行与 V 相乘得到的 X' 中的第一行元素 x'_1 ，可以看作利用所给权重结合 x_1, \dots, x_5 信息后的 x_1 。因此最终运算得到的矩阵即代表按照所给权重结合全局信息后的元素 x'_1, \dots, x'_5 按行排列而成的矩阵 X' 。

3.2.2.4 最终对 X' 再做一次线性变换使之恢复为原来的矩阵形状后输出

这里所乘的线性变换矩阵 W_0 也为可学习的参数。

综上所述，总体的过程如图3.11所示。

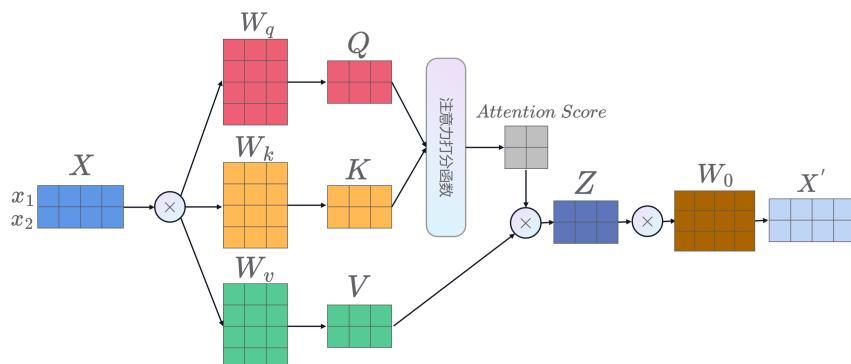


图 3.11 自注意力机制的实现

3.2.3 多头注意力机制

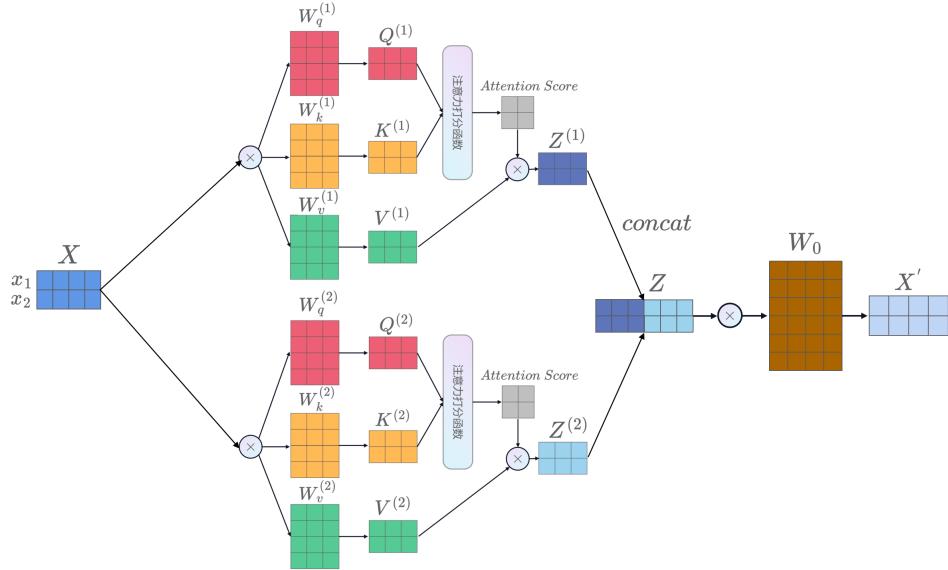


图 3.12 二头注意力机制的实现

正如 CNN 中不同的卷积核能学习到不同的特征，研究人员相信通过增加注意力机制中的注意力打分函数个数，可以让模型学习到不同的信息权重特征，进而优化模型表现。

比如图3.12所示的二头注意力机制，通过两组线性变化 $(W_q^{(1)}, W_k^{(1)}, W_v^{(1)})$ 和 $(W_q^{(2)}, W_k^{(2)}, W_v^{(2)})$ 将 X 映射成两组 Q, K, V 后，通过自注意力机制得到 $z^{(1)}, z^{(2)}$ ；将得到的 $z^{(1)}, z^{(2)}$ 经过 concat 连接后，再对其进行一次线性变换，恢复原形状后输出。

其余多头注意力机制的实现与二头注意力机制的实现是相似的，区别仅在于 n 头注意力机制的注意力打分函数个数为 n 个。

3.3 Transformer 在 CV 领域的应用与改进

将自注意力机制运用于 CV 领域是现今 Transformer 发展的一个重要方向。一张图片由一个个像素点组成，如果直接将 Transformer 应用在图片上就需要每个像素点跟其他所有像素点都算一下权重。那么一张分辨率为 $n * m$ 的图片就要计算 $(n * m) * (n * m)$ 次注意力机制。随着图片像素的增加，运算的复杂度就会呈现平方级增长。

为了应对这个问题，提出了不同的 Transformer 模型的改进方法。下面介绍几种典型的改进方法：

3.3.1 与卷积相结合的 CV Transformer

在这类模型中，先利用卷积层将图像进行降采样后，使其分辨率降低。然后再将其输入到 Transformer 中。

3.3.2 Axial-Attention (轴向注意力)

Axial-Attention 对一个像素进行自注意力机制的计算时，不是让它与其他所有像素做注意力机制，而是先只与同行像素做注意力机制后，再与同列像素做注意力机制。在这种改进方法下，单个像素的运算量从原来的 $O(n * m)$ 变成了现在的 $O(m + n)$ ，使得计算量大大降低。

由于第一步的轴向注意力操作能使每个像素点就蕴含了整行的信息，而第二步的轴向注意力操作能使得这些已经蕴含了整行信息的像素点之间进行同列信息的相互结合，于是研究人员认为这种串联的处理方法不仅使得单个像素点能结合所在行与列的信息，而且还能让单个像素点结合整个图象的全局信息。

轴向注意力机制的具体操作如图3.13。

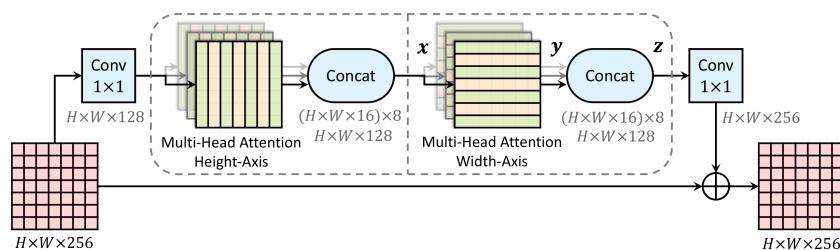
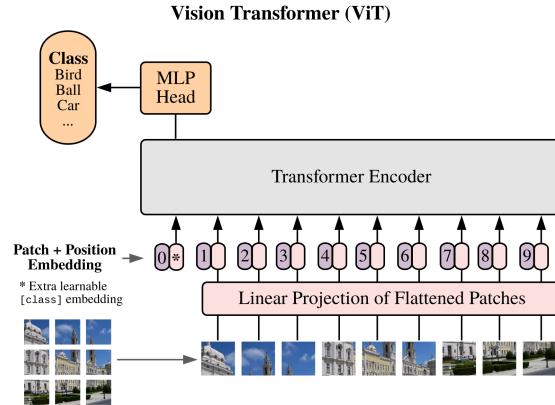


图 3.13 Axial-Attention 模型的结构^[3]

3.3.3 ViT 神经网络对注意力机制的实现

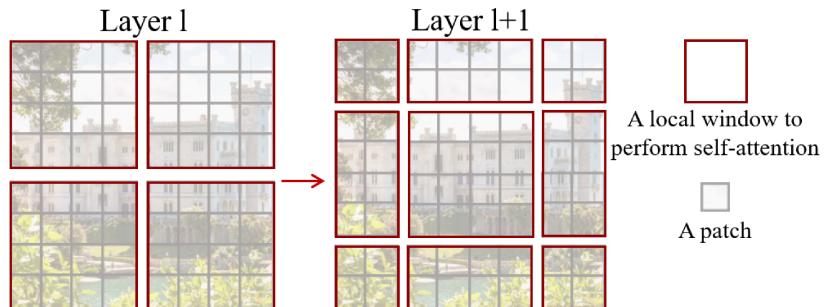
ViT 神经网络对传统注意力机制的改进方法是把先将图像进行分割，分割成一个个固定大小的 patch，将这个 patch 看成一个大像素（每个 patch 展平当作一个向量），然后在让所有这些“大像素”向量之间做自注意力机制。这时就等价于对一张更低像素图片做注意力机制，达到将计算量变少的目的。具体 ViT 神经网络的实现如图3.14所示。

图 3.14 ViT 模型的结构^[4]

3.3.4 Swin Transformer

与 ViT 的改进不同，Swin Transformer 在将图像进行分割成 patch 后，并不是在这些 patch 间运算注意力机制，而是对 patch 内的各像素间做注意力机制。这部分操作是在 W-MSA 内完成的（即 W-MSA (Window Attention)：分成一个个 patch，然后在 patch 内部做自注意力机制，做完后再拼接到一起）。虽然这样运算能因像素点减少而使得运算量降低，但由于 patch 与 patch 是相互独立，缺乏联系的，而导致单个像素无法很好地融合整个图像的信息。

于是 Swin Transformer 在 W-MSA 完成后又引入了 SW-MSA 操作，即采用滑动窗口后的再分割来规避上述缺陷。具体操作见图3.15。

图 3.15 Swin Transformer 中的 SW-MSA 操作^[5]

如图3.15所示，整个过程是首先由 W-MSA 做上图左边四个 patch 的自注意力；然后采用滑动窗口的再分割得到上图右边的 9 个 patch，再在这九个 patch 内部进行自注意力的运算。此时经过 SW-MSA 的操作后就使得单个像素能更好的融合整个图像的信息。

3.4 SwinTransformer 的优势

在上述四种改进方法中，Swin Transformer 模型在图像分类、目标检测、图像分割等常见的 CV 领域都有更好的实现效果。经过分析，Swin Transformer 模型相较于其他现有模型的优势可能存在如下几点：

- 1) Swin Transfomer 的 W-MSA 和 SW-MSA 设计使其计算复杂度为输入图像大小的线性计算复杂度，相较于其他模型显著降低。
- 2) Swin Transfomer 的 Patch Merging 层设计使得输入图像的分辨率随着层数的加深而不断减小，进而进一步降低整个模型的计算复杂度，使得更深层数神经网路模型的实现成为可能。而深层模型相较于浅层模型往往具有更好的效果。
- 3) 从每个 patch 的感知范围的角度，Swin Transformer 中的层次化构建方式类似于 CNN，在逐层缩小图片分辨率的同时，使得每个 patch 的感知范围扩大。而 ViT 等改进方法中每个 patch 的感知范围是固定的。

由于 Swin Transformer 模型在 CV 领域具有良好的表现，本项目采用 Swin Transformer 模型的改进思路来实现将低精度光声重建图像优化为高精度光声重建图像的模型。

4 训练数据的生成与预处理

4.1 数据集的介绍

本次项目采用的数据集为 Skin Cancer MNIST: HAM10000。该数据集由 10000 张来自不同人群的皮肤镜图像组成。

数据集中的几张皮肤癌图片举例如图4.1。

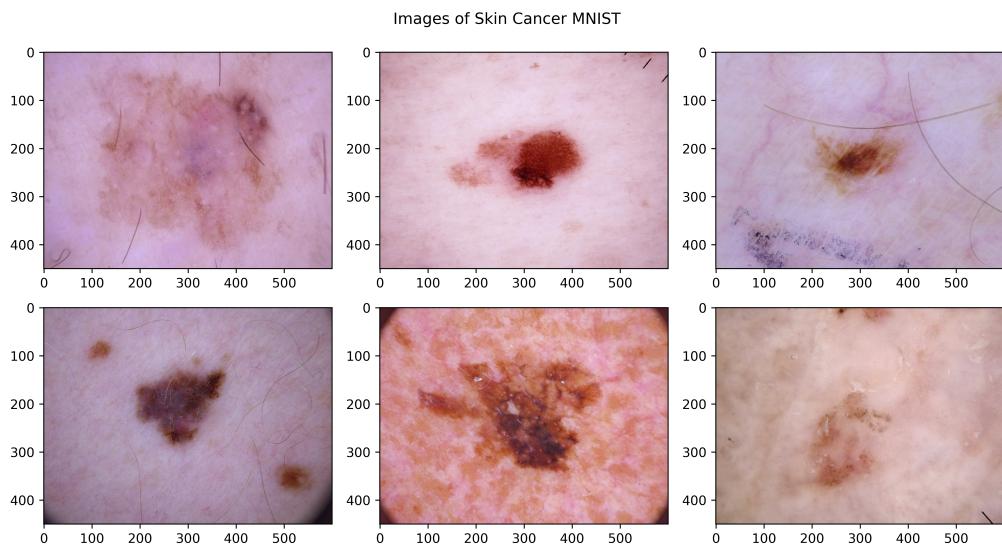


图 4.1 Skin Cancer MNIST 数据集示例

4.2 k-wave 的介绍与 k 空间伪谱法

k-Wave 是 MATLAB 的开源声学工具箱。该软件基于 k 空间伪谱法能实现复杂组织在真实介质中的声学仿真，并且还实现了多种对光声成像的重建方法。本项目采用该库实现对皮肤癌图像的光声成像仿真与重建。

4.3 训练数据的生成与预处理

下面介绍使用 Skin Cancer MNIST 数据集与 k-Wave 中的光声成像仿真与重建算法生成训练集与测试集的具体操作流程。

4.3.1 在仿真前对 Skin Cancer MNIST 中的皮肤癌图像进行预处理

- 1) 读取图片并转为灰度图。

- 2) 将图像归一化。由于用 k-Wave 进行模拟时外围填充的像素一定得是 0，所以要在数据归一化的时候把原图里正常皮肤的部分对应到 0。因此，取原图边界上的一些像素的平均值作为正常皮肤对应的像素值大小。然后在图像上将这个平均值映射到 0；再在图像外围补上 0 像素点。
- 3) 将图像降低分辨率（分辨率降低为原来的二分之一）。降采样的目的是使图像的分辨率变低，有效控制仿真的时间。
- 4) 将图像的外围填充一圈 0，目的是确保在光声成像模拟时，皮肤病灶完全位于圆形传感器阵列之中，使得光声成像模拟结果更加精确。

下面选取 Skin Cancer MNIST 数据集中的一副皮肤癌图像对上述过程进行演示，如图4.2所示。

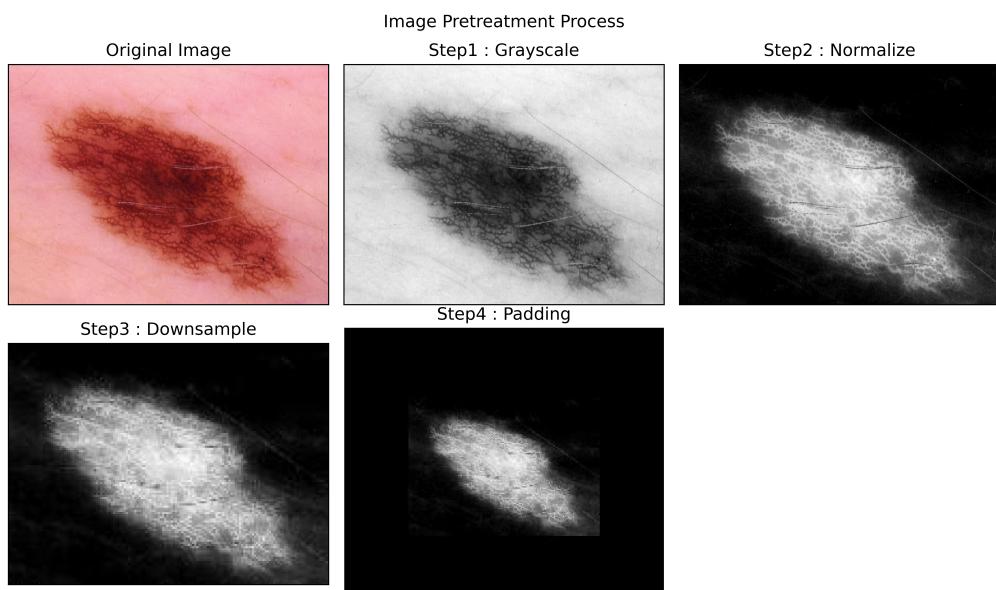


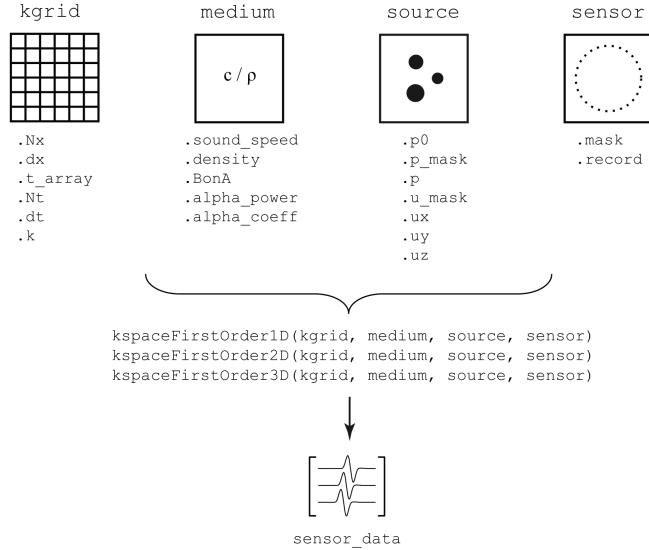
图 4.2 预处理操作流程

4.3.2 利用 k-Wave 对预处理后的图像进行光声成像仿真

首先需要初始化 k-Wave 进行光声成像仿真各参数。在使用 k-Wave 进行光声成像模拟前，要设置的参数如图4.3所示。

即创建计算网格、定义介质属性、定义初始压力、定义传感器掩模。

- 1) 其中由于查资料得知：人体软组织声速接近 1540m/s ； 2006 年全国男人人体密度 $=1.0913 - 0.0016 \cdot (10.8 + 15.8) = 1.0487 \cdot 10^3\text{kg}/(\text{m}^3)$ 、女人人体密度 $=1.0897 - 0.00133 \cdot (17.5 + 17.5) = 1.0431 \cdot 10^3\text{kg}/(\text{m}^3)$ 。于是我设置介质声速为人体软组织声速，将男人人体密度与女人人体密度的平均值设置为介质密度。

图 4.3 k-Wave 仿真的设置参数^[6]

- 2) 将预处理好的图片导入模型中作为初始压力分布。
- 3) 定义具有 50 个传感器元件的中心圆的笛卡尔传感器掩模

在设置完如上参数后，利用 MatLab 运行 `kwave` 仿真。得出的输出数据 `senser data` 是形状为 $[num\ sensor, time\ step]$ 的矩阵，记录了各传感器在各时间步长所接收到的压力数据。将其运用 Matlab 中的 `images` 函数做出图4.4。图片每一行代表着一个传感器在仿真过程中所接收到的压力的大小变化。

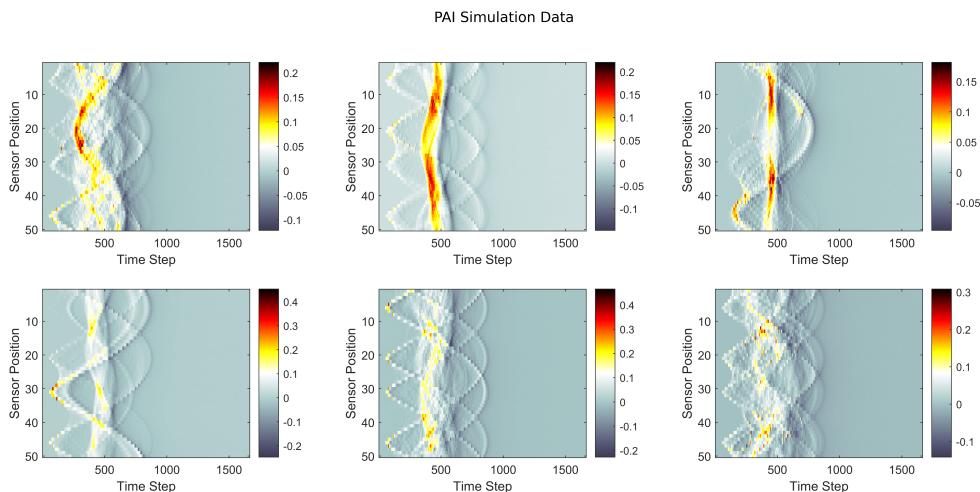


图 4.4 光声传感器接收到的压力数据示例

4.3.3 进行光声成像的重建得到训练数据

在进行光声成像重建前，同样需要初始化 k-Wave 的各项参数，并且确保其与仿真时保持一致；然后才能使用 k-Wave 中的相关函数进行光声成像的重建。重建

图像如图所示4.5。其中左侧为原图像，右侧为相应的重建图像。

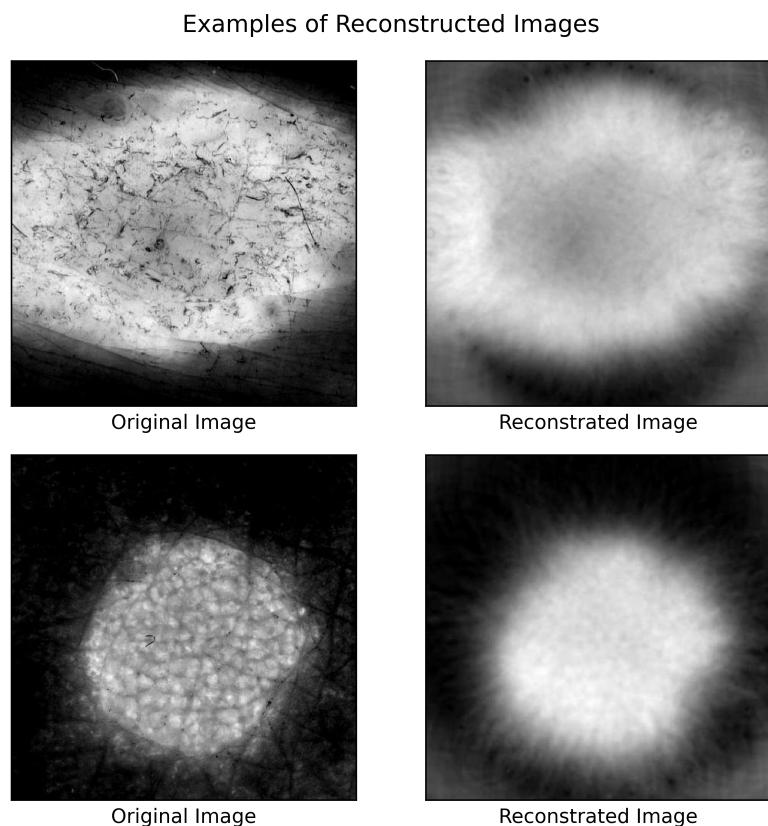


图 4.5 重建图像示例

4.3.4 将得到的数据集划分成训练集及测试集

经过上述光声成像仿真与重建，最终共得到 6000 张重建图像。将其中的 5000 张重建图像及其原图像划分为训练集，将其余的 1000 张重建图像及其原图像划分为测试集。

5 Transformer 模型的搭建与训练

5.1 神经网络的结构

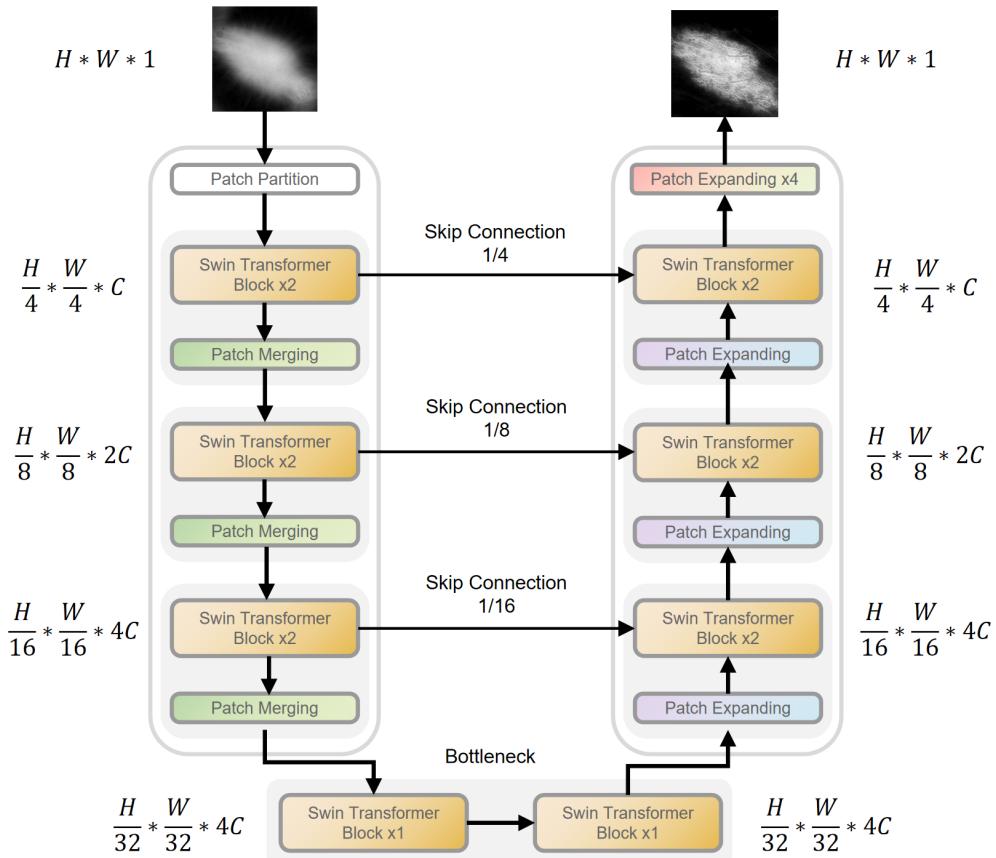


图 5.1 神经网络的结构

如图5.1，整个神经网络由左半部分的 Encoder 和右半部分的 Decoder 两部分组成。其中光声成像重建的灰度图像经过 Encoder，分辨率逐层递减，通道数逐层增加；而后经过 Decoder，分辨率逐层增加，通道数逐层递减，最后恢复到输入图像的通道数与分辨率；其中，Encoder 和 Decoder 相同分辨率的 tensor 使用 concat 操作进行 Skip Connection。

下面将详细阐述神经网络模型各层的实现方式与实现目的。

5.2 Encoder 和 Decoder 结构与 Long Skip Connection

由于在实验初期，浅层神经网络对光声重建图像的优化效果并不理想，于是选择加深网络层数以达到较好的实现效果。但是深层神经网络存在两大问题：

- 1) 加深模型的层数导致的训练模型时训练速度较慢、应用模型时推理速度较慢等问题，从而使得训练过程与部署过程面临较大困难。
- 2) 随着隐藏层数量的增多，深层神经网络模型无法很好的结合浅层与深层的信息。

为此，本项目拟采用 Encoder 和 Decoder 相结合的 U 型神经网络结构与深层和浅层神经网络的 Skip Connection 来解决以上两个问题。

5.2.1 Encoder 和 Decoder 结构

观察到 tensor 的分辨率是影响神经网络模型推理速度的主要原因之一，因此本项目采用在 Transformer Block 之间使用降采样层的方法来降低中间过程 tensor 的分辨率。然后再通过上采样恢复成原图像的分辨率，最终输出优化图片。这种结构称为 Encoder-Decoder 架构。

这种 Encoder-Decoder 架构使得中间隐藏层处理的 tensor 的分辨率显著降低，从而达到加快模型推理速度的目的。

5.2.2 Long Skip Connection

本模型拟采用将 Encoder 和 Decoder 相同分辨率的 tensor 进行 Skip Connection，从而达到保留浅层信息与结合浅层和深层信息的目的，提高对重建图像的优化效果。

5.3 Patch Embedding、Patch Merging 层、Patch Expanding 层和 Patch Expandx4 层

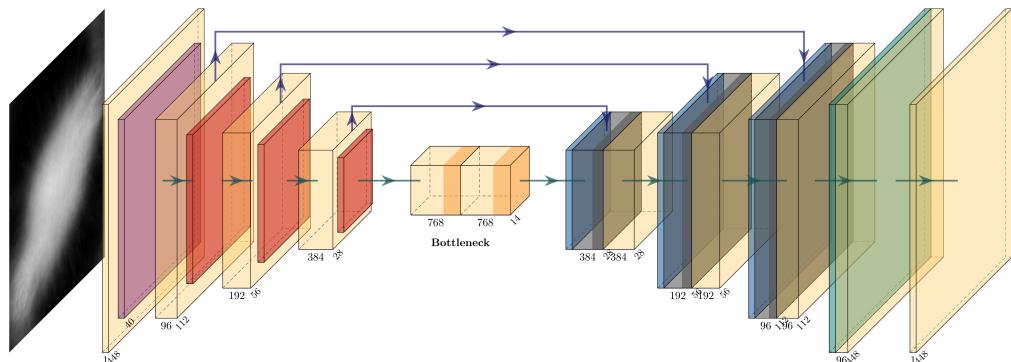


图 5.2 tensor 在神经网络内的形状变化

图5.2黄色层为输入、输出图片或 Transformer block，紫色层为 patch embedding 层，红色层为 Patch Merging 层，蓝色层为 Patch Expanding 层，绿色为 Patch Expandx4 层。

从 Transformer 的推导公式中，我们知道，Transformer 层不改变输入 tensor 的形状。于是，tensor 在神经网络中的 shape 改变来源于 Patch Embeding 层、Patch Merging 层、Patch Expanding 层和 Patch Expandx4 层。图5.2展示了一张图片在神经网络中的形状变化。

- 1) 光声成像重建图片是灰度图片，形状为 $(H_0, W_0, 1) = (448, 448, 1)$ ；
- 2) 在经过 patch embedding 层（图5.2紫色层）后， $H(\text{Height})$ 、 $W(\text{Width})$ 变为原来的四分之一， $C(\text{Channel})$ 变为 96，即形状变为 $(H, W, C) = (112, 112, 96)$ ；
- 3) 将 patch embedding 的输出作为 Transformer block 的输入，然后再经过 Patch Merging 层(图5.2红色层)， $H(\text{Height})$ 、 $W(\text{Width})$ 变为原来的二分之一， $C(\text{Channel})$ 变为原来的二倍。即：

表 5.1 Patch Merging 的形状变化

	Patch Merging 1	Patch Merging 2	Patch Merging 3
Input Shape	(H, W, C) $=(112, 112, 96)$	$(H/2, W/2, 2*C)$ $=(56, 56, 192)$	$(H/4, W/4, 4*C)$ $=(28, 28, 384)$
Output Shape	$(H/2, W/2, 2*C)$ $=(56, 56, 192)$	$(H/4, W/4, 4*C)$ $=(28, 28, 384)$	$(H/8, W/8, 8*C)$ $=(14, 14, 768)$

- 4) 由于 Bottleneck 是两个 Transformer block 串联而成，所以不会改变 tensor 的形状。
- 5) Bottleneck 的输出经过 Patch Expanding 层（图5.2蓝色层）， $H(\text{Height})$ 、 $W(\text{Width})$ 变为原来的二倍， $C(\text{Channel})$ 变为原来的二分之一。即：

表 5.2 Patch Expanding 的形状变化

	Patch Expanding 1	Patch Expanding 2	Patch Expanding 3
Input Shape	$(H/8, W/8, 8*C)$ $=(14, 14, 768)$	$(H/4, W/4, 4*C)$ $=(28, 28, 384)$	$(H/2, W/2, 2*C)$ $=(56, 56, 192)$
Output Shape	$(H/4, W/4, 4*C)$ $=(28, 28, 384)$	$(H/2, W/2, 2*C)$ $=(56, 56, 192)$	(H, W, C) $=(112, 112, 96)$

- 6) 输出 tensor 经过 Patch Expandx4 层（图5.2绿色层）， $H(\text{Height})$ 、 $W(\text{Width})$ 变为原来的四倍， $C(\text{Channel})$ 数不变，即形状由 $(H, W, C) = (112, 112, 96)$ 变为 $(448, 448, 96)$ ；

7) 最后通过一层卷积层，将通道 (Channel) 数恢复为 1，即形状由 (448,448,96) 变为 (448,448,1) 作为神经网络的输出

下文详细阐述 Patch Embedding 层、Patch Merging 层、Patch Expanding 层和 Patch Expandx4 层的实现原理：

5.3.1 Patch Embedding 层

Patch Embedding 层由一个卷积核大小为 4*4，卷积核个数为 96，步长为 4 的卷积层组成。它能将输入的灰度图片的 H(Hight)、W(Width) 变为原来的四分之一，C(Channel) 变为 96。

5.3.2 Patch Merging 层

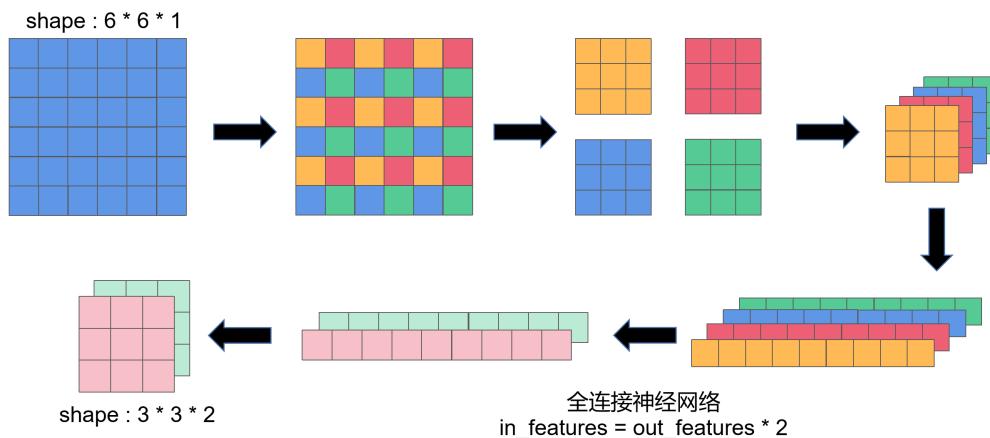


图 5.3 Patch Merging 的实现原理

Patch Merging 层的原理如图5.3所示，它能将输入 tensor 的 H(Hight)、W(Width) 变为原来的二分之一，C(Channel) 变为原来的二倍。

5.3.3 Patch Expanding 层

对于形状为 [B, H, W, C] 的输入 tensor，Patch Expanding 采用的是将通道数 C 翻倍的线性层，将 tensor 的形状改变为 [B, H, W, 2*C]。然后通过使用 einops 库中的 rearrange 函数将输入 tensor 的分辨率 H(Hight)、W(Width) 扩展到原分辨率的 2 倍，并将通道数 C(Channel) 减小到输入维数的四分之一。即：

$$\begin{aligned}
[B, H, W, 2 * C] &= [B, H, W, (p1, p2, c)] \\
&\rightarrow [B, (H, p1), (W, p2), c] \\
&\rightarrow [B, H * p1, W * p2, c] \\
&= [B, 2 * H, 2 * W, \frac{1}{4}C]
\end{aligned} \tag{5.1}$$

where $p1 = 2, p2 = 2, c = \frac{1}{4}C$

5.3.4 Patch Expandingx4 层

与 Patch Expanding 层所做的操作类似，不同的是 Patch Expandingx4 层采用的是将通道数 C 增加为 $16*C$ 的线性层，将 tensor 的形状改变为 $[B, H, W, 16*C]$ 。

然后使用 einops 库中的 rearrange 函数进行如下操作：

$$\begin{aligned}
[B, H, W, 16 * C] &= [B, H, W, (p1, p2, c)] \\
&\rightarrow [B, (H, p1), (W, p2), c] \\
&\rightarrow [B, H * p1, W * p2, c] \\
&= [B, 4 * H, 4 * W, C]
\end{aligned} \tag{5.2}$$

where $p1 = 4, p2 = 4, c = C$

从上面的 tensor 的形状变化过程可以看出，Patch Expanding 层是将 C 拆分为 $C=2*2*(C/4)$ ，而 Patch Expandingx4 层是将 C 拆分为 $4*4*(C/16)$ ，然后使用 rearrange 方法。

5.4 Swin Transformer Block

Swin Transformer Block 的内部结构如图5.4所示。

Swin Transformer Block 中 W-MASA 与 SW-MSA 总是成对出现，先将 input tensor 输入到 W-MSA，经过一个全连接层后，再输入到 SW-MSA，最后再经过一个全连接层后输出。在上面各层之间存在 Layer Normation 层，且存在残差连接 (Short Skip Connection)。

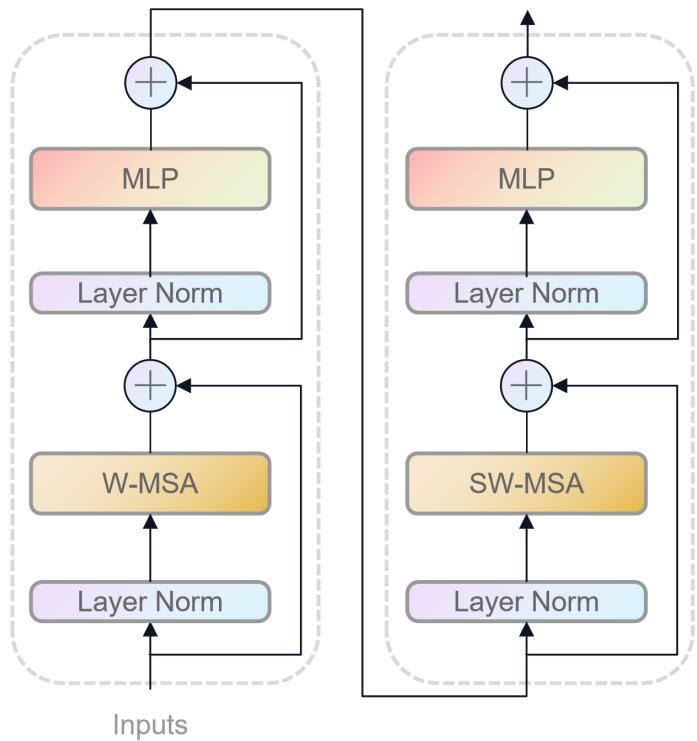


图 5.4 Swin Transformer Block 的结构

5.4.1 Short Skip Connection

在实验的过程中，深层神经网络往往会面临网络退化的问题。一般认为若整个网络中的部分子网络是当前最优的网络，那么更深的神经网络的其它部分只需要拟合为恒等映射 (Identity Mapping) 就可以取得与最优的网络一致的结果，因此深层网络应比浅层网络更优。但事实是由于非线性激活层的存在，深层网络难以拟合为恒等映射，从而导致网络发生退化。

相较之下，神经网络的子网络拟合为 0 值映射是较为容易的，因此残差单元以跳层连接的形式将单元输入直接与单元输出加在一起，神经网络拟合恒等映射等价于残差神经网络拟合 0 映射。因此能较好地解决深度网络发生退化的问题。

其次，与 Long Skip Connection 相对应，ResNet 残差连接可以看成是一种 Short Skip Connection。它同样能使得浅层信息得到更好的保留，且使得深层信息与浅层信息能得到更好的结合。

5.5 模型训练结果

在选取 MSE 作为损失函数后，经过 200 个 epoch 的训练，其训练集的 Loss 与测试集的 Loss 变化如图5.5所示。

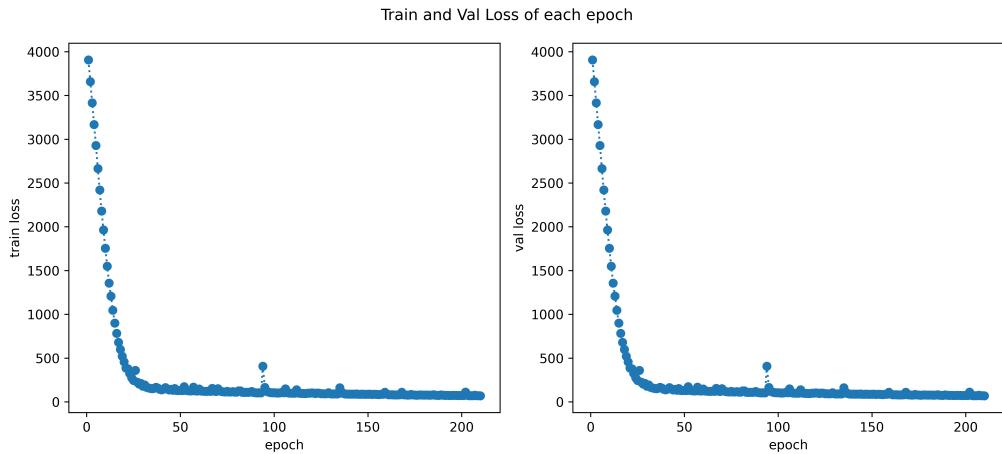


图 5.5 loss 随 epoch 的变化曲线

从 Loss 的变化曲线可知，模型在第 1 个 epoch 到第 25 个 epoch，Train Loss 和 Val Loss 都大幅度下降，最终到第 200 个 epoch 成功收敛。

将验证集 (Validation Set) 中抽取的两张由 50 个光学传感器收集数据反演所得的重建图像输入到已训练好的模型中得到预测图像。将原图像、重建图像和预测图象作图进行对比，如图5.6所示。

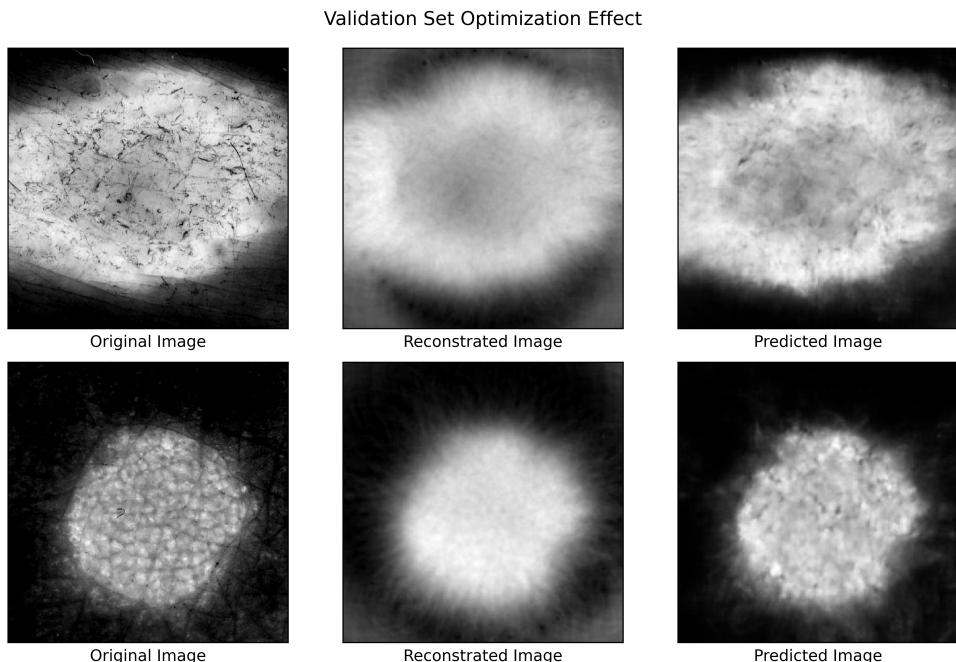


图 5.6 重建图像优化效果示例

图5.6左侧图像为选定图像的原图像，中间图像为选定图像的光声重建图像，右侧图像为将重建图像输入模型所得的预测图像。从直观上可以看出，相比于光声重建图像，模型的预测图像具有更高的分辨率与对比度，且在整体和细节上更接近原图像。下面对模型的优化效果进行定量分析。

6 模型的分析与评估

将验证集中的 1000 张重建图像输入已训练的模型，得到重建图像的预测图像后，计算预测图象与原图像的 MSE、PSNR、SSIM 等评价指标，并对重建图像与原图像之间的 MSE、PSNR、SSIM 等值，对模型的优化效果进行分析与评估。

6.1 MSE

6.1.1 MSE 简介

MSE 是衡量两张图片的相似程度的一种常见方法。两个图片之间的 MSE 即求两张图片各个相对应的像素点的平方差之和的均值。具体公式如下：

$$MSE(I, K) = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (6.1)$$

6.1.2 使用 MSE 衡量模型效果

计算验证集中的 1000 张重建图像与原图像的 MSE 值和预测图像与原图像的 MSE 值，结果如图6.1所示。

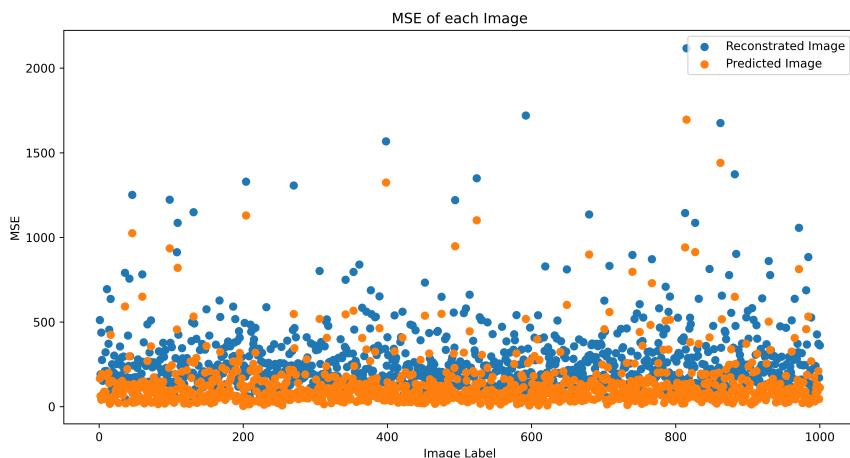


图 6.1 测试集图像及其预测图象与原图像的 MSE

在图6.1中，X 轴为各图像的编号，Y 轴为对应的 MSE 值。从图6.1可以看出，预测图象的 MSE 值在图中的主要分布区域为 0 到 250，而重建图像的 MSE 值的主要分布区域为 100 到 500。且在 MSE 值超过 500 的区域，重建图像的数量显著

多于预测图像。进一步，计算这 1000 张重建图像与预测图像的 MSE 值的平均值，计算结果如图6.2。

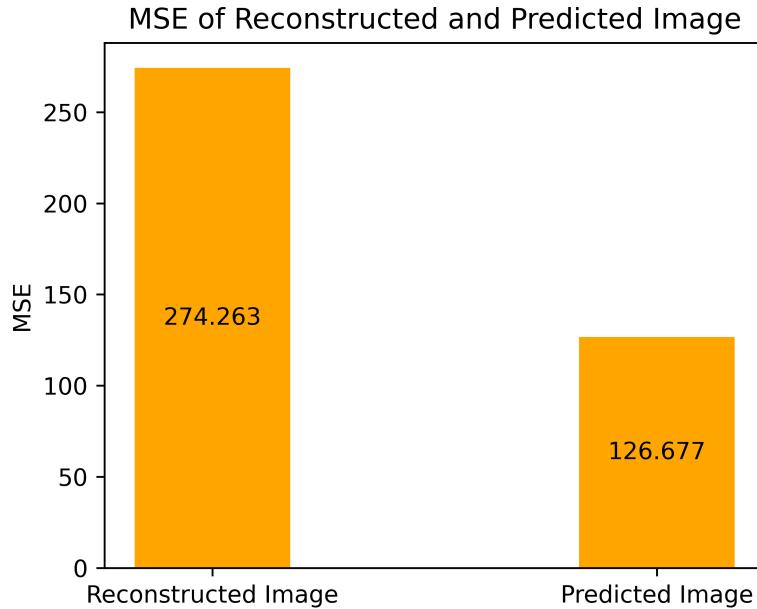


图 6.2 测试集图像及其预测图象与原图像的 MSE 的均值

从图6.2可以得出，在输入模型后所得优化图像，相比原来的重建图像，平均的 MSE 值由 274.263 下降为 126.677，这说明模型对原来的重建图像具有良好的优化效果。

6.2 PSNR

6.2.1 PSNR 简介

PSNR 即峰值信噪比，PSNR 的计算涉及到 MSE。它的公式如下：

$$PSNR(I, K) = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \quad (6.2)$$

对数内分母为均方误差 MSE，分子中的 MAX_I 为最大像素值，即若为 b 位图像，该值为 $2^b - 1$ 。当两张图片的 MSE 差异越小，则对数内的值越大，此时 PSNR 就越大。因此，PSNR 越大就表示图像相似程度越高。

6.2.2 利用 PSNR 衡量模型效果

计算验证集中的 1000 张重建图像与原图像的 PSNR 值和预测图像与原图像的 PSNR 值，结果如图6.3所示。

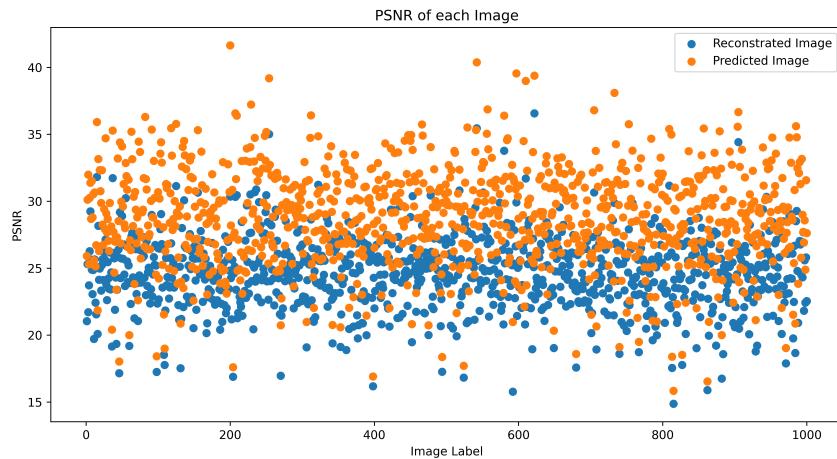


图 6.3 测试集图像及其预测图象与原图像的 PSNR

在图6.3中，X 轴为各图像的编号，Y 轴为对应的 PSNR 值。从图6.3可以看出，预测图象的 PSNR 值在图中的主要分布区域为上半区域，而重建图像的 PSNR 值主要分布在下半区域。计算这 1000 张重建图像与预测图像的 PSNR 值的平均值，计算结果如图6.4。

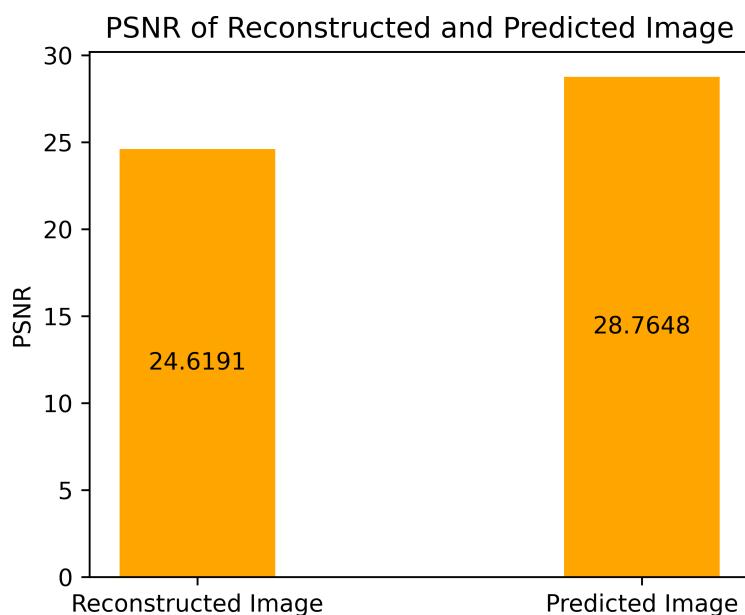


图 6.4 测试集图像及其预测图象与原图像的 PSNR 的均值

从图6.4可以得出，在输入模型后所得优化图像，相比原来的重建图像，平均

的 PSNR 值由 24.6191 上升为 28.7638。由上述 PSNR 公式可得，更高的 PSNR 代表着更高的图像相似程度，这说明模型的预测图象相比于重建图像与原图像具有更高的相似度，模型具有良好的优化效果。

6.3 SSIM

6.3.1 SSIM 简介

SSIM 即结构相似度，它是一种判断两张图片在结构上的差距的一项指标。

与传统的 MSE 不同，对于两张图片基于 MSE 的损失大小不足以表达人类视觉系统对这两张图片所感觉到的差距。比如两张只是亮度不同的图片，由于 MSE 是计算二者像素差的平方之和，因此所计算出的损失很大，但在人眼看来，这两张图片是十分相近的。人类视觉相关的研究普遍认为人类衡量两幅图的差距，更倾向于比较两图的结构相似性，而不是像 MSE 那样逐像素计算两图的差异。

SSIM 的衡量标准有两张图片的亮度相似度，对比度和结构相似度三个指标。

6.3.1.1 亮度相似度

设图像 X 所含的像素点个数为 N，各像素值为 x_i ，则定义其平均亮度为 X 中各像素的均值，即： $\mu_X = \frac{1}{N} \sum_{i=1}^N x_i$

定义衡量两幅图 X 和 Y 的亮度相似度的公式为：

$$l(X, Y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (6.3)$$

其中分母所在的常数是为了后续计算中避免出现分母为 0 的情况。

6.3.1.2 对比度

对比度定义为全体像素值的标准差，代表着图像明暗变化的剧烈程度。一张图像的标准差的计算公式为： $\sigma_X = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu_X)^2}{N - 1}}$ 。衡量两幅图 X 和 Y 的对比度的相似度的公式为：

$$c(X, Y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (6.4)$$

同样地，分母中的常数是为了后续计算中避免出现分母为 0 的情况。

6.3.1.3 结构相似度

由于在上面的推导中已考虑亮度和对比度，因此在研究结构相似度时，应该首先排除这两个指标的影响，即将图像进行归一后以排除均值和标准差的影响。定义归一化后的两个向量 $\frac{X-\mu_X}{\sigma_X}$ 和 $\frac{Y-\mu_Y}{\sigma_Y}$ 之间的结构相似度为：

$$\begin{aligned} s(X, Y) &= \left(\frac{1}{\sqrt{N-1}} \frac{X - \mu_X}{\sigma_X} \right) \cdot \left(\frac{1}{\sqrt{N-1}} \frac{Y - \mu_Y}{\sigma_Y} \right) \\ &= \frac{1}{\sigma_X \sigma_Y} \left(\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \right) \end{aligned} \quad (6.5)$$

上式中第一行“.”表示向量内积；第二行括号内的部分为协方差公式： $\sigma_{XY} = \frac{\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y)}{N-1}$ 。因此得到结构相似度的表达式为：

$$s(X, Y) = \frac{\sigma_{XY} + C_3}{\sigma_X \sigma_Y + C_3} \quad (6.6)$$

同理，为了防止后续计算出现分母为 0，分子分母同时加 C_3 。

6.3.1.4 得出 SSIM 表达式

由上面的推导得出的表达式 (6.3)、(6.4) 和 (6.6)，三个标准相乘作为 SSIM 相似度公式，即：

$$SSIM(x, y) = l(x, y)^\alpha \cdot c(x, y)^\beta \cdot s(x, y)^\gamma \quad (6.7)$$

公式中的 α, β, γ 代表上述三个特征在 SSIM 指标中的占比，当三者都为 1 时，SSIM 的表达式为：

$$SSIM(X, Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)} \quad (6.8)$$

从以上的推导公式可知，两张图像的 SSIM 值是一个介于 0 和 1 之间的常数，且越接近 1 代表两张图像越是接近。从这个角度出发，接下来我们将尝试将 SSIM 作为指标来分析所得模型做出的预测图像的准确程度。

6.3.2 使用 SSIM 衡量模型效果

计算验证集中的 1000 张重建图像与原图像的 SSIM 值和预测图像与原图像的 SSIM 值，结果如图6.5所示。

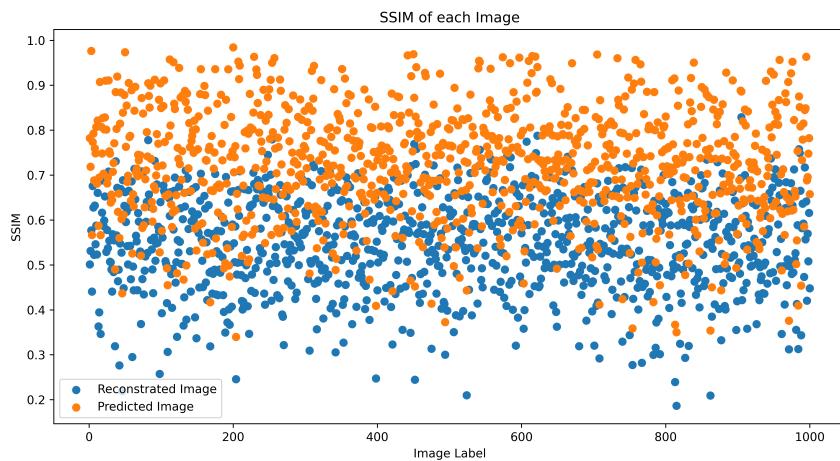


图 6.5 测试集图像及其预测图象与原图像的 SSIM

在图6.5中，X 轴为各图像的编号，Y 轴为 SSIM 值。从图6.5可以看出，预测图象的 SSIM 值在图中大多分布在重建图象的 SSIM 值之上。而且计算这 1000 张重建图象与预测图象的 SSIM 值的平均值，计算结果如图6.6。

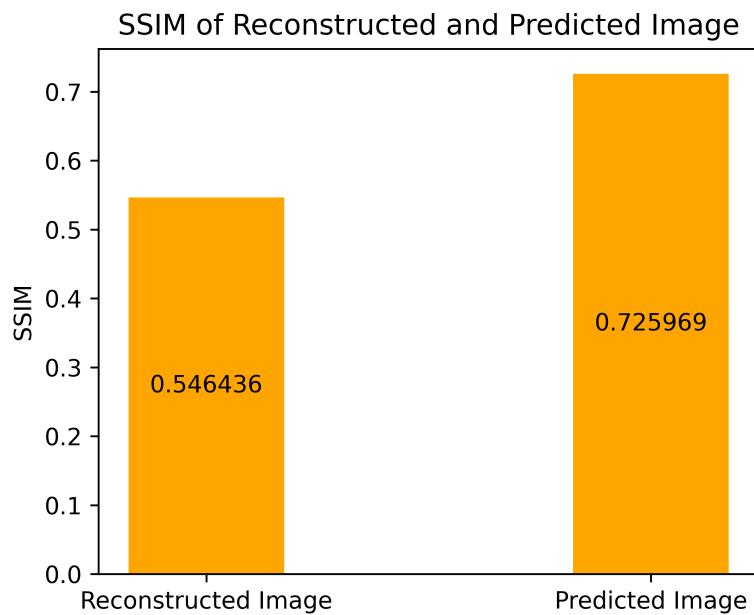


图 6.6 测试集图像及其预测图象与原图像的 SSIM 的均值

从图6.6可以得出，在输入模型后所得优化图像，相比原来的重建图像，平均的 SSIM 值由 0.55 上升为 0.73。可以发现预测图像对应的 SSIM 值比较接近 1，这

说明从 SSIM 的角度上看，原图像与预测出的图像在亮度、对比度与结构相似度上较为接近，所以该模型的预测还是较为准确的。

6.4 余弦相似度

6.4.1 余弦相似度简介

余弦相似度是衡量两个向量相似程度的一项指标，即使用两个向量之间的夹角的余弦值作为衡量两者相似程度的标准。余弦值越接近于 1，表示夹角越接近于 0，表明两者越相似；反之，余弦值越接近于 0，表示夹角约接近于 $\frac{\pi}{2}$ ，表明两者越不相似。对于两个向量 $X = x_i$ 和 $Y = y_i$ 而言，其余弦相似度的计算公式为：

$$\text{Cosine Similarity}(X, Y) = \frac{\sum_{i=1}^n (x_i \times y_i)}{\sqrt{\sum_{i=1}^n (x_i)^2} \times \sqrt{\sum_{i=1}^n (y_i)^2}} \quad (6.9)$$

我们可以将两张图片表示为向量后，使用余弦相似度来衡量两张图片的相似程度。

6.4.2 使用余弦相似度衡量模型效果

对验证集的 1000 张图像，我们计算了原图像与重建图像或预测图像间的余弦相似度，将其结果画成直方图如图6.7所示。



图 6.7 测试集图像及其预测图象与原图像的余弦相似度的均值

从上述结果可以看出，预测图像与原图像的余弦相似度大于重建图像与原图

像间的余弦相似度，这说明在余弦相似度的指标下，预测图象对比重建图像具有更高的相似程度。

6.5 哈希相似度

6.5.1 哈希相似度简介

使用哈希值衡量两张图片的相似度分三步：第一步，计算两张图片各自的哈希值；第二步，计算两个哈希值之间的汉明距离 (Hamming Distance)；第三步，将汉明距离转化为两张图片的相似度。

6.5.1.1 计算哈希值

哈希值有三种定义方法，它们分别是均值哈希值 (aHash)，差值哈希值 (dHash)，感知哈希值 (pHash)。它们的计算方法如下：

1) 均值哈希值

- 将图像缩放成如 8x8 像素大小的小图像并转为灰度图，计算小图像像素的均值。
- 将小图像中比均值高的像素转换为 1，比均值小的像素转换为 0，生成哈希码。

2) 差值哈希值

- 将图像缩放成 8x9 像素大小的小图像并转为灰度图，使得每行 9 个像素之间存在 8 个不同的差异值。
- 对于小图像中的相邻像素，若左像素比其右像素的值大，则左像素转换为 1；否则转换为 0，最终得到一个 8x8 的哈希矩阵。

3) 感知哈希值

- 将图像缩放成如 8x8 像素大小的小图像并转为灰度图
- 对小图像进行 32x32 的 DCT (离散余弦变换) 得到 32x32 的 DCT 系数矩阵，并取左上角的 8x8 的矩阵，计算该矩阵的平均值。
- 将 8x8 矩阵中比均值高的元素转换为 1，比均值小的元素转换为 0，生成哈希码。

6.5.1.2 计算汉明距离

汉明距离即两个哈希值之间不相等的元素个数。

6.5.1.3 将汉明距离转化为两张图片的相似度

转化为相似度的公式为：

$$\text{similarity} = 1 - \frac{\text{dist}}{\text{pixelnum}} \quad (6.10)$$

其中 dist 为汉明距离, pixelnum 为小图像的像素个数(若小图像为 8x8, 则 pixelnum 等于 64)

6.5.2 使用哈希相似度衡量模型效果

对验证集的 1000 张图像, 我们计算了原图像与重建图像或预测图像间的三种哈希相似度, 将其结果画成直方图如图6.8所示。

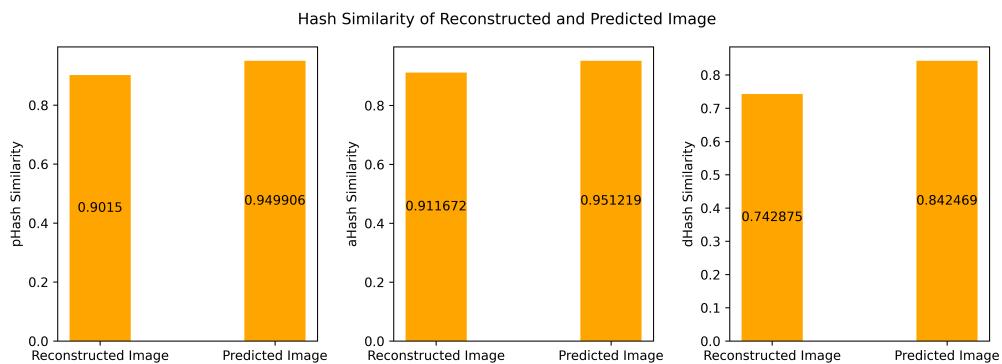


图 6.8 测试集图像及其预测图象与原图像的三种哈希相似度的均值

从图6.8可得, 无论是均值哈希值, 差值哈希值还是感知哈希值, 预测图像的相似度均大于重建图像的相似度。

6.6 直方图相似度

6.6.1 直方图简介

图像的直方图即统计全图各像素值所占的像素个数的直方图。对于 256 值的灰度图而言, 就是统计 0~255 这 256 个值的像素个数所作而成的直方图。例如皮肤癌数据集图像的直方图见图6.9。

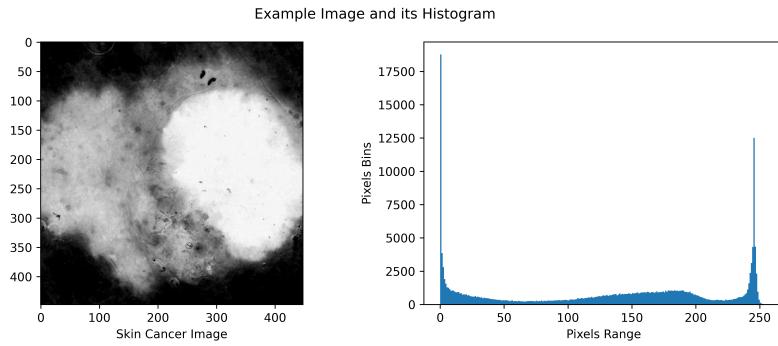


图 6.9 皮肤癌图片的直方图

假设 $hist_1$ 和 $hist_2$ 是存放着 0~255 像素值所占像素个数的数组，则两个图像间的直方图相似度的计算公式为：

$$Degree(hist_1, hist_2) = \sum_i \frac{1 - |hist_1[i] - hist_2[i]|}{\max(hist_1[i], hist_2[i])} \quad (6.11)$$

6.6.2 使用直方图相似度衡量模型效果

对验证集的 1000 张图像，我们计算了原图像与重建图像或预测图像间的直方图相似度，其结果如图6.10所示。

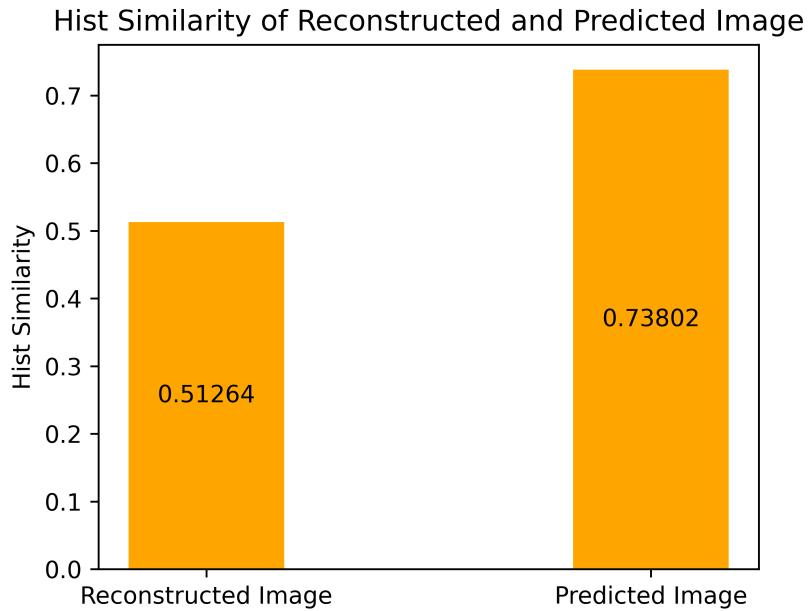


图 6.10 测试集图像及其预测图象与原图像的直方图相似度的均值

由图6.10可知，重建图像与原图像之间的直方图相似度低于预测图像与原图像之间的直方图相似度。这说明预测图像相比于重建图像，在各像素值的统计学标准上更接近于原图像。

6.7 归一化互信息

6.7.1 归一化互信息简介

在信息论中，归一化互信息 (NMI) 是衡量两个随机变量之间的依赖程度大小的一项指标。NMI 的值越大代表两个随机变量间的依赖程度越高。若将其运用于分析两个图片间的相似程度，则 NMI 的值越大代表两张图片具有越高的相似性。

对于两幅图像 A 与 B，其归一化互信息的计算分成三步：

- 1) 分别计算图像 A,B 的信息熵，其计算公式如下。

$$H(A) = - \sum_a P_A(a) \log_2 P_A(a) \quad (6.12)$$

$$H(B) = - \sum_b P_B(b) \log_2 P_B(b) \quad (6.13)$$

其中，公式中 a 为图像 A 中的各像素值（若 A 为 256 值的灰度图像，则 a 的取值范围为 [0,255]）。像素值 a 的概率值 $P_A(a)$ 为对图像进行灰度直方图统计后，像素值 a 所占像素点个数占全图像素点个数的比例值。对于公式 (6.13) 同理。

- 2) 使用 A 与 B 各自的信息熵计算二者的联合信息熵。

$$H(A, B) = - \sum_{a,b} P_{AB}(a, b) \log_2 P_{AB}(a, b) \quad (6.14)$$

其中，公式中的 a,b 为图像 A、B 的各像素值（若 A、B 为 256 值的灰度图像，则 a、b 的取值范围均为 [0,255]）；联合概率密度 $P_{AB}(a, b)$ 指的是两图在相同坐标系下，满足 A 的像素值为 a 且 B 的像素值为 b 的像素点，其个数占全图总像素点个数的比例值。

- 3) 计算归一化信息熵。将第一第二步计算的 A 与 B 的信息熵 $H(A), H(B)$ 及其联合信息熵 $H(A, B)$ 代入如下的公式中得到归一化信息熵。

$$NMI(A, B) = \frac{H(A) + H(B)}{H(A, B)} \quad (6.15)$$

6.7.2 使用互信息衡量模型效果

对验证集的 1000 张图像，我们计算了原图像与重建图像或预测图像间的归一化互信息，其结果如图6.11所示。

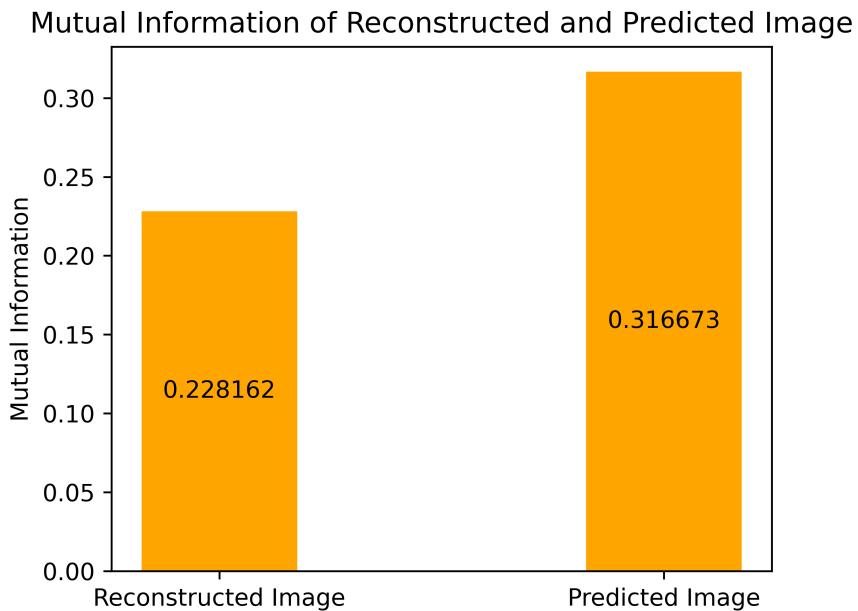


图 6.11 测试集图像及其预测图象与原图像的 NMI 的均值

从计算结果可知，预测图像与原图像之间的 NMI 值大于重建图像与原图像间的 NMI 值。其结果表明，预测图像比重建图像具有更高的相似性。

6.8 模型评估总结

对于验证集中的皮肤癌图片，我们计算出这 1000 张预测图像与原图像、重建图像与原图像的上述各评价指标的均值，记录如下表：

表 6.1 验证集的各项评价指标的均值

	MSE	PSNR	SSIM	Cosine
Reconstructed Image	274.263	24.6191	0.546436	0.934411
Predicted Image	126.677↓	28.7648↑	0.725969↑	0.967805↑
	Hash	Hist	NMI	
Reconstructed Image	pHash:0.9015 aHash:0.911672 dHash:0.742875	0.51264	0.228162	
Predicted Image	pHash:0.949906↑ aHash:0.951219↑ dHash:0.842469↑	0.73802↑	0.316673↑	

根据上述各指标的评估结果可知：训练出的神经网络模型的优化图像在逐像素比较的评估标准 (MSE、PSNR、Hist) 下均能取得相较重建图像更好的分辨率；且在结构相似度的评价指标 SSIM 下也具有更好的表现；在余弦相似度和哈希相似

度等指标下也具有更高的相似度；在信息论的评价标准 (NMI) 下，预测图像相较于重建图像丢失更少的信息，具有更高的还原度。

因此，可以得出结论：Transformer 模型在将低质量光声重建图像预测为高质量光声重建图像的优化任务上，具有较好的优化效果。将该模型应用于光声成像的后处理能显著改善采样信息丢失导致的成像精度低的问题，为低成本下获取高精度光声图像的研究提供了一个解决方法。

6.9 模型不足与发展方向

- 1) 由于项目实施的时间有限，数据集的图片数量和丰富程度还不够充实。在今后的优化中，可以适当增加数据集的来源，或通过在光声成像的仿真及重建中设置不同的传感器数量等方式来获取更丰富的数据集。
- 2) 由于训练次数有限，该模型在优化 50 个传感器下的重建图像上效果显著；而在优化 100、200、400 个传感器下的重建图像的效果并不是很好。这点可以通过增加由 100、200、400 个传感器得出的重建图像的数据集及增加训练 epoch 次数等方式解决。
- 3) 在医学的运用过程中，模型的推理成本也是一项很重要的指标。在今后的优化中，可以通过量化、剪枝^[7]、知识蒸馏^{[8][9]}、参数共享^[10]等方法减少内存消耗与提高模型运行速度。

参考文献

- [1] BRUNKER J, YAO J, LAUFER J, et al. Photoacoustic imaging using genetically encoded reporters: a review[J]. Journal of biomedical optics, 2017, 22(7): 070901-070901.
- [2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [3] WANG H, ZHU Y, GREEN B, et al. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV. Springer, 2020: 108-126.
- [4] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[A]. 2020.
- [5] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
- [6] TREEBY B E, COX B T. k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields[J]. Journal of biomedical optics, 2010, 15(2): 021314-021314.
- [7] RAO Y, ZHAO W, LIU B, et al. Dynamiccvit: Efficient vision transformers with dynamic token sparsification[J]. Advances in neural information processing systems, 2021, 34: 13937-13949.
- [8] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. PMLR, 2021: 10347-10357.
- [9] WU K, ZHANG J, PENG H, et al. Tinyvit: Fast pretraining distillation for small vision transformers[C]//Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI. Springer, 2022: 68-85.
- [10] ZHANG J, PENG H, WU K, et al. Minivit: Compressing vision transformers with weight multiplexing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12145-12154.

致谢

首先我要感谢论文的指导老师时聪教授。在课题立项初期，我对当今的学术热点了解不深，时聪老师以其对整个研究领域的深刻理解和敏锐的洞察力给了我针对性的指导，帮助我完成了课题的选择及研究方向的确定。在项目实践的过程中，由于匮乏独自完成课题研究所需的经验，我常常请教时聪老师，在解决一项项学术问题的过程中，我不断提高着分析和解决问题的能力与养成了一定的科学素养，这使我在项目研究后期能开始独立完成课题的研究工作。这段经历的收获不在于学习到了多少具体的专业知识，而在于培养了我的科研思维和坚定了我选择学术研究道路的决心。

其次我还要感谢我本科的老师们，在此我郑重附上他们的名字：叶小平教授，刘长剑教授，陈秀卿教授，易泰山教授，赵育林教授，李铎教授，刘海峰教授，邵国宽教授，杨燕教授，魏国栋教授。他们不仅教会了我相关的专业知识，还培养了我数学的思维方式。

最后，我还要感谢我的家人，如果没有他们在背后的付出和支持，我就无法取得现在的成果。

黄梓航

2023年5月9日