SURE: SUrvey REcipes for building reliable and robust deep networks

摘要

本研究旨在复现论文《SURE: SUrvey REcipes for building reliable and robust deep networks》中的方法，并验证其在实际任务中的有效性。论文提出了一种名为 SURE 的新框架，通过集成多种模型正则化、分类器和优化技术，显著提高了深度神经网络（DNN）在处理不确定性预测时的准确性。我们首先复现了 SURE 方法，并评估了其在 CIFAR100 和其他数据集上的表现。通过对比基线模型 MSP，逐步将 RegMixup、CRL、SAM、SWA 和 CSC 等技术集成到 SURE 框架中，我们观察到这些技术的增量影响，并评估了每个组件对模型性能的贡献。我们通过多次实验验证了 SURE 方法在故障预测等不确定性估计任务中的优越性，并展示了它在处理现实世界挑战（如数据腐蚀、标签噪声和长尾分类分布）时的鲁棒性。尤其是在 Animal-10N 和 Food-101N 数据集上，SURE 展现出了与最先进方法相当的性能，且无需任何任务特定的调整。我们的实验结果验证了 SURE 方法在各类数据集和模型架构中的有效性，表明 SURE 为不确定性估计领域的稳健性提供了新的基准。

关键词：深度神经网络；鲁棒性；

1 引言

深度神经网络（DNNs）已经在多个领域取得了显著的进展，包括图像分类、自然语言处理和强化学习。然而，这些模型在面对现实世界中不确定性和扰动时，往往表现出不可靠和不鲁棒的特性。SURE（SUrvey REcipes）旨在总结和推荐一些有效的策略与技巧，用于构建更加可靠和鲁棒的深度网络。

本文的目标是复现论文《SURE: SUrvey REcipes for building reliable and robust deep networks》[4] 中的方法，并分析其实现细节与实验结果。我们将详细介绍如何实现论文中的方法，包括训练数据集的选择、模型架构的设计、训练技巧等，并对实验结果进行讨论。

2 相关工作

在深度学习中，模型的可靠性和鲁棒性一直是一个重要的研究方向。过去的研究主要集中在以下几个方面：

模型正则化：如 Dropout、L2 正则化等技巧，旨在防止模型过拟合并提高泛化能力。数据增强：通过数据变换（如翻转、旋转、裁剪等）来扩展训练集，从而增加模型的鲁棒性。对

抗训练：生成对抗样本并在训练中进行优化，以增强模型对抗攻击的鲁棒性。网络架构改进：通过设计更加复杂或特殊的网络结构，提高模型的可靠性和表现。SURE 方法总结了这些技术，并提出了一些新的策略，如如何选择训练数据、网络架构优化等，从而帮助研究人员构建更加可靠和鲁棒的模型。

3 本文方法

3.1 本文方法概述

本文提出了一个 SURE 的方法，该方法结合了五种方法，实验证明 SURE 方法达到比其他方法更好的效果。这五种方法分别是：RegMixup、CSC、CRL、SAM、SWA。为了更好的了解本文的方法，不得不详细介绍这五种方法。

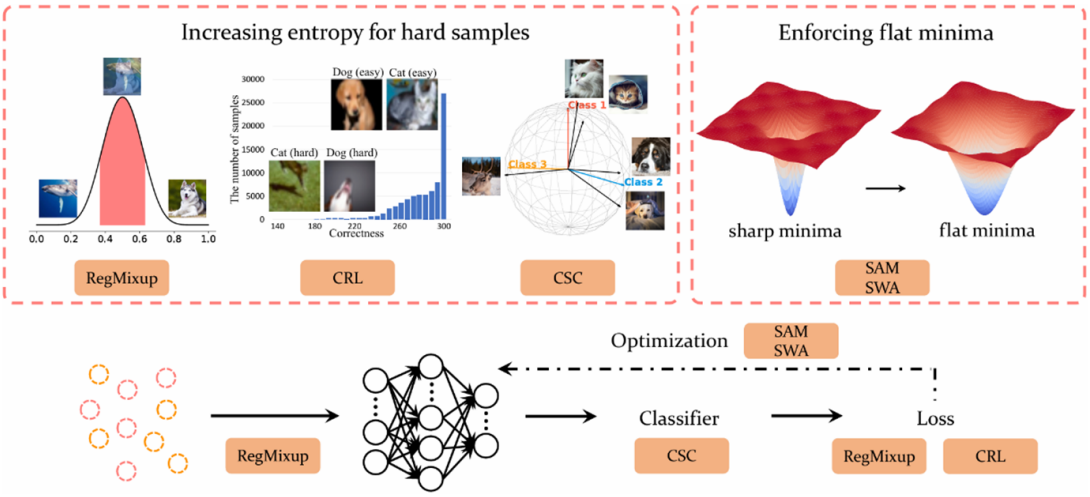

图 1. 方法示意图

3.2 RegMixup

RegMixup [6] 是一种在神经网络训练中用于增强正则化的方法，旨在通过结合不同样本的特征和标签来提高模型的泛化能力。这是 Mixup 的一种扩展，Mixup 本身通过对训练样本进行线性插值来生成新的样本，从而实现数据增强和正则化。RegMixup 在此基础上进一步增加了正则化项，帮助模型在训练过程中更好地避免过拟合。Mixup 是一种数据增强技术，通过对两个样本及其标签进行加权平均，生成新的样本。可以增加模型对不同样本的鲁棒性，从而有效地减少过拟合。

$$x = x_i + (1 -)x_j$$

$$y = y_i + (1 -)y_j$$

其中，和 x_i 是两张输入图像，y_i, y_j 是对应的标签， 是一个从 [0, 1] 区间中随机抽取的系数，用来控制插值的比例。

RegMixup 是对 Mixup 的改进，引入了一个正则化项来增强对不确定性和不一致性的惩罚，进一步提高了模型的泛化能力。正则化网络的输出，RegMixup 可以有效避免模型对训练数据的过拟合。

3.3 CSC

余弦相似度分类器（CSC）通常用于神经网络的最后一层，尤其是在度量学习或特征嵌入的任务中。CSC 主要用于替代传统的全连接层（FC Layer）与 Softmax 输出层，特别是在那些关注相似性度量和特征嵌入的任务中。在度量学习中，任务通常是学习一个映射，使得同一类别的样本在特征空间中距离更近，而不同类别的样本距离更远。为了实现这一目标，CSC 使用余弦相似度来衡量输入特征与类别中心之间的相似性。类别中心可以通过训练获得（如通过计算每个类别的平均特征向量）。CSC 用于最后一层，通过计算输入特征与各个类别中心之间的余弦相似度，从而为每个类别分配一个预测概率或相似度评分。在特征嵌入（如人脸识别、图像检索等任务）中，CSC 也用作神经网络的最后一层。通过对网络输出的特征进行余弦相似度计算，可以直接得出输入特征与数据库中存储的类别中心（嵌入向量）之间的相似度。这种方法常用于图像检索系统，在该系统中输入图像与数据库中的其他图像进行匹配。CSC 适用于小样本数据 [2]

$$\text{CosineSim}(x, w_i) = \frac{\mathbf{x} \cdot \mathbf{w}_i}{\|\mathbf{x}\| \|\mathbf{w}_i\|}$$

$$\hat{y} = \arg \max_{i \in \{1,2,\ldots,C\}} \text{CosineSim}(x, w_i)$$

3.4 Correctness Ranking Loss (CRL)

Correctness Ranking Loss (CRL) [5] 是一种用于优化排序问题的损失函数，常用于需要对多个候选答案、预测或排序列表进行准确性排名的任务中。它的核心思想是通过对正确项和错误项进行对比，学习一个更准确的排序模型。CRL 的目标是通过比较正确和错误预测之间的得分差距，鼓励模型将正确答案的得分排在错误答案之上，从而优化排序的准确性。

$$\mathcal{L}_{\text{CRL}} = \max\left(0, \Delta + f(x, y^-) - f(x, y^+)\right)$$

3.5 Sharpness-AwareMinimization(SAM)

深度学习模型的参数优化通常使用梯度下降方法，但传统优化过程可能会收敛到“尖锐”最小值（sharp minima），这些最小值在训练集上表现良好，但在测试集上的泛化能力较差。SAM [1] 是一种优化技术，旨在提高深度学习模型的泛化能力。SAM 的核心思想是通过显式地优化损失函数在参数空间中的平坦区域，从而增强模型的鲁棒性和对数据分布变化的适应性, 提升模型的泛化性能。

$$\mathcal{L}_{\text{SAM}}(\mathbf{w}) = \max_{\|\boldsymbol{\epsilon}\| \leq \rho} \mathcal{L}(\mathbf{w} + \boldsymbol{\epsilon}),$$

其中：$\mathbf{w}$ 表示模型参数，$\mathcal{L}(\mathbf{w})$ 是损失函数（例如交叉熵损失），$\boldsymbol{\epsilon}$ 是扰动向量，其 ℓ_2 范数被限制为 $\rho > 0$。

该公式的含义是：在当前模型参数 $\mathbf{w}$ 的邻域内，找到使损失最大的扰动 $\boldsymbol{\epsilon}$，并最小化该最大损失值。通过这种方式，SAM 鼓励模型的参数更新趋向于平坦区域（Flat Minima），从而提升模型的鲁棒性和泛化能力。

为了实际优化，SAM 分为两个步骤：

1. **计算扰动向量：**

$$\boldsymbol{\epsilon} = \rho \frac{\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})}{\|\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})\|},$$

即在梯度方向上添加大小为 ρ 的扰动。

2. **更新参数：**使用加了扰动的损失计算梯度，更新参数：

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w} + \boldsymbol{\epsilon}),$$

其中 η 是学习率。

3.6 Stochastic Weight Averaging (SWA)

Stochastic Weight Averaging (SWA) [3] 是一种优化技术，基本思想是，在训练的最后几个周期中，通过对多个周期的模型权重进行平均，来获得一个更加平滑和鲁棒的解。这种做法假设，在训练过程中，模型的权重会沿着损失函数的平坦区域振荡，平均这些权重能够找到一个具有更好泛化性能的解。

$$\mathbf{w}_{\text{SWA}} = \frac{1}{T} \sum_{t=T_1}^{T_2} \mathbf{w}_t,$$

其中：

- $\mathbf{w}_t$ 是第 t 轮训练结束后的模型权重，
- T_1 是开始平均的周期（通常在训练的后期开始），
- T_2 是结束周期（通常是训练的最后一个周期），
- T 是参与平均的周期数量。

4 复现细节

4.1 与已有开源代码对比

复现论文中使用的数据集不够丰富，比如只使用 Animal-10N 和 Food-101N 进行带有噪声标签的学习。只使用具有不平衡因子（10、50、100）的 CIFAR-LT 进行长尾分类。仅使用两个数据集得出的结论在某些情况下可能不够可靠。因此我的工作相比于开源代码，增加了几个数据集进行实验，从而得到更可靠的结论。

4.2 实验环境搭建

实验可以在单个 gpu 中运行，该代码在 Python 3.9 和 PyTorch1.13.0 上进行了测试。

4.3 界面分析与使用说明

需要手动下载数据集，比如 cifar10，并且文件的结构应该是：目录./data/CIFAR10/下分别保存 train、val、test。

在目录./run 文件夹下找对应数据集名字目录下面的对应网络名字，即可找到对应的运行代码，重复实验过程。

5 实验结果分析

5.1 故障预测实验

Backbones	Methods	CIFAR-10 [40]				CIFAR-100 [40]				Tiny-ImageNet [41]			
		Acc. ↑	AURC ↓	AUROC ↑	FPR95 ↓	Acc. ↑	AURC ↓	AUROC ↑	FPR95 ↓	Acc. ↑	AURC ↓	AUROC ↑	FPR95 ↓
ResNet-18 [28]	MSP [31]	94.89±0.20	6.78±0.33	92.20±0.55	38.73±2.89	75.87±0.31	69.44±2.11	87.00±0.21	60.73±1.16	63.39±0.59	136.50±1.08	85.62±0.35	63.99±0.64
	RegMixup [59]	95.69±0.13	4.74±0.27	92.96±0.29	34.26±1.98	77.90±0.37	59.23±1.65	87.61±0.13	58.65±0.43	66.36±0.43	115.08±1.98	86.53±0.27	62.54±0.43
	CRL [54]	94.85±0.10	5.09±0.28	93.64±0.48	35.33±1.73	76.42±0.21	62.78±0.21	88.07±0.17	59.02±0.39	65.50±0.03	117.46±0.56	87.01±0.13	61.15±0.07
	SAM [19]	95.30±0.25	3.97±0.33	94.53±0.31	31.13±3.62	76.60±0.21	62.97±1.02	87.72±0.10	59.35±0.87	64.95±0.21	120.04±2.11	87.19±0.57	**59.98±0.55**
	SWA [35]	95.38±0.09	4.00±0.21	94.40±0.50	35.70±1.44	77.65±0.19	55.87±0.32	88.55±0.25	60.43±1.90	68.09±0.19	102.11±0.51	87.27±0.15	60.63±1.38
	FMFP [81]	95.60±0.09	3.56±0.06	94.74±0.10	33.49±0.33	77.82±0.08	55.03±0.52	88.59±0.07	59.79±0.31	68.18±0.42	100.93±2.12	87.45±0.05	60.18±1.26
	SURE	**96.14±0.16**	**2.97±0.13**	**95.08±0.04**	**28.64±0.66**	**80.49±0.18**	**45.81±0.15**	**88.73±0.24**	**58.91±0.58**	**69.55±0.10**	**93.46±0.82**	**87.67±0.12**	60.13±0.32
VGG [64]	MSP [31]	93.30±0.21	10.41±0.33	90.71±0.04	44.66±1.81	72.43±0.42	91.40±1.95	85.69±0.90	64.41±1.66	59.52±0.62	156.45±2.51	86.33±0.63	63.79±0.95
	RegMixup [59]	94.11±0.28	9.89±0.81	89.90±0.26	39.93±1.58	73.51±0.18	85.98±1.05	86.35±0.32	61.70±1.83	63.04±0.57	146.72±2.59	85.60±0.39	**59.00±1.27**
	CRL [54]	93.42±0.09	7.61±0.44	92.88±0.56	39.66±2.83	72.63±0.27	80.94±0.47	87.37±0.28	61.96±0.77	60.20±0.36	146.76±1.42	**87.42±0.28**	59.26±1.44
	SAM [19]	94.11±0.06	5.97±0.08	93.68±0.13	37.21±2.92	73.33±0.36	77.44±0.75	87.42±0.33	63.19±0.58	61.24±0.07	142.54±1.04	86.82±0.25	62.93±1.12
	SWA [35]	93.76±0.25	6.64±0.24	93.43±0.16	40.44±1.27	73.98±0.16	74.23±0.58	87.30±0.14	62.89±1.80	62.48±0.19	137.01±0.71	86.29±0.16	62.15±1.64
	FMFP [81]	94.26±0.23	5.89±0.16	93.46±0.26	40.67±3.14	74.77±0.31	70.07±1.26	87.58±0.19	60.98±1.16	62.95±0.16	134.04±1.42	86.36±0.12	61.71±1.08
	SURE	**95.00±0.11**	**4.98±0.24**	**93.79±0.62**	**35.92±2.95**	**76.51±0.07**	**65.25±0.17**	**87.59±0.07**	**60.27±0.60**	**63.75±0.11**	**131.40±0.28**	86.12±0.19	63.04±1.05
DenseNet [34]	MSP [31]	94.72±0.23	5.94±0.23	93.00±0.45	37.00±0.31	75.14±0.07	74.68±0.32	86.22±0.22	62.79±0.80	57.90±0.25	180.08±2.52	83.65±0.29	68.61±0.37
	RegMixup [59]	95.13±0.22	6.03±0.50	92.20±0.80	38.63±1.63	77.29±0.16	63.96±1.15	86.57±0.07	63.76±1.10	61.96±0.09	147.22±1.57	**84.91±0.17**	65.92±0.40
	CRL [54]	94.79±0.02	5.58±0.42	93.22±0.61	37.34±2.73	76.09±0.06	65.96±0.62	87.41±0.11	60.67±0.72	58.80±0.56	169.44±3.74	84.49±0.04	66.05±0.60
	SAM [19]	95.31±0.10	4.25±0.17	94.15±0.46	33.33±1.27	78.17±0.26	57.20±0.73	86.99±0.23	61.42±0.74	60.49±0.31	158.94±3.86	84.39±0.57	66.51±1.85
	SWA [35]	94.86±0.09	4.65±0.18	94.27±0.27	35.78±4.61	78.17±0.26	57.20±0.73	87.23±0.22	63.33±0.63	60.74±0.46	159.68±3.12	83.83±0.07	68.03±0.75
	FMFP [81]	95.07±0.15	4.11±0.19	94.74±0.06	34.67±0.48	78.33±0.40	54.88±1.62	87.92±0.46	60.52±1.12	61.18±0.72	154.98±3.72	84.29±0.26	66.66±1.21
	OpenMix [82]§	95.51±0.23	4.68±0.72	93.57±0.81	33.57±3.70	78.97±0.31	53.83±0.93	87.45±0.18	62.22±1.15	-	-	-	-
	SURE	**95.57±0.06**	**3.51±0.09**	**94.91±0.25**	**29.52±0.56**	**80.02±0.13**	**46.69±0.59**	**88.78±0.26**	**58.37±0.39**	**62.61±0.18**	**142.59±2.16**	84.31±0.42	**65.39±2.12**
WRNet [76]	MSP [31]	95.71±0.17	5.90±0.89	92.19±0.82	35.95±3.75	79.15±0.19	53.02±0.89	88.21±0.06	59.46±1.23	67.52±0.18	107.97±0.80	86.78±0.20	61.68±0.99
	RegMixup [59]	97.03±0.04	3.47±0.26	93.10±0.56	26.16±1.17	82.14±0.47	47.01±2.12	87.70±0.17	55.24±1.19	69.63±0.09	95.96±0.21	87.38±0.21	59.09±0.75
	CRL [54]	95.87±0.08	3.85±0.20	94.10±0.06	32.73±1.22	80.10±0.28	47.99±1.08	88.43±0.34	59.44±1.45	69.00±0.22	97.46±0.90	87.42±0.23	61.02±1.71
	SAM [19]	96.47±0.11	2.91±0.38	94.79±0.29	28.05±1.56	80.67±0.31	44.93±0.87	89.01±0.31	56.60±1.30	69.86±0.37	93.66±2.03	87.49±0.30	60.44±1.19
	SWA [35]	94.86±0.09	4.65±0.18	94.27±0.27	35.78±4.61	81.31±0.33	41.15±0.89	89.39±0.16	57.57±1.97	71.27±0.16	84.97±0.12	87.71±0.26	60.00±2.42
	FMFP [81]	96.47±0.12	2.33±0.08	95.73±0.01	26.68±2.62	81.66±0.12	39.60±0.15	89.51±0.10	56.41±1.44	71.62±0.04	83.04±0.16	87.78±0.03	60.09±0.83
	OpenMix [82]§	**97.16±0.10**	2.32±0.15	94.81±0.34	22.08±1.86	82.63±0.06	39.61±0.54	89.06±0.11	55.00±1.29	-	-	-	-
	SURE	97.02±0.20	**1.79±0.16**	**96.18±0.01**	**19.53±1.23**	**83.71±0.10**	**32.10±0.28**	**90.33±0.18**	**54.34±0.29**	**73.34±0.36**	**74.11±0.97**	**88.23±0.31**	**58.17±1.50**
DeiT-B * [70]	MSP [31]	98.28±0.08	0.97±0.02	95.76±0.28	20.47±5.38	89.71±0.03	17.66±0.56	90.40±0.25	50.99±0.61	-	-	-	-
	RegMixup [59]	98.90±0.04	0.89±0.05	94.30±0.25	24.98±3.87	90.79±0.11	15.38±0.51	90.34±0.33	52.01±1.76	-	-	-	-
	CRL [54]	98.27±0.04	0.99±0.11	95.85±0.44	19.65±2.51	89.74±0.16	17.61±0.71	90.30±0.18	51.58±0.23	-	-	-	-
	SAM [19]	98.62±0.10	0.58±0.09	96.89±0.34	**15.74±1.71**	90.43±0.17	15.29±0.19	90.75±0.15	50.02±1.52	-	-	-	-
	SWA [35]	98.44±0.07	0.82±0.03	96.11±0.20	17.78±3.23	90.17±0.34	15.37±0.44	90.86±0.38	50.64±3.37	-	-	-	-
	FMFP [81]	98.76±0.02	**0.46±0.02**	**97.15±0.16**	16.17±0.55	90.53±0.13	14.30±0.18	**91.15±0.32**	51.90±1.50	-	-	-	-
	SURE	**98.92±0.07**	0.86±0.08	94.37±0.69	27.52±3.11	**91.18±0.01**	**13.79±0.29**	90.85±0.05	**48.81±0.39**	-	-	-	-

图 2. CIFAR10、CIFAR100 和 Tiny-ImageNet 上的故障预测性能比较

模型的性能指标，包括准确度（Acc.）、风险覆盖曲线下面积（AURC）、受试者工作特征曲线面积（AU-ROC）、95% 真实阳性率时的假阳性率（FPR95）。具体来说，我们利用与准确度互补的 AURC 来衡量模型的不确定性。AURC 测量通过绘制根据覆盖率的风险而绘制的曲线下面积。给定一个置信度阈值，覆盖率表示置信度估计高于置信度阈值的样本的比例，风险（也称为选择性风险）是使用这些样本计算的错误率。AURC 值越低意味着准确度越高，正确和错误预测可以通过置信度阈值很好地区分。

实验使用不同的网络进行, 包括 ResNet18、VGG16-BN、DenseNetBC、WRNet28 和 DeiT。我们将 10% 的训练数据作为超参数选择的验证集，并报告测试集上的性能。所有实验重复 3 次，我们在表中报告平均值和标准差。实验说明 SURE 在几乎所有指标上都比所有竞争方法在不同数据集和不同架构上取得了更好的性能，这证明了我们提出的方法的有效性和鲁棒性。

5.2 长尾分类实验

Methods	CIFAR10-LT [12]			CIFAR100-LT [12]		
	IF=100	IF=50	IF=10	IF=100	IF=50	IF=10
CE	70.40	74.80	86.40	38.30	43.90	55.70
Mixup [77]	73.06	77.82	87.1	39.54	54.99	58.02
CB-Focal [12]	74.57	79.27	87.10	39.60	45.17	57.99
LDAM-DRW [4]	77.03	81.03	88.16	42.04	46.62	58.71
SSP [73]	77.83	82.13	88.53	43.43	47.11	58.91
BBN [80]	79.82	81.18	88.32	42.56	47.02	59.12
Casual model [69]	80.60	83.60	88.50	44.10	50.30	59.60
MetaSAug-LDAM [45]	80.66	84.34	89.68	48.01	52.27	61.28
Hybrid-SC [71]	81.40	85.36	91.12	46.72	51.87	63.05
ResLT [11]	82.40	85.17	89.70	48.21	52.71	62.01
Dynamic Loss [37]	82.95	88.30	91.24	50.14	54.51	63.99
BCL [83]	84.32	87.24	91.12	51.93	56.59	64.87
GLMC [16]	**87.75**	90.18	94.04	55.88	61.08	70.74
SURE	83.28	87.72	93.73	51.60	58.57	71.13
GLMC + MaxNorm [1]	87.57	**90.22**	94.03	57.11	62.32	72.33
SURE + re-weighting	86.93	**90.22**	**94.96**	**57.34**	**63.13**	**73.24**

图 3. ResNet32 在 CIFAR10-LT 和 CIFAR100-LT 上具有不同不平衡因素 [100, 50, 10] 的准确率（%）

提出的 SURE 最初不是为长尾分类设计的，但它实现了具有竞争力的结果。结果表明，利用不确定性估计进行下游应用是有前景的，尤其是使用 SURE 来训练 DNN。

5.3 带噪声标签的实验

在数据集 Animal-10N 和 Food-101N 上的准确率表明：虽然 SURE 不是为有噪声标签的学习而设计的，但在默认设置下，它的表现明显优于目前所有其他方法。

5.4 实验分析 SURE 的有效性

Method	Loss		Optimzation		Classifer	CIFAR100 [40]			
	λ_{crl}	λ_{mix}	SAM	SWA	CSC	Acc. ↑	AURC ↓	AUROC ↑	FPR95 ↓
Baseline(MSP)	0	0	✗	✗	✗	75.87±0.31	69.44±2.11	87.00±0.21	60.73±1.16
SAM	0	0	✓	✗	✗	76.60±0.21	62.97±1.02	87.72±0.10	59.35±0.87
SWA	0	0	✗	✓	✗	77.65±0.19	55.87±0.32	88.55±0.25	60.43±1.90
CSC	0	0	✗	✗	✓	74.05±0.18	78.14±0.26	86.82±0.24	63.56±1.20
FMFP	0	0	✓	✓	✗	77.82±0.08	55.03±0.52	88.59±0.07	59.79±0.31
SAM + CSC	0	0	✓	✗	✓	75.97±0.39	64.20±1.55	88.06±0.19	59.36±1.21
SWA + CSC	0	0	✓	✓	✗	78.46±0.33	55.68±0.41	87.74±0.44	61.22±2.54
FMFP + CSC	0	0	✓	✓	✓	78.45±0.13	54.18±0.47	88.23±0.20	60.05±1.03
CRL	1	0	✗	✗	✗	76.42±0.21	62.78±0.21	88.07±0.17	59.02±0.39
CRL + SAM	1	0	✓	✗	✗	76.98±0.32	59.71±1.39	88.26±0.07	59.52±1.92
CRL + SWA	1	0	✗	✓	✗	77.56±0.20	56.88±0.28	88.24±0.45	61.73±1.77
CRL + CSC	1	0	✗	✗	✓	75.61±0.46	67.83±1.98	87.84±0.11	59.80±2.16
CRL + FMFP	1	0	✓	✓	✗	77.71±0.54	56.24±0.89	88.21±0.44	61.75±1.74
CRL+ SAM + CSC	1	0	✓	✗	✓	78.21±0.53	53.55±3.28	88.86±0.45	56.37±1.71
CRL+ SWA + CSC	1	0	✓	✓	✗	78.09±0.10	56.61±0.91	87.78±0.21	61.37±1.56
CRL+ FMFP + CSC	1	0	✓	✓	✓	78.24±0.18	55.01±0.44	88.14±0.11	60.48±0.27
Reg	0	1	✗	✗	✗	76.99±1.19	63.09±4.22	87.71±0.13	58.78±0.50
Reg + SAM	0	1	✓	✗	✗	77.45±0.55	60.68±3.75	87.70±0.39	58.72±1.42
Reg + SWA	0	1	✗	✓	✗	78.55±0.62	52.31±2.10	88.71±0.22	58.99±2.07
Reg + CSC	0	1	✗	✗	✓	78.32±0.28	62.40±0.58	86.57±0.34	58.77±2.27
Reg + FMFP	0	1	✓	✓	✗	79.04±0.50	50.09±1.00	88.89±0.20	58.47±0.88
Reg + SAM + CSC	0	1	✓	✗	✓	78.91±0.34	57.43±2.25	87.16±0.23	58.35±0.22
Reg + SWA + CSC	0	1	✓	✓	✗	80.17±0.52	49.87±1.86	87.89±0.10	61.08±1.06
Reg + FMFP + CSC	0	1	✓	✓	✓	79.88±0.07	48.58±0.34	88.50±0.20	58.52±0.75
CRL + Reg	1	1	✗	✗	✗	78.38±0.17	52.93±1.19	**88.97±0.38**	**56.12±1.33**
CRL + Reg + SAM	1	1	✓	✗	✗	78.21±0.53	53.55±3.28	88.86±0.45	56.37±1.71
CRL + Reg + SWA	1	1	✗	✓	✗	78.64±0.16	50.96±1.01	88.96±0.31	59.27±1.47
CRL + Reg + CSC	1	1	✗	✗	✓	79.42±0.11	54.35±0.91	87.59±0.20	59.67±0.53
CRL + Reg + FMFP	1	1	✓	✓	✗	79.17±0.30	49.96±1.63	88.70±0.20	59.85±2.07
CRL + Reg + SAM + CSC	1	1	✓	✗	✓	79.10±0.34	56.39±1.25	87.44±0.16	56.98±0.31
CRL + Reg + SWA + CSC	1	1	✓	✓	✗	79.63±0.27	49.14±0.22	88.51±0.34	59.28±2.14
SURE	1	1	✓	✓	✓	**80.49±0.18**	**45.81±0.15**	88.73±0.24	58.91±0.58

图 4. SURE 中使用的不同组件及其组合在 CIFAR100 上的消融研究

采用控制变量法，根据是否选择这五种方法，从而得到 32 个不同组合，比较 32 个实验结果，分析了每个组件对我们模型在 CIFAR100 上的表现的贡献。我们报告了在使用 ResNet18 进行消融时三次运行的平均值和标准差。从我们的基线模型 MSP 开始，我们观察到将 Reg-Mixup、CRL、SAM、SWA 和 CSC 技术添加到 SURE 框架所带来的增量影响。SURE 方法的每次添加都提高准确率和 AURC，其中完整的 SURE 方法获得了研究中报告的最高分数。

6 总结与展望

本文介绍了 SURE，这是一个集成了多种模型正则化、分类器和优化技术的新框架，旨在提高 DNN 的可靠性和鲁棒性。我们的工作强调了现有方法在处理现实世界数据的复杂性时的不足。这一见解强调了对 SURE 等方法的迫切需求。通过严格的评估，SURE 在故障预测方面始终优于各种数据集和模型架构中的单个方法。此外，它在解决现实世界的挑战（例如长尾分类、使用嘈杂标签和数据损坏进行学习）中的应用不仅在长尾分布数据集中产生了

与最先进方法相当的结果，而且在具有标签噪声的场景中也表现出色。这项工作为不确定性估计方法在各种复杂的现实世界情况中的应用铺平了道路。

参考文献

[1] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[2] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

[3] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[4] Yuting Li, Yingyi Chen, Xuanlong Yu, Dexiong Chen, and Xi Shen. Sure: Survey recipes for building reliable and robust deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17500–17510, 2024.

[5] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pages 7034–7044. PMLR, 2020.

[6] Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. *Advances in Neural Information Processing Systems*, 35:14608–14622, 2022.