




# 基于深度学习的图像对抗 样本防御算法研究



 学号: 2020022104

 专业: 电子信息 (计算机技术)

 报告人: \*\*\*\*\*

 时间: 2023/05/31

# 目录

CONTENTS

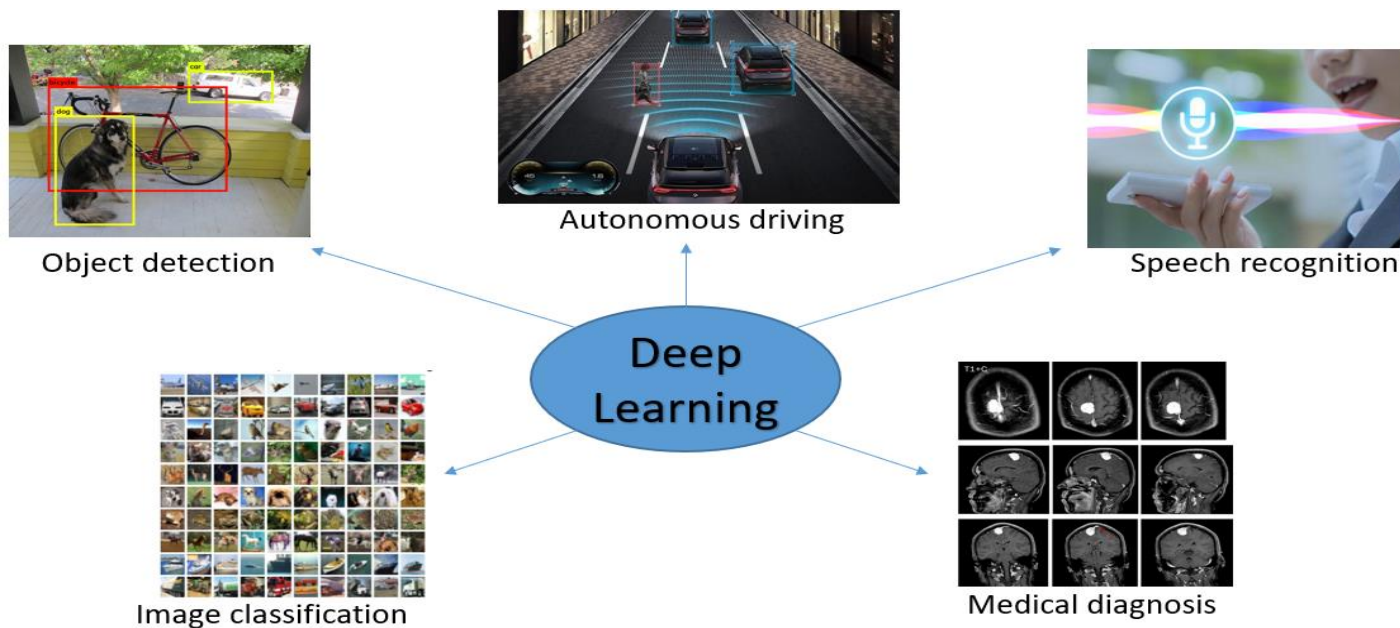
- 01 研究背景与意义
- 02 国内外研究现状
- 03 研究内容
- 04 总结与展望
- 05 科研成果



# 01 研究背景与意义



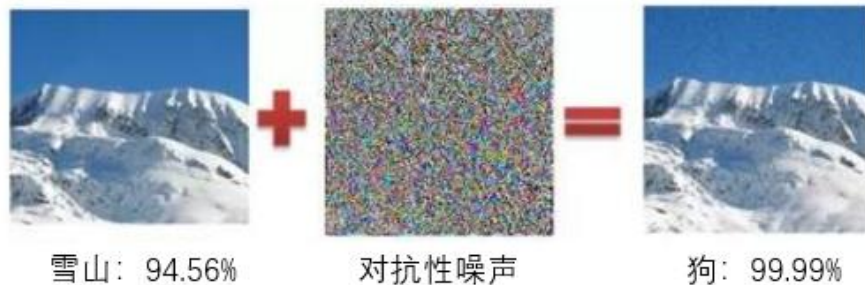
# 研究背景与意义



● 深度学习广泛应用于目标检测、自动驾驶、图像分类、语音识别等。

对抗样本

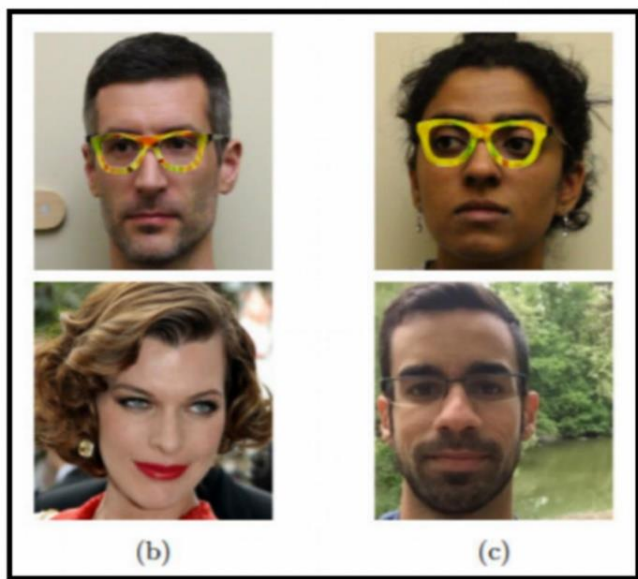
$$x' = x + \eta$$



● 深度学习中的模型极易受到对抗样本攻击。



# 研究背景与意义



通过3D打印的眼镜对抗人脸识别系统

- 目前大量的基于深度学习的技术已经被应用到现实世界中
- 人脸识别、自动驾驶、医疗、军事、基因组、多媒体等等
- 深度神经网络的鲁棒性研究显得极为重要与迫切

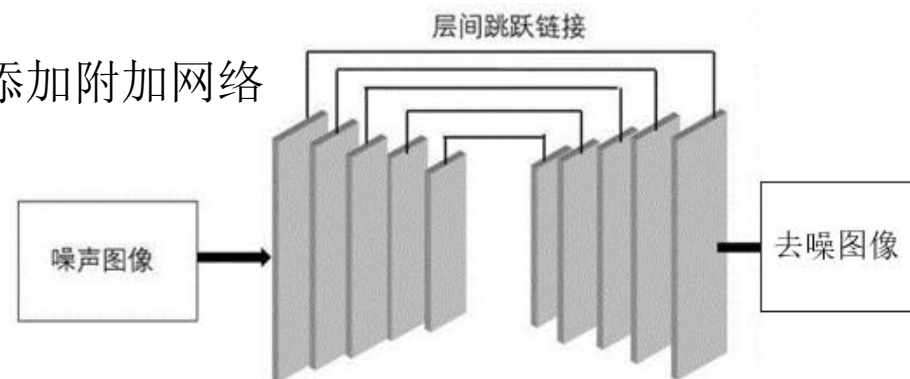
## 防御方法

### ①对抗训练

$$\min_{\theta} \mathbb{E}_{(z,y) \sim \mathcal{D}} \left[ \max_{\|\delta\| \leq \epsilon} L(f_{\theta}(X + \delta), y) \right]$$

### ②防御性蒸馏

### ④添加附加网络





## 02 国内外研究现状



# 国内外研究现状—对抗鲁棒性算法

- [1] Jin G, Shen S, Zhang D, et al. APE-GAN: Adversarial perturbation elimination with GAN[C]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019:3842-3846.
- [2] Li Y T, Ruan S, Qin H F, Deng S J, El-Yacoubi M A. Transformer Based Defense GAN Against Palm-Vein Adversarial Attacks[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1509-1523.

总结：文献[1]与文献[2]利用重构网络能够修复图像特征的特点，在训练过程中学习对抗样本的特征分布从而提高网络对扰动和干净特征的辨识能力，以此达到去除对抗扰动的能力。

- [3] Bai Y., Zeng Y., Jiang Y., Xia S., Ma X., Wang Y. Improving adversarial robustness via channel-wise activation suppressing. In International Conference on Learning Representations (ICLR), 2021.
- [4] Dawei Zhou, Tongliang Liu, Bo Han, Nannan Wang, Chunlei Peng, and Xinbo Gao. Towards defending against adversarial examples via attack-invariant features. In Proceedings of the 38th International Conference on Machine Learning, pages 12835-12845, 2021.

总结：文献[3]与文献[4]利用图像鲁棒特征与扰动特征有区别的特点，通过在网络层中抑制扰动特征的影响或者有效分离出鲁棒特征，从而达到防御效果。



## 03 研究内容





# 研究内容

## 问题

防御**准确率**、防御**泛化性**和**迁移能力**不足，亦或是模型过于**复杂**，防御过程中对原始图像可视质量损坏较大。



## 研究方案

根据对抗样本图像的特点有针对性地设计对抗防御框架，使防御泛化性、准确率与模型复杂度达到要求



### 研究内容1

基于对比学习和频率域的  
对抗样本防御算法

### 研究内容2

基于多层次自适应权重计算的  
知识蒸馏防御算法





## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 针对问题：

- 当前防御模型多是直接以重构干净样本为目标去除对抗扰动，使网络关注度过于泛化，导致网络的整体学习能力下降，可视质量也不理想。
- 鲜有防御模型使用基于决策边界的正则化，导致防御准确率、泛化性不理想。

### 提出算法：

- 提出对抗样本经多层卷积后的特征图进行通道分离，并与干净样本的高、低频特征图进行相似度计算，以有效去除对抗扰动、增强模型学习能力，并提高模型的防御能力。
- 提出基于对抗防御的对比正则化，该正则化使重构样本回归干净样本的流形决策边界，进一步提升网络去除对抗扰动的能力，并提高重构样本的分类准确率。
- 设计了类VAE的轻型网络架构CFNet, 该模型结构简单，易于部署。



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 网络框架

#### 防御算法步骤：

- 步骤1：使用不同攻击算法生成对抗样本；
- 步骤2：将对抗样本在网络层中的特征通道分离，采用干净样本的高、低频特征图进行约束；
- 步骤3：采用对比正则化对重构样本进一步约束；
- 步骤4：将重构样本与干净样本做均方误差损失，提升重构样本质量。

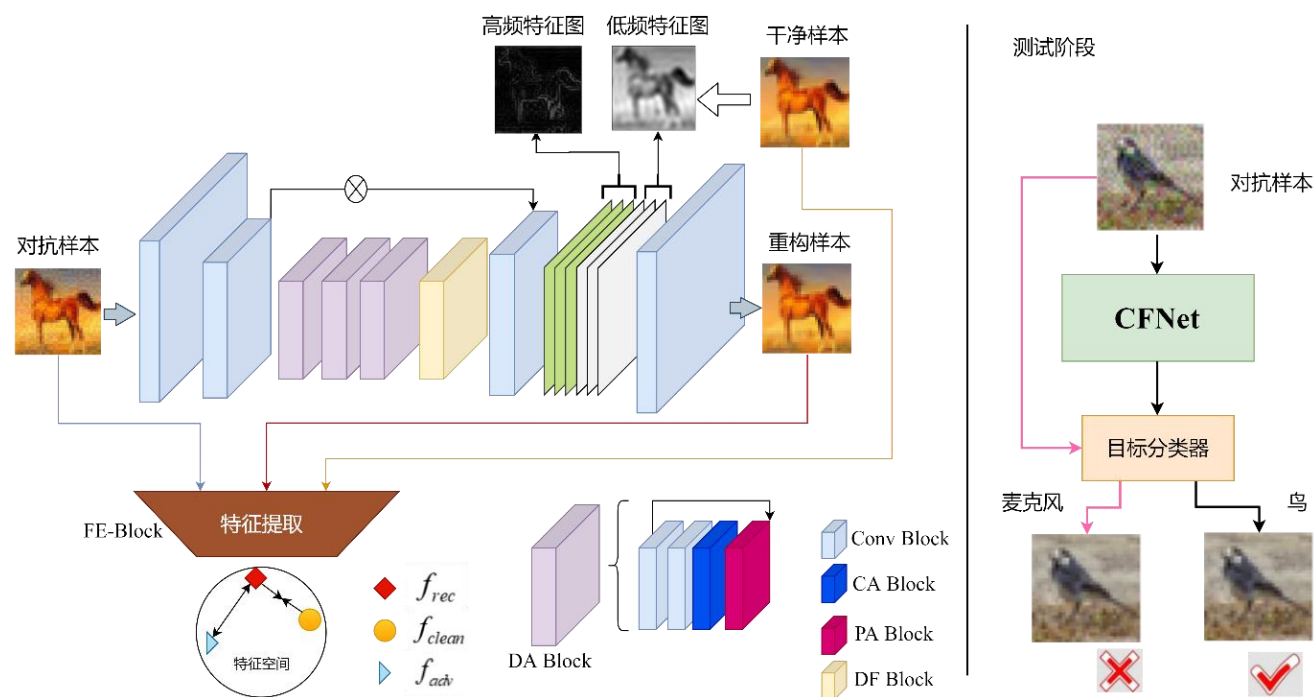


图3-1 防御算法框架



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 频率域去噪

#### 算法要点

- 对干净样本和对抗样本做傅里叶变换得到频率系数图，通过高斯低通滤波器分离得到图像的高频系数和低频系数；
- 将特定隐藏层的通道分离，前一半通道C1作为对抗样本高频信息学习通道，后一半通道C2作为对抗样本低频信息学习通道。损失函数如（公式3.1）所示。

$$Lf = \alpha \frac{1}{n} \sum (C1 - F_{clean}^H)^2 + \beta \frac{1}{n} \sum (C2 - F_{clean}^L)^2 \quad (\text{公式3.1})$$

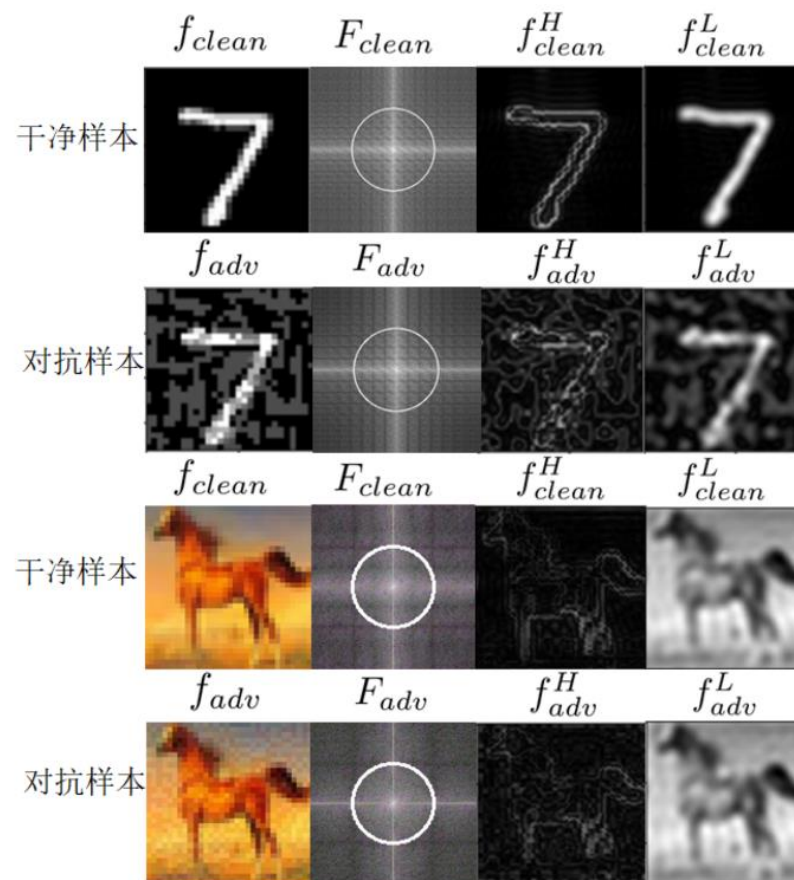


图3-2 干净样本和对抗样本高/低频分量图



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 频率域去噪

#### 算法要点

➤ 将去除对抗扰动的特定隐藏层通道融合，经过卷积后使模型重构出与干净样本极为近似的重构样本；使用感知损失（3.2）；

➤ 进一步提升重构质量，分别将重构样本和干净样本在VGG-16网络模型中得到的浅层特征做感知损失（3.3）。

$$Lre = \sum_{i=1}^n \| N(f_{adv}) - f_{clean} \|_2^2 \quad (\text{公式3.2})$$

$$Lp = \frac{1}{C_j H_j W_j} \| P_j(f_{rec}) - P_j(f_{clean}) \|_2^2 \quad (\text{公式3.3})$$



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 对比正则化

#### 算法要点

- 将重构样本作为锚点，以干净样本为正样本，而对抗样本作为负样本；
- 选择预训练好的模型VGG16作为特征提取网络，特征提取网络被划分为m个子模块，将每个子模块提取得到的特征图放入特征列表中。
- 利用互信息度量两个事件集合之间的相关性，如公式（3.4）。

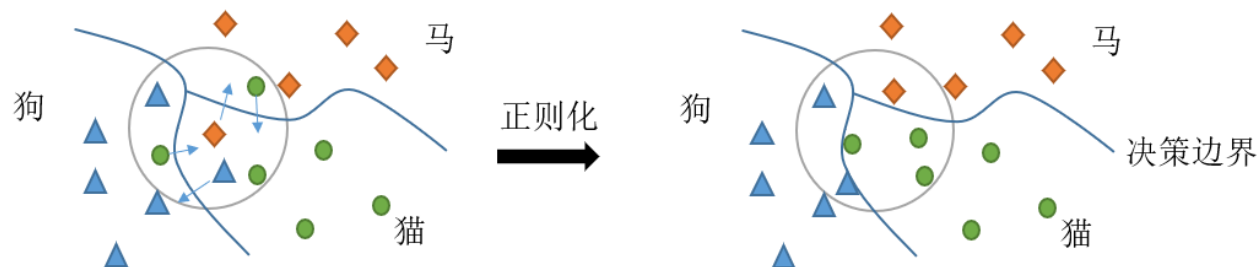


图3-3 对比正则化效果图

$$I(x_i; y_i) = H(x_i) - H(x_i | y_i) \quad (\text{公式3.4})$$

$$H(x_i) = -\sum_{a \in x_i} p(a) \log p(a) \quad (\text{公式3.5})$$

$$H(x_i | y_i) = -\sum_{a \in x_i} \sum_{b \in y_i} p(a, b) \log p(a | b) \quad (\text{公式3.6})$$



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 对比正则化

#### 算法要点

- 由此可知重构样本与干净样本的互信息，如公式（3.7）所示；
- 重构样本与对抗样本的互信息，如公式（3.8）所示。
- **最大化**重构样本与干净样本互信息的相关性，**最小化**重构样本与对抗样本互信息的相关性，如公式（3.9）所示。

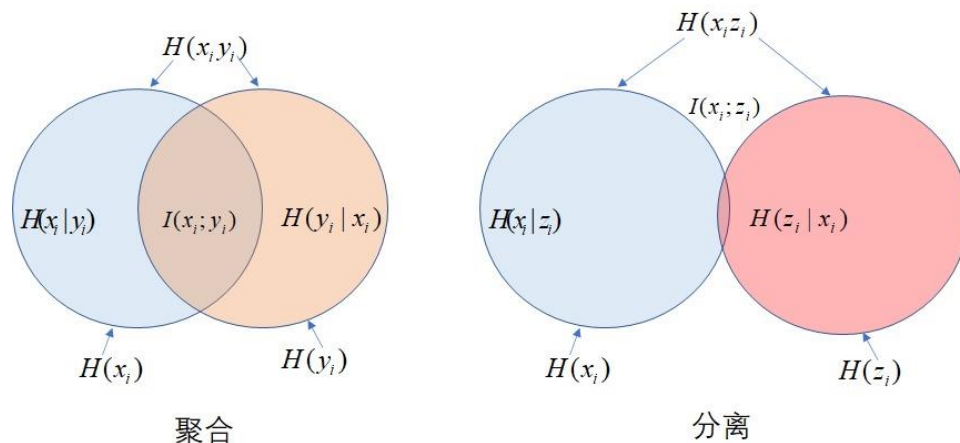


图3-4 重构样本、干净样本和对抗样本特征图互信息学习过程

$$I(f_{rec}; f_{clean}) = \sum_{i=1}^m I(x_i; y_i) \quad (\text{公式3.7})$$

$$I(f_{rec}; f_{adv}) = \sum_{i=1}^m I(x_i; z_i) \quad (\text{公式3.8})$$

$$R_{loss} = \sum_{i=1}^n \text{Max}(I_i(f_{rec}; f_{clean})) + \text{Min}(I_i(f_{rec}; f_{adv})) \quad (\text{公式3.9})$$



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 网络模型

#### 算法要点

- 为了简化模型结构，采用简单的下、上采卷积层Conv、Deconv，注意力机制模块DA-block和可变形卷积模块DF-Block；
- 总损失如公式（3.10）所示。

$$L_{\text{total}} = \alpha L_f + \beta L_{re} + \chi L_p + \delta R_{\text{loss}} \quad (\text{公式3.10})$$

$\alpha$ 、 $\beta$ 、 $\chi$ 、 $\delta$ 表示各项损失所占的权重

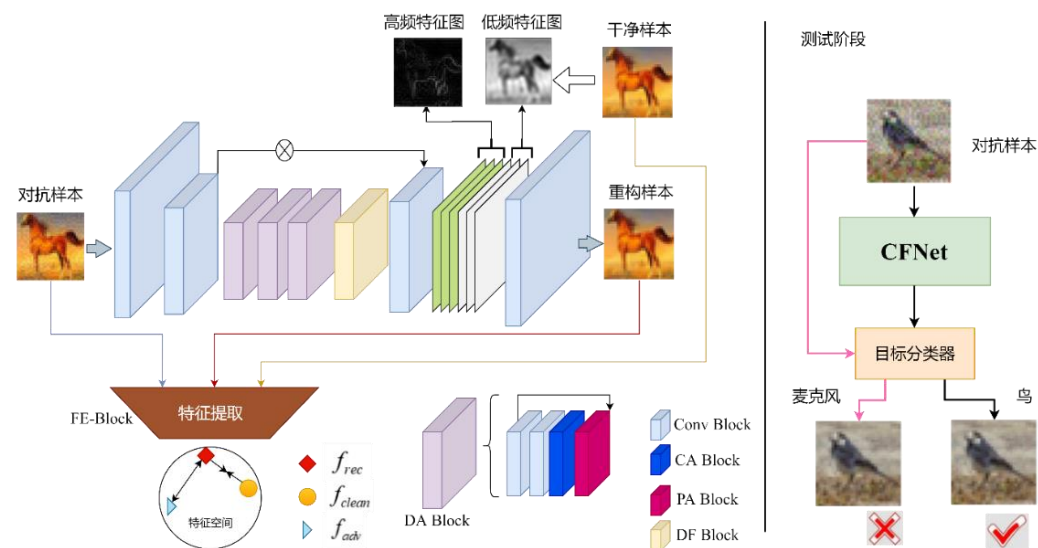


图3-5 CFNet 网络模型结构图





## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 实验

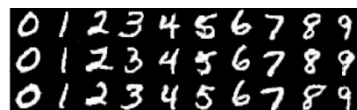
**评价指标：**实验中使用的评价指标是图像在目标分类器的分类错误率(%), 错误率越低表明防御模型的防御效果越好。

$Err = \text{sum\_err} / \text{total}$ ,  $\text{sum\_err}$ 表示错误分类的样本数,  $\text{total}$ 表示样本总数。

$Cor = \text{sum\_cor} / \text{total}$ ,  $\text{sum\_cor}$ 表示正确分类的样本数。

数据集

MNIST数据集



CIFAR-10数据集



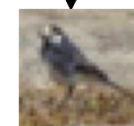
Caltech 101数据集



对抗样本

CFNet

目标分类器





## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### MNIST数据集

#### 实验结果

防御模型CFNet在模型未知的对抗样本攻击下防御错误率达到最低，表现出较强的防御能力，防御泛化能力得到极大提升。

表3-1 MNIST数据集中不同防御模型在不同攻击下的防御错误率

防御模型	$PGD_N$	$PGD_T$	$CW_N$	$DDN_N$	$AA_N$	$JSMA_T$	$PGD_{N\epsilon'}$	$AA_{N\epsilon'}$
None	100	100	100	100	100	100	100	100
AT [1]	9.63	8.38	6.42	5.91	12.60	28.59	54.34	60.06
APE-GAN [2]	8.76	3.20	2.34	1.91	12.40	36.49	34.86	46.72
HGD [3]	1.89	1.30	1.67	1.23	2.43	50.62	75.79	90.34
ARN [4]	1.85	1.29	1.45	1.22	2.38	16.75	15.27	26.84
CFNet(Ours)	<b>1.01</b>	<b>0.39</b>	<b>0.32</b>	<b>0.27</b>	<b>0.97</b>	<b>0.62</b>	<b>2.01</b>	<b>3.58</b>

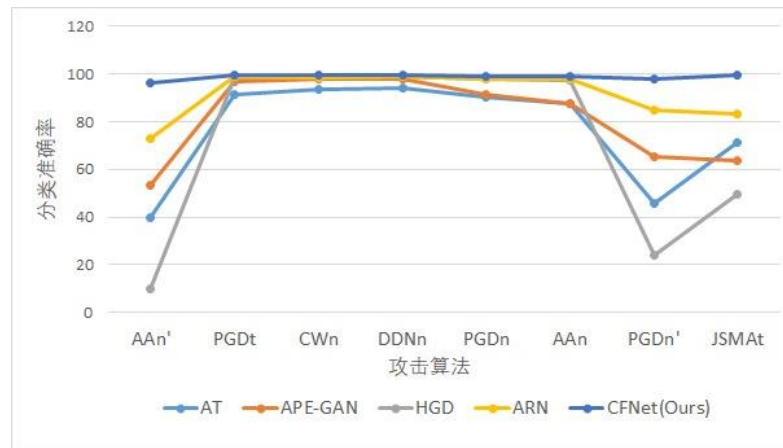


图3-6不同攻击算法下各种防御模型分类准确率



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### MNIST数据集

#### 实验结果

防御模型迁移到其他分类器上仍具有较高的分类准确率。所提防御模型获取的重构样本在不同目标分类器的**迁移性较强**。

表3-2 防御模型在不同目标分类器上的分类错误率

防御模型	Model_A	Model_B	Model_C	ResNet-50	VGG-16
None	0.93	1.14	3.12	0.60	0.73
CFNet(ours)	1.88	2.57	11.03	0.64	0.72

表3-3 模型A、B、C的网络结构

Model_A	Model_B	Model_C
Conv(1, 64, 5, 1, 2) + ReLU	Conv(1, 128, 3, 1, 1) + ReLU	FC(784) + ReLU
Conv(64, 64, 5, 2) + ReLU	Conv(128, 64, 5, 2) + ReLU	Dropout (0.25)
Dropout (0.25)	Dropout (0.25)	FC(200) + ReLU
FC(9216) + ReLU	FC(9216) + ReLU	Dropout (0.25)
Dropout (0.25)	Dropout (0.25)	FC(200) + Softmax
FC(128) + Softmax	FC(128) + Softmax	



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### CIFAR-10数据集

#### 实验结果

防御模型CFNet在模型未知的对抗样本攻击下防御错误率达到最低，表现出较强的防御能力，防御泛化能力得到极大提升。

表3-4 不同攻击实验分类错误率（CIFAR-10）

防御模型	$PGD_N$	$PGD_T$	$CW_N$	$DDN_N$	$AA_N$	$JSMA_T$	$PGD_{N\epsilon'}$	$AA_{N\epsilon'}$
None	100	100	100	99.99	100	100	100	100
AT <sup>[1]</sup>	51.02	49.68	50.17	49.19	53.66	44.59	59.09	61.65
APE-GAN <sup>[2]</sup>	44.38	39.09	23.18	24.73	60.09	39.10	79.34	87.16
HGD <sup>[3]</sup>	39.44	23.03	12.46	10.04	42.34	38.65	57.97	58.41
ARN <sup>[4]</sup>	38.66	20.43	11.47	14.64	38.94	35.49	49.45	52.64
CFNet(Ours)	21.55	13.42	12.57	12.55	18.74	16.74	26.90	21.83

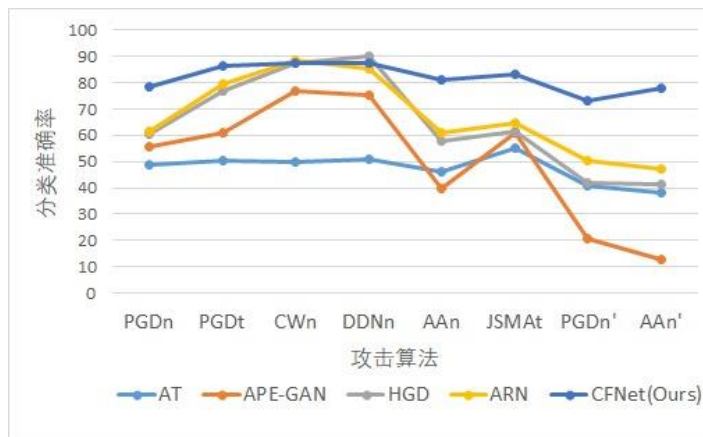


图3-7 不同攻击算法下各种防御模型分类准确率



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### Caltech 101数据集

#### 实验结果

防御模型在**大尺寸数据集**上防御错误率仍然取得不错的结果，表现出较强的防御能力。

表3-5 在 Caltech 101 数据集中防御模型在不同攻击下的防御错误率

防御模型	FGSM	PGD	BIM	DeepFool	CW
APE-GAN <sup>[2]</sup>	40.1	—	—	45.9	26.1
DADL <sup>[5]</sup>	<b>25.09</b>	41.54	—	—	—
Crop-Ens <sup>[6]</sup>	44.54	—	40.35	39.77	40.20
STL <sup>[7]</sup>	44.12	—	36.52	35.32	37.8
CFNet(Ours)	26.05	<b>31.96</b>	<b>26.52</b>	<b>20.92</b>	<b>20.83</b>



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### Caltech 101数据集

#### 实验结果

防御模型的总参数量及参数大小均小于对比模型，单张对抗样本的去噪时间降低至0.03S左右。实验表明提防御模型较为轻量，易于部署。

表3-6 模型参数量大小

模型	总参数量	参数大小 (MB)
HGD <sup>[3]</sup>	11, 037, 699	42. 11
ARN <sup>[4]</sup>	155, 475, 200	593. 09
CFNet(Ours)	<b>5, 781, 928</b>	<b>22. 06</b>

表3-7 单张对抗样本的去噪时间 (S)

模型	FGSM	PGD <sub>N</sub>	PGD <sub>T</sub>	CW <sub>N</sub>	DDN <sub>N</sub>
CFNet	0. 04	0. 02	0. 03	0. 06	0. 04



## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 消融实验

#### 实验结果

取消对抗样本卷积后特征图分离模块，取消干净样本频域变换后防御模型防御准确率下降。取消对比正则化损失后准确率也降低至60%左右。

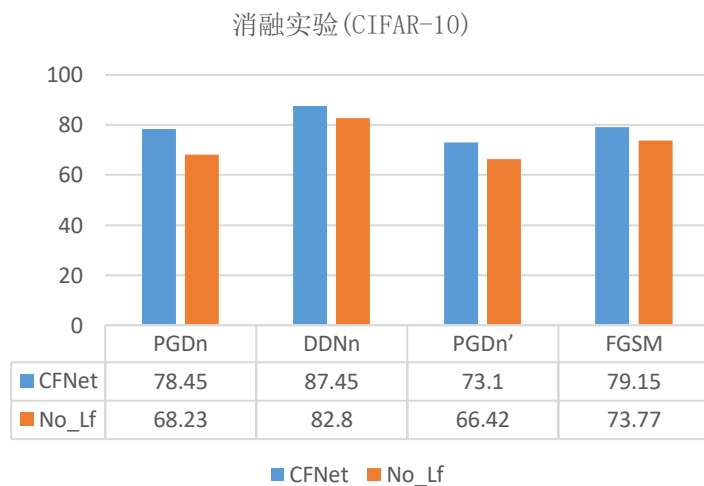


图3-8去掉 Lf 损失后的分类准确率

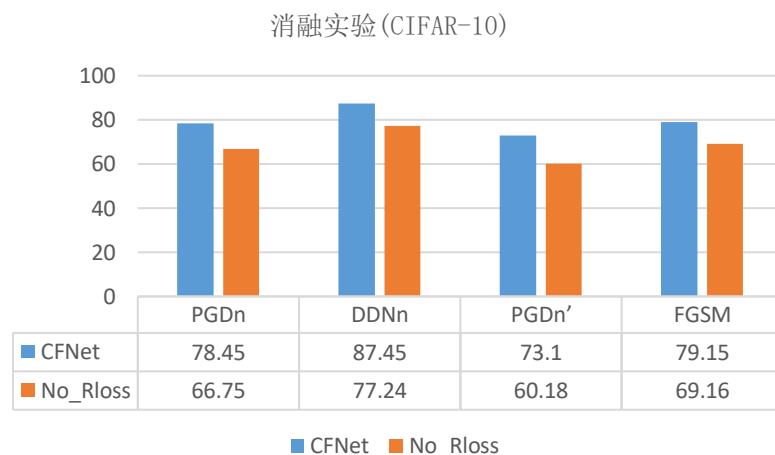


图3-9去掉对比损失后的分类准确率





## 研究内容1——基于对比学习和频率域的对抗样本防御算法

### 可视化实验

#### 实验结果

防御模型重构后的重构样本有效的去除了对抗扰动，并且图像质量较高。防御模型不仅在防御准确率、泛化性等表现优异，在防御过程中几乎不影响图像可视质量。



图3-10 CFNet 修复的 MNIST 图像

图3-11 CFNet 修复的 CIFAR-10 图像

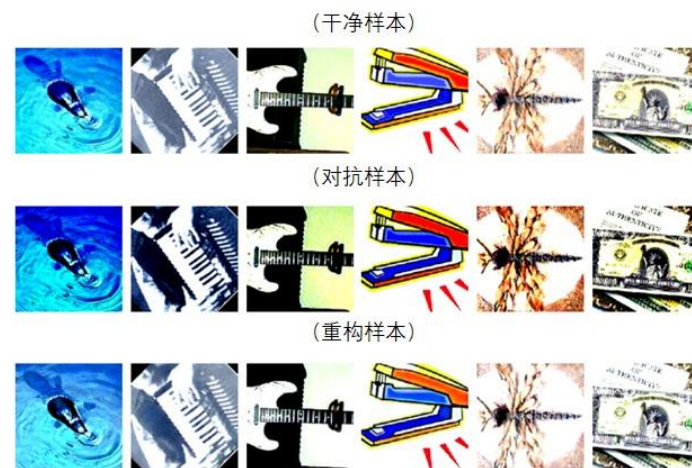


图3-12 CFNet 修复的 Caltech 101 图像





## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 针对问题：

- 当前算法通常存在防御泛化能力以及迁移性较差等特点，网络对特殊特征的关注度不够。
- 防御泛化性有待提高，鲜有文献采用中间特征提升防御能力，导致网络对特征分辨能力不强。

### 提出算法：

- 提出一种多层次自适应权重计算模块，增强网络的学习能力。
- 设计并实现了基于知识蒸馏的去噪防御模型（DDNet），将教师模型学到的知识，迁移到学生模型针对对抗样本的去噪中，约束学生模型深层特征的学习。
- 提出防御对比正则化强化学生模型的重构能力。



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 多层次自适应权重计算模块(multi-level adaptive weight calculation module,MAWM)

#### 算法要点

- 对输入样本采用不同大小的卷积核进行特征提取；
- 将提取的特征分别传入通道注意力模块进行重要特征通道的筛选，并将结果进行通道拼接；
- 再一次利用**通道注意力优化**，最后将筛选出的重要通道进行**空间注意力优化**。

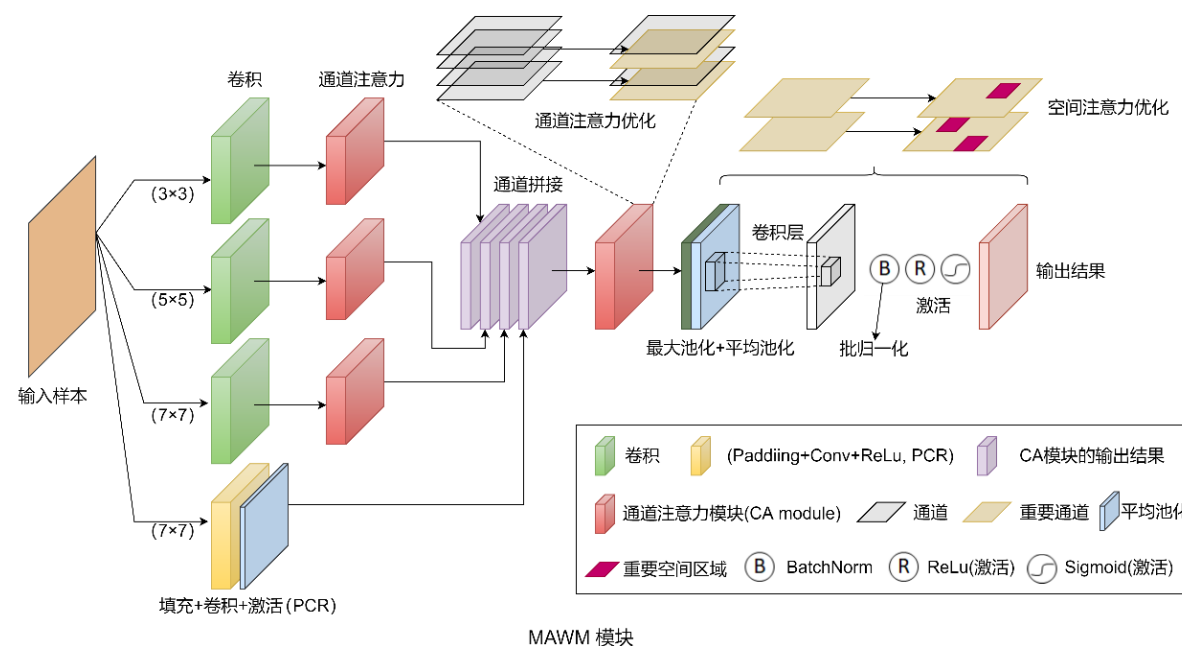


图3-13 多层次自适应权重计算模块(MAWM)



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 多层次自适应权重计算模块

#### 算法要点

- 使用全局平均池化(GAP)获取每个通道的标量特征，进而计算得到每个通道的注意力权重；
- 经过Sigmoid函数激活后，每个通道将得到一个0-1范围内的权重值；
- 将权重值与输入样本点乘，输入样本中重要的特征通道将会凸显。

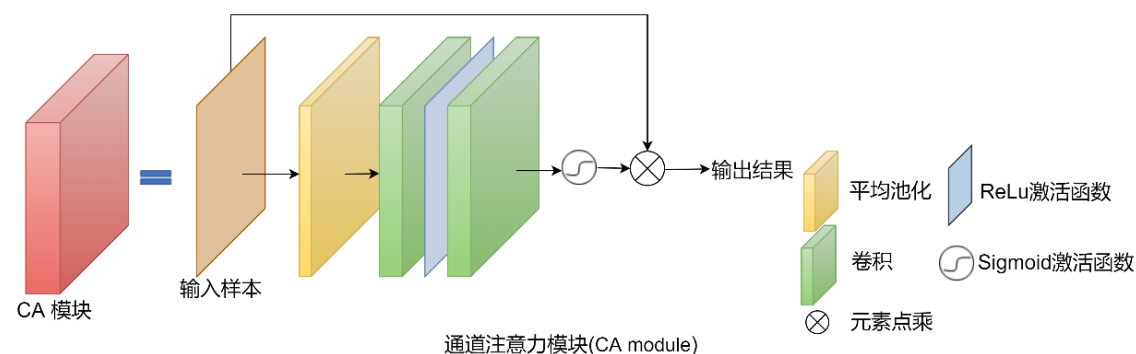


图3-14 通道注意力模块（CA module）



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

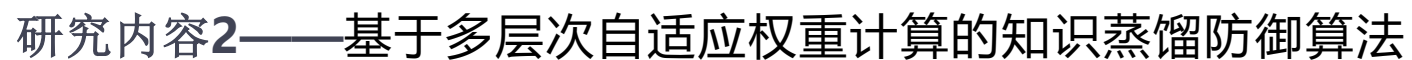
### 知识蒸馏

#### 算法要点

- 首先训练教师模型，其中采用的感知损失如公式(3.11)所示；
- 将教师模型中学到的知识迁移到学生模型，逐层去除对抗扰动，损失如公式(3.12)所示；
- 使用不同的中间层损失权重系数，使中间层特征的学习能力更强，达到去除对抗扰动的效果。

$$L_T = \sum_{i=1}^n \|T(I_{clean}) - I_{clean}\|_2^2 \quad (\text{公式3.11})$$

$$L_G = \sum_{i=1}^n \|T^i(I_{clean}) - S^i(I_{adv})\|_1 \quad (\text{公式3.12})$$



## 算法要点

- 学生模型主干模块由混合注意力模块组成；
- 更好的捕获图像的重要通道信息，再由空间注意力模块来自适应的强调重要区域的特征，从而增强网络对噪声区域的关注度。

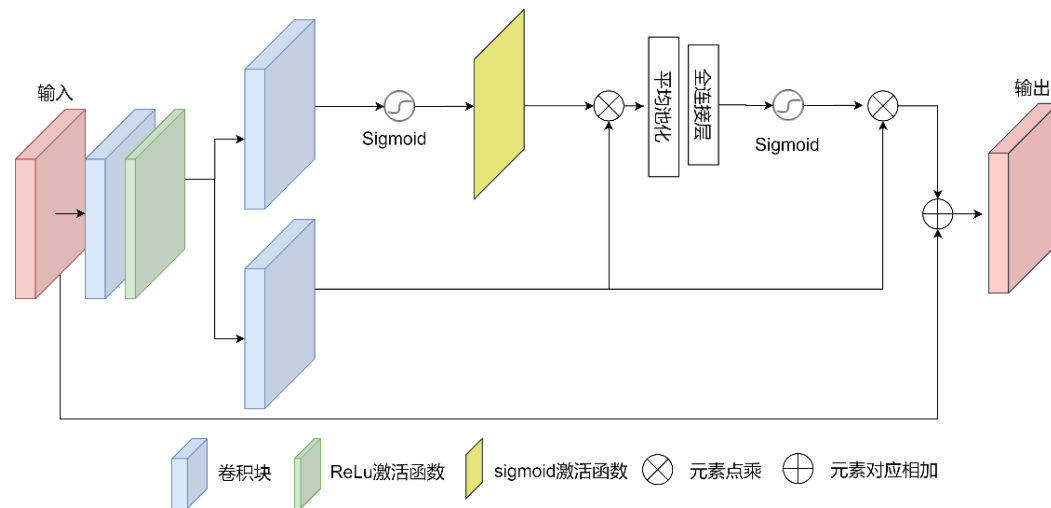


图3-15 主干网络注意力模块



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 防御对比正则化损失(补)

#### 算法要点

- 将重构样本作为目标样本。干净样本及对抗样本作为正、负样本；
- 从网络不同层之间分别提取目标样本，正、负样本的对应特征，并在特征之间做损失得到特征图距离损失 $L_v$ ，如公式(3.13，3-14)所示；
- 利用不同层的特征距离损失，计算得到防御对比正则化损失如公式(3.15)所示；

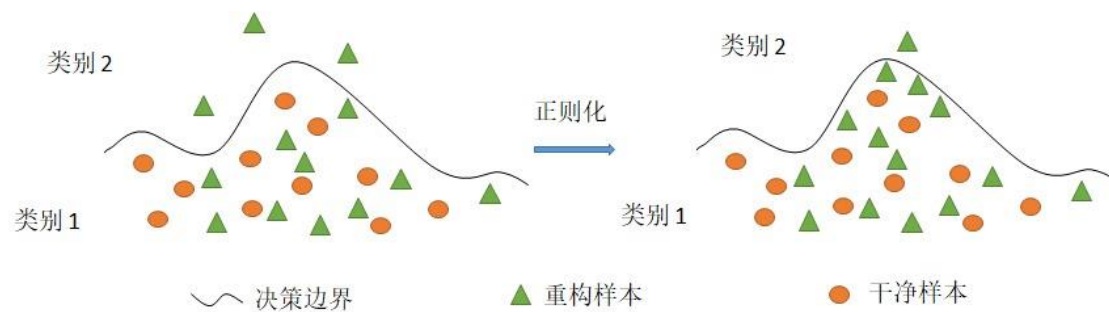


图3-16 防御对比正则化（DCR）

$$F_{rec}^i, F_{clean}^i, F_{adv}^i = VGG^i(I_{rec}, I_{clean}, I_{adv}) \quad (\text{公式3.13})$$

$$L_v = \frac{\|F_{rec}^i - F_{clean}^i\|_1}{\|F_{rec}^i - F_{adv}^i\|_1 + \omega} \quad (\text{公式3.14})$$

$$L_C = \alpha L_v^i + \beta L_v^{i+1} + \chi L_v^{i+2} + \delta L_v^{i+3} \quad (\text{公式3.15})$$



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 网络模型

#### 算法要点

- 教师模型中主要学习干净样本的特征，从而使网络记住干净特征的权重参数；
- 学生模型主要由下采模块、主干模块以及上采模块组成。下采模块采用MAWM，主干网络由transformer模块、注意力模块组成。上采模块采用双线性插值上采，后接卷积及激活层获得重构样本；

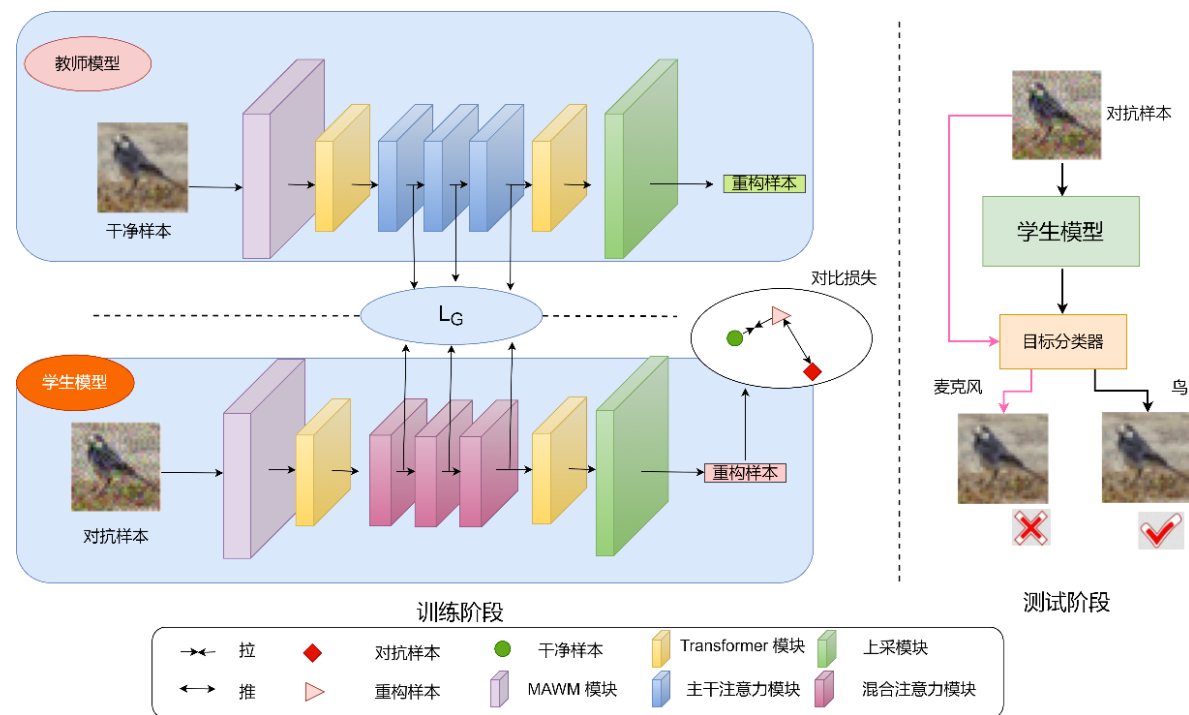


图3-17 网络模型结构图



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 网络模型

#### 算法要点

- 学生模型中的重构损失如公式(3.16)所示;
- 为了提高重构样本的可视质量, 本文采用预训练的VGG-19网络对重构样本及干净样本做特征相似度损失, 公式(3.17);
- 为了保证重构图像与干净样本的结构相似性, 本文将重构样本与干净样本做结构相似性损失, 公式(3.18)
- 学生模型的总损失如公式(3.19)

$$L_{REC} = \sum_{i=1}^n \|S(I_{adv}) - I_{clean}\|_2^2 \quad (\text{公式3.16})$$

$$L_P = \sum_{i=1}^m \|F_{rec}^i - F_{clean}^i\|_1 \quad (\text{公式3.17})$$

$$L_{SSIM} = ssim(I_{rec}, I_{clean}) \quad (\text{公式3.18})$$

$$L_{Total} = \eta L_{REC} + \iota L_P + \kappa L_C + \lambda L_{SSIM} + \mu L_G \quad (\text{公式3.19})$$





## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 对抗防御白盒测试

#### 评价指标

重构样本在目标分类器的分类  
正确率(%),  $Cor = \text{sum\_cor}/\text{total}$ ;

#### 实验结果

防御模型在模型已知对抗攻击  
CW 攻击中, 防御准确率为  
99.30%, 在模型未知的对抗攻  
击中, 防御准确率均超过其它  
对比模型。

表3-8 在MNIST数据集中各类防御模型在不同攻击下的防御准确率 (%)

数据集	防御模型	攻击算法							
		Clean	PGD <sub>N</sub>	PGD <sub>T</sub>	CW <sub>N</sub>	DDN <sub>N</sub>	AA <sub>N</sub>	PGD <sub>N'</sub>	AA <sub>N'</sub>
MNIST	Normal	99.28	0	0	0	0	0	0	0
	AT <sup>[1]</sup>	98.81	90.37	91.62	93.58	94.09	87.4	45.66	39.94
	APE-G <sup>[2]</sup>	98.43	91.24	96.8	97.66	98.09	87.6	65.14	53.28
	HGD <sup>[3]</sup>	98.82	98.11	98.7	98.33	98.77	97.57	24.21	9.66
	ARN <sup>[4]</sup>	98.89	98.15	98.71	98.55	98.78	97.62	84.73	73.16
	Ours	99.26	98.68	99.61	99.30	99.26	98.47	93.66	94.14

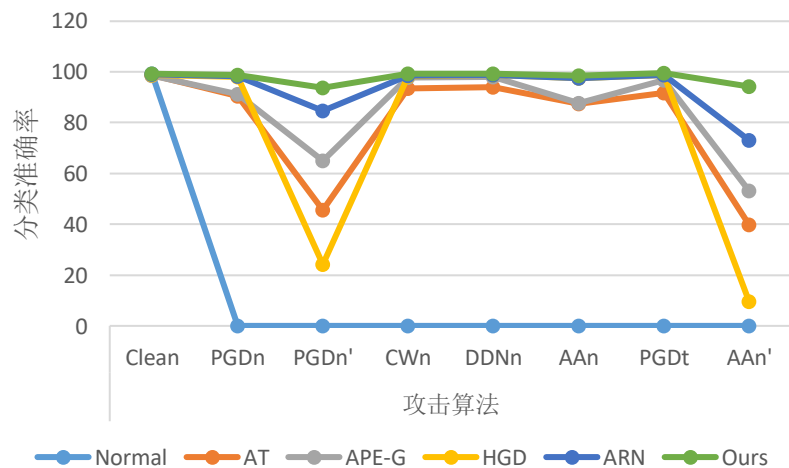


图3-18不同防御模型在不同攻击下  
防御效果折线图



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 对抗防御白盒测试

#### 实验结果

防御模型在模型已知对抗攻击PGD攻击中，防御准确率为84%，在CW攻击中，防御准确率均超过其它对比模型。

表3-9 在CIFAR-10数据集中不同攻击下的分类准确率(%)

数据集	防御模型	攻击算法					
		Clean	PGD- $L_{\infty}$	PGD- $L_2$	CW- $L_{\infty}$	CW- $L_2$	StAdv
CIFAR-10	Normal	92.36	0.0	0.0	0.0	0.0	0.0
	AT <sup>[1]</sup>	86.8	51.7	24.3	52.0	26.0	4.8
	HGD <sup>[3]</sup>	80.75	75.93	75.44	75.84	77.15	23.04
	APE-G <sup>[2]</sup>	90.93	59.28	65.17	59.23	65.30	7.28
	CD-VAE <sup>[8]</sup>	86.81	77.05	78.02	77.04	78.29	19.41
	Ours	88.07	78.90	84	86.21	86.63	11.23



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 基于空间域对抗攻击的防御

#### 实验结果

在 MNIST 数据集中，本文所提防御模型在与 HGD、ARN 等优秀防御模型相比较下防御准确率平均值 Avg 提高了 17%，所提防御模型在面对基于空间域的攻击时也能表现出很好的防御效果。

表3-11 防御模型在基于空间域的对抗样本攻击下的攻击错误率(%)

数据集	防御模型	攻击算法					
		Clean	STAN	STAT	FWAN	FWAN'	Avg
MNIST	Normal	<b>0.72</b>	100	100	98.56	99.91	–
	APE-G <sup>[2]</sup>	1.57	16.57	21.40	33.95	51.81	30.93
	HGD <sup>[3]</sup>	1.36	21.32	36.41	50.43	71.12	44.82
	ARN <sup>[4]</sup>	1.16	9.08	13.73	25.79	43.76	23.09
	Ours	0.78	<b>1.89</b>	<b>3.13</b>	<b>6.92</b>	<b>10.79</b>	<b>5.68</b>
CIFAR-10	Normal	<b>7.64</b>	100	100	99.83	99.98	–
	APE-G <sup>[2]</sup>	23.08	47.19	36.46	42.79	50.53	44.24
	HGD <sup>[3]</sup>	10.41	42.89	31.97	37.67	43.41	38.99
	ARN <sup>[4]</sup>	8.21	36.81	23.62	24.17	31.89	29.12
	Ours	11.93	<b>22.92</b>	<b>11.80</b>	<b>24.10</b>	<b>31.58</b>	<b>22.6</b>



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 对抗防御黑盒攻击实验

表3-12和表3-13中，“No defense”表示在FGSM攻击下，对抗样本在目标分类器的分类正确率。

表3-12 在 MNIST 数据集上的黑盒测试

Classifier	No attack	No defense	With defense (Ours)				
			FGSM	PGD	CW	DDN	Avg
Model A	99.07	13.11	98.25	97.97	98.78	99.16	<b>98.54</b>
Model B	98.87	15.25	97.93	97.73	98.59	98.74	<b>98.25</b>
Model C	97.05	1.32	94.0	92.74	96.67	96.66	<b>95.02</b>
Model D	98.35	8.96	93.75	95.62	98.1	98.09	<b>96.39</b>

表3-13在 CIFAR-10 数据集上的黑盒测试

Classifier	No attack	No defense	With defense (Ours)				
			FGSM	PGD	CW	DDN	Avg
VGG-16	80.38	9.93	61.15	55.76	62.21	54.16	<b>58.32</b>
VGG-19	82.46	9.23	60.48	55.1	65.7	54.06	<b>58.84</b>
ResNet-34	85.10	6.87	62.79	47.39	82.24	83.82	<b>69.06</b>
ResNet-50	85.37	7.01	63.73	52.01	81.06	82.99	<b>69.95</b>



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 对抗防御量化实验

#### 实验结果

攻击下各类别之间相互交错，很难区分类别标签，图右边表示经过防御模型防御后的类别簇状图，**各类别轮廓清晰**。在可视化实验中，修复样本**接近**原始干净样本，未见扰动信息。

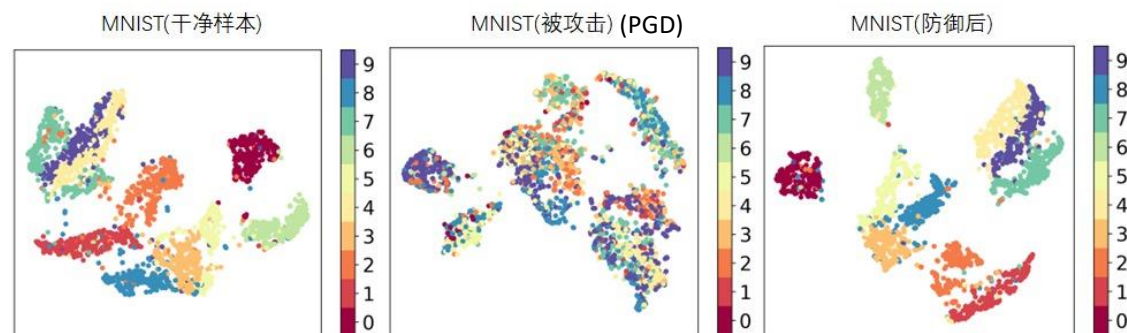


图3-21 MNIST 防御前后的聚类图



图3-22不同对抗样本攻击下的可视化图像



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 大尺寸数据集实验

#### 实验结果

在大尺寸数据集上，防御分类准确率保持较高水准，图像可视质量优良。在大尺寸数据集上，本防御模型仍具有很好的防御能力以及防御泛化能力。

表3-14防御模型在不同攻击下的分类准确率 (Caltech 101)

数据集	防御模型	攻击算法					
		Clean	FGSM	L-BFGS	DeepFool	CW-L <sub>2</sub>	CW-L <sub>∞</sub>
Caltech 101	APE-G <sup>[2]</sup>	76.0	59.9	57.1	54.1	73.9	73
	Ours	83.33	63.43	68.12	76.53	75.11	75.06



图3-23 Caltech 101 去噪效果图





## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 消融实验

#### 实验结果

去掉多层次自适应权重计算模块后，网络整体防御能力有所下降，去掉教师模型指导损失，尤其是在FWA和PGD大扰动攻击下，防御准确率分别下降了23.64%和18.46%。

表3-15 在 CIFAR-10 数据集上的消融实验

消融实验	PGD <sub>N</sub>	PGD <sub>N'</sub>	FWA	FGSM
DDNet	78.90	72.09	75.90	76.62
No_Lc	74.68	62.42	65.79	74.97
No_Lg	73.03	53.63	52.26	72.89
No_module	73.62	70.03	72.27	73.56

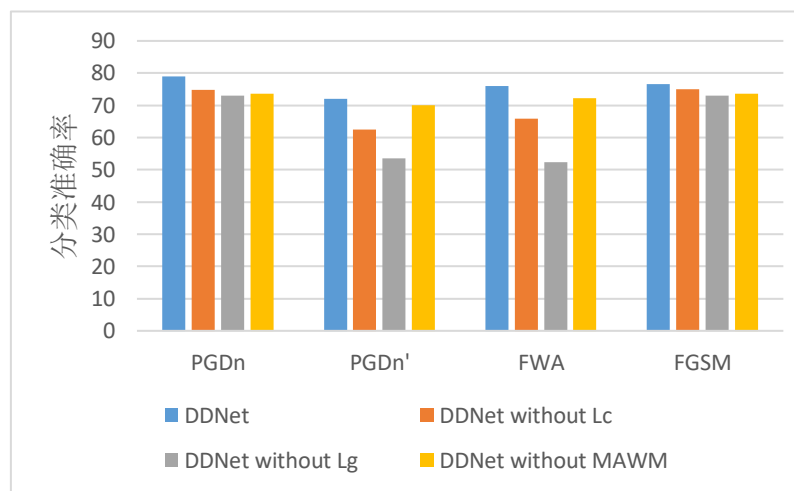


图3-24 消融实验直方图



## 研究内容2——基于多层次自适应权重计算的知识蒸馏防御算法

### 可视化实验

#### 实验结果

由上到下分别是MNIST、CIFAR-10和Caltech 101数据集的可视化实验，从实验结果可知，防御模型在去除对抗扰动的过程中对原图可视质量的影响是微小的



图3-25 不同数据集可视化实验





## 04 总结与展望



## 总结与展望

### (1) 提出基于对比学习和频率域的对抗样本防御算法。

#### 创新点

- 提出了一种基于对比学习和频率域的对抗样本防御算法，设计了防御模型 CFNet。
- 提出对比正则化（Contrastive Regularization, CR），使重构样本回到干净样本的流形决策边界，从而提高网络的分类准确率。

#### 结论

- 提防御模型超过当前优秀的防御模型，在模型未知的（unseen type）对抗样本攻击下仍表现出较高的防御能力。

#### 展望

- 防御性能仍有提升空间，而且在某些特定攻击下可能表现不佳。因此，下一步的研究重点是如何完善现有的对抗防御方法并增强其通用防御能力



## 总结与展望

### (2) 提出基于多层次自适应权重计算的知识蒸馏防御算法。

#### 创新点

- 提出了一种基于多层次自适应的知识蒸馏防御算法，提出多层次自适应权重计算模块MAWM。
- 结合知识蒸馏设计并实现了蒸馏防御模型DDNet。
- 设计防御对比正则化（Defense Contrastive Regularization, DCR）用于强化学生模型的去噪重构效果

#### 结论

- 所提防御模型超过当前最优秀的防御模型，且防御模型的泛化能力、迁移能力得到大幅提升。
- 在模型未知攻击算法的攻击下防御准确率提高 11%，在白、黑盒测试中均表现出很好的防御性能。

#### 展望

- 当前的防御场景仅针对图像分类这一领域，而深度学习在现实中的应用是各方面的，因此要扩展防御方法在不同领域的应用



## 05 科研成果



## 科研成果

### 1.学术成果

- [1] Joint contrastive learning and frequency domain defense against adversarial examples[J]. NEURAL COMPUTING & APPLICATIONS, 2023. (SCI3区, 第一作者, CCF C类)
- [2] 基于可逆网络的对抗样本防御算法的设计与研究 [J].贵大学学报, 2023,40-5. (中文期刊, 第一作者)
- [3] DDNet: A knowledge distillation defense network based on multi-level adaptive weight calculation for adversarial defense, NEURAL NETWORKS Under Review. (SCI1区Top, 第一作者, CCF B类)
- [4] 获授权国家发明专利《一种基于多尺度特征学习的弥散加权图像的鲁棒水印方法》(共同作者)

### 2.参与科研项目

- [1] 国家自然科学基金项目一项
- [2] 贵州省科技计划项目

### 3.荣获奖项

研究生奖学金



## 参考文献

- [1] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks[C]. Proceedings of the International Conference on Learning Representations, 2018:1-27.
- [2] Jin G, Shen S, Zhang D, Dai F, Zhang Y. Ape-gan: Adversarial perturbation elimination with gan[C]//ICASSP 2019- 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 3842-3846.
- [3] Liao F, Liang M, Dong Y, Pang T, Hu X, Zhu J. Defense against adversarial attacks using high-level representation guided denoiser[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1778- 1787.
- [4] Zhou D, Liu T, Han B, Wang N, Peng C, Gao X. Towards defending against adversarial examples via attack-invariant features[C]//International Conference on Machine Learning. PMLR, 2021: 12835-12845.
- [5] Shao R, Perera P, Yuen P C, Patel V M. Open-set adversarial defense with clean-adversarial mutual learning[J]. International Journal of Computer Vision, 2022, 130(4): 1070-1087.
- [6] Guo C, Rana M, Cisse M, Van D M L. Countering adversarial images using input transformations[J]. arXiv preprint arXiv:1711.00117, 2017.
- [7] Sun B, Tsai N, Liu F, Yu R, Su H. Adversarial defense by stratified convolutional sparse coding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11447-11456.
- [8] Yang K, Zhou T, Zhang Y, Tian X, Tao D. Class-disentanglement and applications in adversarial detection and defense[J]. Advances in Neural Information Processing Systems, 2021, 34: 16051-16063.
- [9] Rice L, Wong E, Kolter Z. Overfitting in adversarially robust deep learning[C]//International Conference on Machine Learning. PMLR, 2020: 8093-8104.
- [10] Laidlaw C, Singla S, Feizi S. Perceptual adversarial robustness: Defense against unseen threat models[J]. arXiv preprint arXiv:2006.12655, 2020.

# 谢谢老师，请老师予以指导



答 辩 人：\*\*\*\*\*