Augmented Neural Fine-Tuning for Efficient Backdoor Purification

Nazmul Karim[1*], Abdullah Al Arafat[2*], Umar Khalid[1], Zhishan Guo[2], and Nazanin Rahnavard[1]

[1]University of Central Florida, [2]North Carolina State University

[*]Equal Contribution

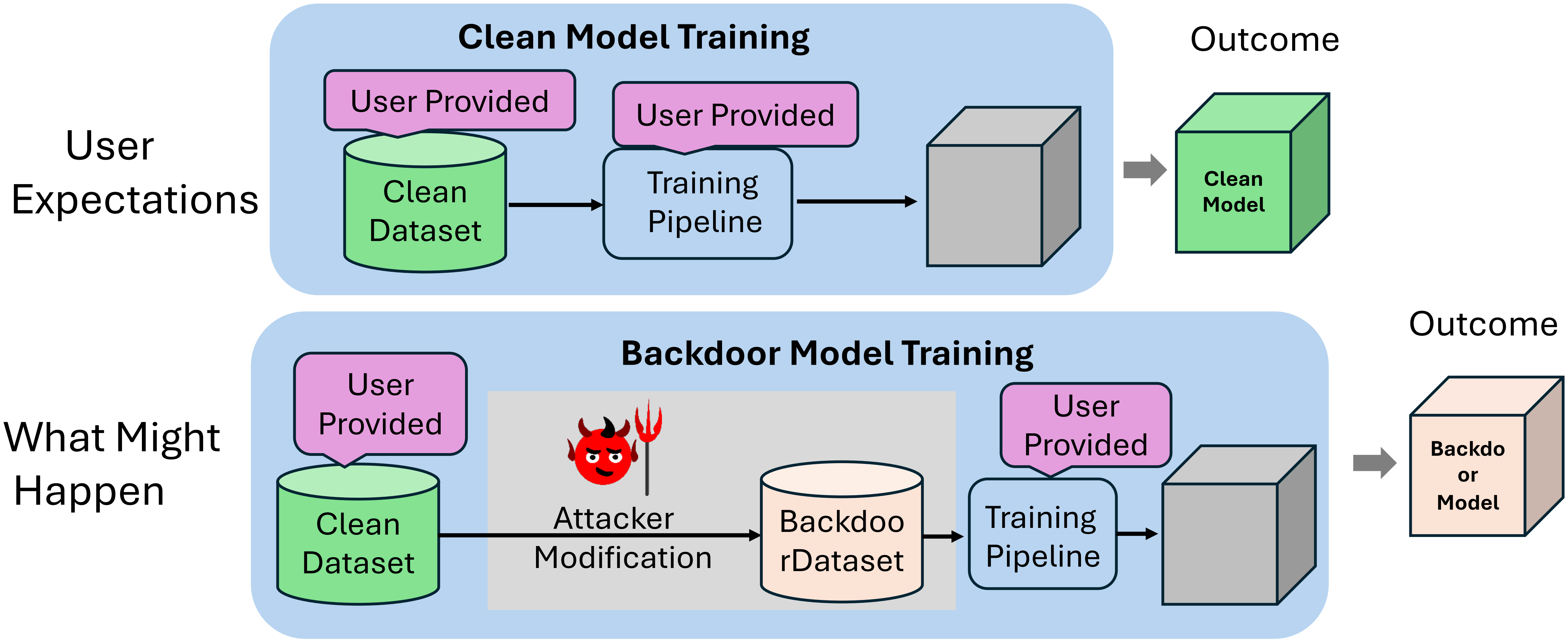
European Conference on Computer Vision (ECCV) 2024

Milan, Italy

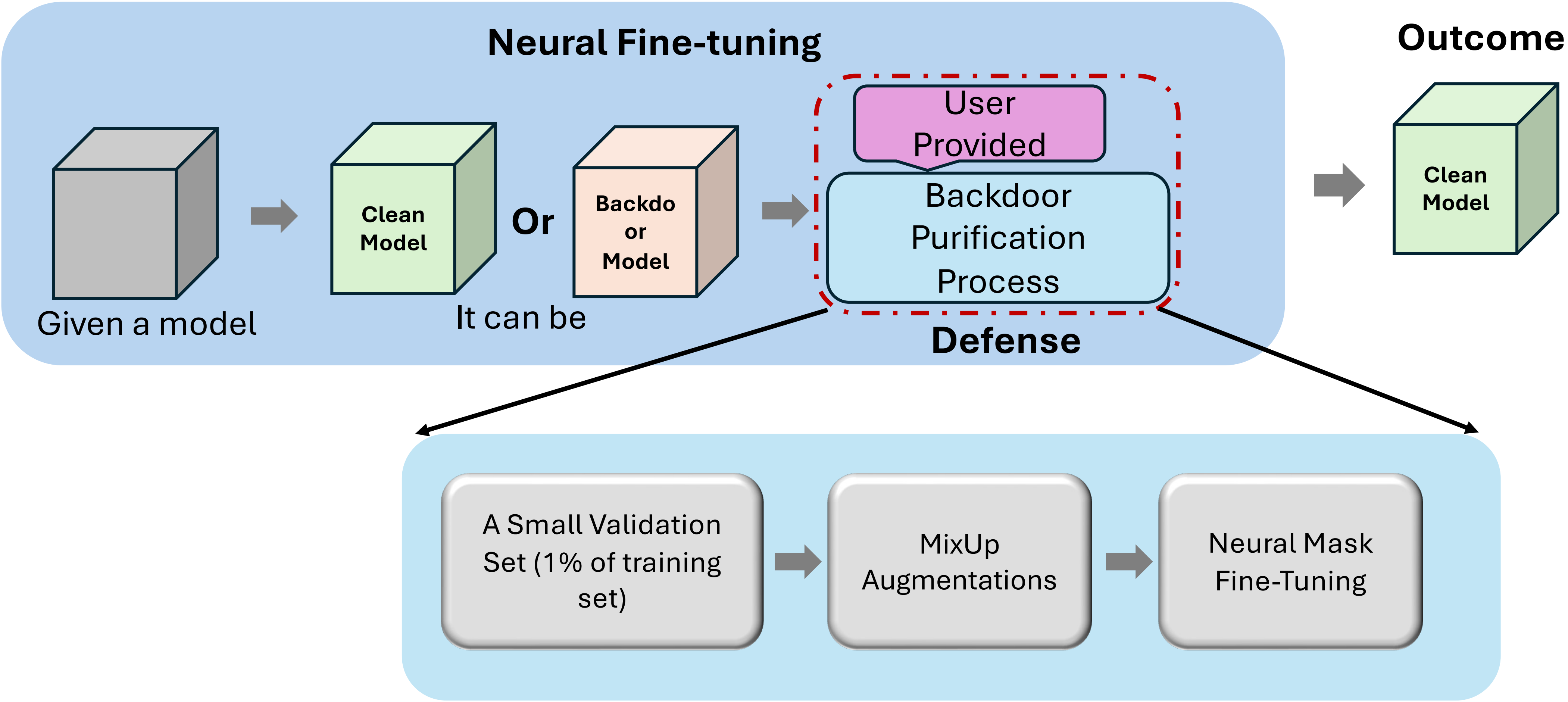
Wed 2 Oct 4:30 a.m. - 6:30 a.m. EDT, Poster# 341

Origin of Backdoor Attack

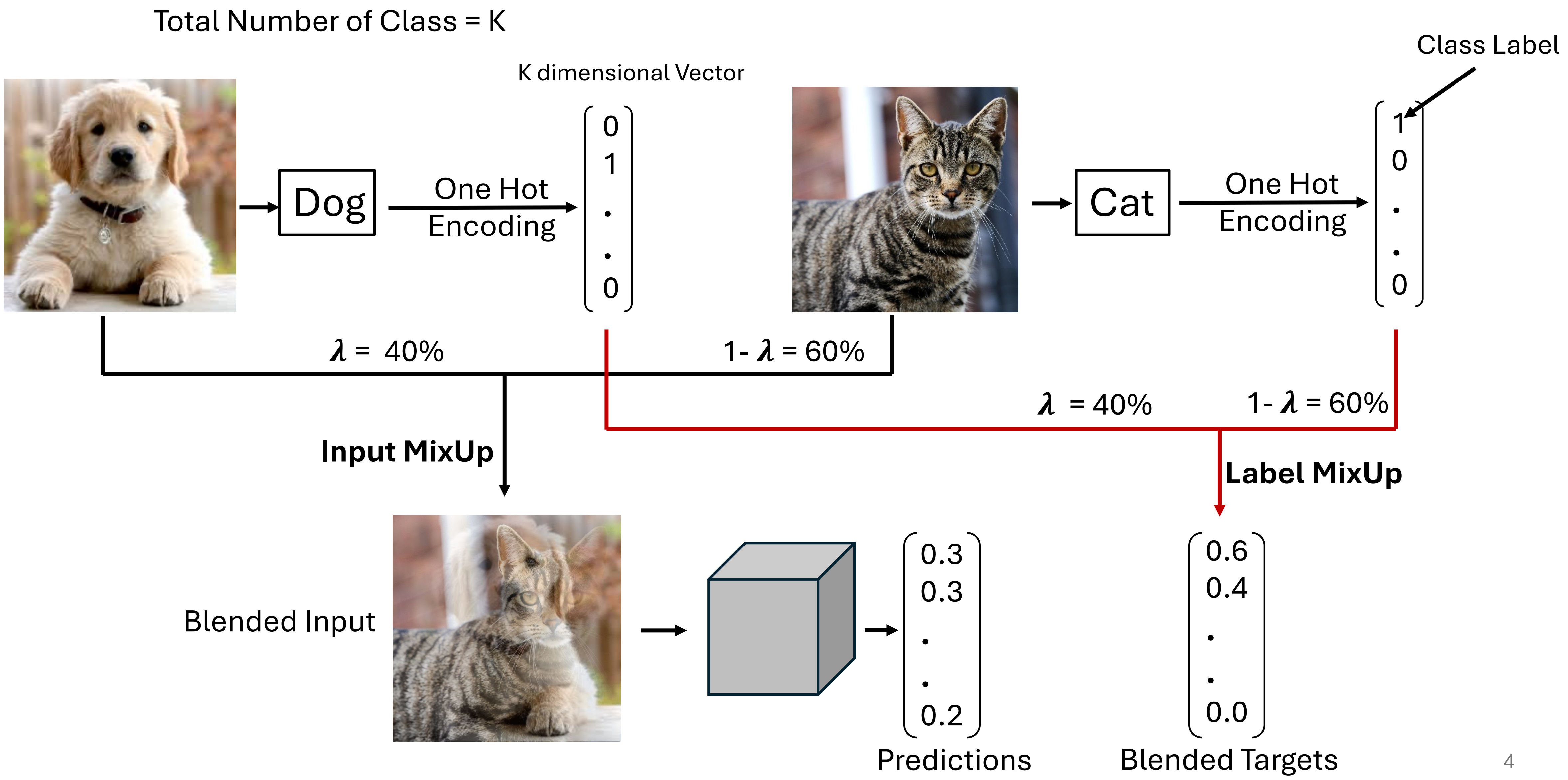
Scenario: Due to resource constraints for large model training, user outsources the training to a third party (who may be an attacker) to train the model



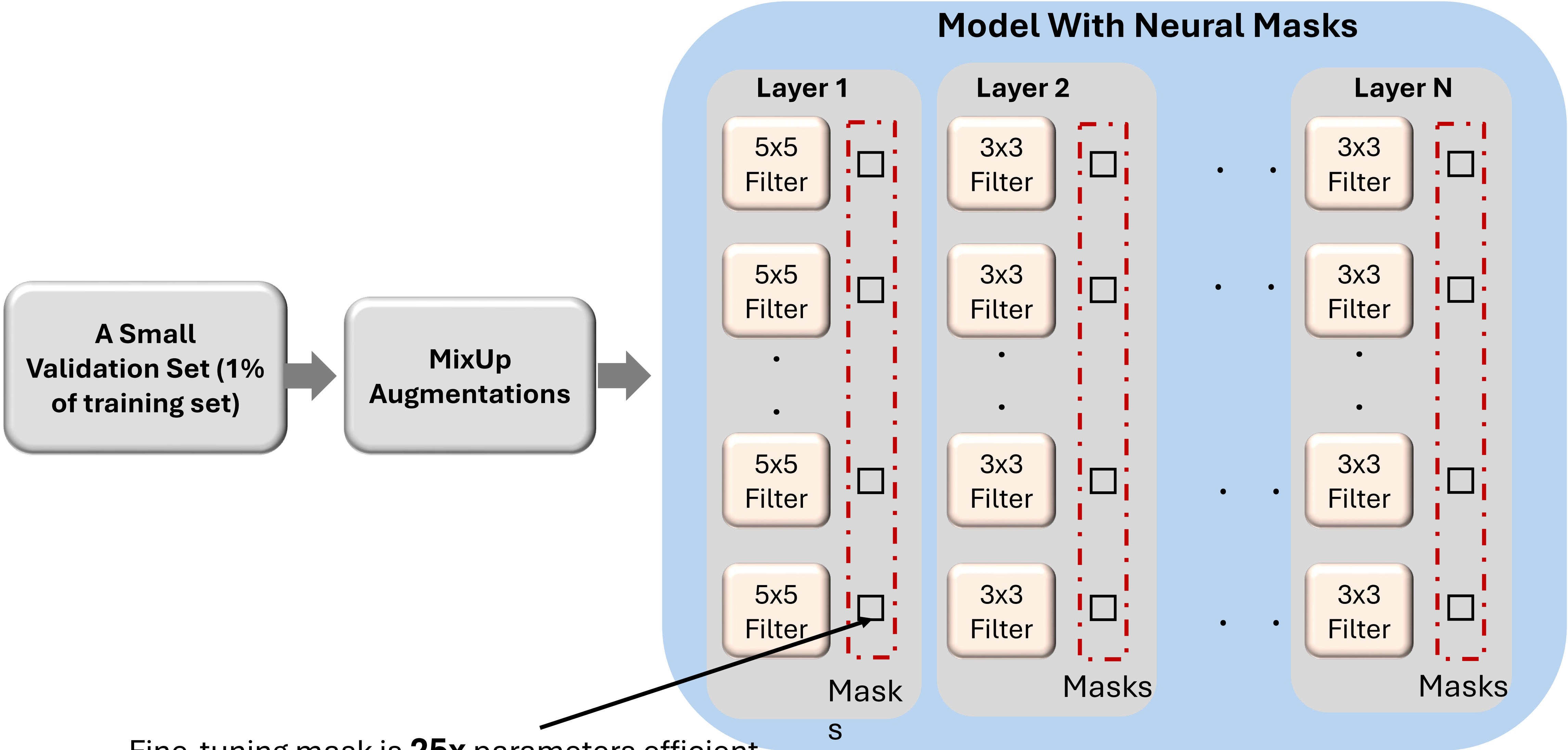
Proposed Defense



How Mixup Works?



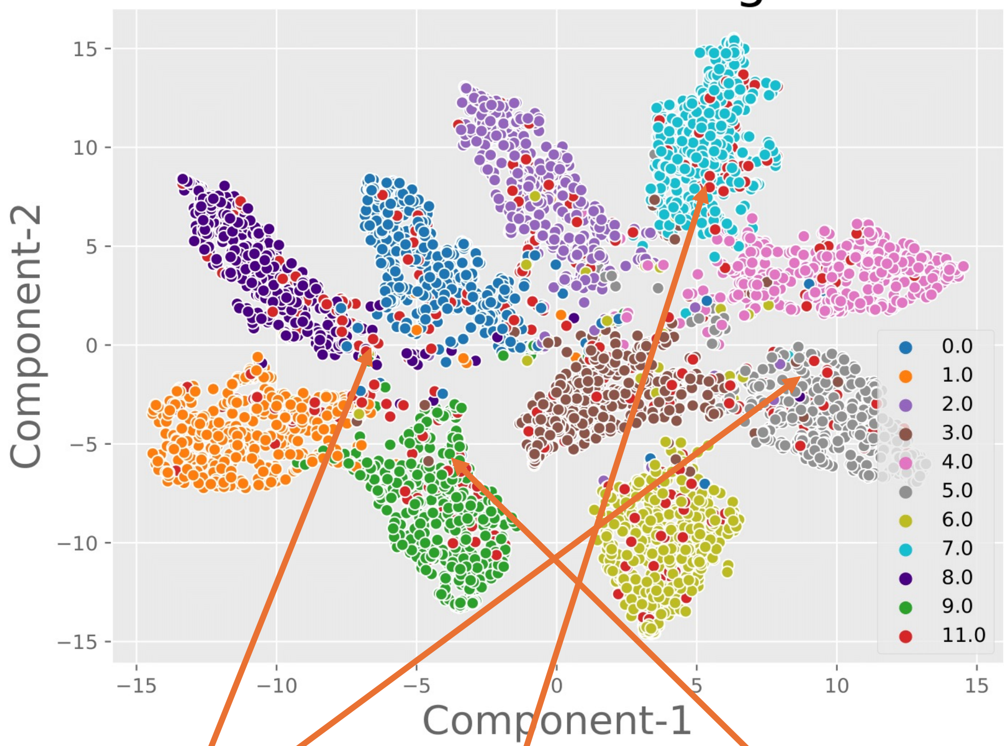
Neural Mask Fine-Tuning (NFT)



Fine-tuning mask is **25x** parameters efficient

t-SNE Visualization

For CIFAR 10 dataset with 10 classes



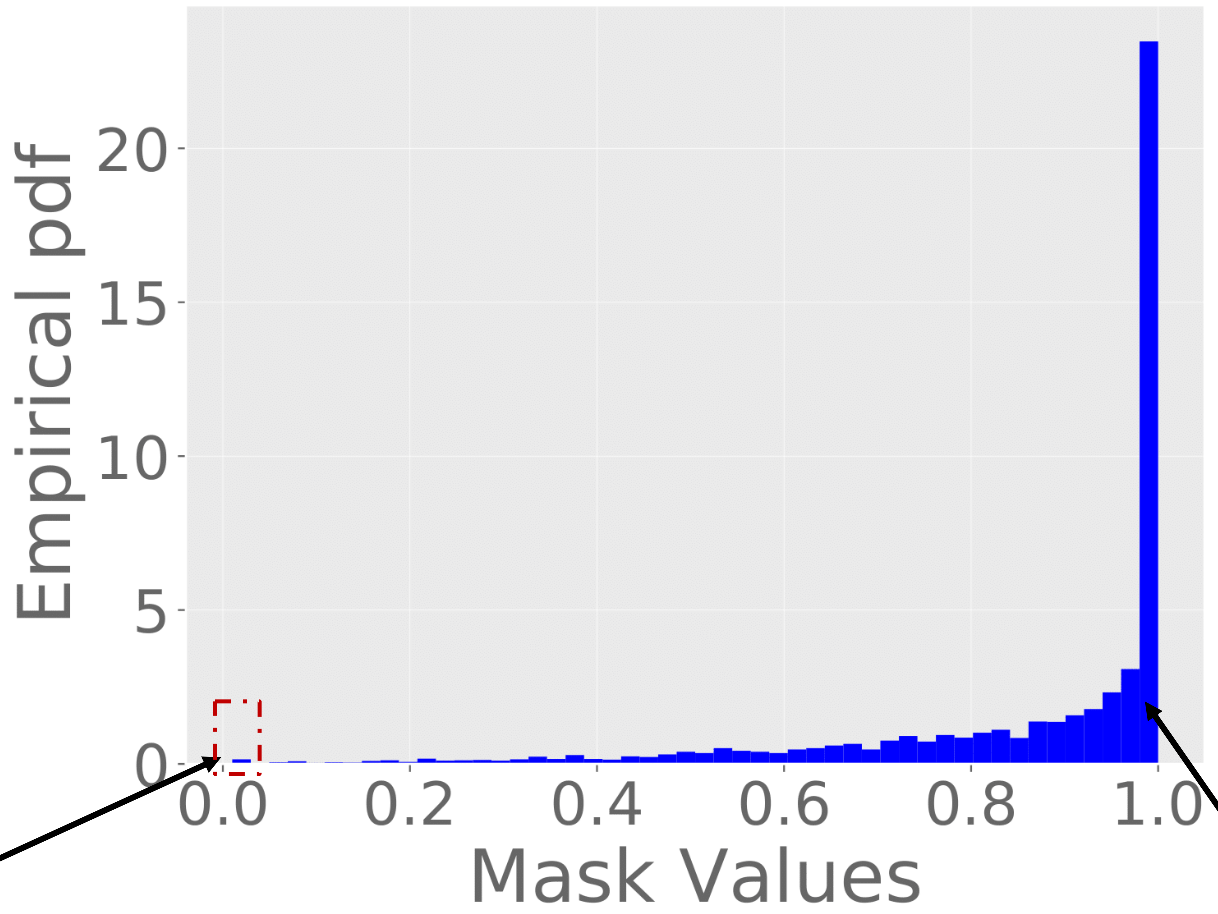
Backdoor Sample Cluster:

i) Take a certain number of clean samples from all classes

ii) Add trigger to those samples and change their label to "0"

After defense, backdoor samples are remapped to their original ground truth

Neural Mask Distribution



We prune few filters (mask value = 0)

We keep most filters intact (mask value = 1)

Experimental Results

- Image Datasets
- Video Action Recognition Datasets
- 3D Point Cloud
- Object Detection
- Natural Language Generation

Image Datasets

4 Image Datasets-

- CIFAR10
- GTSRB
- Tiny-ImageNet
- ImageNet

14 Different Attacks-

- BadNets
- LIRA
- WaNet
- TrojanNet
- ISSBA, etc.

Before Defense: Average Attack Success Rate (ASR) for all dataset is close to 100%

After Defense: Average Attack Success Rate (ASR) for all dataset should be close to 0%

Purification Results

ASR/ACC (%) before and after Backdoor Purification (For BadNets)

Dataset	CIFAR10	GTSRB	Tiny ImageNet	ImageNet
Before Defense	100/92.9	100/ 97.4	100/59.8	99.2/74.5
Previous SOTA Defense [1]	3.95/88.3	2.72/ 94.5	6.29/54.6	2.87/69.4
NFT (Ours)	**1.74/90.8**	**0.24/ 95.1**	**2.34/57.8**	**3.61/70.9**

ACC should be same before and after defense. Higher drop in ACC indicates poor defense

[1] One-shot Neural Backdoor Erasing via Adversarial Weight Masking (NIPS 2022)

Other Datasets

Dataset	No defense		I-BAU		AWM		RNP		FT-SAM		NFT (Ours)	
	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
UCF-101	81.3	75.6	20.4	70.6	20.8	70.1	17.0	70.3	15.9	71.6	**13.3**	**71.2**
HMDB-51	80.2	45.0	17.5	**41.1**	15.2	40.9	12.6	40.4	10.8	41.7	**9.4**	40.8

Video Action Recognition Datasets

Dataset	No defense		ANP		AWM		RNP		FT-SAM		NFT (Ours)	
	ASR	mAP	ASR	mAP	ASR	mAP	ASR	mAP	ASR	mAP	ASR	mAP
VOC07	86.4	92.5	21.7	86.9	26.6	87.3	19.2	87.6	19.3	86.8	**17.3**	**89.1**
VOC12	84.8	91.9	18.6	85.3	19.0	85.9	**13.8**	86.4	14.6	87.1	14.2	**88.4**
MS-COCO	85.6	88.0	19.7	84.1	22.6	83.4	17.1	84.3	19.2	83.8	**16.6**	**85.8**

Object Detection Datasets

Other Datasets

Attack	No Defense		ANP		AWM		RNP		FT-SAM		NFT (Ours)	
	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
PointBA-I	98.6	89.1	13.6	82.6	15.4	83.9	**8.1**	84.0	8.8	84.5	9.6	**85.7**
PointBA-O	94.7	89.8	14.8	82.0	13.1	82.4	9.4	83.8	8.2	85.0	**7.5**	**85.3**
PointCBA	66.0	88.7	21.2	83.3	21.5	83.8	**18.6**	84.6	20.3	84.7	19.4	**86.1**
3DPC-BA	93.8	91.2	16.8	84.7	15.6	85.9	13.9	85.7	13.1	86.3	**12.6**	**87.7**

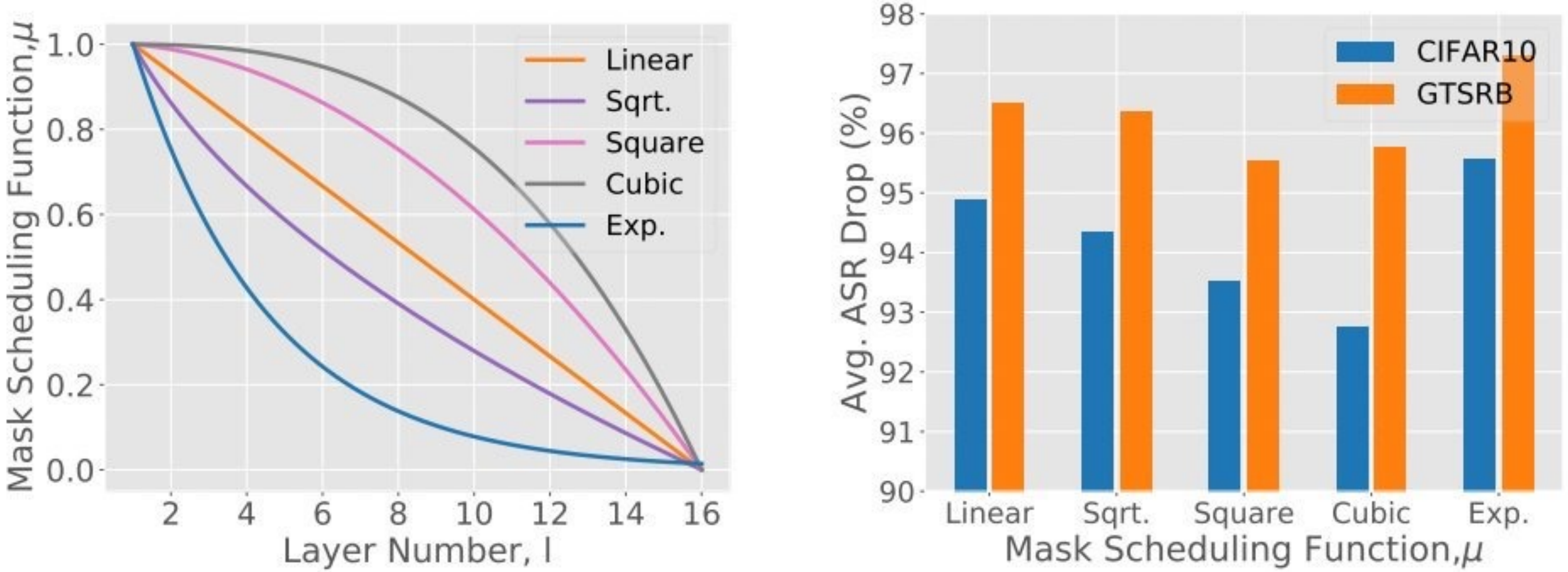
3D Point Cloud Datasets

Ablation Study

Table 7: Purification performance (%) for **various validation data sizes**. NFT performs reasonably well even with as few as 10 samples, *i.e.*, one sample (shot) per class for CIFAR10. We also show the impact of the **mask regularizer**, **mask scheduling function** μ, and **augmentations** on performance, which resonates with Fig. 1. Mask regularizer has the most impact on the clean test accuracy (around 7% worse without the regularizer). Without strong augmentations, we have a better ACC with a slightly worse ASR (around 6% drop).

Attack	Dynamic				WaNet				LIRA			
Samples	10		100		10		100		10		100	
Method	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
No Defense	100	92.52	100	92.52	98.64	92.29	98.64	92.29	99.25	92.15	99.25	92.15
AWM	86.74	55.73	9.16	85.33	83.01	62.21	7.23	84.38	91.45	66.64	10.83	85.87
FT-SAM	8.35	73.49	5.72	84.70	9.35	75.98	5.56	86.63	11.83	72.40	4.85	88.82
NFT w/o Reg.	5.67	76.74	**1.36**	82.21	**4.18**	76.72	3.02	83.31	**4.83**	74.58	2.32	83.61
NFT w/o Aug.	11.91	**81.86**	10.59	89.53	10.36	83.10	7.81	**89.68**	12.23	81.05	9.16	88.74
NFT w/o $\mu(l)$	5.11	80.32	3.04	88.58	5.85	82.46	4.64	88.02	6.48	81.94	4.33	88.75
NFT	**4.83**	80.51	1.72	**90.08**	4.41	**83.58**	**2.96**	89.15	5.18	**82.72**	**2.04**	**89.34**

Ablation Study



Study with different mask scheduling function shows that **Exponential (Exp.) Decay function produces the best performance**

Thank You