编号：信息 032

上海大学

毕业设计（论文）

SHANGHAI UNIVERSITY
GRADUATION PROJECT (THESIS)

题目	**基于频域分析的模型可解释性研究**

学　院　　通信与信息工程学院
专　业　　电子信息工程
学　号　　18122885
学生姓名　张笑语
指导教师　吴汉舟 叶净宇
完成日期　2022 年 6 月

摘要

深度学习不断发展潮流下，神经网络模型的精确率不断提升，但其内部工作机制却由于其非线性特征而难以解释，因而限制了深度学习系统在数据驱动相关分析中的应用。本文旨在从傅里叶分析的视角对神经网络模型进行可解释性研究，挖掘样本频域特征与模型准确度及模型鲁棒性之间的关系，为设计高效的神经网络模型提供可解释性依据。

本文使用卷积神经网络 ResNet-18 模型在数据集 CIFAR-10 中对图像分类任务进行相关分析，共完成两项工作。（一）通过图像傅里叶变换及低通滤波和高通滤波的方式分离图像中不同频率域信息，并使用图像低频信息及高频信息分别进行模型精度测试，探究不同频率域信息对模型预测准确性的影响。实验得出在模型拟合过程中，图像低频信息在模型预测中发挥作用更大，神经网络模型将优先拟合图像低频信息再拟合图像高频信息。（二）本文引入傅里叶热图的分析方式，主要观测低频及高频信息在噪声扰动下的错误率，同时运用 CIFAR-10-C 数据集对不同模型的鲁棒性表现进行测试，以此对数据增强模型的鲁棒性表现进行可解释性分析。本文对高斯增强及对抗攻击两类在噪声扰动下模型鲁棒性表现优异的数据增强模型展开研究，发现两者共性在于提升了集中在高频区域的鲁棒性，降低了集中在低频区域的鲁棒性。同时，本文也运用图像的结构相似性指标结合傅里叶热图的分析方式对图像单通道数据增强模型鲁棒性进行相关可解释性分析。

本文总结了深度学习图像分类任务中频域特征与模型工作机制间的关联，为后续改进神经网络模型结构提供了理论依据。本文所运用的分析方式也可迁移应用至其他领域神经网络模型的可解释性分析。

关键词：卷积神经网络，傅里叶变换，数据增强，鲁棒性问题

Abstract

The accuracy of neural network models has been increasing under the continuous development trend of deep learning. However, due to the nonlinear characteristics of the neural network models, it is difficult to explain their internal working mechanism, thus limiting the application of deep learning systems in data-driven correlation analysis. The purpose of this paper is to investigate the interpretability of neural network models from the perspective of Fourier analysis. Meanwhile, this study surveys the relationship between frequency-domain features of samples and model prediction as well as model robustness, to provide an interpretable basis for designing efficient neural network models.

This essay completes the analysis by using the convolutional neural network ResNet-18 and the dataset CIFAR-10 to solve the image classification task. The work contains two tasks. (a) The task separates different frequency domain information by methods of Fourier transform and low-pass filtering as well as high-pass filtering. Moreover, it tests model accuracy in high-frequency and low-frequency domains respectively to investigate the influence of different frequency domain information on model prediction accuracy. It is concluded that in the process of model refinement, the low-frequency information in the images plays a more important role in model prediction, and the convolutional neural network model prefers fitting the low-frequency information before the high-frequency information. (b) This paper introduces an analysis way called Fourier heat map to observe the different sensitivities of high-frequency and low-frequency information under noise

corruption, so as to analyze the model robustness theory. Moreover, another dataset called CIFAR-10-C is used to analyze the robustness of data enhancement models. By observing Gaussian enhancement and adversarial training which all perform well under noise corruption, it is found that these two ways all improve the robustness concentrated in the high-frequency field and reduce the robustness concentrated in the low-frequency field. What is more, the task also combines the analysis of structural similarity index and Fourier heatmap to provide interpretable analysis for the data enhancement models in different image channels.

This paper summarizes the association between the frequency domain features and the internal principle of the model in deep learning, which provides a theoretical basis for the subsequent improvement of the neural network model. The analysis method used in this paper can also be used for the interpretability analysis of neural network models in other application areas.

Key words: convolutional neural network, Fourier transform, data augmentation, robustness problem

目录

第一章 绪论

1.1 研究背景及意义

1.1.1 深度学习模型的可解释性研究

深度神经网络在近几年迅速发展，被成功地应用于图像分类[1]、 语音识别[2]、自然语言处理[3]等多个领域，取得了与人类相媲美的准确率，因此，其应用范围也不断扩大，在各行各业都举足轻重。但是由于深度神经网络的非线性特征，其内部的工作机理难以解释，所以深度神经网络通常被认为是“黑箱”模型。在例如医疗、自动驾驶等很多的应用场景下，如果模型不能够为自己的决策做出解释，就很难建立有效的人机双向沟通机制，因而限制了这些深度学习系统在数据驱动相关分析中的应用。

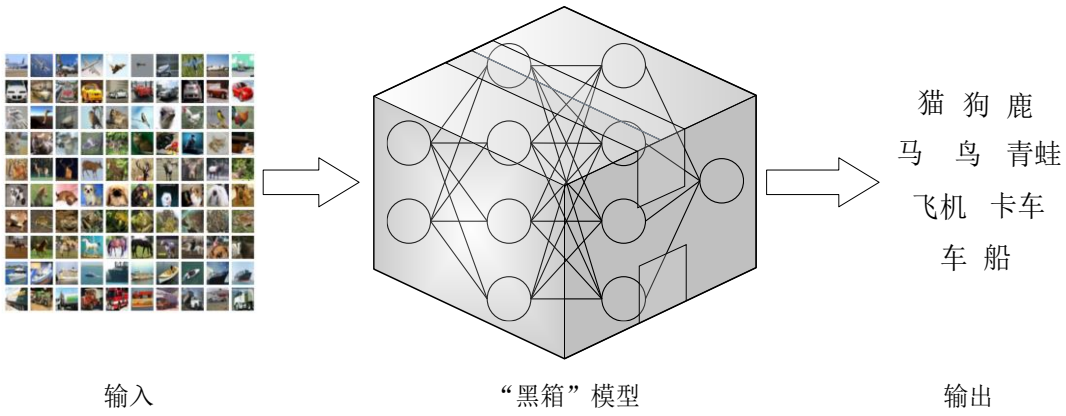


图 1-1 神经网络“黑箱”模型

深度学习的模型可解释性可分为以下三类。第一类，在建模前进行可解释性的分析，主要用于在建模前对数据进行特征分析，包含数据可视化：通过运用交互式方法多角度理解数据分布以及探索性质的数据分析：寻找数据集中代表性样本及非代表性样本[4]等方法。第二类，建立具有可解释性的模型，指模型本身可以通过数学公式或逻辑方法进行解释分析。可归纳为 5 种方法：基于规则，以决策树[5]为例模型本身根据逻辑规则进行解释；基于单个特征，以线性回归、逻辑回归等模型为例，能够根据统计学基础对模型进行理论解释；基于实例，以贝叶斯分类模型[6]为代表，根据典型性样本进行聚类、分类等工作的方法；基于稀疏性，以线性判别分析法(Linear Discriminant Analysis，LDA)[7]为例，利用信息的稀疏性简化模型使其易于理解；基于单调性，在输入输出呈现正相关或负相关时对该单

调性进行可解释性分析。第三类，针对具有黑箱性质的深度学习神经网络模型在建模后运用可解释性方法进行分析并做出解释，可分为三种方法：(一)，隐层分析法[8]，通过反激活、反池化、反卷积的方式，对每一层的神经网络所学到的特征进行学习分析；(二)，敏感性分析法[9-11]，主要研究分析输入变量的变动对输出变量的影响程度，分为输入变量对模型影响程度以及具体样本对模型重要程度两种；(三)，代理模型法，指根据较为复杂模型的输入值及预测值建立较简单能够进行解释的线性回归或决策树模型，用较为简单的模型体现复杂模型的内部机制。

1.1.2 课题研究意义

近些年来，深度神经网络模型不断推陈出新，呈井喷之势发展，依靠强大的计算机算力，通过增加层数、巧妙运用多种激活函数等方法使得模型精确率不断上升。其对于卷积核、激活函数的选取大多以结果为导向，通过反复调参选取性能最优者，相比之下关于神经网络内部工作机制的研究只有寥寥数篇。然而，当仅仅依靠调参得到性能不断提升的模型时，我们无法了解模型本身知识，无法判断模型如何从数据中学到相关的特征、知识并得出最终的决策结果。

如果能够探寻出样本频域特征与模型鲁棒性之间的关系，就能够有针对性的对神经网络模型进行改进，免去反复尝试的繁杂性，为深度学习发展提供有力依据。当深度学习模型具备可解释性时，与其原本的高准确率、高效率特点结合，三者兼顾，将会在许多应用场合发挥极大的优势。

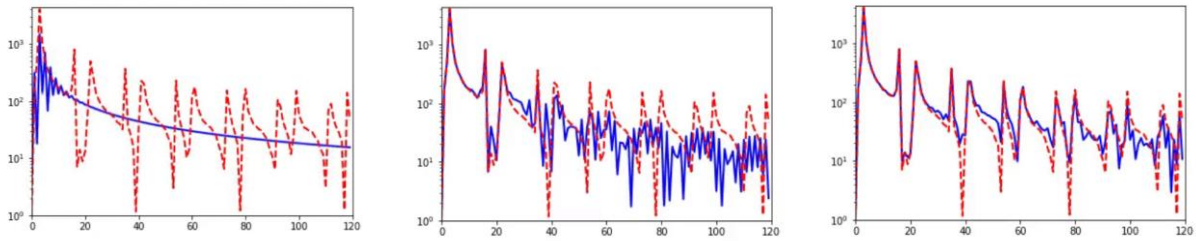
未来发展中，人们将会着力于探索神经网络模型的内部工作机制，对其内部原理进行可解释性研究，揭开内部“黑箱”模型的真实面目。了解其内部工作机制后，将能够有针对性的设计出更为高效的模型结构，为深度学习算法的发展提供便利。

1.2 频域原则

上海交通大学许志钦通过傅里叶分析的视角研究深度神经网络(Deep Natural Networks, DNNs)的训练过程，并由此提出频域原则[12-14]理论：深度学习过程中，通常从低频到高频拟合目标函数。本节，将对频域原则进行理论介绍，并展示如何通过频域原则理论进行模型的可解释性分析。

1.2.1 频域原则理论分析

在深度神经网络频域训练行为的相关研究中，对 DNN 能够拟合的相关一维函数进行快速傅里叶变换(Fast Fourier Transform，FFT)，同时对 DNN 的网络输出同样进行相关 FFT 变换，观察在深度神经网络训练过程中两者的拟合程度变化。如图 1-2 所示，横坐标为频率大小，纵坐标为振幅，红色虚线表示目标函数的傅里叶变换结果，不断变化的蓝色线条表示在训练过程中网络输出的傅里叶变换结果。从左至右为训练轮数由少至多的曲线拟合情况。可以观察得到当深度神经网络演化至频域后，频率由低到高依次收敛。



(a) (b) (c)

图 1-2 神经网络在频域中的拟合过程[1]

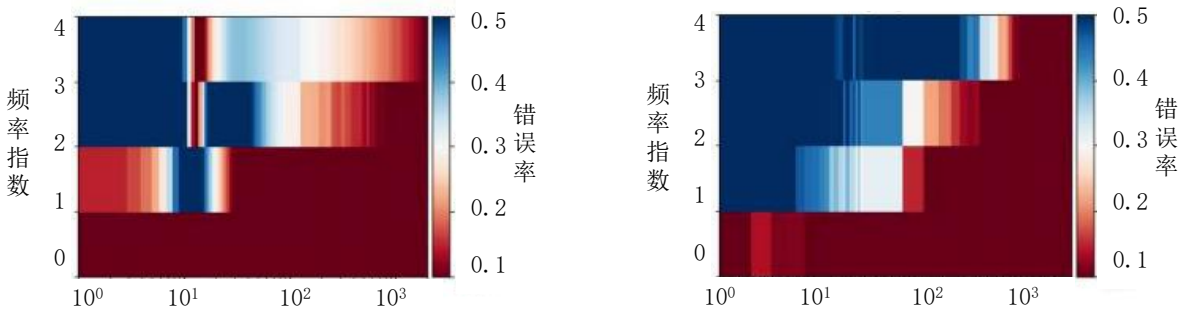


图 1-3 频域原则在真实数据集中的应用[13]

图 1-3 展示了运用不同卷积神经网络模型、激活函数及数据集进行实验时测试集相对错误率与训练轮数之间的关系，以验证频域原则理论。图中纵轴表示频率指数，当频率指数较小时，表示为低频部分，指数较大时表示为高频部分，横轴为训练轮数。图中通过颜色表示测试集相对错误率，红色表示相对错误率较小而蓝色表示相对错误率较大，观察训

[1] 该图像引用自许志钦个人博客主页. https://ins.sjtu.edu.cn/people/xuzhiqin/pub.html

练轮数不断增加的情况下低频及高频信息的错误率变化趋势。选用深度神经网络、tanh 激活函数对 MNIST 数据集测试的结果如(a)所示，选用卷积神经网络(Convolution Neural Networks，CNN)、ReLu 激活函数对 CIFAR-10 数据集进行测试后的结果如(b)所示。由此得出无论是全连接神经网络亦或是卷积神经网络,在训练轮数较低时，其低频错误率较低而高频错误率较高，随着训练轮数的增加，高频错误率慢慢降低。可见在神经网络训练过程中将优先拟合低频数据，再逐渐拟合高频数据。

1.2.2 利用频域原则对模型进行可解释性分析

近年来，深度学习不断发展，在许多实际应用问题如生物特征识别、医学诊断、证券市场分析中获得了不凡的成就。深度神经网络常常能在各类实际问题中有着极强的泛化能力，但由于许多科学领域常常拥有不同的特性，因此深度神经网络模型并不能取得优异的表现效果。例如，拟合奇偶校验函数（Parity function）：

$$f(x)=\prod_{j=1}^{n} x_j, x_j \in \{-1,1\} \tag{1-1}$$

该函数可直观表示为图 1-4 所示，其主要特征为左边部分四张图黑色方块个数皆为奇数，而右边部分四张图黑色方块个数为偶数。因此，Parity 函数值由“-1”的数量决定，如“-1”数量为奇数，则值为-1，反之，则值为 1。对于该函数，如果运用可能映射的子集作为其训练集，深度神经网络能够较好拟合数据，但对于未曾训练的测试集数据，深度神经网络并没有表现出良好的泛化性能。

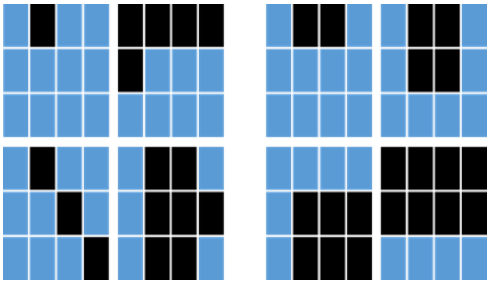


图 1-4 Parity 函数

通过频域原则理论我们可知深度神经网络优先拟合低频数据，从频谱分析的角度，CIFAR-10[15]和 MNIST 数据集都有低频占优的特点，而 Parity 函数则是高频占优。

在神经网络模型的训练过程中，根据频域原则优先拟合低频的机制，深度神经网络倾向于选择低频成分较多的函数拟合训练数据。而当目标函数高频占优时，由于混叠效应的

存在，不充分采样的训练集频谱将会在低频处有异常显著的虚假成分。 因此，当目标函数具有低频占优的特性时，由于高频成分较小，混叠效应带来的虚假低频成分能够忽略不计，并且神经网络训练中优先拟合低频的特性与函数本身低频占优特性相同，从而能够精确获取目标函数中的关键成分，从而拥有良好的泛化性能。然而对于 Parity 函数，由于本身高频占优的特点导致混叠效应引起的虚假低频严重且与频域原则优先拟合低频的特性不匹配，导致输出函数相比于目标函数的差异过大，影响模型本身的泛化性能。

通过频域原则对神经网络模型拟合 Parity 函数问题的可解释性分析可以得出，利用模型的可解释性能够为诸多目前遇到的神经网络无法解决的相关科学领域问题进行解释分析其内部原因。本文将延续从频域视角对深度神经网络模型进行可解释性分析，探寻不同频率域信息与模型精确度及鲁棒性之间的关系。

1.3 本文研究内容

本文将从傅里叶分析的视角对神经网络模型进行可解释性研究，挖掘样本频域特征与模型鲁棒性的关系，为设计高效的神经网络模型提供可解释性依据，为今后能够根据其工作机制设计出对应高效的神经网络模型提供强大的可能性。

本文共分为五章。第一章为绪论部分，主要介绍本文的研究背景、意义以及本文的研究目标。第二章对不同频率信息对模型准确度的影响展开研究，主要研究图像中低频信息与高频信息对模型预测准确性的作用。第三章从频域角度对各数据增强模型鲁棒性进行分析，运用傅里叶热图的分析方式对数据增强模型如何改善模型鲁棒性的问题进行可解释性研究。第四章通过对图像三颜色通道分别进行单通道数据增强，针对单通道下的模型鲁棒性进行分析。第五章总结与展望部分，对本文所获成果进行总结，并对该方向未来可供研究的内容进行可行性展望。

第二章 不同频率信息对模型准确度的影响研究

本章将主要阐述不同频率信息对模型预测准确性的影响。其主要思想为使用神经网络模型进行预训练，对图像进行频率域处理分离出图像的低频及高频信息后作为测试集输入模型，观察模型输出准确率。最终验证得出低频与高频信息对模型准确度的影响。

本章在结构上共分为 5 节，第 1 节介绍本文实验任务及选取的神经网络模型结构及数据集，第 2 节介绍本章所重点实现的任务：分离图像的低频及高频信息，第 3 节展示整个仿真实验所完成的任务并在第 4 节对数据及结果进行分析，最后在第 5 节对本章内容进行总结，并提出下一步的研究方向。

2.1 深度学习图像分类任务

本文将以深度学习中已获巨大成功的计算机视觉领域实际应用问题：图像分类任务为例，选取 ResNet-18 作为神经网络模型，CIFAR-10 作为数据集进行仿真算法实验。

2.1.1 卷积神经网络模型的选取

卷积神经网络（Convolutional Neural Network, CNN）作为人工神经网络的一类，已经成为计算机视觉等领域的研究热点。受生物思考方式启发。其拥有不同的类别层次，各层作用也不尽相同。分别为输入层，对图像进行去均值、归一化、降维等预处理工作。卷积层，提取输入图像的不同特征，通过多层卷积层叠加迭代提取多样化的复杂特征。激励层，主要为激活函数，对卷积层的输出结果进行非线性映射。池化层，进行特征降维，提升模型容错性并减小过拟合。输出层，亦称全连接层，随机剔除神经网络中的一些神经元以预防过拟合，同时进行数据增强、局部归一化等操作以提升模型鲁棒性。常见的卷积神经网络模型有 VGGNet[16]，GoogleNet[17]，ResNet[18]。在本文中，将选用当下高精度、最常用的 ResNet 模型进行可解释性研究。

深度残差网络 ResNet-18 在卷积神经网络领域有着举足轻重的地位，该网络模型的提出也同样刷新了图像识别领域的历史。ResNet 网络的主要贡献为加入了残差单元，以解决堆积网络层数而导致神经网络梯度消失或爆炸等退化问题。

残差学习单元的结构如图 2-1 所示。对于堆积层结构输入x时能学习到的特征$H(x)$，记残差为$F(x)=H(x)-x$,则原始的学习特征为$F(x)+x$。相比于原始特征的直接学习，残差

学习更为容易，当残差为 0 时，堆积层由于完成了恒等映射因此网络性能维持原状，在实际应用中残差大于 0，因此堆积层将从输入特征中学习到新特征，从而拥有更好的性能。

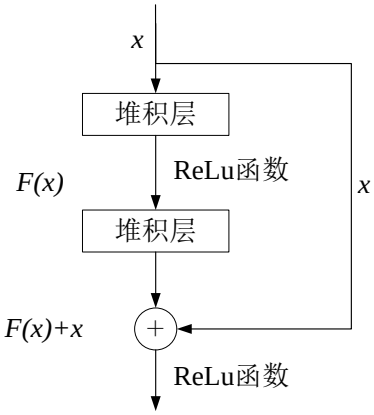


图 2-1 ResNet 网络残差结构

本文仿真实验研究中主要选用 ResNet18 神经网络结构进行实验，其主要参考 VGG 网络结构并在其基础上加入残差单元。其设计原则在于当特征图大小降低一半时，特征图的数量增加一倍以保持网络层复杂度。同时在每两层直接增加短路机制形成残差学习。

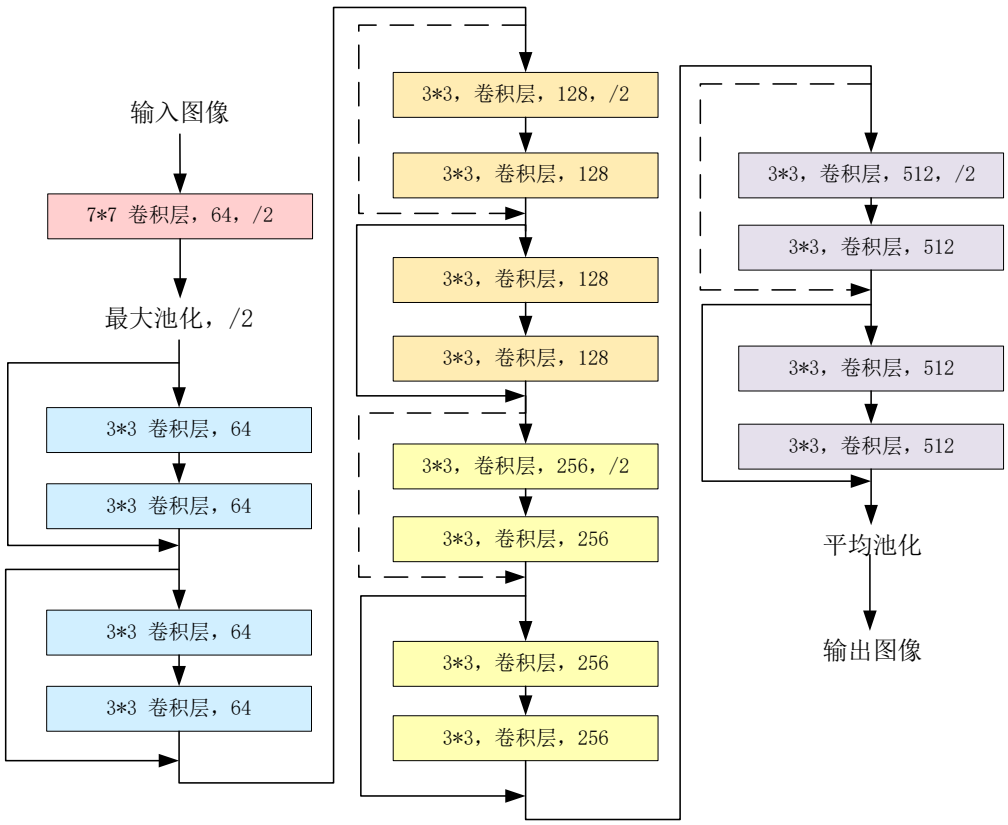


图 2-2 ResNet18 网络结构

2.1.2 图像数据集的选取

本文选用 CIFAR-10[15]数据集作为图像分类任务数据集。该数据集由 50000 张训练图片及 10000 张测试图片构成，图片为 3 通道的彩色 RGB 图像，像素大小为 32×32。该数据集共包含猫(cat)，狗(dog)，马(horse)，鹿(deer)，蛙类(frog)，鸟(bird)，汽车(car)，飞机(plane)，船(ship)，卡车(truck)共十类现实世界中真实物体的图像。

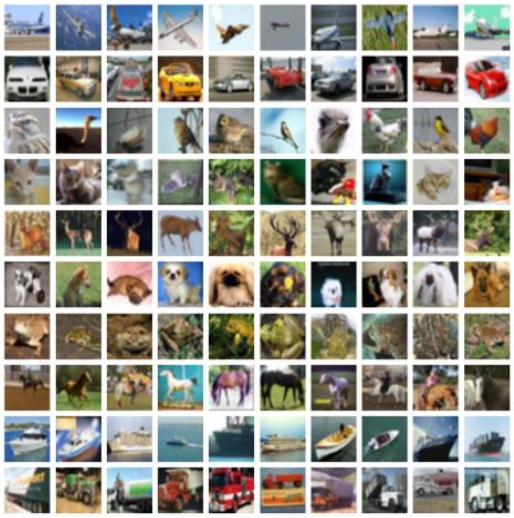


图 2-3 CIFAR-10 图像数据集示例

2.2 频率域处理

图像的频率域处理是本节完成仿真实验设计的重要步骤。该部分主要通过二维傅里叶变换将图像从空域转化为频域，接着通过低通滤波及高通滤波分离出图像的低频及高频部分信息，最后逆变换回空域中即可。通过图像频率域处理步骤将图像中低频信息与高频信息分离，为后续分析不同频率域对模型预测准确性的影响奠定基础。

2.2.1 图像傅里叶变换

通过图像傅里叶变换，可以利用图像与频率成分之间的关系，将在空域上难以解决的问题迁移至频域中解决。首先，需对数据集图像进行二维离散傅里叶变换（Discrete Fourier Transform，DFT），对 M 行 N 列的图像，二维离散傅里叶正变换为：

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)e^{-j2\pi(ux/M+vy/N)} \tag{2-1}$$

由于傅里叶变换的共轭对称性，可由傅里叶变换后的结果继续进行傅里叶逆变换，将频谱图还原回原图像。二维离散傅里叶逆变换为：

$$f(x,y)=\frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}F(u,v)e^{j2\pi(ux/M+vy/N)} \tag{2-2}$$

经过傅里叶正变换后的图像频谱如图 2-4 所示，

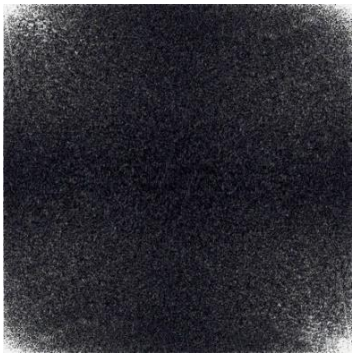


图 2-4 傅里叶变换后的图像频谱

此时经过傅里叶变换的频谱尚未进行中心化处理，低频部分信息分散在(u,v)坐标系四角。为方便后续进行频域滤波处理，需将低频中心位置偏移至(u,v)坐标系图像中心，移频操作如下图 2-5 所示：

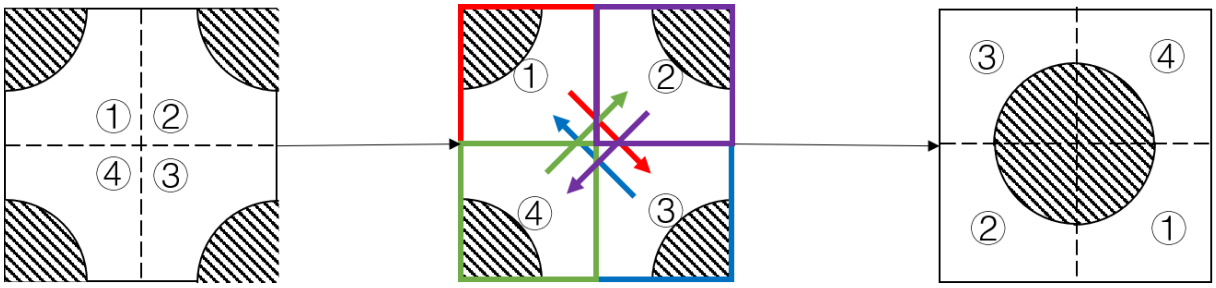


图 2-5 移频操作

经过中心化处理后，图像的低频部分集中于频谱中央，如图 2-6 所示，为后续通过理想低通滤波及理想高通滤波分离低高频信息提供便利。

在仿真实验中，可由快速傅里叶变换(Fast Fourier Transform，FFT)代替离散傅里叶变换(Discrete Fourier Transform，DFT)。FFT 利用了离散傅里叶变换中对称性及周期性的优势，从而缩小了 DFT 的计算量，从 N^2 减少至 $N\times\log_2 N$ 。当 N 的数量越大，则计算优势越明显，运算效率越高。

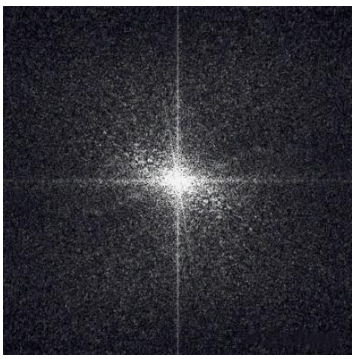


图 2-6 移频后图像频谱图

2.2.2 低通滤波及高通滤波

在对图像进行傅里叶频域处理后，接着能够对图像频谱图进行理想低通滤波及理想高通滤波处理，滤波后通过傅里叶逆变换即可提取出图像的高频及低频信息。理想低通滤波能够将高频信息滤除从而保留频谱中的低频信息，理想低通滤波器公式如下：

$$H(u,v)=\begin{cases}1, & D(u,v)\le D_0\\ 0, & D(u,v)>D_0\end{cases} \tag{2-3}$$

其中，D_0为通带半径，$D(u,v)$为到频谱中心点的距离，可由以下公式进行计算：

$$D(u,v)=\sqrt{(u-M/2)^2+(v-N/2)^2} \tag{2-4}$$

因此，对经过傅里叶变化后的频谱图像内部做一个圆形掩膜，掩膜内部取 1，外部取 0 即可保留图像的低频部分信息。对其进行傅里叶逆变换回空域即可获得图像的低频信息。

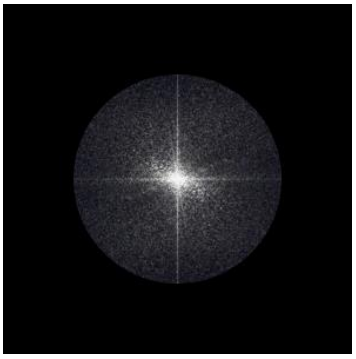


图 2-7 获取图像低频信息

同理，进行相反操作能够获取图像高频信息。理想高通滤波器能够滤除去低频部分信息，保留所需频谱图中高频部分信息，公式如下所示：

$$H(u,v)=\begin{cases}1, & D(u,v)>D_0\\ 0, & D(u,v)\le D_0\end{cases} \tag{2-5}$$

同样对经傅里叶变换后图像内部加入一个圆形全 0 掩膜，即可获得图像高频信息部分，对其进行逆变换后可得空域中图像的高频信息。

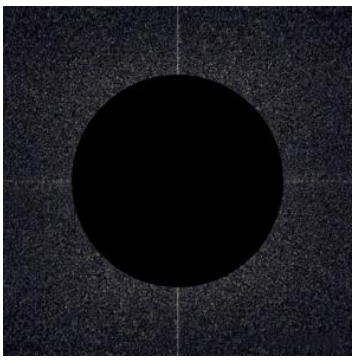


图 2-8 获取图像高频信息

仿真实验中，将图像的低频及高频信息置于可调范围内，调节比例为圆形半径面积与图像$\frac{1}{4}$边长的比值，可调比例大小为$0\sim 2\sqrt{2}$。由图 2-9 图像低频信息示例可得，图像低频信息为图像中平滑区域，主要为色块信息，而图像高频信息则形成了图像的边缘和细节。在低频部分，当圆形掩膜半径较小，即可调比例大小较小时，可知所含低频信息很少，图片呈现模糊色块，随着比例值的增加，低频信息的增大，图片逐渐清晰。

图 2-9 图像低频信息示例

图 2-10 图像高频信息示例

反之，当圆形掩膜半径与图像$\frac{1}{4}$边长的比值从0变化至$2\sqrt{2}$过程中，图像高频信息将逐步减少。由图 2-10 图像高频信息示例从右至左观察可得，可调比例值较大时，图像边缘轮廓几乎不可见，而随比例值减小，高频信息的增大，边缘线条轮廓逐渐明显，最终呈现整幅图片完整样貌。

综上，图像的频率域整体处理整体步骤如图 2-11 所示。

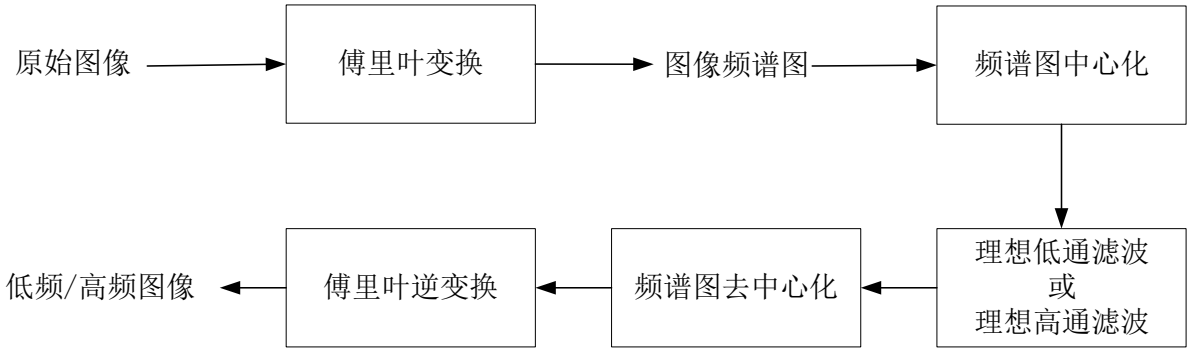


图 2-11 图像频率域整体处理步骤

2.3 不同频率信息对模型准确度的影响分析算法

本仿真实验流程如图 2-12 所示。仿真实验输入为滤波器类型：理想低通滤波器或理想高通滤波器以及所测试的比例，输出为测试集精度。首先，在实验初始化过程中，将生成一个圆形掩膜用于后续对图像进行傅里叶变换并分离低高频信息。接着开始模型的训练过程，使用 CIFAR-10 训练集对 ResNet-18 模型进行训练，提升模型效率，当一轮训练结束后进入一次测试。在测试阶段，首先在加载测试集数据时调用图像频率域处理相关函数，根据输入时的滤波器类型及比例分离出测试集图像所需的低高频信息，将其作为测试集图像输入已经训练完成的 ResNet-18 模型进行精度测试。当测试精度大于目前已保存的最佳测试精度，则保存数据及该模型。重复上述训练及测试步骤直至完成实验设置的 200 轮迭代。最后，创建文本文件，将该滤波器类型及比例下的最佳精度记录于文本文件中以便后续分析。本实验的超参数配置为：训练过程采用 RMSProp 优化算法，批处理大小（batch_size）为 256，共迭代 200 轮，测试过程中批处理大小调整为 100。不断调整低通滤波器及高通滤波器比例，观测不同频率下的低频及高频信息对模型预测性能的影响，具体数值及结果将在下一节详细展示。

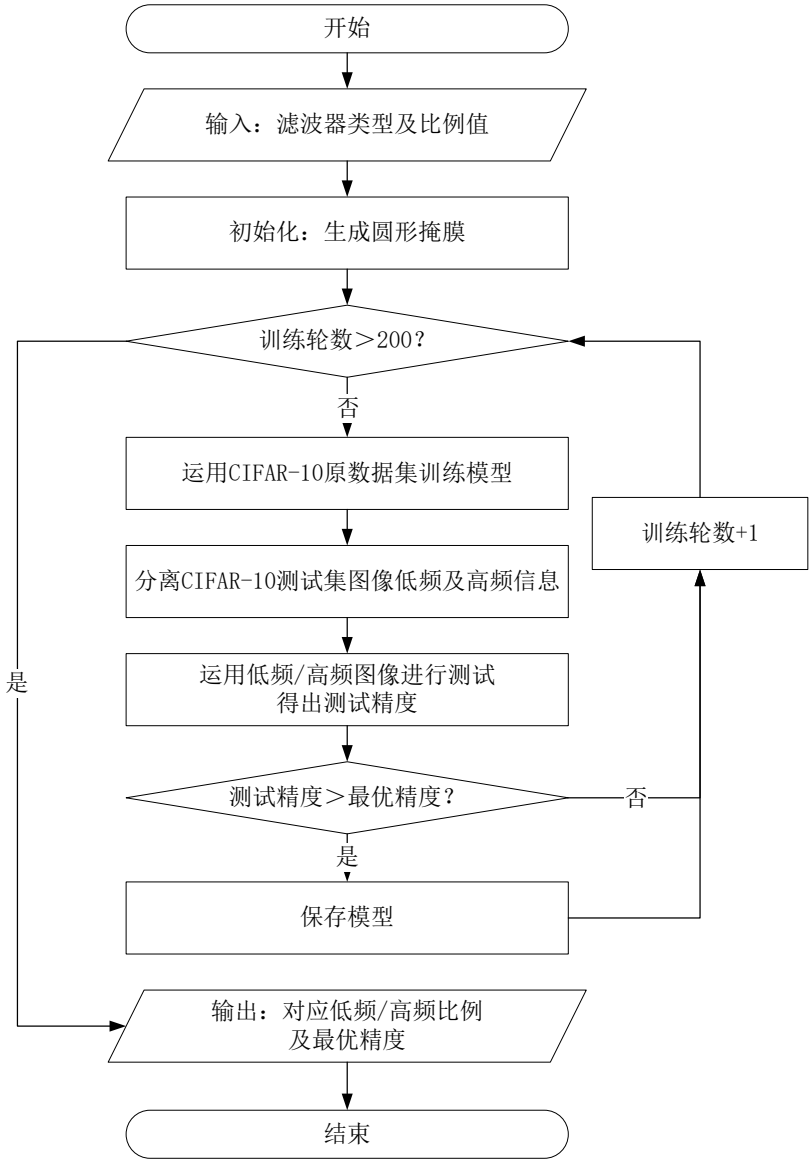


图 2-12 不同频率信息对模型预测准确性算法流程图

2.4 实验结果及分析

本实验共测试不同比例下低频信息及高频信息数据各 20 组。对于低频信息数据来说，由于低频信息集中于频谱图中心部分，因此所输入比例越大，圆形掩膜面积越大，所提取到的低频信息量越大。同时由于输入的可调比例范围为$2\sqrt{2}$，为方便直观感受模型精确度与低频信息比例的关系，对数据进行归一化处理，将比例关系设置为$0\sim1$内。对于高频信息数据而言，由于高频信息分散于频谱图四周，因此需要对高频信息数据进行相关数据处理。由图像频率域处理相关分析可得，高频信息为频谱图像内容减去输入的低频数据，即

$2\sqrt{2}$ –输入比例，随后与低频数据进行相同的数据归一化操作。具体模型预测精度与所输入低频信息比例关系如表 2-1 所示：

表 2-1 输入低频信息及高频信息比例与模型预测精度测试数据

滤波器类型	信息比例	精度(%)	滤波器类型	信息比例	精度(%)
理想低通滤波	0.02	10.00	理想高通滤波	0.01	10.00
理想低通滤波	0.03	10.00	理想高通滤波	0.08	10.02
理想低通滤波	0.04	21.24	理想高通滤波	0.15	10.35
理想低通滤波	0.07	21.27	理想高通滤波	0.26	10.82
理想低通滤波	0.11	22.95	理想高通滤波	0.36	13.65
理想低通滤波	0.14	27.27	理想高通滤波	0.50	14.03
理想低通滤波	0.18	29.58	理想高通滤波	0.58	14.55
理想低通滤波	0.21	31.89	理想高通滤波	0.61	14.30
理想低通滤波	0.28	37.60	理想高通滤波	0.65	14.93
理想低通滤波	0.32	42.39	理想高通滤波	0.68	18.70
理想低通滤波	0.35	53.51	理想高通滤波	0.72	20.38
理想低通滤波	0.42	63.02	理想高通滤波	0.75	32.60
理想低通滤波	0.46	69.47	理想高通滤波	0.79	33.17
理想低通滤波	0.50	69.19	理想高通滤波	0.82	36.83
理想低通滤波	0.57	81.31	理想高通滤波	0.86	45.35
理想低通滤波	0.64	83.39	理想高通滤波	0.90	57.46
理想低通滤波	0.71	86.29	理想高通滤波	0.93	57.72
理想低通滤波	0.78	87.22	理想高通滤波	0.96	60.14
理想低通滤波	0.92	88.50	理想高通滤波	0.97	88.56
理想低通滤波	0.99	88.56	理想高通滤波	0.98	88.56

对上述两项实验所获得的相关数据在 MATLAB 中进行曲线拟合可得，在自然训练的模型中低频信息及高频信息对模型预测准确性的影响如图 2-13 所示。

由数据分析及图像展示可知：

（1） 深度神经网络模型的模型预测性能不仅受到图像中低频信息影响，也受到高频信息影响。

（2） 在自然训练的神经网络模型 ResNet-18 中，当低频信息仅占比 30%时，图像分类模型即可获得 40%至 50%的准确率，说明少量低频信息即可在模型预测中获得较高的准确率。同时，适量增加低频信息能够快速提供准确性，当低频信息占比 50%时，模型预测性能已达到 70%的较高精度，说明低频信息对模型预测准确性做出较大贡献。

（3） 高频信息对准确性的影响则是逐步提升的，当图像中高频信息较少时，模型并不能提供较高的精度，即使高频信息上升到占比 60%时，模型精度仍只有 15%。同时随着高频信息占比的上升，模型的预测性能稳步提升，但高频信息对模型预测性能的贡献小于低频信息。

（4） 由理想低通滤波及理想高通滤波后信息与模型精度的关系可得，经过理想低通滤波后的低频信息再模型预测上的表现优于通过理想高通滤波的高频信息。可验证得出深度学习神经网络模型在训练过程中优先拟合低频信息再拟合高频信息。

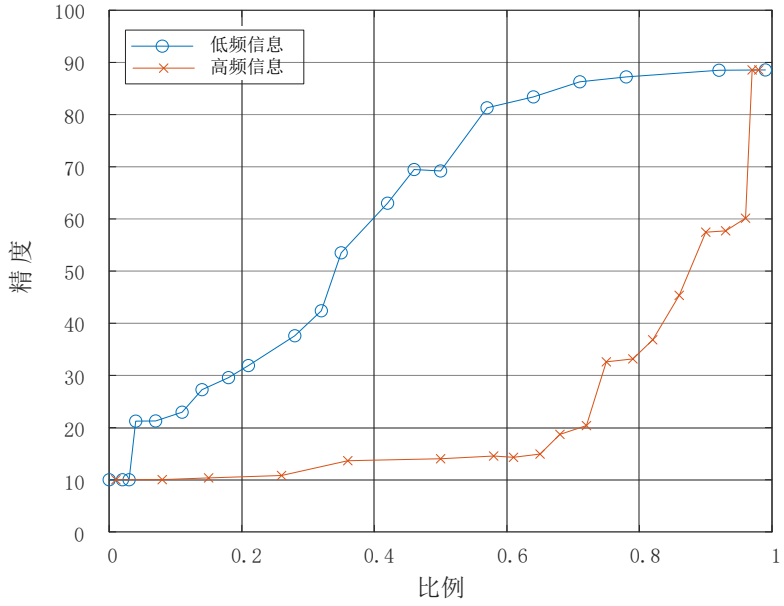


图 2-13 低频及高频信息对模型预测精度影响

2.5 本章小结

本章通过对经过理想低通滤波及理想高通滤波的图像信息进行分析，研究低频信息及高频信息对模型预测准确性的影响。第一节详细阐述了本文实验运用到的神经网络模型 ResNet-18 的网络结构和其在计算机领域的重要意义以及图像分类任务数据集 CIFAR-10。第二节介绍了图像频率域处理用到的相关方法，通过图像傅里叶变化和低通滤波及高通滤

波的方式将图像中的低高频信息分离。第三节将整体实验流程进行具体介绍，方便读者了解实验过程。第四节对实验结果进行处理、分析及总结，得出神经网络模型优先拟合低频信息，再拟合高频信息的结论，对深度学习神经网络模型进行相关可解释性分析。并在本节进行最终总结并提出更进一步的研究方向。

本章通过分离数据集图像中高频及低频信息的相关数据对神经网络模型训练过程有初步了解，阐述如何从频域视角对神经网络模型进行可解释性分析。更进一步的可解释性分析工作将在下一章逐步展开，在下一章，将对不同数据增强模型的鲁棒性性能问题从频域视角进行进一步的分析。

第三章 数据增强提升模型鲁棒性的可解释性分析

本章主要研究内容为通过傅里叶热图的方式对各类数据增强模型如何提升模型鲁棒性进行可解释性分析，主要分为三步，首先，对需要进行可解释性分析的神经网络模型进行 200 轮训练及测试。接着利用 CIFAR-10-C[19,20]数据集测试训练模型的鲁棒性，观测所训练的数据增强模型是否有对于噪声、模糊等相关攻击的能力。最后，将图形变换至频域上，在频域上的每个点随机施加噪声扰动，通过傅里叶热图的形式判断数据增强模型对低频及高频的影响，并对结果进行可解释性分析。

本章共分为五节。第一节对机器学习中的模型鲁棒性定义进行了解释，同时对本实验中测定模型鲁棒性的数据集 CIFAR-10-C 进行了介绍。第二节主要进行原理解释，对傅里叶热图的生成原理进行了详细说明。第三及第四节为实验内容及结果分析，主要完成了高斯增强、几何增强及对抗训练三类数据增强模型的实验，根据模型鲁棒性数据及观察傅里叶热图从频域角度对数据增强方式如何提升其模型鲁棒性进行了可解释性分析。第五节为本章小结，总结本章内容及未来发展方向。

3.1 模型鲁棒性问题

神经网络模型鲁棒性是目前研究的大热问题。在科学领域，鲁棒性可以被定义为组织系统或物体抵御对其不利因素的能力。而在计算机领域中，鲁棒性表示该类算法能够应对不同的干扰，即当轻微改变某些参数或使控制量偏离最优值时，算法是否仍能保持其有效性。稳健统计学[21]中，Huber 提出了从三个层面定义模型鲁棒性。第一，基于机器学习的基本要求，模型需要具有较高的精度及有效性；第二，针对噪声一类对模型假设出现较小偏差的攻击，仅能对算法性能产生较小影响；第三，针对离群点等会对模型假设出现较大偏差的攻击，不能产生对算法性能灾难性的影响。

在本仿真实验中，将采用新数据集 CIFAR-10-C 对 ResNet-18 模型各类数据增强算法的鲁棒性进行检测。该数据集根据 ImageNet-C 数据集[19,20]中相关知识迁移得出。CIFAR-10-C 数据集用于衡量分类器对损害扰动的鲁棒性。该数据集包含对 CIFAR-10 数据集的 15 中扰动类型构成，包含各类噪声、模糊、天气变化等，具体如图 3-1 所示。同时，由于现实中扰动现象的多样性，每种扰动类型都包含 5 个不同严重程度的对图像造成的损害，能够较好的模拟现实中不同现象下训练模型的鲁棒性性能。

图 3-1 CIFAR-10-C 数据集：不同噪声攻击下的图像形态[20]

表 3-1 自然训练的 ResNet-18 模型在 CIFAR-10-C 数据集下的精度测试结果

噪声类型\损坏程度	1	2	3	4	5
高斯噪声	73.30	58.03	43.93	38.23	33.88
散粒噪声	80.45	73.60	55.10	48.51	39.59
脉冲噪声	79.57	71.34	64.66	49.80	38.73
散焦模糊	87.84	84.66	76.68	65.32	45.12
玻璃模糊	43.72	45.01	47.66	36.90	37.74
动态模糊	80.09	71.43	62.82	62.81	55.44
缩放模糊	75.75	72.08	65.40	59.72	50.75
雪	79.89	67.31	72.24	69.69	63.72
霜	82.97	75.82	64.80	85.37	52.73
雾	87.46	85.37	82.19	77.96	62.47
明度变化	88.20	87.46	86.21	85.05	80.82
对比度变化	87.09	80.55	73.69	58.40	25.95
弹性变换	79.87	79.75	74.67	68.24	67.01
像素化	84.77	77.84	72.07	54.58	39.42
Jpeg 压缩	81.28	76.48	74.94	72.74	69.02

以自然训练的 ResNet-18 模型为例，在训练 200 轮的 ResNet-18 模型精度达到 90%的情况下，该模型的鲁棒性测试数据精度（单位：%）如表 3-1 所示。由表中所测数据可得，针对不同类型的图像损害，模型的鲁棒性能表现各不相同，当图像损害较小时，模型仍能取得较高测试精度，而随着损害程度的加大，测试精度逐渐减小，这是由于图像的损害严重程度提升到了破坏图像语义本身信息的地步，导致模型无法对其进行准确预测。

在本章所进行的仿真实验中，将重点关注针对噪声干扰的模型鲁棒性研究。在后续实验中，将主要针对高斯噪声、散粒噪声及脉冲噪声三项噪声攻击展开模型鲁棒性分析。为防止强度过大导致图像语义信息被破坏，因此采用程度最低等级的攻击进行测试。通过使用损害程度 1 的三类噪声扰动，对模型进行鲁棒性测试，并与原模型数据进行比较，判断各类数据增强模型对抵抗噪声扰动的模型鲁棒性影响。

3.2 傅里叶热图分析

在本节中，将引入傅里叶热图[22]的分析方式从频域的角度对神经网络模型进行可解释性分析。傅里叶热图将不同频率域上的点对噪声扰动下的精确度通过不同颜色进行标注，能够简洁直观的感受其不同频率域信息的鲁棒性能。本节第一部分将介绍傅里叶热图的生成原理，第二部分将运用自然训练下生成的傅里叶热图进行示例展示。

3.2.1 傅里叶热图生成原理

傅里叶热图的主要思想为将完成图像分类任务的神经网络模型从频域的角度对其进行相关数据分析，并通过在频域上施加随机值扰动，以研究模型高频与低频信息的敏感性，获取不同模型在噪声鲁棒性上的差异表现。生成傅里叶热图的步骤将分为四个阶段。首先，将生成一个二维傅里叶基矩阵如图 3-2 所示。在第一阶段，运用函数返回二维傅里叶基的矩阵生成器，再对数据集图像进行循环遍历，利用傅里叶基的形式，将图像进行二维傅里叶变换至频域图像上，同时与第二章所述内容相同，在变换至频率域过程中需将低频分量移至频谱中心，以便后续进一步分析。

在第二阶段，在频域内嵌入噪声。在已变换完成的二维傅里叶基矩阵中将噪声嵌入至矩阵。若图像为单通道，则直接嵌入；若图像为三通道，则在所有通道内都嵌入噪声。噪声嵌入参考以下公式：

$$X_{i,j} = X + \gamma \nu U_{i,j} \tag{3-1}$$

其中，γ 表示噪声系数，将于 $(-1,1)$ 内均匀随机进行选取。ν 为扰动的范数，通常 $\nu > 0$ 。

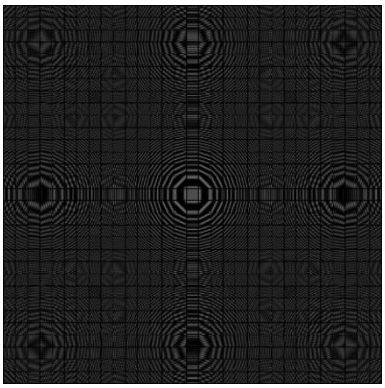


图 3-2 二维傅里叶基矩阵

第三阶段，将加载所需分析的图片数据集及神经网络模型，在本仿真实验中采用 CIFAR-10 数据集及 ResNet-18 神经网络模型，评估给定神经网络模型架构及数据集下的傅里叶热图。首先设置所需要评估的体系结构中错误值的标准，本实验选择分析 top1 情况下的错误率，即完全进行正确分类的错误率大小。接着在给定数据集上进行错误率评估并生成相应的误差矩阵。

在第四阶段，根据上一阶段的误差矩阵生成傅里叶热图，其中，傅里叶热图关于原点对称，因此根据象限 1 及象限 4 的误差矩阵，通过原点进行逆运算，由误差矩阵生成傅里叶热图。当错误率较高时，颜色偏红，错误率较低时，颜色偏蓝。最后，将傅里叶热图保存在设置路径下。

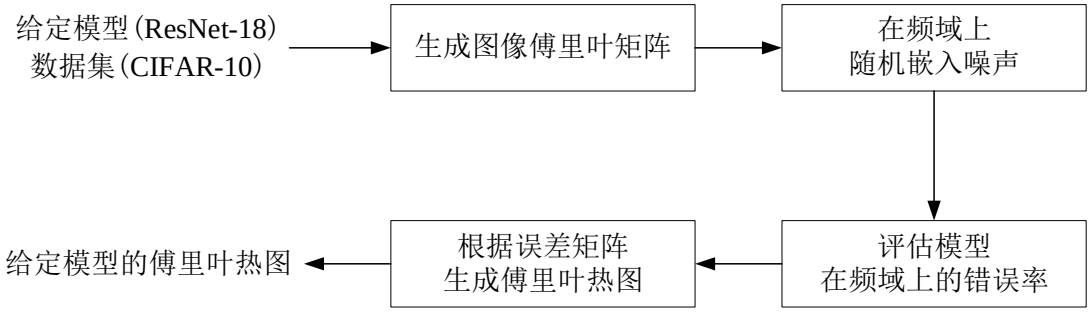


图 3-3 傅里叶热图生成步骤

在本章实验中，将运用该方式对多种数据增强模型进行分析，将在后两节中一一介绍。在实验阶段，首先，对神经网络模型进行训练及测试，确定模型结构、数据集、模型路径、

再对已生成的神经网络模型进行傅里叶热图的评估分析，将生成的傅里叶热图结果保存。

3.2.2 傅里叶热图示例分析

经过测试可得自然训练的模型傅里叶热图如图 3-4 所示。右方图例表示在各点的错误率，颜色偏黄色及红色则表示错误率高于 0.6，错误率较高，鲁棒性较弱；颜色偏蓝色则错误率较低，鲁棒性能较强。在 ResNet-18 原模型上进行测试能够得出，图像中心位置错误率较低，而四周位置错误率较高。根据第二章图像频率域分析可得，在经过傅里叶变换后，图像中心位置为低频部分而四周为高频部分，越向图像边缘部分，频率越高。可见高频区域的敏感性较强而低频区域的敏感性较弱。可得出结论，对于神经网络原模型而言，除对图像的低频信息噪声扰动之外，对其余频域的噪声扰动都非常敏感，鲁棒性能较差。

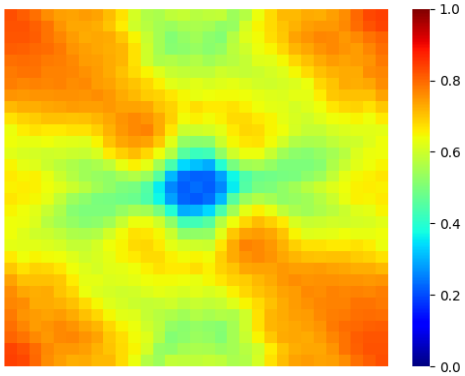


图 3-4 ResNet-18 自然训练模型傅里叶热图

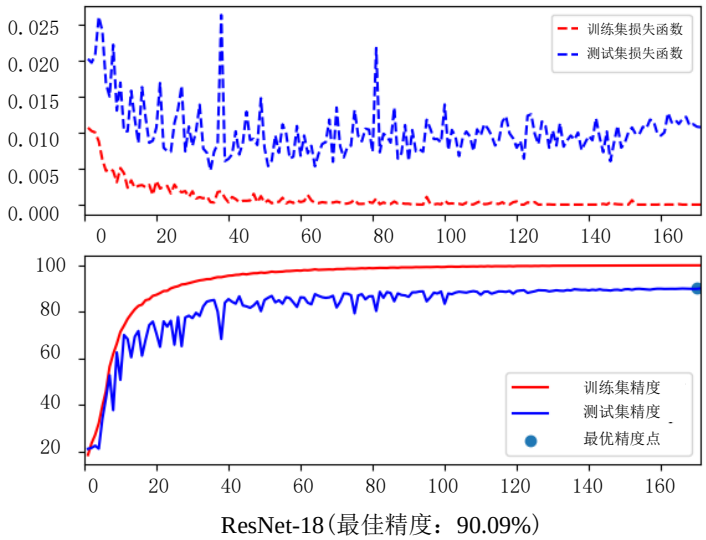


图 3-5 ResNet-18 模型损失函数及其精度变化曲线

在后续两节实验中，将对各模型的傅里叶热图进行分析，观测各类数据增强模型对低频信息及高频信息敏感度的改善程度。在实验中，将对模型本身精确度首先进行测定，模型精确度通过绘制神经网络精度与损失曲线实现，将每一轮的损失函数及测试集精度运用 plot 函数标注，以便直观观测其在 100 轮中的精度变化。如图 3-5 所示，ResNet-18 模型的精度将稳步上升至 90%左右并逐渐趋于平稳状态。在后续实验中，将控制数据增强模型的模型精度控制在 87%以上，保证模型本身性能不受影响的情况下分析模型鲁棒性。

3.3 数据增强模型原理及结果分析

在本节中将主要通过傅里叶热图对几类基础数据增强模型的鲁棒性能进行分析。将分为高斯增强、几何增强，其中包含旋转、翻转、亮度变化，以及高斯增强与几何增强两者进行结合共三类数据增强模型。运用傅里叶热图分析数据增强模型对高频信息及低频信息的敏感度，观测傅里叶热图中高频信息及低频信息的错误率与原模型之间的变化，并运用 CIFAR-10-C 数据集中高斯噪声、散粒噪声及脉冲噪声三类扰动因子分析数据增强模型中噪声部分的鲁棒性能，结合傅里叶热图及鲁棒性能的改善对数据增强模型进行分析。同时，调用神经网络精度与损失函数曲线验证模型本身预测性能是否存在下降过于严重不具备分析参考价值的情况。

3.3.1 高斯增强

噪声通常在图像中表现为孤立的像素点或像素块，能够引起人体的视觉反应。一般情况下，噪声信号与图像本身研究对象并不相关，作为一类扰动因子扰乱图像本身。对于数字图像信号，噪声将作为极大或极小值加减与图像像素真实灰度值中，降低图像质量。同时在神经网络中影响图像的识别分类工作，导致模型精度下降。通过在模型训练阶段将噪声嵌入输入图像中的方式，能够有效的减少过拟合问题，提升模型的泛化性能。由于真实世界中噪声是较为常见的一类问题，嵌入噪声的训练方式能够使得神经网络模型更加鲁棒，以适应现实生活中的各类场景应用。同时在训练过程的图像中嵌入噪声有助于解决小数据集的问题，在训练时增加更多数据，提升模型性能。噪声增强能够嵌入多种噪声，例如椒盐噪声、高斯噪声、泊松噪声等，而在本实验中，选择嵌入高斯噪声进行相关仿真并进行数据分析。

高斯噪声，即其概率密度函数服从高斯正态分布的一类噪声。高斯噪声将会几乎在图

像的每一个点出现噪声大小及噪点深度随机的噪声。高斯噪声的概率密度服从高斯分布，包含平均值及标准方差两个参数，其概率密度函数如下。μ 为平均值，亦称期望值，σ 表示标准差，标准差的平方 σ^2 称为方差。

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \tag{3-2}$$

可调参数为高斯增强的振幅，通过输入值控制整个图像上的噪声强度，控制嵌入图像中的噪声强度不会对图像语义信息造成损害。对于每个输入像素，通过与嵌入值相加即可得到输出像素：

$$P_{out}=P_{in}+F\left(\mu,\sigma\right) \tag{3-3}$$

对数字图像嵌入高斯噪声的处理步骤为，首先设定参数平均值及方差，同时输入高斯增强振幅值，根据其输入像素计算输出像素值，接着对像素值进行处理，对像素值进行缩放处理，将其限制于(0,255)之间，循环所有像素并最终输出图像。如图 3-6 所示，为设置不同嵌入高斯振幅值情况下 CIFAR-10 图像输出结果。在神经网络模型训练过程中，在数据处理步骤时对输入数据集进行高斯噪声增强相关数据处理，由此得出经过高斯增强后的模型，并在后续实验中进行傅里叶热图及鲁棒性能相关检测。

图 3-6 高斯增强图像示例

对已嵌入高斯噪声的数据增强模型基于频域视角进行傅里叶热图分析及模型鲁棒性测试可得结果如图 3-7 及表 3-2 所示。

（1） 测试集精度及模型鲁棒性分析

嵌入高斯噪声后，由于图像本身出现噪声扰动，经数据增强训练后的神经网络模型测试集精度由于在训练过程中掺杂部分噪声导致略有下降，但仍能维持较高水准，不影响整体模型性能。运用 CIFAR-10-C 数据集进行模型鲁棒性检测后发现，高斯噪声、散粒噪声

及脉冲噪声三者在高斯增强模型上的表现都要优于原模型所测得的精确度，可见经过高斯增强的模型在对抗噪声攻击的鲁棒性能上有了较大的提升。

表 3-2 高斯增强模型精度及模型鲁棒性测试

数据增强类型	参数	测试集精确度（%）	鲁棒性（%）		
			高斯噪声	散粒噪声	脉冲噪声
原模型	/	90.09	73.30	80.45	79.57
高斯增强	a=3	88.79	81.94	85.13	86.04
高斯增强	a=5	88.92	86.44	87.75	87.08
高斯增强	a=10	87.47	82.25	85.45	85.47

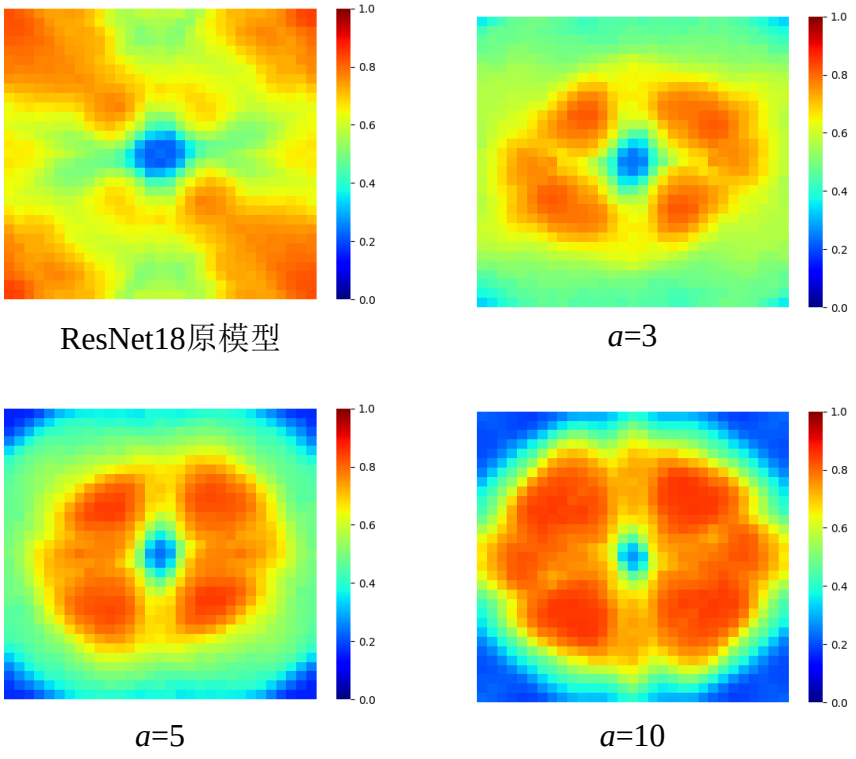


图 3-7 高斯增强模型的傅里叶热图

（2） 傅里叶热图分析

对原模型及高斯增强模型的傅里叶热图进行相关比较分析，观察其中的不同，易得相比于未进行数据增强训练的原始模型而言，高斯增强模型的外围部分错误率有所下降，且随着嵌入噪声值的增加，外围错误率逐渐趋近于 0，而傅里叶热图中间部分原错误率较低

的范围随着高斯嵌入噪声的增多而减少。可见经过高斯数据增强后的模型其高频敏感区域减少，而低频敏感区域增加。

综上，当数据增强模型在针对噪声的鲁棒性性能上有所改善时，可以从频域视角分析得，数据增强改善了图像中高频相关信息的鲁棒性，而图像低频相关信息的鲁棒性却与之相比略微有所下降。

3.3.2 几何增强

在深度学习时代，大量的数据集能够使模型拥有较强的泛化能力，但在实际的工程应用中，采样数据难以覆盖生活中的全部场景，例如图像的光线明暗变化将会造成大量差异，因此需要在训练时进行色度变幻等数据增强。这里所用到的是各类基础数据增强方法，例如旋转、翻转、色度变换等手段。在代码仿真时将使用 transforms 库进行相关处理，transforms 库中集成了大量数据增强函数，比如裁剪(Crop)，旋转(Rotation)，翻转(Flip)，随机隐藏(RandomErasing)，填充(Pad)，转换灰度图像(Grayscale)等。在训练过程中每一次对原始图像进行相关的扰动处理，经过大量轮数的训练后，将等同于模型数据增强的效果。在本节实验中，将主要使用翻转、旋转及色度变换相关几何数据增强手段对模型进行数据增强操作，并对模型鲁棒性及傅里叶热图中高频及低频敏感区域进行相关分析。

翻转及旋转两类操作主要用于将原始图像像素在位置空间上进行相关变换。图像翻转为对原始图像进行镜像操作，包含水平镜像翻转及垂直镜像翻转。在本仿真实验中选择 transforms 库中 RandomHorizontalFlip(*p=x)*，RandomVerticalFlip(*p=x*)分别进行水平镜像翻转及垂直镜像翻转数据增强操作，根据输入概率水平或垂直翻转数据集图像，其中可调参数为 x， $p = x$ 代表图像的翻转概率。

角度旋转操作与图像翻转功能相对，主要用途为沿中心点位置对图像进行任意角度变换。任意角度变换将通过原图像与仿射变换矩阵相乘实现，需计算平移量使仿射变换效果等效于旋转轴画面中心。在图像旋转后，相关图像将会出现黑边，如需减除图像黑边，则对旋转后图像将取最大内接矩阵，最大内接矩阵长宽比与原始图像相同。在仿真实验中，选用 transforms 库中 RandomRotation(*degrees*, *resample=False*, *expand=False*, *center=None*)函数，该函数功能为随机旋转图片，其中可调参数为 *degree*，表示输入旋转角度范围，当输入值为a时，图像将在$[-a,a]$之间进行任意角度旋转，当输入值为(a,b)时，图像将于$[a,b]$

间旋转任意角度。此外，*resample* 代表重采样方法，*expand* 代表是否通过扩大图片以保持原图信息，*center* 表示为以何坐标点为中心进行旋转，默认情况下采用中心点旋转。在本节仿真实验中，采用最基础的旋转功能，除调整角度变化以观察模型精度及鲁棒性变化外，其余参数均使用默认数值。

色度变换主要用于消除在实际应用过程中图像在不同背景下的差异性，光照、对比度的影响通常会影响模型预测的准确性，因此通常进行相关色度变换操作以扩充数据集合。色度变换主要内容为在图像颜色方面进行数据增强操作，例如调整图像的亮度、饱和度及对比度。在 transforms 库中将使用 ColorJitter(*brightness*=x_1,*contrast*=x_2,*saturation*=x_3, *hue*=x_4)函数调整数据集图像的亮度、对比度、饱和度及色相。其中可调参数x_1代表亮度调整因子，当输入值为 *a* 时，图像亮度变换将从$\left[\max\left(0,1-a\right),1+a\right]$中随机选择数值，当输入值为$\left(a,b\right)$时，则从$\left[a,b\right]$中选择参数。*Contrast* 对比度调整因子与 *saturation* 饱和度调整因子的输入值调整与亮度调整因子相同。*Hue* 为色相调整因子，当输入值为 *a* 时，将从$\left[-a,a\right]$中选择任意数值进行图像增强，此时输入值 *a* 的调整范围为$0\le a\le 0.5$；当输入值为$\left(a,b\right)$时，从$\left[a,b\right]$中选择数值，注意此时a,b值的调整范围为$0\le a\le b\le 0.5$。在本节实验中，设定亮度、对比度及饱和度三者的输入值为 0.5，色相输入值为 0.3。

在训练模型数据处理阶段分别进行翻转、色度变换及旋转三类测试，同样对模型精度、对抗噪声鲁棒性及傅里叶热图进行结果分析。模型精度及鲁棒性相关数据如表 3-3 所示。

（1） 测试集精度及模型鲁棒性分析

由表中数据可得运用 transforms 库中的函数对图像进行几何增强相关变换后，较大部分模型精确度维持在 87%至 89%附近。垂直翻转、旋转的测试集精度则出现明显下降，当图像进行垂直翻转及随机旋转 90 度两类增强时，精度下降至 85%以下。这是由于错误的选用了数据增强方式，由于 CIFAR-10 数据集大多为正常拍摄的动物及物体图像，若加入垂直镜像翻转，则会对原始图像产生干扰，导致测试集精度的下降。

运用 transforms 库进行数据增强时，在针对噪声干扰的模型鲁棒性能上并不会有较大提升，在经过高斯噪声、散粒噪声及脉冲噪声处理过后的图像精度均与原模型精度相差无几。分析原因可得图像几何变换所提升的模型鲁棒性能主要针对水平镜像翻转或日常生活中光照色度条件等干扰下，能够维持模型测试精度在较高精度，而并不提升各类噪声扰动下的模型鲁棒性。

表 3-3 几何增强模型精度及模型鲁棒性测试

数据增强类型	参数	测试集精确度（%）	鲁棒性（%）		
			高斯噪声	散粒噪声	脉冲噪声
原模型	/	90.09	73.30	80.45	79.57
水平翻转	*p*=0.3	89.09	76.66	82.24	82.02
垂直翻转	*p*=0.5	84.71	67.11	74.86	73.33
亮度	*brightness*=0.5	88.84	76.11	81.98	82.78
对比度	*contrast*=0.5	88.65	75.44	81.91	81.68
饱和度	*saturation*=0.5	88.36	72.45	79.18	81.59
色相	*hue*=0.3	87.09	75.57	81.04	80.5
旋转	*degree*=45	86.29	71.01	77.58	77.1
旋转	*degree*=90	83.66	59.89	68.14	64.23

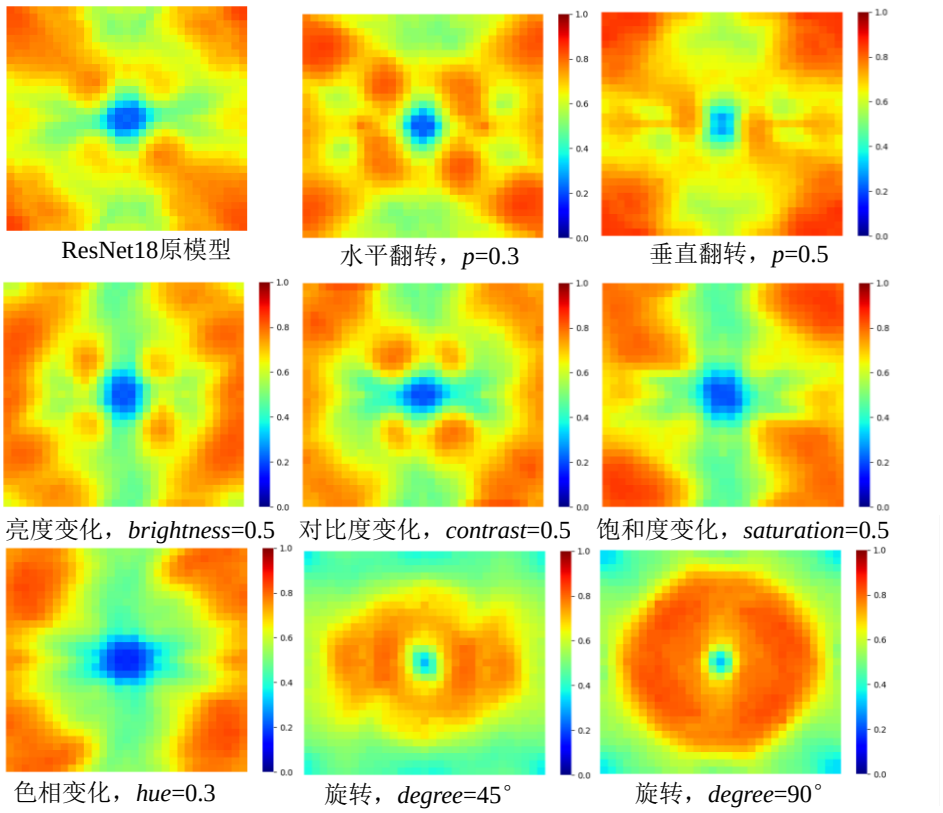


图 3-8 几何增强模型的傅里叶热图

（2） 傅里叶热图分析

观察图 3-8 各数据增强模型的傅里叶热图与原模型对比，翻转及色度变换各图的敏感区域与原模型基本相同，都是高频区域错误率较高，敏感性较强，低频区域的错误率较低，敏感性较弱。在翻转数据增强模型实验过程中，由于翻转的特性使得图像将在水平或垂直镜像上对称，因此其傅里叶热图相比于原模型在 4 个象限上也更为对称。在本实验中，通过旋转进行数据增强的方式是其中的特例。当输入数据集经过旋转后，虽然旋转数据增强方式并没有使得模型鲁棒性获得提升，但其傅里叶热图产生了变化，高频部分的敏感性降低而低频部分的敏感性大幅提升，且随着旋转度数的增加，敏感区域逐渐扩大，猜测由于在旋转这一数据增强过程中，图像的频域空间将会随图像空域的旋转而同时进行旋转，由于 transforms 库中每一张图的旋转角度是在设定范围内随机选择，导致经过旋转后频域图像的位置同样发生了改变，这一点值得在未来研究中采用图神经网络等分析方式对其进行更进一步的实验研究及分析。

3.3.3 高斯增强与几何增强的结合

为进一步验证相关结论，选用高斯增强与几何增强两者结合的方式进一步分析。在数据处理时同时进行高斯增强及几何增强，分析其精度、鲁棒性及傅里叶热图结果。

（1） 测试集精度及模型鲁棒性分析

经测试集测试，当高斯增强与几何增强结合时，测试集精度稳定于 87%至 88%，模型拥有较好的图像分类性能。各模型经 CIFAR-10-C 测试后鲁棒性数据如表 3-4 所示，当两类几何增强模型结合时，噪声扰动下精度低于自然训练下模型，其模型鲁棒性最差。而当高斯增强与其他增强方式结合时，高斯增强将发挥其主导性能，在三类噪声攻击下精度都达到了 80%以上，相比于几何增强模型而言，其模型鲁棒性有了明显提升。

（2） 傅里叶热图分析

通过图 3-9 傅里叶热图对比能够看出，当高斯增强与几何增强相结合时，出现高频区域鲁棒性能明显提升而低频区域的鲁棒性能有所下降趋势。当仅几何增强之间相互结合时，高频与低频敏感区域相比于自然训练模型而言，其整体区域的敏感性都有大幅提升。以垂直翻转与亮度变化两者结合的模型为例，高频区域的敏感区域逐步扩大，外围最高频区域错误率已高于 0.8，低频区域的敏感区域也有所减小。

综上，采用高斯增强的方式将提升模型在噪声扰动下的鲁棒性，且由频域视角对高斯

增强模型可解释性分析可得高斯数据增强模型改善了高频信息的对抗扰动的鲁棒性，但牺牲部分低频信息的对抗扰动的鲁棒性。同时，两类几何模型相结合后，其针对噪声扰动的鲁棒性将比自然训练模型更低，其傅里叶热图高频及低频敏感区域也都出现扩大趋势。

表 3-4 高斯增强与几何增强结合下模型精度及模型鲁棒性测试

数据增强类型	参数	测试集精度（%）	鲁棒性（%）		
			高斯噪声	散粒噪声	脉冲噪声
原模型	/	90.09	73.30	80.45	79.57
垂直翻转与亮度变化	p=0.3, *brightness*=0.5	87.05	61.26	72.17	78.77
高斯增强与水平翻转	a=5, p=0.5	88.07	83.25	86.23	85.28
高斯增强与对比度变化	a=5, *contrast*=0.5	87.26	83.11	85.37	85.00

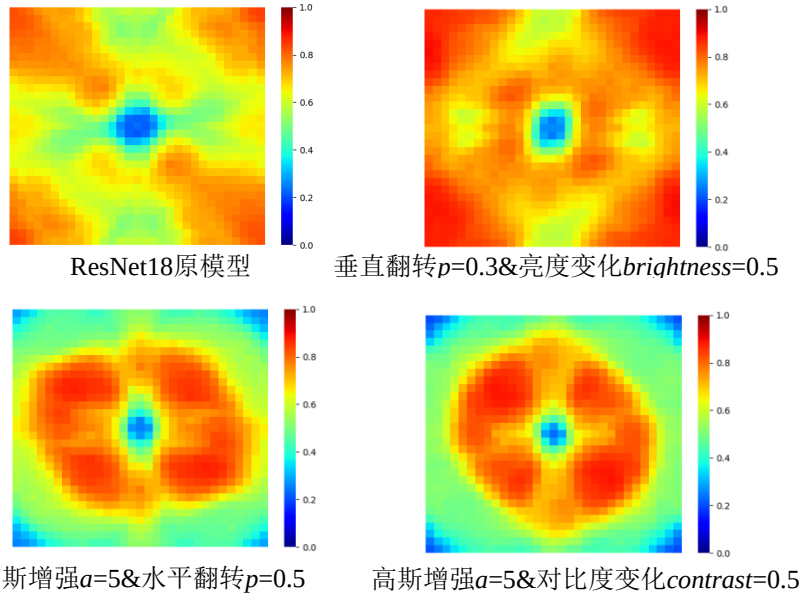


图 3-9 高斯增强与几何增强结合的模型傅里叶热图

3.4 对抗训练模型原理及结果分析

对抗训练[23,24]在深度学习中也是一类增强神经网络模型鲁棒性的重要方式。相比于高

斯增强与几何增强，对抗训练增强方式在原理上稍显复杂，主要采用一类博弈训练的方式，在最大化扰动的同时最小化对抗期望风险。简言之，通过在对抗训练过程中样本掺入一些细微人类无法辨别的扰动，但此类微小的扰动将会导致模型错误分类。在训练时加入相关样本扰动使神经网络适应，从而提升模型鲁棒性。由数学公式角度描述对抗样本，设输入图像数据为 x，分类器为 f，则相对应的图像分类结果表示为 $f(x)$，若存在一个细微扰动 ε 使得满足下式：

$$f(x+\varepsilon) \neq f(x) \tag{3-4}$$

则称 $f(x+\varepsilon)$ 为对抗样本，该对抗样本导致模型发生了相关改变。通过对抗样本训练模型的方式称为对抗训练。

本节将使用经典 FGSM（快速梯度符号方法，Fast Gradient Sign Method）对抗攻击[25]算法对模型进行对抗训练。在模型训练完成后对其同样进行模型鲁棒性测试及傅里叶热图分析，并与前一节中的基本数据增强方式进行对比，进一步验证结论并比较对抗训练的增强方式与高斯增强、几何增强几类数据增强方式之间的不同之处。

3.4.1 FGSM 对抗攻击原理

FGSM（快速梯度符号方法，Fast Gradient Sign Method）对抗攻击方法由 Ian J. Goodfellow 于论文《Explaining and Harnessing Adversarial Examples》中提出。通过 FGSM 算法产生一些攻击样本，通过增加训练时长，使神经网络既能拟合正常数据样本也能够拟合有扰动的对抗攻击样本。FGSM 攻击属于基于梯度的对抗样本生成算法。设图像原始数据为 x，识别结果为 y，叠加的细微扰动为 η，则可用数学公式对其进行表示：

$$x = x + \eta \tag{3-5}$$

修改后图像输入至分类模型中，x 与参数矩阵 ω^T 相乘得：

$$\omega^T x = \omega^T x + \omega^T \eta \tag{3-6}$$

对抗样本在生成过程中目的在于以最细微的修改通过激活函数后对分类结果产生较大变化导致分类结果的错误。通过令扰动 η 的方向沿梯度提升能使损失增大到最大，从而产生较大变化。通过赋予 η=sign$\pi(\omega)$，能够在 η 受到最大范数约束情况下将该增量最大化。当

x 维数为 n，模型参数在每一个维度上平均值为 m，η 无穷范数为 ε 由于每一个维度的细微修改与梯度函数的方向相一致，因此其累计的效果可以用 $nm\varepsilon$ 表示。虽然 $\|\eta\|_{\inf}$ 不会变化，然而随着数据维度的增大、非线性激活函数以及矩阵权重的变换，累计所达到的效果则较大，对分类结果产生较大影响。这成为了对抗样本与原始样本看似相似但产生不同输出的原因。综上，FGSM 对抗攻击产生对抗样本的方式为：

$$\eta=\varepsilon\text{sign}\left(\nabla_x J\left(\theta,x,y\right)\right) \tag{3-7}$$

J 为分类损失函数，通过梯度上升的方式最大化损失函数，使原始输入数据 x 的分类结果最终不属于 y 类。

3.4.2 FGSM 对抗攻击下的模型可解释性分析实验及结果

在训练模型过程中，加入 FGSM 对抗攻击的训练方式，通过改变输入参数值 epsilon 调整加入细微扰动的大小，并进行模型鲁棒性测试集傅里叶热图观测。

（1） 测试集精度及鲁棒性分析

由表 3-5 可见，当模型经过对抗训练数据增强方式后，其针对噪声攻击的模型鲁棒性有明显提升，对于具有噪声扰动的图像测试集，精度仍能到 85%以上，在噪声扰动下，对抗训练模型表现非常优异。同时需要注意 FGSM 对抗攻击中 *epsilon* 参数的选取，能够看到当 *epsilon*=0.01 时，本身模型精度高且鲁棒性表现优异。而当 *epsilon* 增加至 0.02 时，由于模型有可能造成过拟合，因此模型精度和鲁棒性表现都有所下降。

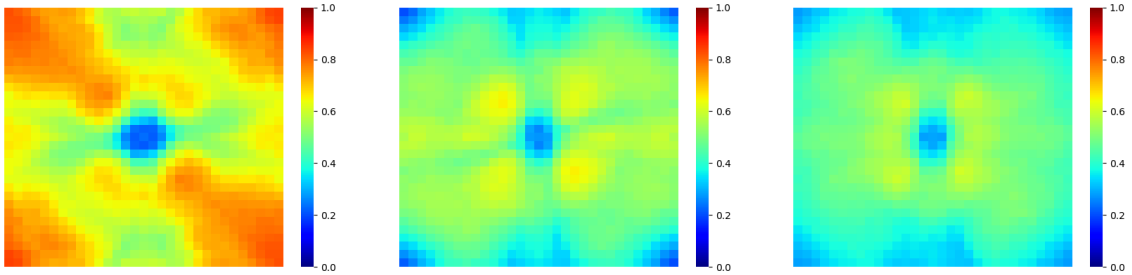
表 3-5 FGSM 对抗攻击模型精度及模型鲁棒性测试

数据增强类型	参数	测试集精确度（%）	鲁棒性（%）		
			高斯噪声	散粒噪声	脉冲噪声
原模型	/	90.09	73.30	80.45	79.57
对抗训练	*epsilon*=0.01	89.11	87.16	88.39	85.16
对抗训练	*epsilon*=0.02	86.98	85.83	86.45	84.00

（2） 傅里叶热图分析

如图 3-10 所示，除最中心极小位置低频部分其错误率提高，其余位置精度都有了较大提升。高频区域的敏感区域有大幅缩减，在高频位置，其精度都有大幅提升。与前一节

结论相结合，验证得出自然训练下的模型对除了较低频率外的所有频率所加入的加性扰动都高度敏感，而经过对抗训练后提升对抗噪声鲁棒性能的模型在高频区域上的加性扰动敏感程度都大幅下降，相比与自然训练的模型而言，仅在最低频率处出现了一个微小的鲁棒性能缺失。



ResNet18原模型　　*epsilon*=0.01　　*epsilon*=0.02

图 3-10 FGSM 对抗攻击的傅里叶热图

3.5 本章小结

本章主要通过使用 CIFAR-10-C 数据集对相关数据增强模型进行鲁棒性检验，主要针对噪声扰动进行分析。同时运用傅里叶热图的分析方式从频域视角对数据增强模型进行可解释性分析，得出数据增强模型提升了高频信息的鲁棒性能，但在低频部分会造成微小的鲁棒性能缺失。本章第一节介绍了一个新数据集 CIFAR-10-C，该数据集通过在 CIFAR-10 数据集基础上进行衍生，增加雨雪等天气、各类噪声及模糊扰动，用于判断各类训练后的模型鲁棒性。第二节讲述了傅里叶热图的生成原理，即介绍了从频域视角对模型鲁棒性进行可解释性分析的方式，并展示了自然训练的 ResNet-18 模型傅里叶热图及调取损失函数及精度图的方式，为后续仿真实验分析做准备。第三节开始着手对各类数据增强模型进行模型鲁棒性及傅里叶热图相关分析，对高斯增强、旋转、翻转及各类数据增强模型进行原理解释，并进行相关仿真实验，得出高斯增强相比于其它增强而言，在对抗噪声扰动的模型鲁棒性性能上有极大突破，在傅里叶热图上反应出该数据增强模型提升了高频区域的鲁棒性能但在最低频域处有略微的性能下降。第四节运用更为复杂的对抗训练模型分析验证结论，首先介绍经典 FGSM 对抗攻击模型原理，并运用该模型测试得出对抗训练在噪声干扰下的鲁棒性表现优越，且傅里叶热图在除最低频外的其余区域鲁棒性表现都十分优越，仍然仅仅在最低频域位置有一微小的鲁棒性缺失。最终在第五节中进行相关总结并提出下一章研究目标及计划。总而言之，高斯噪声及对抗训练两类数据增强模型，在噪声扰动下

其模型鲁棒性都有优异表现，两者在傅里叶热图上皆提升了集中在高频区域的鲁棒性，降低了集中在低频区域的鲁棒性。

在下一章内，将对高斯数据增强及对抗训练两类在噪声扰动的情况下有优越模型鲁棒性表现的模型进行三颜色通道可解释性分析。了解更进一步情况下模型对图像不同通道的学习情况，解密神经网络的黑箱模型。

第四章 图像三颜色通道的模型可解释性分析

本章将在第三章基础上进一步延伸，分离图像的红(Red)、绿(Green)、蓝(Blue)三颜色通道，并对图像中的 R、G、B 三通道单独进行模型的可解释性分析。本章也将运用一个新的评价指标结构相似性指标[26](Structural Similarity，SSIM)。对图像三颜色通道分别进行数据增强操作后得到的傅里叶热图与自然训练的模型和对所有通道进行数据增强的傅里叶热图分别进行对比分析。

本章第一节将对图像三颜色通道进行介绍并给出在模型训练过程中对图像三颜色通道进行单通道数据增强模型训练的方法。第二节介绍图像结构相似性指标原理，并将其在 MATLAB 软件中进行仿真。第三节分别对高斯增强及 FGSM 对抗攻击模型进行单通道增强并分析傅里叶热图，计算 SSIM 值并进行相关分析。第四节总结本章内容。

4.1 图像三颜色通道

彩色图像通常能够表示为三颜色通道图，即每一个像素点都能用 3 个值对其进行表示，生活中常用的彩色图像也为三通道图像，分为红(Red)、绿(Green)、蓝(Blue)三通道。RGB 色彩为工业界的颜色标准之一，通过对 R、G、B 三个颜色通道相互叠加以获得生活中的多种颜色。RGB 标准包含人类视力范围内能够感知到的所有颜色，也是目前运用最广的颜色模式之一，根据其基于发光的模式，多用于面向硬件的色彩模式，例如显示器、彩色视频摄像等发光设备中经常使用 RGB 模式作为其色彩模式。RGB 模式通过三原色合成的颜色其亮度为三色亮度组合，因此混合原色越多则亮度越强。在图像中 RGB 三通道为(0,0,0)时亮度最暗，为黑色，而当 RGB 三通道为(255,255,255)时亮度最亮为白色。

在本实验中，在高斯增强的数据处理阶段，对作为训练集输入至模型中的图像进行 R、G、B 三通道分离。导入 numpy 库，运用 np.stack 函数，保留 R、G、B 所需的一个通道并运用 zero 函数将另外两通道置零，由此所添加的高斯噪声或 FGSM 对抗攻击中的梯度反向传播函数将只作用于指定一个通道中进行训练，为后续实验结果分析做准备。

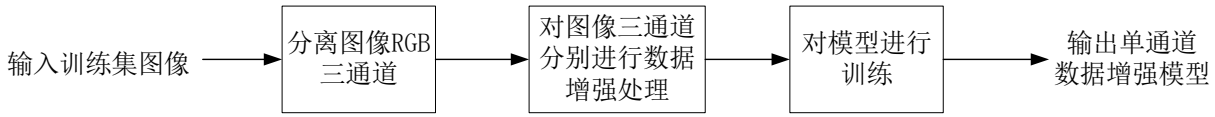


图 4-1 对单通道进行数据增强步骤

4.2 图像结构相似性评价指标

对于不同在不同通道内进行数据增强后进行傅里叶热图分析时，同时引入一项新的分析指标，即结构相似性指数[26]（Structural Similarity Index Measure, SSIM），该指标用于度量两个给定图像间的相似性，当两个图像相似度越高，则其指数越高。结构相似性指标主要提取图像中 3 个关键特征：亮度、对比度及结构。对于输入图像 x 及 y，首先对两张图像进行亮度测量及计算并进行比对，得到第一项结构相似性评价指标。接着减去亮度评价指标的影响，测量并计算第二项对比度指标并进行比较。最后除去第二项对比度指标的影响，进行第三项结构评价指标的对比。三项指标测量对比完成后，综合数据并得出最后的评价结果。

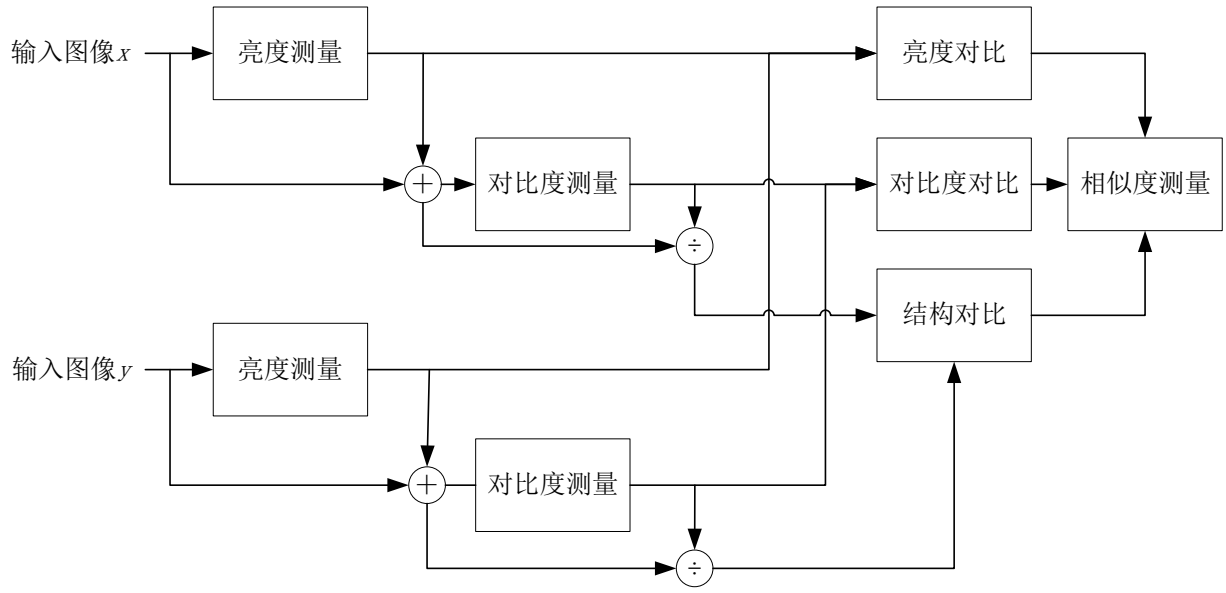


图 4-2 结构相似度 SSIM 计算流程

在实际应用中，常用高斯卷积的方式计算图像均值、方差及协方差已获得更高的效率。其中，对图像亮度部分进行比较的测量公式为：

$$l(x,y)=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1} \tag{4-1}$$

对比度测量公式为：

$$c(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2} \tag{4-2}$$

结构测量公式为：

$$s(x,y)=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3} \tag{4-3}$$

其中，亮度公式本质为μ_x与μ_y的函数，对比度公式本质为σ_x与σ_y的函数，结构公式本质为σ_{xy}与σ_x，σ_y的函数。$\sigma_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)$。最终结构相似性公式如下：

$$SSIM(x,y)=[l(x,y)]^{\alpha}\cdot[c(x,y)]^{\beta}\cdot[s(x,y)]^{\gamma} \tag{4-4}$$

α，β，γ三个参数用于表示三类模块的重要性。为简化公式，令α=β=γ=1且$C_3=C_2/2$。经简化后的公式为：

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{4-5}$$

在 MATLAB 中对所需分析的图像进行结构相似度指标的计算。首先在输入所需分析的图像后筛选图片是否符合计算条件，接着进行预处理，对图像进行采样并做归一化处理。随后运用公式求C_1和C_2值，使用高斯滤波器进行滤波，求两张图像的均值μ_x及μ_y，同时求协方差期望，σ_x，σ_y及σ_{xy}。对于C_1和C_2皆大于 0 的情况，则直接可用简化公式求 SSIM 值，若C_1和C_2值并不满足上述情况，则需单独计算并得出最终结果。

4.3 实验结果与分析

在本实验中，首先在数据处理阶段分离图像三通道，在 R、G、B 三个通道中分别进行高斯增强与对抗训练单通道数据增强。获取对应单通道数据增强模型后，对 R、G、B 三个单通道数据增强模型分别进行数据分析。具体分析内容包含：经训练后模型本身精度，模型针对噪声扰动的鲁棒性检验，傅里叶热图分析，同时包含在 MATLAB 中将单通道数据增强模型的傅里叶热图与自然训练模型及经由三通道同时增强的模型的傅里叶热图进行比较并计算 SSIM 值。具体分析步骤如图 4-3 所示。

本仿真实验采用 ResNet-18 模型对其三通道分别进行高斯增强及对抗训练实验。具体模型测试集精度及模型鲁棒性表现如表 4-1 所示，数据增强模型及 FGSM 对抗攻击模型傅里叶热图如图 4-4 及图 4-5 所示。

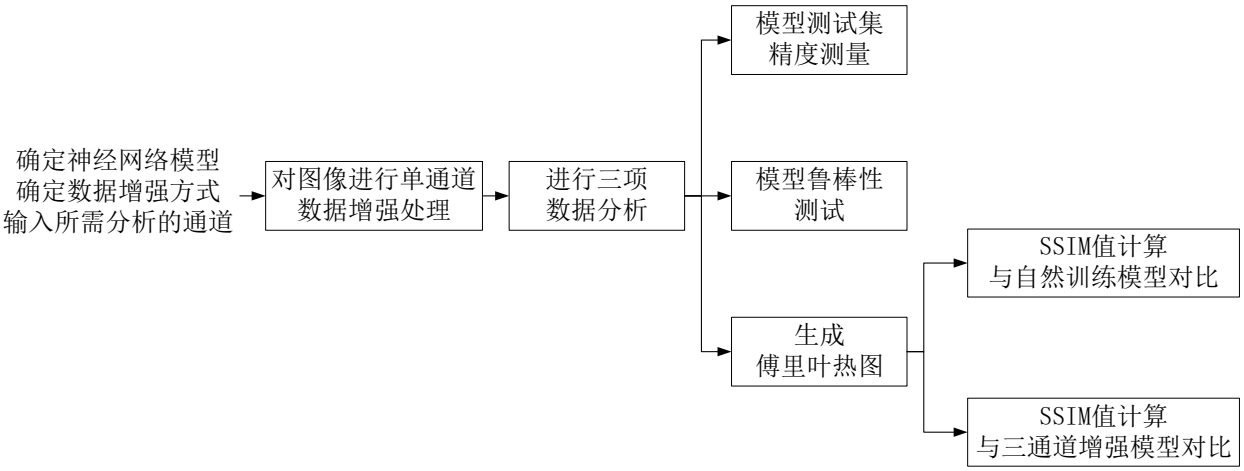


图 4-3 RGB 三通道的模型可解释性分析步骤

（1） 高斯增强

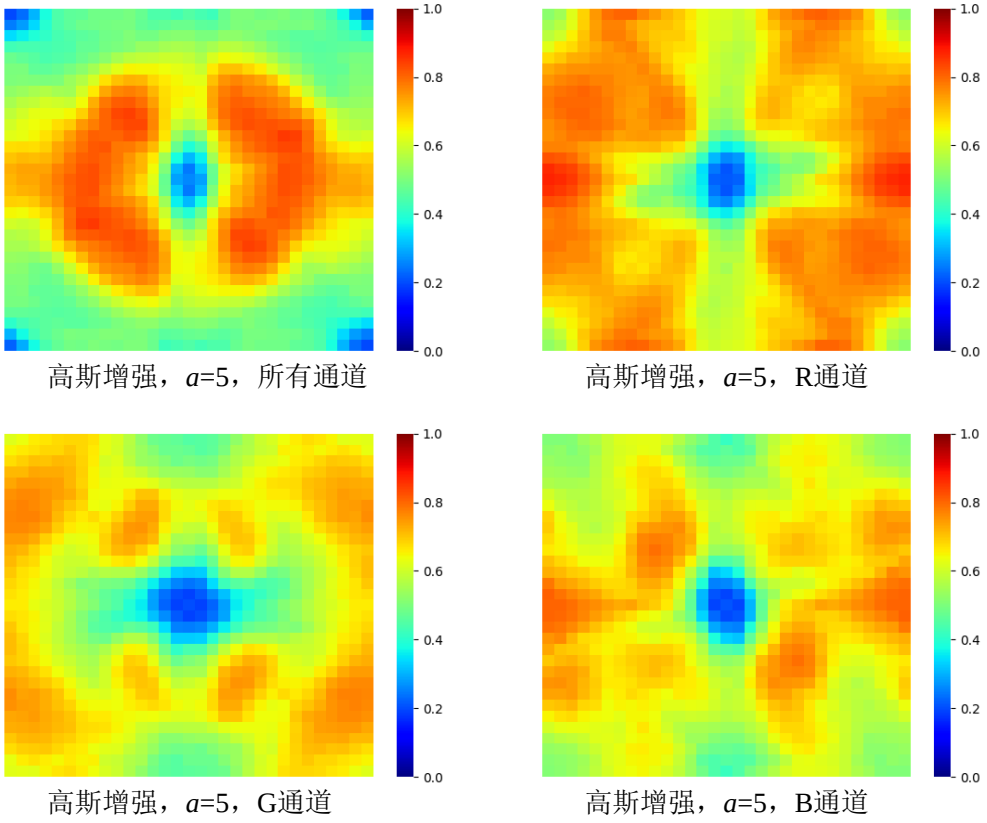
在高斯增强方式中，模型测试集精度不变情况下，分别对 R、G、B 通道单独进行高斯噪声增强后的模型鲁棒性表现都劣于在三颜色通道内同时嵌入高斯噪声的结果。尤其在高斯噪声及散粒噪声扰动下，仅对单通道进行数据增强后运用 CIFAR-10-C 数据集测得精度较低，模型鲁棒性性能差之甚远，可见高斯增强的优异表现是由 R、G、B 三通道共同作用下的结果。通过傅里叶热图及 SSIM 值对比同样能够验证该结论。在 R、G、B 进行单通道增强时，傅里叶热图敏感区域分布更接近于自然训练模型的傅里叶热图敏感区域分布：高频敏感区域较多而低频敏感区域较少。通过 SSIM 值对比可得单通道下数据增强模型傅里叶热图与自然训练模型所得傅里叶热图相似度更高，且三通道的表现更为平均。

（2） 对抗训练

在 FGSM 对抗攻击下，单独对 R 与 G 通道进行单通道增强都获得了较好的模型鲁棒性。在测试集精度相当的情况下，R、G 两通道在噪声扰动下的模型鲁棒性与直接对图像进行 FGSM 对抗攻击训练的表现相当。在傅里叶热图及 SSIM 值对比上亦可看出 R、G 两类单通道增强的傅里叶热图在结构相似度上已十分接近 FGSM 对抗攻击的图像三通道增强。相较而言，图像 B 通道数据增强所获取的效果则较弱，模型鲁棒性表现较差且傅里叶热图与自然训练模型相比缺少明显高频敏感区域的变化，该傅里叶热图与 FGSM 对抗攻击三通道增强模型傅里叶热图的结构相似度也较小。可见在 FGSM 对抗攻击中，图像 R、G 两通道在数据增强改善模型鲁棒性性能上作用更强，在频域上也同样体现出 R、G 两通道提升高频区域鲁棒性并牺牲部分低频区域鲁棒性的特性。

表 4-1 RGB 单通道数据增强的模型精度及模型鲁棒性测试

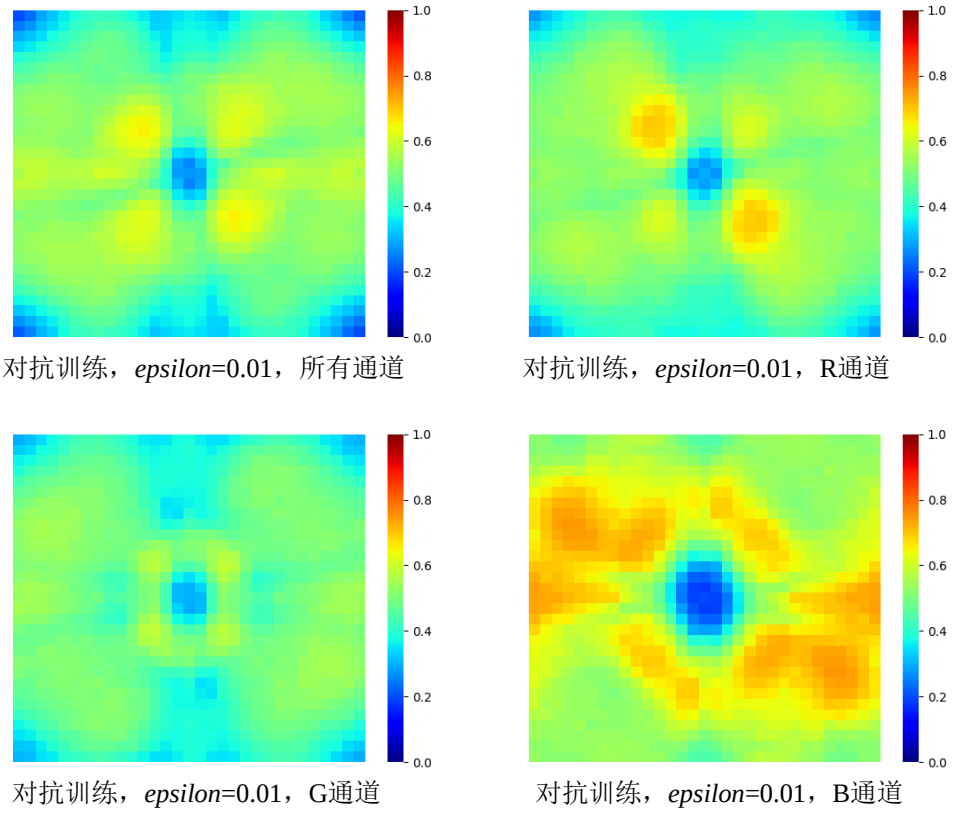
模型名称	参数	通道	测试集精度（%）	鲁棒性（%）			SSIM与三通道对比	SSIM与原模型对比
				高斯噪声	散粒噪声	脉冲噪声		
高斯增强	*a*=5	all	88.92	86.44	87.75	87.08	1.0000	0.9023
高斯增强	*a*=5	R	89.98	66.43	75.56	82.89	0.8658	0.9070
高斯增强	*a*=5	G	90.26	69.07	78.69	83.09	0.8811	0.9341
高斯增强	*a*=5	B	90.13	63.22	74.87	81.14	0.8944	0.9245
对抗训练	*epsilon*=0.01	all	89.11	87.16	88.39	85.16	1.0000	0.9470
对抗训练	*epsilon*=0.01	R	87.85	85.58	86.76	84.10	0.9858	0.9550
对抗训练	*epsilon*=0.01	G	88.11	85.72	86.91	85.16	0.9845	0.9471
对抗训练	*epsilon*=0.01	B	90.28	66.17	76.16	81.49	0.9382	0.9293



高斯增强，*a*=5，所有通道　　高斯增强，*a*=5，R通道

高斯增强，*a*=5，G通道　　高斯增强，*a*=5，B通道

图 4-4 单通道高斯增强模型的傅里叶热图



对抗训练，*epsilon*=0.01，所有通道　　对抗训练，*epsilon*=0.01，R通道

对抗训练，*epsilon*=0.01，G通道　　对抗训练，*epsilon*=0.01，B通道

图 4-5 单通道 FGSM 对抗攻击模型的傅里叶热图

4.4 本章小结

本章针对图像的 RGB 三通道进行相关分析。延续上一章对数据增强改善模型鲁棒性的可解释性分析，同时引入 SSIM 指标进一步对分别作用于 R、G、B 三个不同通道内的数据增强模型进行相关模型可解释性研究。相较而言，在高斯增强中，R、G、B 三通道对于模型鲁棒性能的贡献程度大致相当，当三通道同时增强时，方能展现优良的模型鲁棒性。而对抗训练过程中，相较于 B 通道而言，R、G 两通道的表现更为优异，在对抗训练过程中对模型鲁棒性能的提升贡献更大。同时，在本章通过 FGSM 对抗攻击实验中 R、G 两通道模型鲁棒性的优越表现及傅里叶热图分析进一步验证得出数据增强模型主要改善了高频信息鲁棒性，但降低了部分低频信息的鲁棒性。

本章第一节对图像 RGB 颜色空间进行简要介绍并给出在仿真实验中分离 RGB 三通道的方式。第二节针对本章新引入的分析指标 SSIM 进行了相关介绍并简述其原理。第三节展示了对图像 RGB 三通道进行数据增强的模型可解释性分析的流程及方法，并给出最终的实验数据及结论。第四节总结整章内容并进行总结归纳。

第五章 总结与展望

神经网络的可解释性模型通过探寻深度学习相关模型内部的工作原理及机制，尝试解密其黑箱模型。当了解神经网络模型内部的工作原理，便能够理解其做出决策的原因并建立良好的人机沟通机制。相比于传统的以结果为导向不断试错并获得最终具有优越表现的模型结构及激活函数，通过对已有神经网络模型进行可解释性研究分析后能够更好地了解其优势及劣势所在并更好对模型进行优化改进，以高效地取得模型精度及效率的提升。

5.1 总结

本文以神经网络的频域原则为基础展开分析，从频域的视角探寻模型内部特性。从傅里叶变换的角度着手，对低频部分信息及高频部分信息展开仿真实验，以推测图像中低频及高频信息与模型预测准确性和数据增强改善模型鲁棒性之间的关系。

本文紧扣当前人工智能深度学习的潮流，并从傅里叶视角对深度神经网络内部机制进行分析。比起大量不断优化的深度神经网络模型来说，当下，仅有少部分研究着手关注其模型内部工作机制。然而，一旦深度学习本身工作的“黑箱”模型被破解，今后关于神经网络模型的研究将能够以此为依据展开研究及改进，开发出更高效的模型。

本文的具体工作如下：

（1） 运用图像的频率域处理相关知识分离测试集图像中的低频及高频信息，得到图像的低频信息为色块等平缓信息，高频信息为边缘处轮廓信息。将图像的低频及高频信息设置比例大小作为测试集输入已训练完成的 ResNet-18 神经网络中进行精度测量。能够得出不同频率信息对模型预测准确性的影响，即高低频信息都会对模型预测产生贡献，其中低频信息对模型预测的贡献更大。并且神经网络在模型拟合中将优先拟合图像的低频信息部分再拟合图像的高频信息部分。

（2） 运用 CIFAR-10-C 数据集对高斯增强、旋转、翻转、色度变换及对抗训练几类数据增强的模型鲁棒性表现进行分析，主要针对其噪声扰动下模型鲁棒性进行数据的对比研究，得出高斯增强及对抗训练两者在数据增强过程中能够改善模型针对噪声扰动的鲁棒性能。相较而言，旋转、翻转、色度变换等则在噪声扰动下表现平平，说明该类增强方式可能在亮度对比度变换下有一定的鲁棒性能提升，但并无针对噪声扰动的鲁棒性能提升。

（3） 运用傅里叶热图的方式对数据增强改善模型鲁棒性进行可解释性分析。在频域

上对图像施加噪声扰动，通过测量低频及高频信息在噪声扰动下的精度由此判断不同频率域下的模型鲁棒性。得出结论，在低频区域，敏感程度较弱，鲁棒性较强；高频区域，敏感程度较强，鲁棒性较弱。

（4） 运用高斯增强、几何增强及对抗训练三类数据增强模型分别进行傅里叶热图测试，能够看出高斯增强及对抗训练两类模型的傅里叶热图高频敏感区域大幅减少而低频敏感区域有细微提升。几何增强后的傅里叶热图则与原图整体分布大致相同。得出结论针对噪声扰动的数据增强模型改善了高频部分的模型鲁棒性，但同时也会牺牲一小部分在低频部分的模型鲁棒性。这也与低频信息在图像分类过程中起到较大作用，因此更难以提升其性能有关。

（5） 将图像拆分为 R、G、B 三通道，对三通道内分别进行数据增强，并与自然训练模型和数据增强模型进行结构相似度对比，探寻其中不同。相较而言，单独对三通道进行高斯增强训练后其针对噪声的鲁棒性能都不如直接进行高斯增强训练，三通道对高斯增强模型鲁棒性提升贡献大致相当。而对抗训练过程中 R 与 G 通道对提升模型鲁棒性的贡献明显比 B 通道更大。

5.2 展望

尽管在本篇论文中，已从频域角度对神经网络模型在图像分类上进行了一定的可解释性分析，并探寻出模型与图像低频及高频信息的关系，对于这一课题，仍具有较大的研究前景，包含大量可待继续研究的方向。接下来，我将针对该课题未来能够进行研究分析的方向提出以下一些思路供以参考。

（1） 在针对几何增强类进行研究时，发现经过随机角度旋转后的图像在针对噪声扰动的鲁棒性性能上并没能有所突破，但其在傅里叶热图上却呈现出高频敏感区域减少而低频敏感区域增加的特征，与目前所得结论不符。猜测该结果可能由于在增强过程中通过旋转进行数据增强时，频域空间将会随空域空间的旋转而同样进行旋转，导致在频域当中其本身的位置产生了偏移。后续或许可使用图神经网络对其进行进一步的分析研究。

（2） 对抗训练作为人工智能中非常重要的一环，具有很大的研究价值及意义。在本文中，仅仅采用了 FGSM 攻击一种对抗攻击模型对模型鲁棒性展开分析。在未来发展中，能够运用傅里叶分析的理论对对抗样本进行模型的可解释性分析，例如

在对抗样本的扰动上增加一个滤波器将扰动限制于某一低频或高频频率段中，并分析网络模型对该频率段扰动的敏感性。

（3） 在对图像分类领域神经网络模型的可解释性分析有一定掌握后能够将该方法迁移至其它领域进行相关研究。例如能够利用频域分析的方式对隐写分析、数字取证领域进行相关可解释性研究。同时，相较于图像分类任务而言，隐写分析领域由于隐写信号十分微弱，在采用傅里叶变换以外，能够尝试运用离散余弦变换（Discrete Cosine Transform，DCT）等将其变换至变换域，对隐写分析领域的神经网络模型进行可解释性研究。

致谢

时光蹁跹，大学的四年已至尾声。虽然疫情阻隔，大二下的半年和临近毕业的这半年都失去了大学原本的色彩，但回望整个大学生活，在面临诸多困难时也夹杂着许多欢笑。在四年的时光里，身边的老师、朋友都为我提供了许多的帮助，让我能够越来越好。

首先，我要感谢上海大学通信与信息工程学院为我本科生活提供了非常优良的学习平台，以及在过去三年中授课的所有老师，为我在未来发展中打下了扎实的基础。

其次，感谢吴汉舟老师以及组内研究生刘勇学长、柳琦云学姐及李晨学长在我整个毕设实施推进中的指导。能够遇上吴老师担任我的毕设指导老师是我非常大的荣幸，不仅在留英时作为我的推荐老师撰写推荐信，同时在毕设期间关注我的毕设进展，及时进行相关指导工作。课题组的各位学长学姐也会为我答疑解惑，对我帮助良多。同时，也感谢企业导师叶净宇对我的帮助与指导。

同时，感谢我的室友们，在一起相处的三年里，大家共同努力，不断奋进，也会互相排遣考试周压力大时心中的焦虑与难过。那些一起熬过的夜和一起做过的项目，都将成为一笔宝贵的回忆和财富。

在大学生活中，还有一群笛箫社的伙伴们，携手同行。很开心能在大学阶段遇到这样一群兴趣相同的朋友们，在社团 4 年，从大一的专场，大二的任职，大三的指导，再到大四一起毕业，我们一起经历的故事，于我而言都无比珍贵。

最后，我想要感谢我的家人们，在我选择理工，选择通信这条道路时，对我义无反顾的支持。进入大学后，无论学习的成果好坏，都给予我鼓励，让我能够在四年内专心学习，同时，也无比支持我未来的发展与规划，在我大二决定出国留学时，就一直支持我，希望我学有所成，往自己更感兴趣的方向发展。

本科至此，已经快到终点。但未来的发展才刚刚开始，也在此希望我能够在未来的留英旅途上能够拥有更好的结果！

参考文献

[1] 季长清,高志勇,秦静,汪祖民.基于卷积神经网络的图像分类算法综述[J].计算机应用,2022,42(04):1044-1049.

[2] 王嘉伟.基于卷积神经网络的语音识别研究[J].科学技术创新,2019(31):71-73.

[3] 罗枭.基于深度学习的自然语言处理研究综述[J].智能计算机与应用,2020,10(04):133-137.

[4] Kim B, Koyejo O, Khanna R, et al. Examples are not enough, learn to criticize! Criticism for Interpretability[C]. neural information processing systems, 2016: 2280-2288.

[5] 季桂树,陈沛玲,宋航.决策树分类算法研究综述[J].科技广场,2007(01):9-12.

[6] Kim B, Rudin C, Shah J. The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification[J]. Computer Science, 2015, 3:1952-1960.

[7] Doshi-Velez F, Wallace B C, Adams R. Graph-sparse LDA: a topic model with structured sparsity[C]//Twenty-Ninth AAAI conference on artificial intelligence, 2015:2575-2581.

[8] Dosovitskiy A, Brox T. Inverting Visual Representations with Convolutional Networks[C]// Computer Vision and Pattern Recognition. IEEE, 2016:4829-4837

[9] Olden J D, Jackson D A. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks[J]. Ecological Modelling, 2002, 154(1–2):135-150.

[10] Dimopoulos Y, Bourret P, Lek S. Use of some sensitivity criteria for choosing networks with good generalization ability[J]. Neural Processing Letters, 1995, 2(6):1-4.

[11] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]// International conference on machine learning. PMLR, 2017: 1885-1894.

[12] Xu Z, Zhang Y, Xiao Y. Training behavior of deep neural network in frequency domain[C]//International Conference on Neural Information Processing. Springer, Cham, 2019: 264-274.

[13] Xu Z, Zhang Y, Luo T, et al. Frequency principle: Fourier analysis sheds light on deep neural networks[J]. Communications in Computational Physics,2020, 28 (5): 1746-1767.

[14] Xu Z, Zhou H. Deep Frequency Principle Towards Understanding Why Deeper Learning Is Faster[C]// AAAI-21 Technical Tracks 12, 2021: 10541-10550

[15] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. 2009.

[16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[17] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[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] Hendrycks D, Dietterich T G. Benchmarking neural network robustness to common corruptions and surface variations[J]. arXiv preprint arXiv:1807.01697, 2018.

[20] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations[J]. arXiv preprint arXiv:1903.12261, 2019.

[21] Huber P J. Robust statistics[M]//International encyclopedia of statistical science. Springer, Berlin, Heidelberg, 2011: 1248-1251.

[22] Yin D, Gontijo Lopes R, Shlens J, et al. A fourier perspective on model robustness in computer vision[J]. Advances in Neural Information Processing Systems, 2019, 32.

[23] Gilmer J, Ford N, Carlini N, et al. Adversarial examples are a natural consequence of test error in noise[C]//International Conference on Machine Learning. PMLR, 2019: 2280-2289.

[24] Qian Z, Huang K, Wang Q F, et al. A Survey of Robust Adversarial Training in Pattern Recognition: Fundamental, Theory, and Methodologies[J]. arXiv preprint arXiv:2203.14046, 2022.

[25] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

[26] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.

附录一 英译汉

Data Augmentation for JPEG Steganalysis

Abstract：Deep Convolutional Neural Networks (CNNs) have performed remarkably well in JPEG steganalysis. However, they heavily rely on large datasets to avoid overfitting. Data augmentation is a popular technique to inflate the datasets available without collecting new images. For JPEG steganalysis, the augmentations predominantly used by researchers are limited to rotations and flips (D4 augmentations). This is due to the fact that the stego signal is erased by most augmentations used in computer vision. In this paper, we systematically survey a large number of other augmentation techniques and assess their benefit in JPEG steganalysis.

Index Terms：Steganography, steganalysis, convolutional neural network, data augmentation

I. INTRODUCTION

Convolutional Neural Networks (CNNs) are the superior detectors of steganography today [1]. They replace the so-called rich media models, which are high-dimensional feature representations hand-designed for specific purposes in steganalysis as a well as the related field of digital forensics. In contrast to rich models, CNNs learn the best (internal) image representation as well as the detector itself via a training process, which is usually a form of a Stochastic Gradient Descent (SGD).

Data augmentation is a way to increase the training set size by including in training transformed versions of the images. Typical augmentations used in computer vision are rotations, resizing, cropping, channel shuffle, dropout, and in general any transformation that fundamentally preserves the label assigned to the image. Larger training sets usually lead to better detectors / classifiers because they are exposed to more diverse content. Augmentations can be domain-specific, depending on what task the CNN is trained on. For steganalysis, the signal of interest is rather fragile, formed by slight perturbations of cover image pixels or quantized DCT coefficients (for JPEG images). Thus, augmentations that remove or suppress this signal are undesirable and cannot improve the detection performance. Steganalysts typically employ the so-called dihedral D4 augmentation, which consists of rotations by integer multiples of 90 degrees and mirrorrings.

Indeed, such transformations do not disturb the stego signal while exposing the network to a more diverse dataset. On the other hand, resizing and rotations by non-integer multiples of 90 degrees are not desirable as the resampling that is inherently part of these transformations disturbs the stego signal to a large degree.

Previous work in steganalysis [2], [3] addressed the need for an increased dataset size by acquiring more images using similar devices or using other datasets that are close in development to the test dataset. Not only this solution does not follow our definition of data augmenting (i.e. not requiring acquisition of new images), but it is unclear how one would replicate a cover source as noted by the winners of the BOSS competition who failed to duplicate the test set's cover source even when knowing the camera model and the development script used[4]. Other steganalysis augmentation techniques have been introduced, such as BitMix [5] (Section III-C) and Pixels-off[6]. The latter was not included in this study because it was not developed for the JPEG domain.

In this paper, we look beyond the usual random D4 augmentation group in search for new augmentations that can improve the detector performance for steganalysis of digital images. We use the Albumentations Library [7] as well as custom augmentations specifically designed for steganalysis. In particular, we take a look at various forms of drop out augmentations, which make good sense for computer vision tasks because they simulate "occlusions" that may naturally occur and thus robustify the classifier. We also study color channel shuffle, bitmix, convex combinations of cover and stego images (with soft labels), and multiple stego image sampling to expose the network to multiple versions of the stego image embedded with different stego keys.

Interestingly, the idea to use symmetries of natural images to robustify the detector is already present in rich models [8], [9]. Their "features" are formed by co-occurrences of adjacent quantized and truncated noise residuals obtained via pixel predictors. These features are typically "symmetrized" or robustified by leveraging directional and sign symmetries of natural images. In particular, co-occurrences computed from the original image as well as their versions rotated by integer multiples of 90 degrees and their mirrored forms were typically added to one, better populated co-occurrence matrix (feature). Because noise residuals exhibit symmetrical marginals centered around zero, additional co-occurrences can be added by flipping the signs of the noise residuals, a process that required some caution when applied to the so-called "min" and "max" non-linear residuals in the Spatial Rich Model (SRM) [8].

In Section II, we describe the setup of all our experiments, the datasets, and performance measures to evaluate the effectiveness of various augmentations. All tested augmentations are explained in Section III. Section IV-A shows all experimental results in a graphic form together with a discussion. Finally, the paper is closed in Section V.

II. EXPERIMENTAL SETTING

A. Datasets

We use the ALASKA II 256 × 256 dataset [10] which contains 75,000 different cover images compressed with quality factors 75 and 95. The covers were randomly divided into three sets with 66,000, 3,000, and 6,000 images, for training, validation, and testing, respectively. The images were embedded using J-UNIWARD [11], J-MiPOD [12], and F5 [13] with payloads 0.5 , 0.4, 0.3, 0.2, and 0.1 bpnzac. For J-UNIWARD, the payload was spread into the chrominance channels using Color Channels Merging (CCM), which concatenates the color cost maps before minimizing the additive distortion. For J- MiPOD and F5, we only embedded the payload in the luminance channel.

B. Detectors

We use the EfficientNet B3 [14] pre-trained on Ima- geNet [15] and refined for JPEG domain steganalysis [16], [17] with the training hyper-parameters described in Section 4.2 in [17]. No modifications were done to the architecture besides changing the Fully Connected (FC) layer.

We use the following three performance measures to compare detectors: $P_E = min(P_D(P_{FA}) + P_{FA})$,wAUC[10], $MD5 = P_{MD}(P_{FA} = 0.05)$, and $FA80 = P_{FA}(P_{MD} = 0.8)$.

III. AUGMENTATIONS

In this section, we describe all augmentation techniques surveyed in this paper.

A. Dropout augmentations

Dropout style augmentations simulate occlusions.

a) CoarseDropout: CoarseDropout is a dropout augmentation that randomly zeros out rectangular regions of the image. It evolved from the cutout augmentation [18], which drops a single

square region. The location of the dropped regions (holes) is randomized, while their size is set to 8 × 8 and their number to 32 holes. Figure 1 shows an example of a CoarseDropout augmented image. Note that the holes can overlap as well as not completely fit in the image. They also do not respect the 8 × 8 grid of JPEG blocks.

b) GridDropout: GridDropout [19] is another dropout augmentation that drops out rectangular regions of an image in a grid fashion. The grid shape is a hyper-parameter of the augmentation. We set the grid to correspond to JPEG 8 × 8 squares, and vary the dropout ratio parameter which controls the number of dropped blocks, the number of dropped blocks was set to 36. Figure 2 shows an example of a GridDropout augmented image.

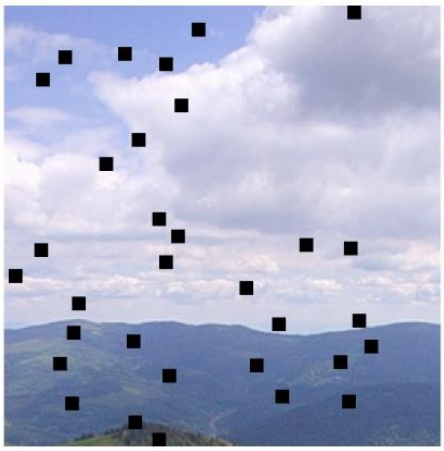


Figure 1

c) RandomGridDropout: This augmentation combines the GridDropout and CoarseDropout. It drops a number of non overlapping 8 × 8 squares while respecting the 8 × 8 JPEG grid; the number of holes was also set to 32. Figure 3 shows an example of a RandomGridDropout augmented image.

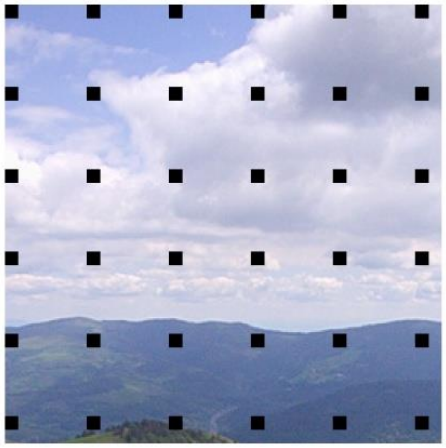


Figure 2

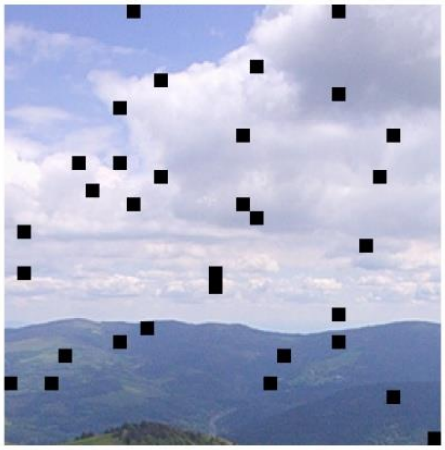


Figure 3

B. Channel augmentations

This section describes augmentations operating on the channel dimension of the input image. Such augmentations are only useful for color steganography.

a) ChannelShuffle: This is a channel-style augmentation that randomizes the order of channels in a color image. For example, an RGB image could become a GBR image. Note that this augmentation is only used when training on RGB in- puts since swapping the channels in the YCbCr representation is detrimental because of the heterogeneity of these channels.

b) ToGray: ToGray converts the sampled image to grayscale. This augmentation does not completely destroy the stego signal since a vast majority of the payload is typically in the luminance channel. This augmentation was also used with networks trained with RGB inputs.

C. Mixing augmentations

Next, we describe augmentations which mix two images, a cover image C and a stego image S, to create an augmented image X. Such augmentations evolved from the Mixup augmentation[20], which saw a great success in computer vision applications. These augmentations often require changing the label vector to reflect the amount of mixing between the classes. The loss used is the cross-entropy with soft targets.

a) BitMix: BitMix [5] takes a cover image and replaces a randomly sampled patch with the stego image and vice versa. This patch is chosen by randomly sampling a rectangular area whose maximum size is determined by a maximum mix ratio parameter. The patch is represented using a binary mask M. For simplicity, we assume the mask is applied to a cover image C but in practice it is applied to cover and stego images. The label vector y is changed using the system of Equations 1–3:

$$X = M \odot C + (1 - M) \odot S$$

$$\lambda = \frac{\|M \odot C - M \odot S\|_1}{\|C - S\|_1}$$

$$\mathcal{Y}x = (\lambda, 1 - \lambda)$$

Figure 4

Figure 4 shows an example of a BitMix augmented image as well as its corresponding soft label.

b) ConvexMix: This augmentation forms a convex combination of a cover-stego image pair by sampling the mixing parameter $\lambda \sim U(0, 1)$:

$$X = \lambda C + (1 - \lambda) S$$

$$x = (\lambda, 1 - \lambda)$$

D. Sampling augmentations

a) StegoSampling: This augmentation is special to the steganalysis task. In fact, each stego image S is a random sample from the steganographic simulator, which simulates embedding changes operating on the rate–distortion bound. This enables sampling different stego images from the same cover and with the same payload while getting different stego samples at each iteration. In practice, we use the change rates β+, β− and the cover image to sample a stego image on the fly while training the network. This augmentation will be called StegoSampling.

IV. RESULTS

A. Successful augmentations

Figure 5 shows the results for all tested augmentations, three embedding algorithms, two payloads, and two quality factors. The baseline detectors were trained with the standard D4 augmentation. For J-UNIWARD, every tested augmentation was successful in improving on the baseline. For larger payloads, the StegoSampling and dropout augmentations performed very well. This is clear for QF75 and even more so for QF95. For smaller payloads, StegoSampling and dropout augmentations are still very capable but this changes for the larger quality 95. Despite the drop in performance from QF75, there are still meaningful improvements for small payloads at QF95. The results suggest that detection of steganography in images with low QFs will benefit from the assistance of augmenta- tions the most, whereas there is a limit on how helpful the augmentations can be for higher QFs images as the payload size decreases.

For J-MiPOD, the additional tested augmentations were less impactful. The one that stood out was ConvexMix, which helped detection of J-MiPOD much more than J-UNIWARD. Furthermore, at QF75 the RandGridDropout and GridDropout augmentations actually hurt the testing accuracy. Even when the augmentations provide improvements, the effectiveness of additional augmentations falls of much more quickly in comparison to J-UNIWARD. As the payload size decreases and the QF increases, the benefits of augmentations in detecting images embedded with J-MiPOD become less pronounced.

Table I
BASELINE PERFORMANCE AND COARSEDROPOUT FOR EFFICIENTNET B3 TRAINED ON 10,000 PAIRS OF COVER AND J-UNIWARD IMAGES, QF75 AT 0.4 BPNZAC.

Data Augmentation	Accuracy	MD5	FA80	wAUC
	QF75 J-UNIWARD 0.4 bpnzac			
Baseline	0.8881	0.1701	0.0335	0.9797
CoarseDropout, 16 blocks	0.9029	0.1488	0.0293	0.9812

For the F5 algorithm, the 0.1 bpnzac payload was different compared to the lowest payloads for J-MiPOD and J- UNIWARD. Every augmentation except for the RandGrid- Dropout augmentation resulted in worse accuracies for both QFs. For the 0.3 bpnzac payload, every augmentation was able to provide a slight improvement upon the baseline. Moving from QF75 to QF95 seemed to produce nearly equivalent results. For F5, the payload influenced the amount of improvement the most.

B. Low data regime

Another experiment that we ran was taking an augmentation and testing it against a smaller dataset. This was done using the CoarseDropout augmentation with the results shown in Table I. The settings were identical to those described in Section II except for the training data set size, which was reduced from 66,000 to 10,000. The original number of 32 dropout blocks used for CoarseDropout as described in Section III-A did not work with this smaller dataset. However, when the number of blocks was halved there was a substantial gain in accuracy and MD5. This suggests that smaller datasets are more delicate but still benefit from toned down version of the augmentations used in our experiments.

C. Unsuccessful augmentations

Here, we report on the augmentations that failed to improve upon the baseline. The channel augmentations for color images (ChannelShuffle and ToGray) failed to give better results than the baseline as shown in Table II.

Table II
CHANNELSHUFFLE AND TOGRAY PERFORMANCE FOR J-UNIWARD QF75 AND 95 AT 0.4 BPNZAC.

Data Augmentation	Accuracy	MD5	FA80	wAUC
QF75 J-UNIWARD 0.4 bpnzac				
Baseline	0.9571	0.0308	0.0012	0.9961
ChannelShuffle	0.9509	0.0297	0.0018	0.9961
ToGray	0.9515	0.0272	0.0015	0.9962
QF95 J-UNIWARD 0.4 bpnzac				
Baseline	0.8308	0.3264	0.1300	0.9498
ChannelShuffle	0.8194	0.3571	0.1538	0.9459
ToGray	0.8292	0.3212	0.1366	0.9491

Additionally, we tried to combine all augmentations that produced a gain to see if their combined effect would provide further benefit. Surprisingly, this was not the case as shown in Table III. The augmentations were combined using a “OneOf” strategy: for each sample, a randomly sampled augmentation technique was applied.

Table III
BASELINE, BEST SINGLE AUGMENTATION, AND ALL AUGMENTATIONS PERFORMANCE FOR J-UNIWARD AND J-MIPOD AT QF75 AND 95.

Data Augmentation	Accuracy	MD5	FA80	wAUC
QF75 J-UNIWARD 0.4 bpnzac				
Baseline	0.9571	0.0308	0.0012	0.9961
GridDropout	0.9669	0.0173	0.0012	0.9974
All	0.9603	0.0155	0.0020	0.9973
QF95 J-UNIWARD 0.4 bpnzac				
Baseline	0.8309	0.3264	0.1300	0.9498
GridDropout	0.8490	0.2935	0.1044	0.9562
All	0.8398	0.3128	0.1200	0.9531
QF75 J-MiPOD 0.5 bpnzac				
Baseline	0.9128	0.1243	0.0166	0.9854
ConvexMix	0.9180	0.1215	0.0178	0.9853
All	0.9193	0.1156	0.0173	0.9863
QF95 J-MiPOD 0.5 bpnzac				
Baseline	0.7132	0.5946	0.3706	0.8662
CoarseDropout	0.7186	0.5879	0.3724	0.8679
All	0.7164	0.5917	0.3641	0.8680

V. CONCLUSIONS AND FUTURE WORK

Our study of augmentations shows that there are ways to successfully augment data for steganographic deep learning applications beyond the standard D4 augmentations. With some care, the correct selection of additional augmentations can result in a substantial boost in performance (up to 3% in accuracy and 5% in MD5). We observed that smaller data sets are more likely to benefit from using the proposed data augmentations than large datasets because the augmentations

effectively increase the size of the training set and make deep learning models more robust.

For possible future directions, we recommend further investigation of the effect of the cover or stego source on the gains of each augmentation. For example, one could study a more diverse stego source comprised of different stego schemes with different payloads. Additionally, it is probably worth looking at how much the augmentations boost deeper CNN architectures, such as the EfficientNet B7.

JPEG 隐写分析的数据增强

摘要：

深度卷积神经网络（CNN）在 JPEG 隐写分析中有着优异的表现，然而，很大程度上它们依赖大型数据集以避免过拟合。数据增强是常用的用于扩大数据集而不需要收集新图像的技术手段。对于 JPEG 隐写分析来说，研究人员主要使用的增强技术仅限于旋转和翻转（D4 扩增）。这是由于计算机视觉中使用的大多数数据增强方法都会抹去隐写信号。在本文中，我们系统地调查了大量其他的数据增强技术并评估它们在 JPEG 隐写分析中的好处。

关键词： 隐写术，隐写分析，卷积神经网络，数据增强

I. 简介

卷积神经网络(CNN)是如今最优秀的用于检测隐写分析的技术[1]。它们取代了所谓的为隐写分析以及数字取证的相关领域的特定目的而手工设计的高维特征表示的丰富媒体模型。与丰富的模型相比，CNN 通过训练过程学习到最佳（内部）图像表示以及检测器本身，这通常是一种随机梯度下降（SGD）的形式。

数据增强是一种通过训练图像的转换版本来增加训练集的方法。在计算机视觉领域中，传统的数据增强方法包含旋转、调整大小、裁剪、通道洗牌（打乱原特征图通道顺序）、擦除，以及任何从根本上保留了分配给图像的标签的变换都是如此。较大的训练集会接触到更多样化的内容，因此通常意味着较好的分类效果。数据增强可以是特定领域的，这取决于 CNN 所训练的任务。对于隐写分析领域，我们所需要的信号相当脆弱，是由覆盖图像像素的轻微扰动或量化的 DCT 系数（JPEG 图像）形成，因此，如果数据增强删除或者抑制了这些信号那么就不能提高它的检测性能。隐写分析人员通常采用由 90 度的整数倍旋转和镜像组成的二面体 D4 增强的方式。这种转换在不扰乱隐写信号的同时将网络置于一个更加多样化的数据集中。另外，旋转及非整数倍的 90 度旋转的方法并不能取得理想效果，这是因为重采样作为转换的内在组成部分，会在很大程度上干扰隐写信号。

以前的隐写分析工作[2], [3]会通过使用类似设备获取更多的图像或使用与测试数据集发展接近的其他数据集来解决扩大数据集规模的需要。这种解决方案不仅不符合我们对数据

增强的定义（不需要获取新的图像），也不清楚如何复制覆盖源，正如 BOSS 比赛的获胜者即使知道相机模型和使用的开发脚本，也无法复制测试集的覆盖源[4]。其他的隐写分析的数剧增强技术比如 BitMix[5]（第三节 C）和 Pixels-off[6]已经被引入。本次研究并没有包含后者 Pixels-off，因为它不是为 JPEG 领域开发的数据增强技术。

在本文中，我们超越了通常的随机 D4 增强组，寻找能够提高数字图像隐写分析检测性能的新的增强方式。我们使用了 Albumentations 库[7]以及专门为隐写分析设计的自定义增强器。尤其我们关注到了各种形式的擦除增强，因为它们模拟了可能自然发生的 "闭塞 "，从而增强了分类的鲁棒性，对计算机视觉任务有重大意义。我们还研究了彩色通道洗牌、比特混合、覆盖图像和隐写图像的凸形组合（带软标签），以及多个隐写图像采样，使网络能接触到嵌入不同隐写密钥密钥的隐写图像的多个版本。

有趣的是，利用自然图像的对称性来提升检测器鲁棒性的想法已经存在于丰富的模型中[8], [9]。他们的 "特征 "是由通过像素预测器获得的相邻量化和截断的噪声残差的共同出现形成的。这些特征通常通过利用自然图像的方向性和符号对称性而被 "对称化 "或鲁棒化。特别是，从原始图像中计算出的同现，以及它们被旋转 90 度的整数倍的版本和它们的镜像形式通常被添加到一个更好的共同发生矩阵（特征）。由于噪声残差展现出以零为中心的对称边际，额外的共同发生可以通过翻转噪声残差的符号来添加，这个过程在需要在应用于空间丰富模型（SRM）中所谓的 "最小 "和 "最大 "非线性残差时谨慎一些[8] 。

在第二节中，我们描述了实验的设置、数据集和性能指标，用于评估不同数据增强的有效性。第三节阐述了所有经过测试的增强功能。第四节 A 部分以图表和讨论的形式展示了所有的实验结果。第五节为本文的终章。

II. 实验环境及设置

A. 数据集

我们使用 ALASKA II 256256 数据集[10]，该数据集包含 75000 张不同的封面图像，以质量系数 75 和 95 进行压缩。这些封面图像被随机分为三组，分别有 66,000、3,000 和 6,000 张图像，用于训练、验证和测试。使用 J-UNIWARD[11]、J-MiPOD[12]和 F5[13]算法嵌入图像，嵌入率为 0.5，0.4，0.3，0.2，0.1bpn。对于 J-UNIWARD，使用颜色通道合并（CCM）将信息分散到色度通道中，在最小化加性失真前将颜色成本图连接起来。对于 JMiPOD 和 F5，我们只在亮度通道中嵌入信息。

B. 检测器

我们使用 EfficientNet B3 算法[14]在 ImageNet[15]数据集上进行预训练，并针对 JPEG 领域的隐写分析[16],[17]进行了改进，训练的超参数在[17]的第 4.2 节中描述。除了对全连接（FC）层进行了改变之外，没有对结构做任何修改。

我们使用以下三个性能指标来比较检测器。$P_E = \min(P_D(P_{FA}) + P_{FA})$ ，$wAUC$，$MD5 = P_{MD}(P_{FA} = 0.05)$，$FA80 = P_{FA}(P_{MD} = 0.8)$。

III. 数据增强

在这一节中，将描述本文调查的所有的数据增强技术。

A. 擦除式增强

擦除式的增强模拟了闭塞。

a) CoarseDropout：CoarseDropout 是一种擦除式增强，随机地将图像的矩形区域擦除。它是由删除单一方形区域的剪切增强法[18]演变而来。擦除的区域（孔）的位置是随机决定的，大小为 8×8，数量为 32。图 1 为一个 CoarseDropout 增强图像的例子。值得注意的是，孔可以重叠，也可以不完全适合图像。他们也不遵循 JPEG 块的 8×8 网格。

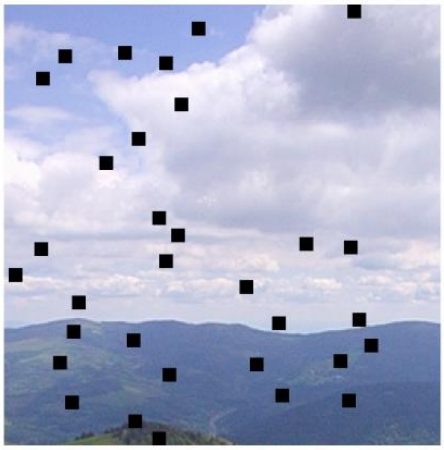


图 6 CoarseDropout 数据增强

b) GridDropout：GridDropout[19]是另一种擦除式增强方法，它以网格的形式擦除图像的矩形区域。网格形状是该增强的一个超参数。我们将网格设置为符合 JPEG 8×8 的方块，并改变控制擦除块数量的擦除率参数，擦除块的数量被设置为 36。图 2 显示了一个 GridDropout 增强图像的例子。

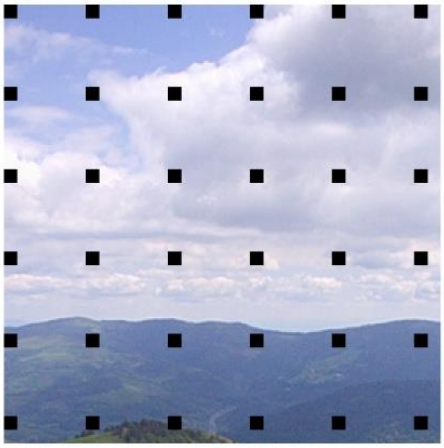


图 7 GridDropout 数据增强

c) RandomGridDropout：这个增强方式结合了 GridDropout 和 CoarseDropout 两种方式。它在遵循 8×8 的 JPEG 网格的同时，擦除一些不重叠的 8×8 的方块；孔的数量同样被设置为 32。图 3 显示了一个 RandomGridDropout 增强图像的例子。

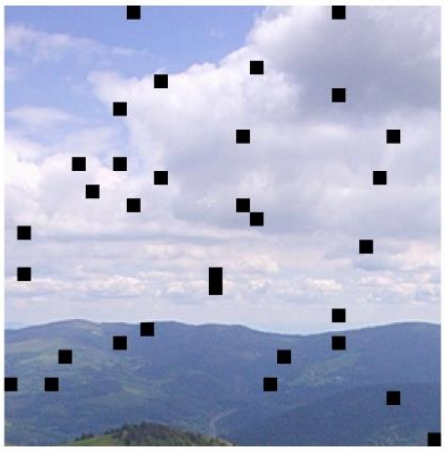


图 8 RandomGridDropout 数据增强

B. 通道增强

本节描述了对输入图像的通道维度进行的增强。这种增强方法只对彩色隐写术有用。

(a) ChannelShuffle：这是一个能够随机化彩色图像中的通道顺序的通道式的增强功能。例如，一个 RGB 图像可以变成一个 GBR 图像。但需要注意的是，这个增强功能只在对 RGB 进行训练时使用，由于通道的异质性，在交换 YCbCr 的通道是有损害的。

b) ToGray：ToGray 方式能够将采样后的图像转换为灰度。这种增强并不完全破坏隐写信号，因为绝大部分的有效载荷通常在亮度通道中。这种增强方法也被用于用 RGB 输入训练的网络。

C. 混合增强

接下来，我们描述混合两幅图像的增强方法，即封面图像 C 和隐写图像 S，以创建一个增强图像 X。这种增强方法是从 Mixup 增强方法[20]发展而来的，它在计算机视觉应用中取得了巨大的成功。这些增强通常需要改变标签向量以反映类之间的混合量。使用的损失函数是具有软目标的交叉熵。

a) BitMix：BitMix[5]使用一个覆盖图像，并将一个随机抽样的补丁替换为隐写图像，反之亦然。这个补丁是通过随机抽样选择一个矩形区域，其最大尺寸由最大混合比例参数决定。该补丁用一个二进制掩码 M 来表示。为简单起见，我们假设该掩码适用于封面图像 C，但在实践中，它适用于封面和隐写图像。标签向量 y 是用方程 1-3 的系统来改变的。

$$X = M \odot C + (1 - M) \odot S$$

$$\lambda = \frac{\|M \odot C - M \odot S\|_1}{\|C - S\|_1}$$

$$\mathcal{Y}x = (\lambda, 1 - \lambda)$$

图 4 显示了一个 BitMix 增强的图像的例子以及其相应的软标签。

图 9 BitMix 数据增强

b) 凸形混合（ConvexMix）：这个增量通过对混合参数 λ ~ U(0, 1):进行抽样，形成了封面-隐写图像对的凸形组合。

$$X = \lambda C + (1 - \lambda) S$$

$$x = (\lambda, 1 - \lambda)$$

D. 采样增量

a) StegoSampling:对隐写分析任务来说，这是一种特殊的增强手段。事实上，每个隐写

图像 S 是来自隐写模拟器的一个随机样本，它模拟了在速率-失真边界上操作的嵌入变化。这使得从相同的封面和相同的有效载荷中抽取不同的隐写图像，同时在每次迭代中获得不同的隐写样本。在实践中，在训练网络时，我们用变化率β^+，β^-和封面图像对隐写图像进行实时采样。这种增强作用将被称为 StegoSampling。

IV. 结果

A. 成功的数据增强方式

图 5 展示了了所有增强、三种嵌入算法、两种有效载荷和两种质量因素的测试结果。基线检测器是用标准的 D4 增强法训练的。对于 J-UNIWARD，每一个测试的增强都能成功地改善基线。对于较大的有效载荷，StegoSampling 和擦除增强功能表现很好。这对 QF75 来说很明显，对 QF95 来说更是如此。对于较小的有效载荷，StegoSampling 和擦除增强功能仍然非常有效，但对于较大的质量 95 来说就有所不同。尽管从 QF75 开始性能下降，但在 QF95 时对小的有效载荷仍有有效的改进。结果表明，在低 QFs 的图像中检测隐写术将从数据增强的帮助中受益最大，而随着有效载荷大小的减少，数据增强对高 QFs 图像的帮助是有限的。

对于 J-MiPOD 来说，额外添加增强功能并不那么有效。最突出的是 ConvexMix，它对 J-MiPOD 的检测性能优越性比 J-UNIWARD 大得多。此外，在 QF75 中，RandGridDropout 和 GridDropout 增强功能实际上降低了测试的准确性。即使有些增强方式有所提升，与 J-UNIWARD 相比，增强功能的有效性下降得更快。随着有效载荷大小的减少和 QF 的增加，增强功能在检测嵌入 J-MiPOD 的图像方面的好处变得不那么明显了。

对于 F5 算法，与 J-MiPOD 和 JUNIWARD 的最低有效载荷相比，0.1 bpnzac 有效载荷是不同的。除了 RandGridDropout 之外的数据增强方式都导致两个 QF 的准确率下降。对于 0.3 bpnzac 的有效载荷，每一个增强都能够在基线上有一些微小的提升。从 QF75 到 QF95 似乎产生了几乎相同的结果。对于 F5 来说，有效载荷对改进量的影响最大。

B. 低数据制度

我们进行的另一个实验是采取一种增强措施，并针对较小的数据集对其进行测试。这是用 CoarseDropout 增强法进行的，结果见表 1。除了训练数据集的大小从 66,000 个减少到 10,000 个，其他设置与第二节中描述的相同。第 III-A 节中描述的用于 CoarseDropout 的

最初的 32 个 dropout 块的数量在这个较小的数据集上不起作用。然而，当块的数量减半时，准确性和 MD5 都有很大的提高。这表明，较小的数据集更微妙，但仍然受益于我们实验中使用的淡化版本的增强措施。

表 2

数据增加	精度	MD5	FA80	wAUC
		QF75 J-UNIWARD 0.4 bpzac		
基线	0.8881	0.1701	0.0335	0.9797
CoarseDropout ,16 块	0.9029	0.1488	0.0293	0.9812

C.效果欠佳的增强方式

我们展示了未能在基线上得到改善的增强。如表二所示，彩色图像的通道增强（ChannelShuffle 和 ToGray）未能提供比基线更好的结果。

此外，我们试图将所有产生增益的增强方法结合起来，看看它们的综合效果是否有更加优异的表现。然而出乎意料的是，情况并非如此，如表三所示。我们使用 "OneOf"策略来组合增强技术：对于每个样本，随机抽样的增强技术被应用。

V. 结论和未来工作

我们对数据增强方式的研究表明，除了标准的 D4 扩增之外，还有一些增强方法可以成功地为隐写深度学习应用扩增数据。只要稍加注意，正确地选择额外的增强方法可以使性能得到大幅提升（准确率高达 3%，MD5 为 5%）。可以观察到较小的数据集比大数据集更有可能从我们提出的数据增强方法中取得优越效果，因为增强可以有效地增加训练集的规模，使深度学习模型更加强大。

对于未来研究方向，我们建议进一步研究封面或隐写源对每个增强的收益的影响。例如，我们可以研究由不同有效载荷的不同窃密方案组成的更多样化的隐写源。此外，这些增强措施对更深层次的 CNN 架构，如 EfficientNet B7 有多大的提升作用也是值得研究的方向。

表 3

数据增加	精度	MD5	FA80	wAUC
		QF75 J-UNIWARD 0.4 bpzac		
基线	0.9571	0.0308	0.0012	0.9961
ChannelShuffle	0.9509	0.0297	0.0018	0.9961
ToGray	0.9515	0.0272	0.0015	0.9962
		QF95 J-UNIWARD 0.4 bpzac		
基线	0.8308	0.3264	0.1300	0.9498
ChannelShuffle	0.8194	0.3571	0.1538	0.9459
ToGray	0.8292	0.3212	0.1366	0.9491

表 4

数据增加	精度	MD5	FA80	wAUC
		QF75 J-UNIWARD 0.4 bpzac		
基线	0.9571	0.0308	0.0012	0.9961
GridDropout	0.9669	0.0173	0.0012	0.9974
所有方式	0.9603	0.0155	0.0020	0.9973
		QF95 J-UNIWARD 0.4 bpzac		
基线	0.8309	0.3264	0.1300	0.9498
GridDropout	0.8490	0.2935	0.1044	0.9562
所有方式	0.8398	0.3128	0.1200	0.9531
		QF75 J-MiPOD 0.5 bpzac		
基线	0.9128	0.1243	0.0166	0.9854
GridDropout	0.9180	0.1215	0.0178	0.9853
所有方式	0.9193	0.1156	0.0173	0.9863
		QF95 J-MiPOD 0.5 bpzac		
基线	0.7132	0.5946	0.3706	0.8662
GridDropout	0.7186	0.5879	0.3724	0.8679
所有方式	0.7164	0.5917	0.3641	0.8680

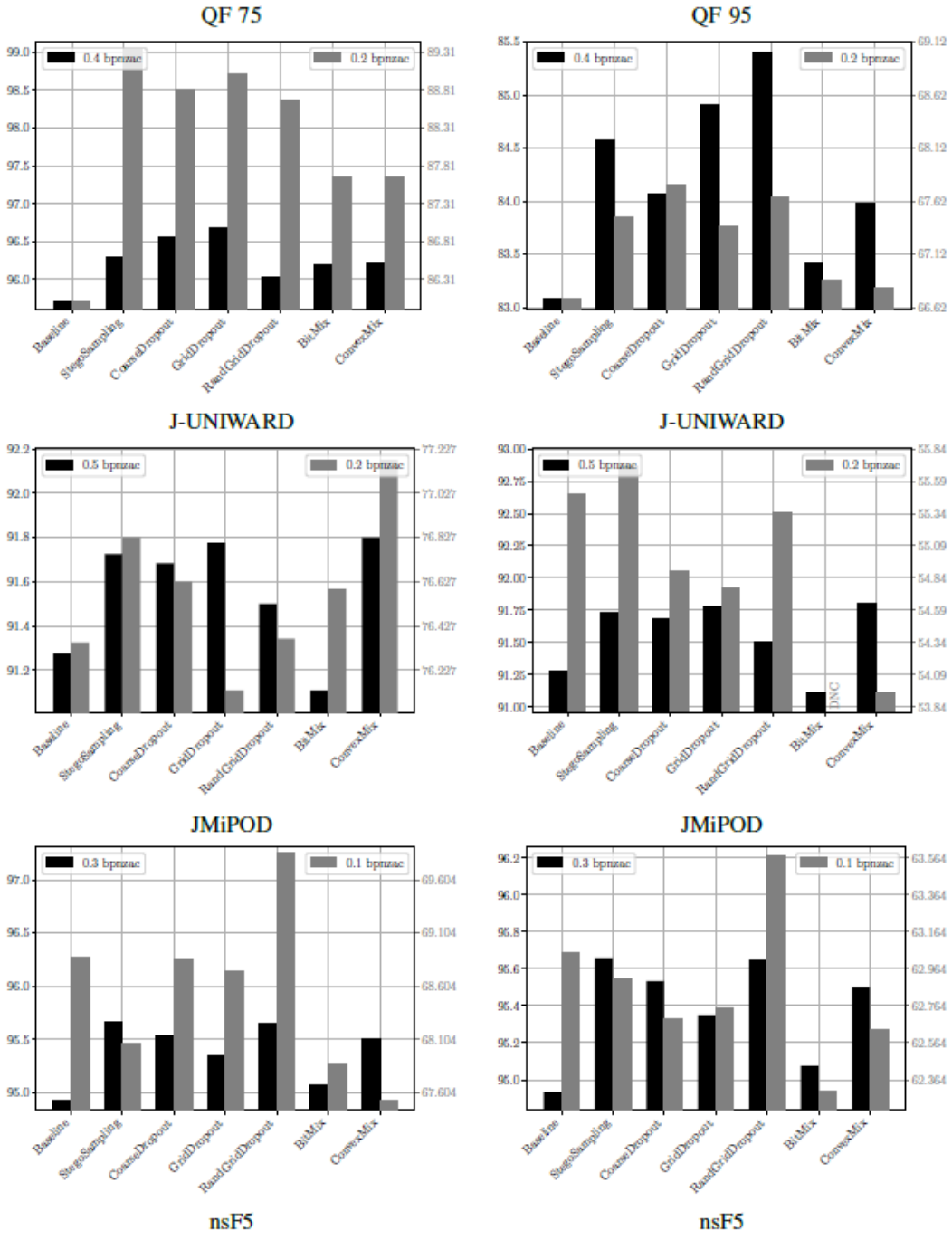


图 10 所有增强、三种嵌入算法、两种有效载荷和两种质量因素的测试结果

附录二 课题调研报告

本课题主要从频域视角对神经网络模型进行相关可解释性分析。运用傅里叶变换相关理论，将深度学习中的图像分类任务从频域角度进行可解释性研究，得出在神经网络学习的过程中，图像的低频区域信息对模型的预测将起到更为重要的作用。而当数据增强模型在改善针对噪声下的鲁棒性能时，其主要改善图像高频信息的鲁棒性能，在高频区域信息的鲁棒性能有较大提升时，低频区域信息鲁棒性将有细微下降趋势。本报告将从课题主要所研究的图像分类任务出发，分析其目的及意义，所发挥的作用及对社会的影响以及关于课题对图像分类任务的建议构思三方面对课题进行进一步的分析，探究其可行性及在实践中的意义。

一、 课题目的和意义

从学术研究角度而言，本课题紧扣当前人工智能深度学习的潮流，并发掘了一个新颖的角度——从傅里叶视角对深度神经网络进行分析。比起大量不断优化的深度神经网络模型来说，极少人关注其模型内部工作机制。在例如医疗、自动驾驶等很多的应用场景下，如果模型不能够为自己的决策做出解释，就很难建立有效的人机双向沟通机制，因而限制了这些深度学习系统在数据驱动相关分析中的应用。一旦深度学习本身工作的“黑箱”模型被破解，今后关于神经网络模型的研究将能够以此为依据展开研究及改进，开发出更优化更高效的模型。

从社会发展角度而言，图像识别及图像分类任务在当今生活中有着极其广泛的应用。随着人工智能热潮的不断发展，许多企业同样尝试将计算机视觉与其本身工作相结合，以提高工作效率，建设人工智能数字化平台。通过机器对采集的数据图像进行智能分类，能够大大减轻人工压力，减少成本。同时，机器的正确率能够通过调整参数在可控制范围内，减少人工情况下失误造成的安全隐患及后果，以更少的成本，更高的准确性及效率促进社会的可持续发展。

二、 课题的作用与思考

通过课题研究，能够通过神经网络模型的可解释性原理，对模型进行有针对性的改善，以更好的符合当下社会的需求。疫情以来，学校、企事业单位、公共场所等地为把控人流量都增加了人脸识别闸门，但人脸识别闸门常常由于光线天气等因素干扰导致人脸无法正常识别，同时将人脸识别错误的情况也时有发生，其安全性能将受到挑战。通过学术分析后，能够提出改善神经网络图像分类模型的方法，免去大量繁琐的试验环节，从原理的角

度出发，将人脸识别系统进行升级改善，既减轻了反复试错的经济压力，也能够提升系统的安全性。

三、 关于课题的建议与构想

在课题中，对神经网络可解释性模型进行了未来发展展望。从频域视角进行分析是模型可解释性的方法之一，而该方法也能够迁移应用至各个领域当中，例如尚未能成功对其进行分析的信息隐藏领域，如果能成功运用频域视角对隐写分析模型进行可解释性分析，则在信息安全领域将会有卓越提升，将对我们身边网络安全、信息安全更有保障。

同时，神经网络模型的可解释性分析方式多种多样，该课题可塑性较强、迁移应用范围较广，能够运用多种方式对挖掘模型内部工作机制，也能够迁移至多个领域进行实际应用研究。总而言之，该课题以其较高的分析价值且在实际应用上有较大的提升空间，能够适应各个领域不同需求的分析，并以此在社会各行业的应用系统上进行相关改善，提升性能。