毕业设计(论文)报告

题 目: 基于被动声呐的非合作目标探测与识别技术

学 号: 71119138

姓 名: 王骏

学 院: 软件学院

专 业: 软件工程

指导教师: 姜龙玉

起止日期: 2023.1 2023.6

东南大学毕业（设计）论文独创性声明

本人声明所呈交的毕业（设计）论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名：_____　　日期：_____年_____月_____日

东南大学毕业（设计）论文使用授权声明

东南大学有权保留本人所送交毕业（设计）论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学教务处办理。

论文作者签名：_____　　导师签名：_____

日期：_____年_____月_____日　　日期：_____年_____月_____日

摘　要

水声目标的探测识别一直是困难且必要的任务。通过被动声呐采集目标发出的声音信号，从而推理出目标的身份，这项任务对于国防安全以及水下众多水下工作的展开有着重要意义。

以往的工作通过数学统计的方法对于少量的水声数据进行归纳建模，来尝试形成一个水声目标分类器，近几年通过机器学习等大数据相关算法来训练分类器。但它们都未能克服水声目标数据集小的问题，获得的模型在准确率以及泛化性上都不尽人意。

本文通过领域迁移方法，先通过双向匹配机制生成迁移学习所需的数据对，然后利用预训练的模型为目标域的数据生成聚类的初始中心。然后利用 K-means 算法来为数据对中的目标域数据进行聚类并生成相应伪标签。最后将伪标签与数据对中源域的标签进行一致性对比，从而筛去不适合迁移的数据对，净化数据集。最后通过一个三分支 Transformer 架构的网络来将源域与目标域的特征进行对齐。其中两个分支分别对源域和目标域应用自注意力机制，而中间的分支是一个由目标域向源域查询的交叉注意力机制，它可以有效消除误配数据对特征迁移的噪声影响。这三个分支直接权重共享，最终将预训练模型微调至适应目标域数据分布。

我们的算法在 ShipsEar 数据集进行预训练，并成功迁移到 Deepmind 数据集，具有 97.53% 的准确率。在所有的对照实验中取得了最好的结果。

关键词：水声目标，目标识别，迁移学习，交叉注意力机制

ABSTRACT

The detection and identification of hydroacoustic targets has always been a difficult and necessary task. The task of inferring the identity of a target by capturing the acoustic signals emitted by the target through passive sonar is of great importance for national defense security and the development of many underwater tasks.

Previous work has attempted to form a hydroacoustic target classifier by inductive modeling of a small amount of hydroacoustic data through mathematical statistics, and in recent years by training classifiers with big data related algorithms such as machine learning. However, they have failed to overcome the problem of small hydroacoustic target data sets, and the obtained models are not satisfactory in terms of accuracy and generalizability.

In this paper, through a domain adaptation approach, the data pairs required for transfer learning are first generated by a bidirectional matching mechanism, and then the pre-trained model to generate the initial centers of clusters for the data in the target domain. The K-means algorithm is then used to cluster the target domain data in the data pairs and generate the corresponding pseudo-labels. The pseudo-labels are compared with the labels of the source domains in the data pairs, so that the data pairs that are not suitable for transfer are dropped out and the data set is purified. Finally, a three-branch Transformer network to align the features of the source and target domains. Two of the branches apply a self-attentive mechanism to the source domain and target domain, while the middle branch is a cross-attention mechanism that queries from the target domain to the source domain, which can effectively eliminate the noise effect of mismatched data on feature transfer. These three branches share the weights directly and finally fine-tune the pre-trained model to fit the target domain data distribution. Our algorithm is pre-trained on the ShipsEar dataset and successfully transfered to the Deepmind dataset with an accuracy of 97.53%. The best result was achieved among all the controlled experiments.

KEY WORDS: Hydroacoustic Targets, Target Identification, Transfer Learning, Cross-attention Mechanism

目　　录

第一章 绪论

1.1 课题背景和意义

1906 年英国海军 Lewis Nixon 第一次发明了被动声呐用以监测冰山，后来这项技术被广泛应用于第一次世界大战。由于电磁波在水中衰减的速率非常快，无法作为侦测信号使用，故在探测水下物体时利用的是声波。声呐按照运作方式主要分为两种，一类是主动发出声波，然后利用回波，通过多普勒效应来探测周围环境的，这种被称为主动声纳。另一种则不会主动发出声波，只接受来自于周遭的各种音频讯号来判断与识别水下物体，与传统的水听器很类似。

水下非合作目标则是指不会主动与观测者发生信息交互的目标，比如敌舰，海洋生物等。如果对这些目标主动发出声波，则会暴露自身位置，或者惊扰对方，故只能被动地等待目标发出的声波，从而对目标进行推测。

相比于陆地和空中环境中的目标检测，水声目标检测面临着许多独特的挑战和难点。首先水下环境复杂多变，水下信道的传播路径不确定，会导致信号传输过程中发生衰减、多径效应、折射等现象，进一步影响声呐信号的接收和处理。另外水下环境中存在大量的噪声干扰，包括来自其它船只、海浪、海洋生物等的声波信号，这些噪声会掩盖目标信号，使得目标检测难度加大。更困难的是，我们依赖的水下声波的传播会受限于水的性质和海底地形等因素，声波在传播过程中会发生衰减、折射、反射等现象，这会对声波信号的强度、相位、频率等参数产生影响，从而影响目标检测的精度和准确性。另外水声数据的获取本身成本就很高，需要借助水下声呐、潜水器、遥控器等设备进行采集和处理，而这些设备的成本和复杂度较高，对水声目标检测的研究和应用造成了很大的限制。

鉴于以上原因，导致水声数据集非常少，并且很多数据集涉及军工信息，不公开，让对这个领域的研究更加困难。为了解决数据集少的问题，本文特地从迁移学习入手，成功构建一个依赖少量目标域数据，但鲁棒性很高的模型。虽然水声目标检测虽然困难重重，但其意义非凡。对于民生海底探测，全球性的海洋生态保护工作以及国防军事都具有重大意义。

水声目标检测能极大程度上推进海洋资源勘探、海底地形测绘、水下文物考古等事业。在海洋资源勘探方面，水声目标检测可以帮助寻找海底矿产资源，如石油、天然气、金属矿产等，从而为国家的能源开发和经济发展提供支持。在海底地形测绘方面，水声目标检

测可以帮助测量海底地形和深度，为海底通信、海底管线布置等提供重要的基础数据。在水下文物考古方面，水声目标检测可以帮助探测和识别水下文物，从而为历史文化的保护和研究提供支持。在海洋生态保护领域。海洋生态环境的保护对于人类的生存和发展都非常重要。水声目标检测可以帮助监测海洋生物和海洋环境的变化，如鱼群数量、种类、分布，水质、温度、盐度等，从而为海洋保护和管理提供重要的数据支持。同时，水声目标检测也可以帮助监测和保护珊瑚礁、海草床等海洋生态系统，从而促进生态平衡和可持续发展。最重要的是，水声被动目标检测能极大提高我国的制海权，在海上作战中，水声目标检测可以帮助识别敌方舰艇、潜艇等水下目标，从而提升海上作战的情报收集和反制能力。

1.2 国内外研究现状

基于被动声呐的水下目标识别主要通过声呐进行水下信号采集，由于水下信号复杂，噪音多，需要进行预处理，包括去噪，滤波，时频分析等，从处理后的信号中提取有效的特征，例如频率、能量、波形等，以区分不同的水下目标。最后利用机器学习或其他算法将特征向量与已知目标的特征向量进行比较，以将信号分类为不同的水下目标。

被动声呐主要采集到的是声波在传播时随时间变化的振幅值，可以用来确定声波的强度和持续时间等参数。2000 年初的方法一般基于单通道数据，欧世峰[1]等于两千年初通过实验证明了使用多个水听器获得多通道数据能更好的来识别目标。

对于采集到的数据，由于水下环境中存在各种背景噪声，如水流、海洋生物、水下船只等，这些噪声会干扰水声目标信号的识别和分析。因此，首先需要对采集到的信号进行去噪处理，通常采用滤波器、降噪算法等方法来实现。常见的滤波器有低通，高通滤波器，分别过滤掉信号中的高频和低频噪音。而对于降噪算法，有经典的中值滤波，均值滤波，最小均方差滤波等；亦有基于小波变换的降噪算法，如章新华等[2]在对舰船辐射噪声进行小波变换、提取目标特征方面做了许多工作。利用小波变换对信号进行分解，然后通过阈值处理来实现降噪。典型的算法有基于硬阈值和软阈值的小波降噪算法等。另外还有利用倒谱系数来进行抗噪处理，如梅尔标度频率倒谱系数 (MFCC), 线性预测倒谱系数 (LPCC), 以及 Wang 等在 2019 年提出的改进的抗噪声功率归一化倒谱系数 (ia-PNCC[3])，它通过归一化 Gama 滤波来对单通道信号进行降噪处理，Wang 等[4]则将 Gammatone 频率倒谱系数 (GFCC) 和改进的经验模态分解 (MEMD) 联合起来提取多维特征。除此之外，Ke[5]等于

2018 年提到了基于共振的稀疏信号分解来进行数据的预处理。

对于预处理后的信号，经过特征提取获得有效信息是做出准确识别的关键。如 Kummert[6]提出的利用模糊算法来分析 DEMON 频谱来自动提取螺旋桨的轴速和叶片数量特征。近年来随着深度学习大范围的应用，亦有将深度卷积网络应用于水声信号的特征中来的例子，Hu[7]等于 2018 年发布了相关的论文。

在完成特征提取后，就是最关键的分类识别算法了。传统的基于统计分类的方法，依赖于已有标记好的大量数据进行分析和基于距离变量的模式匹配，结果是特征向量被判定为各个参考模式的一组概率。由于模板固定，泛化性与准确率都很差，但速度优秀。而更为流行的是机器学习方法，如支持向量机等模式识别算法。

作为机器学习的一个分支，深度学习今年来已经成为主流的分类识别算法，大部分文章采用简单的单层卷积网络，或者深层卷积网络来对特征进行分类识别，最后通过全连接层来获得最终的分类结果。这其中 Yang 等[8]提出了通过一组多尺度深度滤波器子网络对船舶辐射噪声的复杂频率分量进行分解和建模来分解声音，并将卷积网络加深来模拟的听觉系统以提高分类的准确率。另外对于卷积层的改进有从最后的全连接层入手，Hu[7]等等提出使用极限学习机 (ELM) 来完全取代最后的全连接层来加快训练前馈神经网络的时间并提高模型的泛化性。另外的一大类深度模型采用自动编码模式，如 Yang 等[9]人 2019 年通过无监督学习在基于 LSTM 的 DAE 网络中预训练 DLSTM 模型。然后，利用预训练的 DLSTM 模型和 softmax 分类器对船舶辐射噪声进行分类。类似的 Ke 等[5]使用无监督特征提取进行预训练，提出了一个一维卷积自编码器-解码器模型，然后对模型进行预训练以从高共振分量中提取特征。这类算法有着相似的思想，通过非监督的预训练先生成一个模型，然后通过少量的标注数据进行微调，这样成功解决了标记数据少的问题。这样的思想在 Yang 等[10] 2018 年提出的竞争深度置信网络中亦有应用，他利用大量未标记数据进行预训练来初始化参数，将隐藏单元先分类用于竞争学习的初始参数，然后逐层训练网络，有监督的微调。

1.3 本文研究内容

设计动机

从上述的研究现状中可以看出，水声目标的探测识别也随着算力的进步而从传统的统计数学角度，转而聚焦于神经网络的研究。而鉴于利用深度学习方法解决问题除了巨量的

算力外，还需要海量的数据支撑。而水声的数据集的缺失是我们需要克服的最大难关。而我们面临的大量场景中的目标是相同的，大部分是各种船只与大型海洋生物，而由于在不同的海域环境中导致它们所发出的声音有着不同的特征。如果在每一个区域都去训练一个全新的独立的模型需要消耗大量人力物力，但如果使用迁移的思想，将已知的信息应用到其它未知的水下环境中，可以实现使用较低的代价完成较高水平的探测识别任务。故而本文引入迁移学习来实现少量数据场景下稳健模型的构建。而对于域适应领域的迁移噪音问题，我们则采用对迁移数据对进行一致性筛选并通过交叉注意力机制来消除误配对的情况。

文章内容概要

总的来说，本文从传统算法中汲取特征提取相关知识，将音频转为梅尔频谱图。并利用前沿的迁移学习方案来处理我们的目标数据集较小的问题，利用无监督的域适应也有效减少人工标注数据的成本。为能成功迁移，我们使用双向匹配的策略来构建源域到目标域的数据对，并利用预训练模型为目标域生成的聚类中心，进行 K-means[11] 聚类来生成伪标签，并将它们与源域标签进行一致性对比，净化迁移噪音。同时迁移算法用到交叉注意力网络，进一步消除误配对数据对目标域模型的干扰。本文在这样一套模型中，聚焦于解决迁移中噪音干扰问题，通过注意力机制对两域进行目标对齐，强化关键特征，消除负面噪音。

第二章 相关技术基础

2.1 迁移学习简介

当我们在一个耗费了大量人力物力的数据集上，尝试了很多次最终获得了一个结果可观的模型。而我们面对一个相似的场景却又不得不从头重复之前的全部工作，因为模型的重用总会伴随着众多困难，比如虽然是相同的任务但数据的分布却完全不同。为了解决大数据场景下少量标注，计算资源紧缺以及用户对于模型个性化定制等问题，迁移学习应运而生。

迁移学习没有统一的定义，一个可靠的解释是我们将能解决源任务的方案中迁移一部分知识去解决目标任务。比如从下中国象棋中学习一部分知识去应对下国际象棋这个任务场景。更加广泛的场景中，对于医学，教育学等众多领域都可以应用迁移学习。最早的迁移学习例子可以追溯到 1901 年，在心理学领域。当时，心理学家 Edward Thorndike 进行了一系列关于猫的实验[12]，试图研究猫在解决新问题时如何利用之前学习的知识和经验。他发现，猫在解决新问题时，可以利用之前学习的知识和经验，即使这些知识和经验与新问题并不直接相关。而在计算机领域，可以追溯到 1995 年，当时，Caruana 等人在研究神经网络的迁移性时提出了这个概念。他们发现，将一个预先训练好的神经网络模型应用到一个新的任务中，可以提高训练效率和泛化性能。这个研究开启了神经网络迁移学习的研究之路，而迁移学习也逐渐成为了机器学习领域中一个重要的研究方向。而如今大热的 GPT 模型也应用了迁移学习知识，在 GPT 训练之前，使用了大量的未标注文本数据对模型进行预训练。这个过程叫做自监督学习（self-supervised learning）或者无监督学习（unsupervised learning），它可以让模型学习到更加普适的语言模式和规律。预训练的模型也可以迁移到不同的任务中，例如文本分类、问答和机器翻译等。另外，在预训练之后，将预训练好的模型应用到具体任务中，通过微调（fine-tuning）调整模型参数来适应特定的任务。微调可以使用有标注的文本数据，例如情感分类、文本生成和机器翻译等。微调可以使得模型在特定任务上表现更好，同时也可以提高模型的泛化性能。

接下来我们通过严谨的数学公式来对迁移学习问题进行数学建模。首先我们假设有一个源域 $D_s : (x_i, y_i)_{i=1}^N \sim P(x, y)$ 表示源域服从 P 分布，而我们目标域的 D_t 服从 $Q(x, y)$ 分布，我们的任务 T 是找到一个合适的函数 f 使得 $y = f(x)$ 而我们面临的限制来自于两域

的数据分布不同以及它们的任务不同，即：

$$\begin{cases} P(x,y) \neq Q(x,y) \\ T_s \neq T_t \end{cases} \tag{2.1}$$

我们的目标是在给定目标域分布 D_s, 源域分布 D_t 以及目标域任务 T_s 的情况下，面临着目标域与源域分布不同或者目标域任务与源域任务不同的困难，去求得合适的源域任务 T_t。深度学习的过程是我们在寻求一个 f，使得我们预测的 $f(x)$ 与实际的 y 之间的 Loss 函数 L 最小

$$f^* = arg\min f \frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i) \tag{2.2}$$

而迁移学习则是在公式2.1的限制下，去求得一个正则项 $R(x_i, y_i)$，来使 f 可以适应新的场景。

$$f^* = arg\min f \frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i) + \lambda R(x_i, y_i) \tag{2.3}$$

而对于如何实现迁移学习，可以从三个角度来出发。第一种对于迁移目标是源域的子集的情况下，则我们可以忽略公式2.3中的正则项 R，我们可以对源域中样本进行加权，将属于目标域的样本进行强化，最终我们即可实现任务的迁移，这也叫基于实例的迁移。而更为广泛的情况是 $D_S \neq D_T$，这时我们需要优化 R 来时这两个域之间的距离减小甚至消除，我们称这样的迁移算法为基于特征的迁移，它的工作主要是通过显式或者隐式的特征变换来使两域的特征进行更大程度上的对齐，如图2-1所示，在经过变换后，源域的特征与目标域的特征空间有了更大的重合。

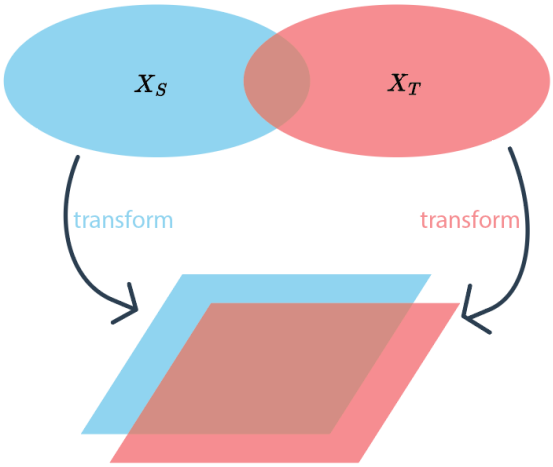


图 2-1 域变换图示

聚焦于这方面工作的有显式特征变换算法如 Joint Distribution Adaptation[13]，Dynamic Distribution Adaptation[14]等，它们都是基于将特征映射到不同的空间，然后利用一个度量值来衡量两域之间的距离，不断优化这个距离来实现迁移。基于特征迁移的另一个想法则是基于生成对抗网络的，在特征提取后，添加了一个特征判别器模块，在分类任务之前，先去判断特征来源于源域还是目标域，当这个网络无法区分特征来自哪个域的适合，则实现迁移，相关的工作有 Domain Adversarial Neural Network[15]等。除此之外，当我们源任务上的 f 效果非常好的时候，可以考虑基于参数的迁移，它的基本思想是，将一个预训练的模型的参数（如神经网络的权重）作为初始化参数，然后将其应用于新任务的模型中，以加速模型的训练和提高模型的性能。基于参数的迁移学习通常需要先使用大量数据训练一个模型，例如在自然语言处理中，可以使用大规模语料库训练一个语言模型。然后再微调模型，将预训练模型的参数作为初始化参数，然后在新任务的数据上进行微调，以适应新任务的特定要求。在微调过程中，通常固定预训练模型的前几层，只更新后面的一些层或者添加一些新的层。比如一个分类动物的模型，其中浅层大概率捕捉的特征比如是否具有眼睛，是否具有四肢，而最终区分出这个动物的特征可能是脸的形状，眼睛的颜色等更为深层的信息。而当我们将猫的分类器迁移到狗的分类器时，就可以固定住前面的浅层的参数，然后对最后的几层进行微调。

2.2 基于注意力机制的图像特征提取

Transformer[16]最初是在自然语言处理领域中提出的，但近年来也被广泛应用于图像分类领域。在传统的卷积神经网络 (CNN) 中，卷积操作被用来捕捉图像中的局部空间相关性。然而，卷积操作是固定的，因此对于不同大小的输入图像，需要使用不同大小的卷积核来进行卷积操作，这导致了网络的复杂性和计算成本的增加。相比之下，Transformer 是一种基于自注意力机制的神经网络，它不需要卷积操作，可以处理任意大小的输入序列，并且能够在输入序列中自动学习相关性。在图像分类领域中，Transformer 可以使用自注意力机制来学习图像中不同区域之间的关系，从而实现局部和全局信息的融合。具体来说，可以将图像分成若干个均匀大小的区域，将每个区域的像素值作为序列输入到 Transformer 中，然后使用 Transformer 来学习区域之间的关系，并输出图像的分类结果。

自注意力机制

自注意力机制可以用于计算输入序列中每个位置与其他位置之间的关系，同时给不同位置分配不同的权重。在计算权重时，它会根据查询向量（query vector）、键向量（key vector）和值向量（value vector）之间的相似度来决定不同位置的权重。我们假设输入是一个长度为 N 的序列 $X = (x_1, ...x_n)$

首先我们需要计算查询向量 q，键向量 k 以及值向量 v 随机初始化三个矩阵 W^Q, W^K, W^V，对于第 i 个输入，将其输入与它们相乘得到对应的 q_i，k_i 以及 v_i

$$\begin{cases} q_i = x_i \cdot W^Q \\ k_i = x_i \cdot W^K \\ v_i = x_i \cdot W^V \end{cases} \tag{2.4}$$

然后对于第 i 个输入来说，用它的查询向量 q_i 与其它所有元素的键向量 k 点乘，除以其 k_i 的维度的开根号 $\sqrt{d_{k_i}}$，并用 $Softmax$ 函数进行归一化，得到其对应的注意力权重：

$$\alpha_i = Softmax(\frac{q_i \cdot k_1, ..., q_i \cdot k_n}{\sqrt{d_{k_i}}}) \tag{2.5}$$

这些权重代表了每个位置对于查询向量的重要性，即 α_{ij} 的值表明了第 x_j 对于 x_i 的重要性。然后我们利用这些权重加权 v_i 并求和，得到我们的输出 $output_i$

$$output_i = \sum_{j=1}^{N} \alpha_{ij} v_i \tag{2.6}$$

在我们面临的水声目标识别中，我们需要先将音频转换为频谱图从而提取其频率与能量特征，然后利用深度神经网络来对频谱图进行分类识别。而将 Transformer 应用到图像中去，首先如图2-2中第一步所示，先将图像进行切分成多个片，然后每个片在数据上来说为 $Height * Width * Channel$ 的三维数组，而为了能投入 Transformer 网络中，将这些数组按照图中第二步拉平为一维数组，最后将拉平后的数据投入网络，同时为了捕获更多特征，会设置多个注意力头，即会有多个 W^Q, W^K, W^V 矩阵，接着就按照之前所描述的那样去求权重，加权和，来获得每个片之间关注度。正如图2-2所示，最终能捕获出图片中值得注意的特征。图中展示的是一列火车行驶于绿地中的场景。可以明显看出图像在经过切片拉平并经过自注意力机制后，只保留了对火车部分的关注，而对草地与雪山这些非关键信息进行了忽略。

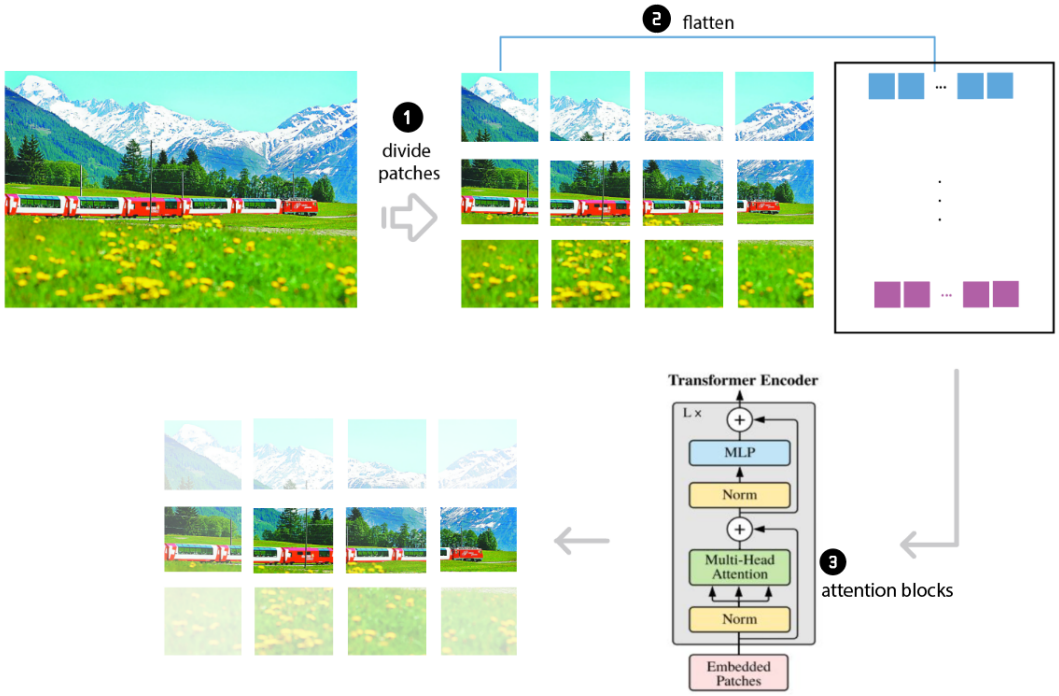


图 2-2 注意力机制在图片中的应用

交叉注意力机制

而对于迁移学习来说，我们需要考量的是两个域之间的特征对齐，而交叉注意力机制可以帮助模型在源领域和目标领域之间建立联系。它由自注意力机制引申而来，自注意力中的输入只来自一个数据集，而交叉注意力机制中，输入分别来自源域 D_S 与目标域 D_T，假设它们的输入经过 q, k, v 映射后为 Q_S, K_S, V_S 和 Q_T, K_T, V_T。对于自注意力来说，它的权重矩阵由2.7给出，可以看出其只使用了一个域中的数据。

$$Attn_{self}(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.7}$$

而对于交叉注意力机制，它对于每个目标域中的元素，都会去源域进行查询，以获取对源域中特征关注度的矩阵，从而迁移关注度高的部分。它由2.8给出，其与自注意力机制最大的不同在于查询的目标

$$Attn_{cross}(Q_S, K_t, V_t) = softmax(\frac{Q_s K_t^T}{\sqrt{d_k}})V_t \tag{2.8}$$

通过交叉注意力机制，可以有效避免迁移学习中的噪音，比如当我们的输入是一个错误的配对，如图2-3所示，源域中的输入是汽车，而目标域中是卡车。我们期待的迁移学习的数据输入是源域与目标域的类别应该是相同的，而错误的目标对必然会导致迁移了部分错误的特征。而由于交叉注意力机制的存在，我们可以从这副热力图中明显看出，对于目标

域的卡车来说，它的轮子部分只对源域中汽车的轮子产生了较大的关注，而对其它汽车的专有特征如扁平的车头等只有很少的关注权重，这就有效去除了无关噪音对迁移效果的影响。

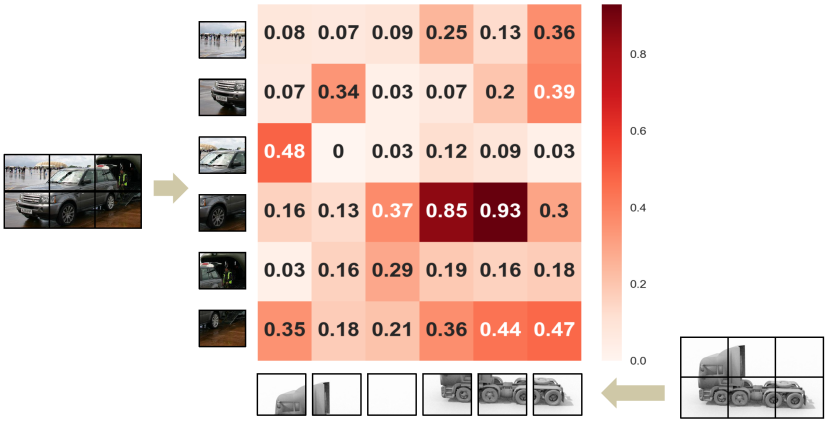


图 2-3 交叉注意力对噪音的消除图示[17]

2.3 基于深度学习的水声目标识别

在深度神经网络中，当网络层数增加时，梯度在反向传播过程中会逐渐变小，甚至消失。这会导致网络的训练变得困难，甚至无法收敛。残差网络的提出者通过引入残差块来解决这个问题。残差块中的每一层都包含了一个残差连接（Residual Connection），这个连接直接将输入的信息传递给输出，使得网络可以学习到残差信息，而不是单纯地学习输入输出之间的映射关系。这样可以避免梯度消失问题，同时也可以避免网络的退化问题。ResNet50[18]（Residual Network-50）是其中一个很经典的网络，它在 ImageNet 数据集上取得了非常好的性能，被广泛应用于图像分类、目标检测、语义分割等计算机视觉任务。在图片多分类问题中，ResNet50 可以作为一个强大的特征提取器，并结合全连接层和 softmax 层进行分类。ResNet50 的主要特点是使用了残差连接，有效地解决深度神经网络中的梯度消失和梯度爆炸问题，使得网络可以更深更容易训练。在图片多分类任务中，可以使用预训练的 ResNet50 网络对图片进行特征提取，然后连接全连接层进行分类。具体地，可以将 ResNet50 的最后一个卷积层输出的特征图作为输入，将其展平成一个一维向量，然后通过多个全连接层进行分类。最后，使用 softmax 层将网络输出转化为每个类别的概率值，得到最终的分类结果。

另外随着注意力机制在自然语言领域取得了巨大的进展，同样有一批工作将他们应用于频谱图的分类中，ViT (Vision Transformer)[19]是一种基于 Transformer 架构的深度学习模

型，用于处理计算机视觉任务。它的核心思想是将图像分成一组小的图像块，然后将这些图像块转换为向量，并通过 Transformer 来学习这些向量之间的关系，以实现图像分类任务。具体来说，ViT 首先将输入的图像分成一组小的图像块，然后将每个图像块通过一个线性变换映射为一个向量。这些向量组成了一个序列，然后通过 Transformer 模型进行处理。在 Transformer 模型中，每个向量都可以看作一个“token”，并与其他“token”进行交互，以学习它们之间的关系。最后，ViT 通过一个线性变换将 Transformer 的输出映射到类别分数上，从而完成图像分类任务。ViT 的优点在于可以处理任意大小的图像，而不需要进行特定的图像预处理或调整图像大小。此外，ViT 还可以利用大规模的预训练数据来提高模型的泛化能力，从而在许多图像分类任务中取得了与传统卷积神经网络相当甚至更好的性能。ViT 也有一些缺点，比如对于一些视觉任务，如目标检测和语义分割等，ViT 的性能可能不如传统的卷积神经网络。此外，ViT 需要更多的计算资源和更长的训练时间，因为它需要处理更大的输入序列。

而除了从头去训练一个独立的模型之外，通过已有模型来进行迁移也是最新兴起的研究领域。GTA[20](Generate To Adapt) 就是其中一种新颖的方法来解决计算机视觉中的域自适应问题。它属于隐式特征变换的迁移。该方法利用无监督数据，在学习到的联合特征空间中将源域和目标域分布靠近。同时，该方法还引入了生成对抗网络，使得学习到的嵌入与生成器之间形成了一种共生关系。通过在三个不同任务上进行实验，证明了该方法的强大性和通用性。该方法的主要思路是通过学习一个联合特征空间，将源域和目标域的分布靠近。具体来说，该方法使用了一个编码器网络将图像映射到特征空间中，并使用一个生成器网络来从该特征空间中生成图像。同时，该方法还引入了一个判别器网络，用于区分生成的图像是否来自目标域。通过训练这三个网络，可以使得编码器学习到一个能够将源域和目标域的图像映射到相似特征空间中的嵌入，并且使得生成器能够从这个特征空间中生成逼真的目标域图像。在训练过程中，该方法使用了对抗损失函数和重构损失函数来优化编码器、生成器和判别器网络。对抗损失函数用于训练判别器网络，使其能够区分源域和目标域的图像，并且鼓励生成器产生更逼真的目标域图像。重构损失函数则用于训练编码器和生成器网络，使其能够在联合特征空间中重构输入图像。总之，该方法通过在联合特征空间中进行编码、解码和判别操作，实现了源域和目标域的分布靠近，并且在多个任务上取得了优秀的实验结果。

另外，自 2018 年 Hadsell 等人提出的 Contrastive Predictive Coding (CPC)[21]模型后，对

比学习也引起人们的广泛关注。对比学习是一种无监督学习方法，其目标是将相似的样本映射到相近的空间位置，将不相似的样本映射到较远的空间位置。对比学习的主要思想是通过比较不同样本之间的相似度，学习出数据的分布特征和模式。传统的无监督学习方法通常是通过学习数据的统计特征来进行建模。但是，这种方法忽略了数据中不同样本之间的关系，因此可能无法很好地捕捉数据的分布特征。相比之下，对比学习通过比较不同样本之间的相似度，可以学习到数据中更丰富的信息，从而更好地捕捉数据的分布特征。对比学习的核心是构建一个对比损失函数。在训练过程中，对于每个样本，我们随机选择另外一个样本作为对比对象，然后计算它们之间的相似度。如果这两个样本是相似的，则它们应该被映射到相邻的空间位置，反之则应该被映射到较远的空间位置。通过最小化对比损失函数，我们可以学习到一个具有区分性的特征表示，使得相似的样本之间距离更近，不相似的样本之间距离更远。基于对比学习的方法很多，如开山之作 CPC，它从高维数据中提取有用的表示。该方法通过将高维数据压缩成更紧凑的潜在嵌入空间来使条件预测更容易建模，并在这个潜在空间中使用强大的自回归模型来进行多步未来预测。具体而言，CPC 框架包括两个主要组件：编码器和解码器。编码器将输入数据映射到潜在嵌入空间中，并且解码器使用自回归模型从潜在嵌入空间中重构原始输入。为了训练这个模型，CPC 使用了一种称为“对比损失”的损失函数，该函数通过比较正确的未来样本和错误的未来样本之间的相似性来鼓励模型学习有用的表示。作者测试了这些表示在各种领域中的效果：音频、图像、自然语言和强化学习，并在作为独立特征时实现了强或最先进性能。

将对比学习应用于音频分类中的还有 COLA(Contrastive Learning of General-purpose Audio Representations)[22] 它是一种自监督预训练方法，用于学习音频的通用表示。它基于对比学习，通过学习将来自同一录音的音频片段分配高相似度，而将来自不同录音的片段分配较低相似度。具体来说，COLA 方法使用三元组损失函数来训练模型。在每个三元组中，模型接收两个正样本（即来自同一录音的两个音频片段）和一个负样本（即来自不同录音的一个音频片段）。模型被要求将两个正样本之间的距离缩小，并将正样本与负样本之间的距离扩大。COLA 方法还使用了一些技巧来提高性能和效率。例如，它使用了多尺度卷积和池化操作以处理不同长度的音频片段，并使用了随机掩码以增加数据多样性。此外，COLA 方法还使用了一种称为“swapped negatives”的技术，在训练过程中随机交换正负样本对以增加数据多样性。在实验中，作者在大规模 Audioset 数据库上预训练嵌入，并将这些表示转移到包括语音、音乐、动物声音和声学场景在内的 9 个不同的分类任务中。

作者表明，尽管其简单，但 COLA 方法显著优于以前的自监督系统。此外，作者还进行了消融研究以确定关键设计选择，并发布了一个库来预训练和微调 COLA 模型。

2.4 本章小结

本章介绍了本文最核心的迁移学习算法的理论部分，同时介绍了目前最强大的特征提取网络之一 Transformer，并介绍了如何利用交叉注意力机制来进行源域与目标域特征的对齐与去噪。除此之外，简要介绍了利用残差网络解决深层网络问题，以及注意力机制，迁移学习，对比学习在音频分类问题中的具体应用方法。接下来将介绍如何将迁移学习与注意力机制算法结合，并应用到水声目标识别中去。

第三章 基于域适应的水声目标探测与识别

本模型从音频数据出发，如总览图 3-1所示，我们首先将源域数据与目标域数据通过音频切片来扩容样本量，然后利用 librosa 库对他们进行特征提取，转换为梅尔频谱图。接着通过双向匹配来构建迁移所用的数据对。具体而言，先通过预训练模型来计算目标域每个类的中心，然后通过 K-means 算法来为目标域的数据进行中心聚类而生成伪标签。最后将这个标签与源域中的标签进行对比，筛去不一致的输入对。最后将提纯的数据对投入由两个自注意力分支，一个交叉注意力分支的三分支 Transformer 网络，三个分支之间共享着参数，并通过交叉注意力来消除之前误配对中的噪音，最后微调出适应于目标域任务的新模型。

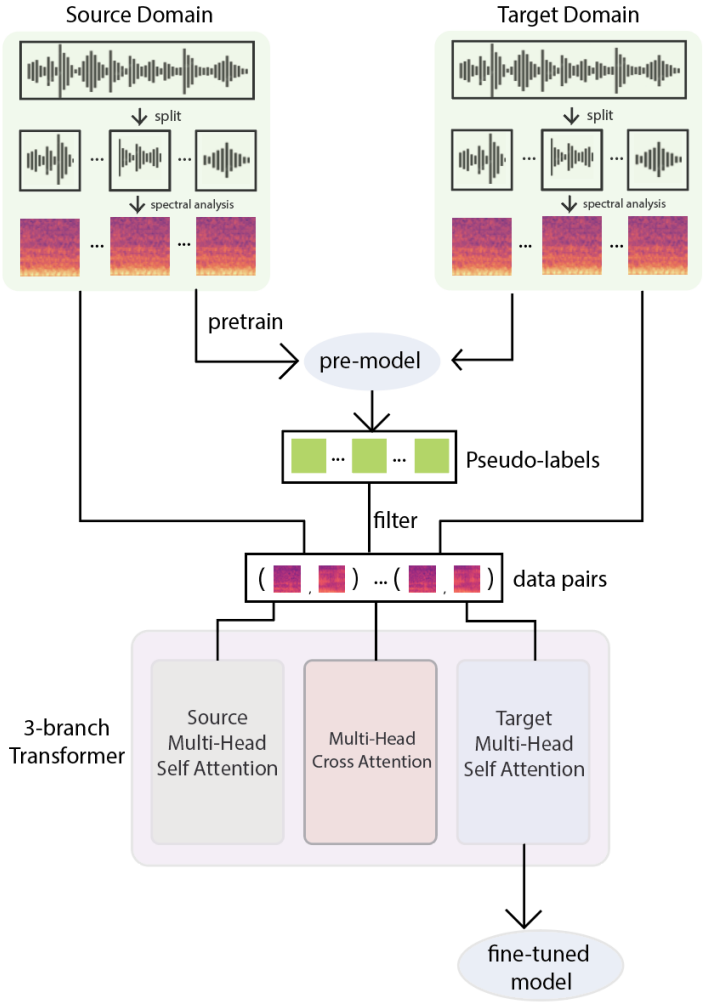


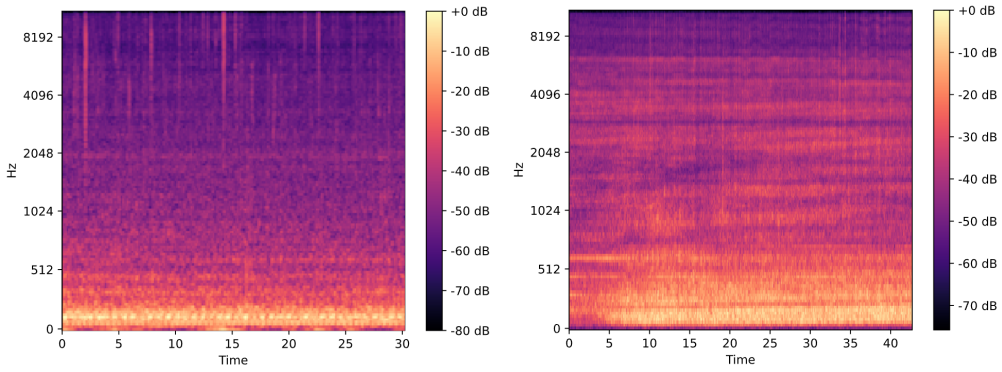
图 3-1 模型总览

3.1 数据预处理

ShipsEar[23]是一个水下船舶噪声数据库，包含了来自 11 种不同船只类型的 90 个录音。这些录音是在西班牙大西洋海岸不同地点进行的，包括港口、港口附近和自然环境中。研究人员通过将高精度的水听器放于不同的水深下来收集声音。这个数据库旨在为水下声学研究人员提供真实的声音数据，以便用于训练船只检测器和分类器或监测海上交通。每个录音都附带有详细的技术和环境信息，例如船只类型、录音地点和日期、水深和天气条件等。而我们的目标数据集是体量更小的水声数据集 DeepShip[24] 与实验室自建的二战数据集混合后的数据集。由于音频的采集大多是连续的时长较长的数据，比如 ShipsEar 中虽然只有 90 个录音，但却由 1.6GB 的大小。每个录音时长在两分钟左右。为了能生成更多的样本，我们对这些音频做了切片操作，将它们统一的裁剪为长度为 10 秒的短音频，然后利用 librosa 库（一个常用的音频处理库）将这些音频转换为梅尔谱图。梅尔频谱图（Mel Spectrogram）是一种常用于音频信号处理的特征表示方法。它将原始音频信号分解成一系列频率区间，然后将每个频率区间的信号能量转换为对应的梅尔频率，最后将所有频率区间的梅尔频率能量组成一个矩阵表示音频特征。梅尔频谱图的计算过程包括以下几个步骤：

1. 将原始音频信号分成若干帧，通常每帧的长度为 20-40 毫秒，且相邻帧之间有一定的重叠。
2. 对每一帧的音频信号进行傅里叶变换，得到其频域表示。
3. 将频域信号分成若干个等宽的频率区间，每个频率区间的宽度通常是线性的。然后，将每个频率区间的信号能量加起来，得到该频率区间的能量。
4. 将每个频率区间的中心频率转换为对应的梅尔频率。
5. 将所有频率区间的梅尔频率能量组成一个矩阵，即梅尔频谱图。梅尔频谱图的优点在于，它能够将音频信号的频谱信息转换为人类听觉系统更能感知的梅尔频率，从而更好地模拟人类听觉系统的工作方式。

下图 3-2是我们将 ShipsEar 中数据转为频谱图后的部分结果。可以看出货船在各个频率下比起客轮都有更高的能量，即颜色更偏紫色。



(a) 货船梅尔频谱图　　(b) 客轮梅尔频谱图

图 3-2 梅尔频谱图展示

在经过切分音频并生成每段音频的频谱图后我们获得了如表3.1的数据分布。实际上 ShipsEar 中还有更多的种类没有办法利用上，由于我们的目标域只有这四种类别。虽然数据总量不大，但对于每个类来说样本数目充分。在广泛用来测试的域迁移算法的 office31[25] 数据集中其源域中每个类平均有 90.87 个样本，而目标中平均有 16 个样本。相对而言，我们的源域数据每个类平均有 504.5 个样本远高于 office31 中源域平均类样本数。而我们的目标域每个类平均样本数为 273 个，也远高于 office31 中的值。总的来说，实验所用的数据集在源域上更具有数据优势。

表 3.1 实验数据统计

来源	货船	客轮	拖船	环境噪音	平均数目
源域	912	430	234	442	504.5
目标域	599	367	80	46	273.0

3.2 迁移数据对构建

在构建完数据集后，我们需要构建用于迁移学习的输入对。首先从源域出发，在目标域中寻找可靠的配对者，记这样寻找到的输入对集合为 P_S，我们期待对任意一个 s 它在目标域的配对者 t 它们之间的距离 d 最小。

$$P_S = (s,t)|t = \min_k d(f_s, f_k), \forall k \in T, \forall s \in S \tag{3.1}$$

同样的我们再从目标域出发来为每个 t 去源域寻找使两者距离最小的 s，并将这样的输入对集合记为 P_T

$$P_T = (s,t)|s = \min_k d(f_t, f_k), \forall k \in S, \forall t \in T \tag{3.2}$$

最后我们再取它们的并集 P 为输入对。

$$P = P_S \cup P_T \tag{3.3}$$

之后我们需要为目标域数据生成伪标签，这里采用的是简单的 k-means 聚类算法，首先将目标域中数据通过源域的预训练模型，获得其在源域上的概率分布 δ，然后用这些分布来分别计算每个目标域上类别的中心，如 k 类别的中心 $center_k$ 由下式给出：

$$center_k = \frac{\sum_{t \in T} \delta_t^k f_t}{\sum_{t \in T} \delta_t^K} \tag{3.4}$$

δ_t^k 表示 t 属于 k 类的概率。而对于 t 来说它的伪标签 y_t 则完全取决于它与哪个类别的中心更近

$$y_t = arg \min_k d(c_k, f_t) \tag{3.5}$$

然后我们再将刚刚打上标签的 t 考虑进去重新生成类的中心。最终我们为所有的目标域元素打上伪标签。最后对于每个 P 中的输入对，如果它目标域的伪标签与源域的标签不一致，我们将其剔除数据集，从而得到最终的输入对集合。

3.3 模型搭建与实现

本小节主要介绍实验中模型的构建，我们实验基于 PyTorch 平台开发，它使用动态计算图，使得模型的构建和调试更加灵活和直观。此外，PyTorch 使用 GPU 加速计算，可以快速处理大规模数据。此外，它还内置了各种优化技术，如自动微分、动态图优化、异步计算等，可以提高模型的训练和推理效率。

本章内容主要分为两个部分，一个是预训练模块，它只使用了源域数据，是一个基本的 Vision Transformer 模型的变种 Data-efficient Image Transformers[26]，目的是降低模型所需要的硬件条件。之后介绍的是非监督域适应模块，它通过交叉注意力机制实现了源域模型向目标域的迁移，成功微调了源域模型使其能适应目标域的数据分布。

3.3.1 预训练模块

预训练模块主要使用了 Vision Transformer(ViT) 算法，它是一种基于 Transformer 架构的深度学习模型，用于计算机视觉任务，特别是图像分类。传统的卷积神经网络（CNN）在

图像分类方面表现出色，但在处理长期依赖性和全局上下文信息方面可能存在缺陷。ViT 旨在解决这个问题，通过将输入图像划分为一组小的图像块，并将这些图像块作为序列输入到 Transformer 中，从而实现对全局上下文信息的建模。

ViT 的核心思想是使用 Transformer 的自注意力机制来学习图像中的特征表示。在 ViT 中，图像被分成固定数量的小图像块，并且每个块被视为序列中的一个令牌。这些图像块的序列输入到 Transformer 编码器中，其中每个编码器层使用多头自注意力机制来学习图像块之间的关系，并且在每个块之间共享权重。最后，通过将序列的最后一个令牌的表示传递给一个全连接层，ViT 能够输出图像的类别标签。

ViT 模型的优点是可以处理不同尺寸的图像，因为输入图像被分成固定数量的小块。此外，ViT 可以学习到图像中的全局上下文信息，而不仅仅是局部特征，这对于某些视觉任务非常重要。然而，ViT 的缺点是在处理大型图像时可能会出现性能问题，因为需要将图像块序列输入到 Transformer 中，这会导致计算成本的增加。因此本文实际采用的是 DeiT（Data-efficient Image Transformers）它是对 ViT 的改进，旨在提高模型的数据效率和计算效率。DeiT 的改进主要包括以下两个方面：数据效率的提高以及计算效率的提高。

DeiT 通过引入蒸馏技术 (Distillation)，使用较大的老师模型 (Teacher) 指导较小的学生模型 (Student) 进行训练，从而降低了模型在大规模数据上训练的要求。具体而言，DeiT 使用一个较大的 Transformer 作为 Teacher 模型，在 ImageNet 等大规模数据集上进行训练，然后使用一个较小的 Transformer 作为 Student 模型，在较小的数据集上进行训练。最后，通过将 Teacher 模型的知识转移给 Student 模型，DeiT 可以在较小的数据集上取得与在大规模数据集上训练的模型相当的性能。

DeiT 通过引入一些计算上的优化，提高了模型的计算效率。具体而言，DeiT 使用了一些轻量级的模型结构，如 Depthwise Separable Convolution[27] 和 Skip Connection[28] 等，来减少模型的参数量和计算量。此外，DeiT 还使用了一种叫做“Token Mixing”的技术，将一些相邻的图像块合并为一个更大的块，从而减少了 Transformer 编码器中的序列长度，进一步提高了计算效率。

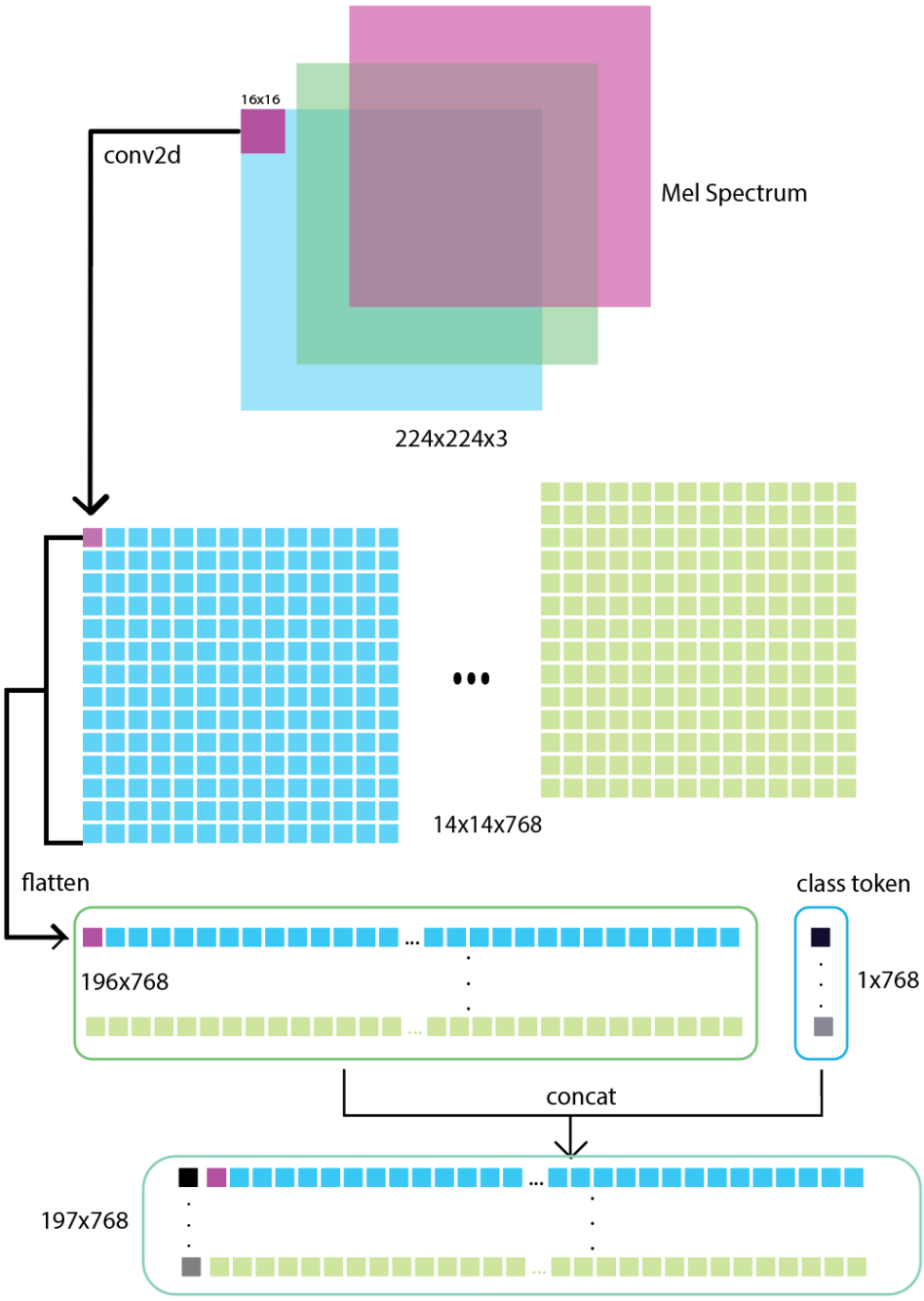


图 3-3 嵌入层图示

下表3.2所展示为本次实验的预训练的模型结构，输入为高宽均为 224 的频谱图，batch-size 为 32。即输入 x 的维度为 $(32, 3, 224, 224)$，x 首先进入嵌入层 $patch_{embed}$，嵌入层由三部分组成，分别担任卷积提取特征，拉平以及连接类标记的作用。如图3-3，输入先通过一个卷积核大小为 $(16, 16)$，步长也为 $(16, 16)$ 的卷积层，并将其 channel 维度从 3 拓展到 768，经过卷积核的运算后，$(224, 224)$ 的图像被分为 $14 * 14$ 块（patch），然后在将这些 $14 * 14$

的块拉平为 196 的一维向量。此时再引入一个类令牌 (class token)，它是一个额外的特殊令牌，它被添加到输入序列的开头，并且它的表示被用作最后的分类器输入。由于它是从外部引入的，所有可以完全避免它对图片中某一部分产生先天的偏移关注。Class Token 的作用类似于传统 CNN 中的全局平均池化层，可以将整个图像的信息压缩成一个向量，供分类器使用。在将它与拉平后的向量进行连接操作后，便形成了可以投入 Transformer 网络的序列。

表 3.2 预训练模块网络架构

模块名称	层类型	层参数	输出维度
patch_embed	Conv2d	input_dim:3, ouput_dim: 768, kernel: (16,16), stride:(16,16)	(32,768,14,14)
	Flatten	none	(32, 196, 768)
	Concat	(x, cls_token)	(32, 197, 768)
atten_block	LayerNorm	none	(32, 197, 768)
	Attention	qkv	(32, 197, 768)
	Linear	in:768, out:3072	(32, 197, 3072)
	GELU	none	(32, 197, 3072)
	Linear	in:3072, out:768	(32, 197, 768)
	Split	return x[:, 0]	(32, 768)
classifer	Linear	in:768, out:4	(32, 4)
	BatchNorm1d	none	(32, 4)

在进入注意力模块后，实验中使用了多头注意力机制，它通过并行计算多个注意力头，来提高模型的表达能力和泛化性能。相比单个自注意力，它同时从不同的角度对输入进行注意力计算，从而提高了模型对输入的表达能力。每个注意力头都可以学习到不同的特征表示，因此它们可以捕捉到输入的不同方面的信息。同时，通过对多个注意力头的输出进行拼接，可以得到一个更丰富的特征表示，从而提高了模型的表达能力。同时可以减少单个注意力头的过度拟合，从而提高了模型的稳定性和鲁棒性。由于每个注意力头都可以学习到不同的特征表示，它们之间可能存在差异，这可以减少模型对单个注意力头的过度依赖，从而提高模型的泛化能力。且它可以并行计算多个注意力头，从而提高了模型的

计算效率。由于每个注意力头之间是独立的，因此可以使用并行计算的方式来加速模型的训练和推理过程。本实验使用了 12 个头，对于 768 个信道，则每个头处理 $\frac{768}{12}=64$ 个信道，q,k,v 的维度都为 (32, 12, 197, 64)。经过这十二个多头注意力后，每个头的输出都是 $(32,12,197,64)$，然后再将它们进行连接变为 (32, 197, 768)。为了增强模型的表达能力和泛化性能，会通过线性层将特征映射到更高维度的空间，把 x 从 (32, 197, 768) 映射到 (32, 197, 3072) 然后经过 GELU（Gaussian Error Linear Units）激活函数，它可以用于神经网络的隐藏层和输出层。GELU 的主要功能是将输入的线性变换映射到一个非线性空间中，从而增强模型的表达能力和泛化性能。具体而言，GELU 的公式为：

$$\text{GELU}(x)=\frac{x}{2}\left(1+\text{erf}\left(\frac{x}{\sqrt{2}}\right)\right) \tag{3.6}$$

其中，erf(x) 为误差函数，可以将输入 x 映射到一个概率分布中。

$$\text{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{3.7}$$

为了避免过拟合和计算负担，再次通过线性层将其维度降低回 (32, 197, 768)。最后剥离出之前的类令牌，即图3-3中最后的黑色部分，我们直接返回 $x[:,0]$。它包含了整张图片所有的信息。

最后我们将类令牌通过一个全连接层从 768 维映射到 4，以进行我们要进行的四分类问题。为了提高模型的稳定性和泛化性能，我们将得到的向量进行 BatchNorm[29] 归一化。通过对中间层进行归一化，BatchNorm 可以使得每个特征维度的数据分布更加平稳，从而降低了网络的内部协变量偏移（Internal Covariate Shift）问题。这样可以使得网络更容易收敛，并且可以使用更大的学习率进行训练，加快网络的训练速度。同时它可以减少特征之间的相互依赖，使得每个特征维度的重要性更加平等，并且可以提高模型在测试集上的性能。它的公式如下：

$$y_i=\frac{x_i-\text{E}[x]}{\sqrt{\text{Var}[x]+\epsilon}}*\gamma+\beta$$

其中，x_i 表示 BN 层输入的第 i 个元素，$\text{E}[x]$ 表示输入 x 的均值，$\text{Var}[x]$ 表示输入 x 的方差，ϵ 是一个很小的数，用于避免方差为 0 的情况，γ 和 β 是可学习的参数，用于缩放和平移归一化后的数据。而由于它是按每个 batch 来计算的均值与方差，所以当 batchsize 过小时会造成每个 batch 中的数据不能够表达整体的数据分布，从而导致网络难以收敛。实验中我们发现在 batchsize 小于 16 后已经有较为明显的影响，会很难收敛。此时这个向量中的四个元素表达的就是图片分别属于四类的概率。最高的一项则为其分类结果。

3.3.2 非监督域适应模块

而在预训练后，我们利用 K-means 算法以及预训练模型按照3.2所提的算为目标域数据进行伪标签构造后，就可以进行迁移学习了。整个流程与预训练模块大体类似，只在注意力模块与最后的损失函数计算中有些许改动。如图 3-4所示，由于预训练要同时输入源域与目标域两组数据，在预训练实验中将 batch size 设为 32 已经是 8GB 显卡容量的极限。

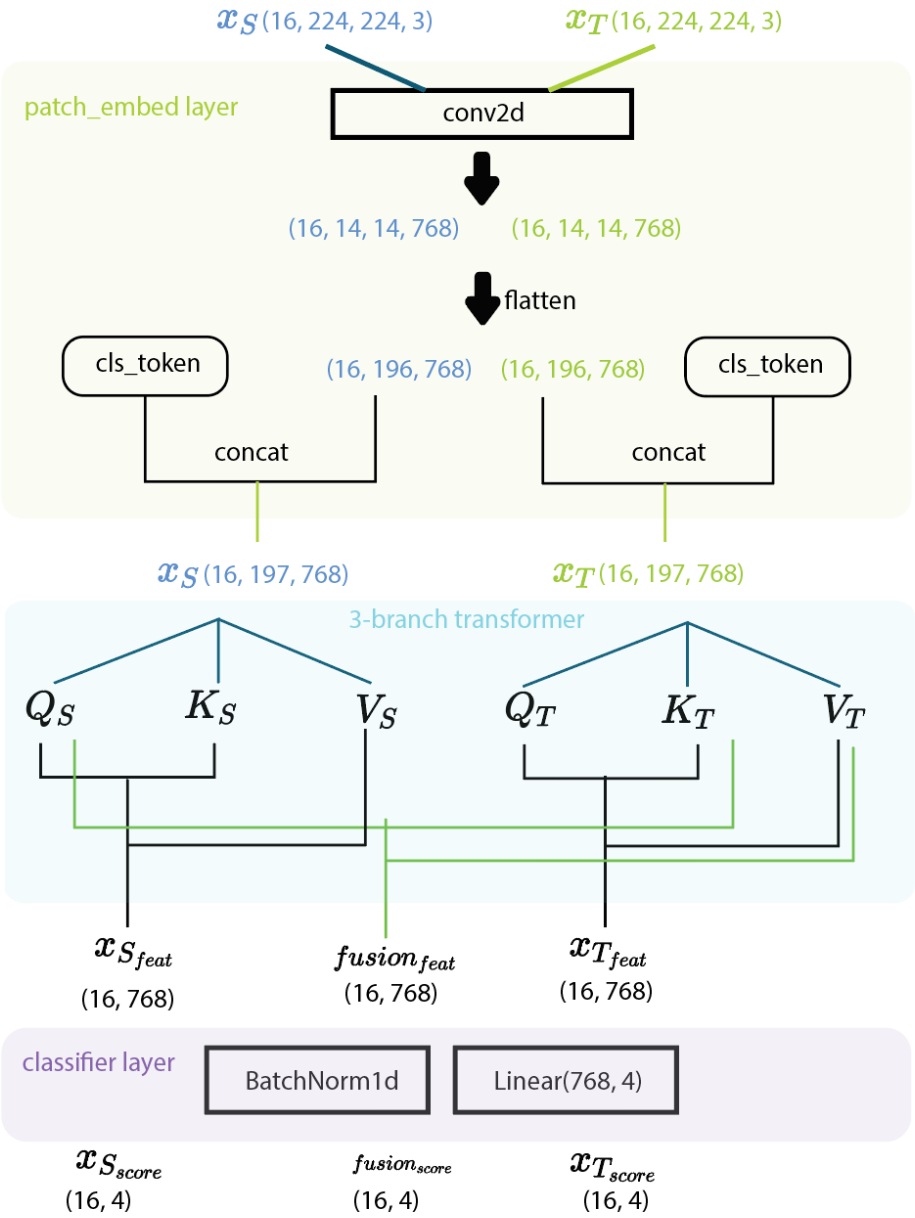


图 3-4 域自适应模块

所以在迁移过程中我们将 batch size 设为 16。输入的数据对 (x_S, x_T) 分别来自源域与目标域，它们的维度均为 $(16, 224, 224, 3)$。在经过之前所提的 patch_embed 层，先通过卷

积核为 16x16 的卷积层将特征从 3 信道映射到 768 信道变为 (16,14,14,768)，然后拉平变为 (16, 196, 768)，最后再与类令牌 cls_token 进行连接变为两组 $(16, 197, 768)$ 的输入对。

然后进入一个具有三个分支的 Transformer 网络。首先它们分别经过自注意力分支，这一步与预训练中的数据处理完全相同，记它们分别通过自注意力模块求得的自注意力得分为 $x_{S_{feat}}, x_{T_{feat}}$。然后使用用目标域中数据 x_T 对应的键向量 K_T 与源域中 x_S 对应的查询向量 Q_S 相乘得到交叉注意力矩阵，用这个矩阵为目标域对应的值向量 V_T 加权并求和，得到最终的交叉注意力得分 $fusion_{feat}$。这样我们我们得到了三个都是 (16, 768) 的关注特征向量。然后我们将它们都通过分类器，一个线性层将它们从 768 映射到 4，表达对四个分类的得分，分别记为 $x_{S_{score}}, x_{T_{score}}, fusion_{score}$，它们的维度都为 (16, 4)。然后我们将两个来自自注意力机制的得分 $x_{S_{score}}, x_{T_{score}}$ 与它们对应的目标 $target_S$ 以及 $pseudo_T$ 交叉熵来计算损失。交叉熵损失函数（Cross Entropy Loss）是一种常用的损失函数，通常用于分类问题中。在本实验场景中，交叉熵损失函数由下式给出：

$$\begin{cases} Loss_S = -\sum target_S \log(x_{S_{score}}) \\ Loss_T = -\sum pseudo_T \log(x_{T_{score}}) \end{cases} \tag{3.8}$$

而得到的交叉注意力机制的得分通过知识蒸馏来计算损失函数，蒸馏是一种模型压缩技术，通过将一个大型的模型（教师模型）的知识传递给一个小型的模型（学生模型），从而达到减少模型大小和加速推理的目的。在蒸馏中，通常使用一个损失函数来量化教师模型和学生模型之间的差异，本实验中仍然采用交叉熵损失。将交叉注意力学习的作为老师，将目标域自注意力得到的作为学生，通过式3.9来计算损失。

$$Loss_{distill} = -fusion_{score} \log(x_{T_{score}}) \tag{3.9}$$

然后将它们三个一起加起来得到总的损失

$$Loss = Loss_S + Loss_T + Loss_{distill} \tag{3.10}$$

最后通过随机梯度下降来更新模型参数。

3.4 实验结果及分析

3.4.1 实验结果

在生成数据对中，我们最终使用了源域中 87.2% 的数据，以及目标域中 72.7% 的数据。配对成功率为 81.3%。说明源域与目标域的特征差异不大，适合迁移。我们用预训练模型之间对目标域进行测试，得到的准确率为 84.4%。接着我们对此模型进行微调来使之适应目标域中的数据分布。

实验中我们将预训练设置为 10 个 epoch，迁移过程设置为 40 个 epoch，图3-5展示了实验过程中三分支 Transformer 下降的趋势，横轴代表每次迭代的 step，它被缩放了 40 倍。(因为每个 step 都记录很影响训练速度，我采取了每 40 次记录一轮值)，纵轴为损失函数的值可以看出，源域中的交叉熵损失 $Loss_S$ ，在迁移过程中始终是处于比较低且稳定的状态，这是因为它已经经过预训练了。而橙色曲线表示的是交叉注意力机制产生的蒸馏损失，它在前 40(x40) 个 step 下降的很快，后面下降较为缓慢，甚至出现波动。而我们的目标域损失为绿色曲线，它在蒸馏损失快速下降时很快的下降，而后缓慢下降且没有波动趋势，表示我们的模型能很好的迁移到目标域上。

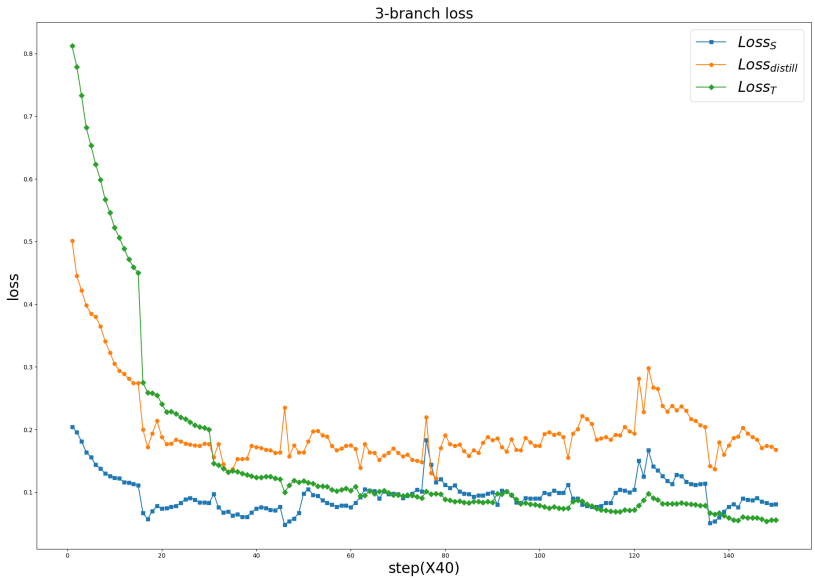


图 3-5 三分支 Transformer 损失下降图

如图3-6, 目标域的准确率从 84.4% 最终达到 97.53%，达到了一个足够健壮的水平。虽然准确率是最常见的评估指标之一，它表示模型正确预测的样本数占总样本数的比例，但

在数据分布不均时并不足以验证一个多分类模型的性能。

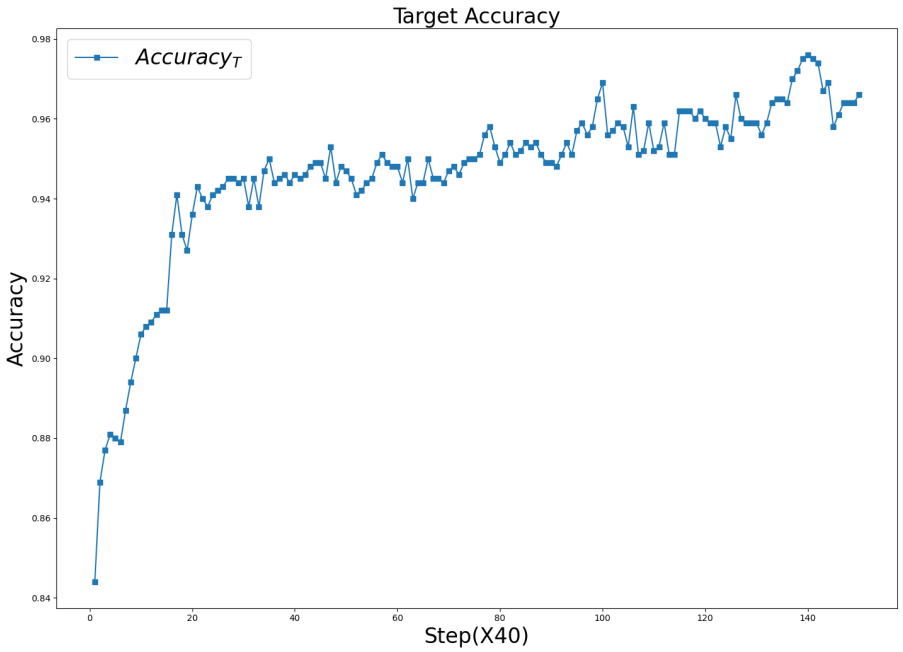


图 3-6 目标域准确率趋势图

当不同类别的样本数量差别很大时，准确率可能会高估模型性能。例如，对于一个有 100 个数据点，其中 90 个属于类别 A，10 个属于类别 B 的数据集，一个简单的模型预测所有数据点都属于类别 A，准确率为 90%。但是这个模型实际上没有学到任何有用的信息，因为它只是简单地预测了数据集中最常见的类别。另外它也无法区分不同类型的错误，只能告诉我们模型在整体上的性能如何，但它并不能告诉我们模型在不同类别上的表现如何。例如，在我们面对的四类别分类问题中，模型可能很好地预测了一个类别的样本，但是对于另外两个类别的样本却预测不准确。使用准确率无法区分这种情况。为此，我们计算了它的混淆矩阵，它将模型预测的结果与真实结果进行对比并分类统计。对于一个 K 类别的分类问题，混淆矩阵是一个 $K * K$ 的矩阵，其中行代表真实类别，列代表预测类别。矩阵的每个元素表示真实类别为 i，但是模型预测为 j 的样本数量。我们实验中有四类，故会生成一个 $4 * 4$ 的矩阵，它的对角线表示的正是预测完全正确的数据量。表3.3为本实验中目标域的混淆矩阵，可以看出几乎所有数据都落在了对角线上，并不存在不同类别中分类效果的偏差问题。在矩阵中货船的主要误判断来自客轮，只有极少样本误判为拖船，没有样本被误判为环境噪音。这说明货船与客轮的相似性更高一些。相似的客轮的误判主要也是集中在货船，只有两艘被分别误判为拖船和环境噪音。而拖船的误差来自环境噪音与

客轮，环境噪音只有一个被误判为拖船。

表 3.3 目标域混淆矩阵

真实\预测	**货船**	**客轮**	**拖船**	**环境噪音**
货船	**580**	17	2	0
客轮	6	**359**	1	1
拖船	0	1	**77**	2
环境噪音	0	0	1	**45**

3.4.2 消融与对照实验

我们首先做了去除掉交叉注意力网络的消融实验。另外作为对照，我们实现图像分类领域常用的 ResNet50 以及 Vision Transformer 网络，同时我们也查阅了最新的对比学习在音频分类中的工作并将它们应用到目标域数据集。最后还选择了一个迁移学习模型来与我们的模型进行对比。结果如表3.4所示，表明我们的模型在目标域上有着最好的表现，达到了 97.53% 的准确率。

消融实验

对于消融实验，我们直接去除了迁移过程中的交叉注意力网络，在代码上的反应只需在计算总损失的时候忽略掉 $Loss_{distill}$，这样在梯度更新时模型参数就不会被交叉注意力网络所影响。去除掉这个模块后，理论上它相当于对目标域进行自注意力机制加持的多分类任务，此时源域数据相当于噪声传入模型。实验结果表明其在目标域数据集上只能取得 83.24% 的准确率，比引入交叉注意力网络后的效果相差 14 个百分点，说明我们的迁移模块能有效使预训练模型适应目标域的数据分布。

ResNet50

ResNet50 是最常用的图像分类模型，我们的实验流程为利用之前所处理得到的目标域频谱图，对它直接做有监督的训练。利用 PyTorch 中的 ResNet 库，进行四分类操作，最终到达 80.13% 的准确率。是测试的几个模型中性能最差的，不能很好的完成目标域任务。

Vision Transformer

ViT 是我们所使用的预训练模型，我们将预训练模块直接运用进目标域即可，但需要传入目标域的准确标签，最后的准确率在 85.71%。这也说明我们的迁移模块可以在此基础上提升近 10 个百分点的准确率。

表 3.4 对照实验结果

模型名称	模型类型	准确率 (%)
ViT	图像分类	85.71
ResNet50	图像分类	80.13
CPC	对比式学习	94.38
COLA	对比式学习	95.11
GTA	迁移学习	90.61
Ours	**迁移学习**	**97.53**

CPC

“对比预测编码（Contrastive Predictive Coding）”是重要的对比学习方法。我们将 CPC 模型的编码器部分作为特征提取器，并在其之上添加一个分类器，然后使用标注的目标域数据对整个模型进行微调。在我们的目标域数据上，其取得了 94.38% 的准确率，性能优异。

COLA

COLA 是对比学习应用于音频分类的典型工作，它发布于 2021 年，是一种新颖且有效的方法。在我们的目标域上取得了 95.11% 的准确率，是除了我们模型外性能最好的。

Generate To Adapt

GTA 是利用生成对抗网络来进行迁移学习的经典工作，我们使用与我们实验中完全一致的数据集结构，最终可以达到 90.61% 的准确率，能很好的完成迁移任务。

3.5 本章小结

本章介绍了我们利用迁移学习完成了从源域 ShipsEar 向目标域 Deepship 数据集的模型迁移，并做了消融实验验证了迁移模块对于模型性能的提升约为 10%。最后将我们的实验结果与其它图像领域常见或新颖优秀的模型进行了对比，证明了我们模型的性能优异，是在此场景下表现最好的。

第四章 总结与展望

4.1 工作总结

针对水下声觉信号稀缺，水下环境复杂多变的问题，以及水声目标探测与识别的重要军事，民用意义，我们提出了一套将迁移学习引入水声被动目标识别的算法模型。总的来说它有以下 4 点贡献：

1. 少数将迁移学习应用到水声识别的工作，可以有效克服目标域数据量少的问题，节省了人工标注的成本。
2. 使用双向匹配机制构建迁移数据对，并通过 K-means 与预训练生成伪标签来和源域进行一致性对比来提纯迁移数据集。
3. 使用交叉注意力机制来对齐源域与目标域特征，有效消除少量误配对数据的噪音影响。
4. 我们的模型在众多常用及新颖的高性能算法中取得了最好的成绩。

4.2 工作展望

由于整个毕业设计时间有限，我觉得我们的工作还可以从以下方面拓展：

1. 在整个研究工作中，主要聚焦于神经网络的研究，而淡化对于音频信号的研究。它本身也可以采用更多的处理方式来提取更广泛的特征。
2. 实验中数据类偏少，且分布不够均匀，整个水声信号的公开数据集相对于图像，人声等热门领域是严重偏少的，我认为构建一个更大的，种类更多，有着统一标准的公开水声数据集能极大程度促进这个领域的发展。这项工作虽然技术难度上并不大，但需要投入较多的设备资金，以及会面对一些国家安全的问题。

参考文献

[1] 欧世峰, 赵晓晖, 顾海军. 改进的基于信号子空间的多通道语音增强算法[J]. 电子学报, 2005, 33(10): 1786-1789.

[2] 章新华, 张晓明, 林良骥. 船舶辐射噪声的混沌现象研究[J]. 声学学报, 1998, 23(2): 134-140.

[3] WANG N, HE M, SUN J, et al. Ia-pncc: noise processing method for underwater target recognition convolutional neural network[J]. Computers, Materials & Continua, 2019, 58(1): 169-181.

[4] WANG X, LIU A, ZHANG Y, et al. Underwater acoustic target recognition: a combination of multi-dimensional fusion features and modified deep neural network[J]. Remote Sensing, 2019, 11(16): 1888.

[5] KE X, YUAN F, CHENG E. Underwater acoustic target recognition based on supervised feature-separation algorithm[J]. Sensors, 2018, 18(12): 4318.

[6] KUMMERT A. Fuzzy technology implemented in sonar systems[J]. IEEE Journal of Oceanic Engineering, 1993, 18(4): 483-490.

[7] HU G, WANG K, PENG Y, et al. Deep learning methods for underwater target feature extraction and recognition[J]. Computational intelligence and neuroscience, 2018, 2018.

[8] YANG H, LI J, SHEN S, et al. A deep convolutional neural network inspired by auditory perception for underwater acoustic target recognition[J]. Sensors, 2019, 19(5): 1104.

[9] YANG H, XU G, YI S, et al. A new cooperative deep learning method for underwater acoustic target recognition[C]//OCEANS 2019-Marseille. IEEE, 2019: 1-4.

[10] YANG H, SHEN S, YAO X, et al. Competitive deep-belief networks for underwater acoustic target recognition[J]. Sensors, 2018, 18(4): 952.

[11] HARTIGAN J A, WONG M A. Algorithm as 136: A k-means clustering algorithm[J]. Journal of the royal statistical society. series c (applied statistics), 1979, 28(1): 100-108.

[12] WOODWORTH R S, THORNDIKE E L. The influence of improvement in one mental function upon the efficiency of other functions.(i).[J]. Psychological review, 1901, 8(3): 247.

[13] LONG M, WANG J, DING G, et al. Transfer feature learning with joint distribution adaptation[C]// Proceedings of the IEEE international conference on computer vision. 2013: 2200-2207.

[14] WANG J, CHEN Y, HAO S, et al. Balanced distribution adaptation for transfer learning[C]//2017 IEEE international conference on data mining (ICDM). IEEE, 2017: 1129-1134.

[15] GANIN Y, LEMPITSKY V. Unsupervised domain adaptation by backpropagation[C]//International conference on machine learning. PMLR, 2015: 1180-1189.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[17] XU T, CHEN W, WANG P, et al. Cdtrans: Cross-domain transformer for unsupervised domain adaptation [A]. 2021.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[19] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[A]. 2020.

[20] SANKARANARAYANAN S, BALAJI Y, CASTILLO C D, et al. Generate to adapt: Aligning domains using generative adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8503-8512.

[21] OORD A V D, LI Y, VINYALS O. Representation learning with contrastive predictive coding[A]. 2018.

[22] SAEED A, GRANGIER D, ZEGHIDOUR N. Contrastive learning of general-purpose audio representations[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 3875-3879.

[23] SANTOS-DOMÍNGUEZ D, TORRES-GUIJARRO S, CARDENAL-LÓPEZ A, et al. Shipsear: An underwater vessel noise database[J]. Applied Acoustics, 2016, 113: 64-69.

[24] IRFAN M, JIANGBIN Z, ALI S, et al. Deepship: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification[J]. Expert Systems with Applications, 2021, 183: 115270.

[25] SAENKO K, KULIS B, FRITZ M, et al. Adapting visual category models to new domains[C]//Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11. Springer, 2010: 213-226.

[26] 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.

[27] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[28] DROZDZAL M, VORONTSOV E, CHARTRAND G, et al. The importance of skip connections in biomedical image segmentation[C]//International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, 2016: 179-187.

[29] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International conference on machine learning. pmlr, 2015: 448-456.

致　　谢

感谢整个工作中姜龙玉老师的悉心指导，实验室安典坤，吕建坤学长的帮助，与他们的交流让从完全没有接触过信号的我快速了解了课题的背景，并在周会中产生了很多灵感。

感谢我的表哥周琪琪对我提出的众多初级琐碎问题的详细解答。

感谢我的室友在我感到疲惫难以前行时的陪伴，虽然他们也不是自愿的，但缘分如此。

最后感谢我的父母，奶奶和姐姐，他们见证了我从一个不识字的顽皮小孩成长为一个会一点技术的大学生。

与君同舟渡，达岸各自归。希望所有帮助过和没有帮助过我的人都安好。