A APPENDIX

A.1 ETHIC STATEMENT

Our work reveals that backdoor attacks can bypass the existing finetuning-based defenses, even the most advanced ones. It also shows that a backdoor can be deeply implanted in the pretrained models and withstand even full-model finetuning on an entirely different downstream task. Hence, our work helps to extend the understanding of the potential capability of backdoor attacks, benefiting both the research community and real-life AI systems. By being informed about the risk, AI system developers will be more careful when using third-party models. The work also stimulates future backdoor defense studies in the quest of searching for safe and trustful AI development.

A.2 REPRODUCIBILITY STATEMENT

Our work is highly reproducible. All datasets used in the paper are popular and publicly available. We include in the supplementary materials our code and pre-trained models. The code and pre-trained models will also be publicly released upon paper acceptance.

A.3 FMN ALGORITHM

We present detailed algorithm of FMN in Algorithm 1. The cyclical learning rate schedule is as described in Figure 1: the learning rate is initialized with the minimum value and linearly increases to the maximum value in n iterations, then linearly decreases back to the minimum value for another n iterations.

A.4 SYSTEM DETAILS

A.4.1 DATASETS

We conduct our experiments on three popular datasets, which are widely used in various previous works, in both backdoor attacks and defenses.

CIFAR10

CIFAR10, introduced by Krizhevsky et al. (2009), is a labeled subset of the 80-millions-tiny-images dataset, collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The dataset consists of 60,000 color images in 10 classes, with 6,000 images per class. The image resolution is 32×32. CIFAR10 is split into 2 subsets: 50,000 images in the training set and 10,000 images in the test set. It is publicly available at `https://www.cs.toronto.edu/~kriz/cifar.html`.

Data augmentation techniques including random crop, random rotation, and random horizontal flip are applied during training. No augmentation is applied during evaluation.

CelebA

CelebFaces Attributes Dataset (CelebA) (Liu et al., 2015) is a large-scale face attributes dataset with more than 202,599 celebrity images from 10,177 identities. There are five landmark locations and 40 binary attribute annotations per image. The dataset is available for use at `http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`. In this work, we select 3 out of 40 attributes, namely Heavy Makeup, Mouth Slightly Open and Smiling, and then concatenate them into 8 compound classes to create a multiple-label classification task, as recommended by Salem et al. (2020).

All input images are resized to 64×64 in both the training and evaluating procedures. Random crop and random rotation are applied to training data. No augmentation is applied during the evaluation.

ImageNet-10

Deng et al. (2009) is a large-scale dataset that contains more than 14 million images of over 20,000 categories. The most widely used subset of this dataset is ImageNet-1K, which consists of 1000 object classes with 1,281,167 training images and 50,000 validation images, and can be found at `https://image-net.org/download.php`. In this work, we randomly select 10 classes

Algorithm 1: FMN

Input: training dataset $\mathcal{S}$, target label $\mathfrak{c}$, backdoor injection function $\mathcal{B}$, poisoning rate p, number of epochs *num_epochs*, number of iterations per epoch *num_iters* first max learning rate LR MAX 1, second max learning rate LR MAX 2, base learning rate LR BASE, learning rate ϵ.

initialize θ
$\gamma \leftarrow$ LR BASE, $epoch \leftarrow 0$, $iter \leftarrow 0$
for $epoch <$ *num_epochs* **do**
  for $iter <$ *num_iters* **do**
    Randomly sample $\mathcal{S}_{\text{mini}}$
    if $(epoch, 2) = 0$ **then**
      Randomly sample $\mathcal{P}_{\text{mini}}$ with ratio p
      Update θ: $\min_{\theta} \sum_{(x_j, y_j) \in \mathcal{S}_{\text{mini}} \setminus \mathcal{P}_{\text{mini}}} \mathcal{L}(f_\theta(x_j), y_j) + \sum_{(x_j, \mathfrak{c}) \in \mathcal{P}_{\text{mini}}} \mathcal{L}\big(f_\theta(\mathcal{B}(x_j)), \mathfrak{c}\big)$
    else
      Update θ: $\min_{\theta} \sum_{(x_j, y_j) \in \mathcal{S}_{\text{mini}}} \mathcal{L}(f_\theta(x_j), y_j)$
    if $epoch <$ *num_epochs*$/2$ **then**
      Update ϵ using cyclical learning rate scheduler with max learning rate LR MAX 1
    else
      Update ϵ using cyclical learning rate scheduler with max learning rate LR MAX 2

Table 5: **Performance of conventional backdoor training and FMN against finetuning-based defenses on CelebA.** For each attack, we report the BA (%) in teal and ASR (%) in purple. The asterisk (*) denotes that the attack is trained with FMN. The ASRs below 50% are underlined.

Attack	No defense	FT (lr = 0.01)	FT (lr = 0.05)	Super-FT	FT-SAM	NAD
Blend	79.26/99.24	78.29/74.72	65.92/19.28	78.75/21.84	78.12/20.14	78.64/32.46
Trojaning	78.89/99.65	76.98/86.24	71.98/9.45	78.07/11.23	78.57/23.16	78.65/31.15
Input-aware	78.7598.97	77.99/92.69	62.92/4.74	78.50/14.95	76.52/12.61	76.74/35.33
LIRA	79.02/99.86	78.88/95.11	67.72/24.44	78.01/32.55	77.49/41.96	78.19/42.59
Narcisuss	79.25/94.96	78.20/89.95	61.15/20.49	79.13/28.26	78.92/18.94	79.03/42.38
Blend*	79.17/97.94	77.52/96.55	67.84/69.37	77.97/88.15	78.90/89.69	78.76/71.70
Trojaning*	78.96/96.84	78.80/94.99	70.89/77.65	78.02/89.52	77.57/90.44	77.45/86.98
Input-aware*	78.92/96.77	78.52/91.34	64.87/76.79	78.48/86.21	78.52/85.62	77.85/80.47
LIRA*	79.40/98.98	78.64/96.78	72.15/80.32	79.21/90.59	78.79/91.10	79.02/92.84
Narcisuss*	79.30/93.22	77.98/89.65	65.12/79.11	78.91/86.95	78.64/91.21	78.72/84.92

from ImageNet-1K to create the ImageNet-10 dataset. Images of this dataset vary in resolution, and we resize them to 224×224 in both the training and testing process.
We apply random crop and random rotation to the training dataset. No augmentation is applied during the evaluation.

A.4.2 Networks

For the CIFAR10, we use Pre-activation ResNet18 (He et al., 2016) as the classifier architecture. For the CelebA and ImageNet10 datasets, we use ResNet18 (He et al., 2016) as the classifier architecture.

A.5 FMN performance against fine-tuning-based defenses on CelebA and ImageNet-10

We show the attack results of conventional backdoor training and training with FMN with different attack methods on CelebA and ImageNet-10, as well as their performance against fine-tuning-based defenses in Table 5 and Tabel 6.

Table 6: **Performance of conventional backdoor training and FMN against finetuning-based defenses on ImageNet-10.** For each attack, we report the BA (%) in teal and ASR (%) in purple. The asterisk (*) denotes that the attack is trained with FMN. The ASRs below 50% are underlined.

Attack	No defense	FT (lr = 0.01)	FT (lr = 0.05)	Super-FT	FT-SAM	NAD
Blend	85.65/98.74	84.07/66.85	71.11/2.09	84.50/16.74	83.45/21.23	83.40/21.99
Trojaning	84.91/97.79	82.32/57.53	72.24/4.29	80.75/19.35	81.41/20.96	83.98/44.26
Input-aware	85.64/91.88	84.27/62.96	73.41/6.63	83.76/10.55	83.54/14.77	82.42/19.64
LIRA	86.02/97.47	84.21/72.36	71.98/19.42	83.87/19.79	83.25/24.85	85.01/41.52
Narcisuss	85.17/97.22	84.54/70.40	74.02/11.25	84.59/21.41	84.66/20.95	83.64/39.87
Blend*	85.64/91.68	84.04/89.42	72.96/56.85	84.25/78.67	84.32/80.75	84.22/62.64
Trojaning*	87.62/97.72	84.98/92.19	79.35/66.67	85.42/86.20	85.65/90.14	84.32/82.52
Input-aware*	87.60/95.98	86.49/92.16	76.82/62.24	87.01/75.52	84.42/89.55	86.86/79.48
LIRA*	86.75/96.26	86.32/94.72	82.59/70.25	85.99/89.03	85.86/91.71	86.27/91.92
Narcisuss*	86.59/92.95	85.25/87.61	76.92/68.94	86.63/85.11	85.27/85.42	84.24/81.77

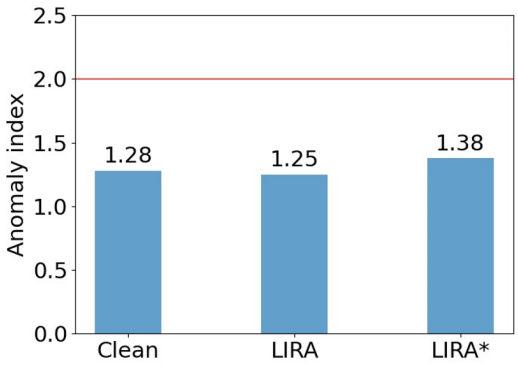


Figure 3: Performance of LIRA and LIRA* against Neural Cleanse.

A.6 Defense experiments with LIRA*

Since our training design aims to counter finetuning-based defense, the robustness against other types of defenses should depend on the choice of the trigger. In this section, we investigate if FMN can maintain that robustness from the conventional attack. For the following experiments, we choose LIRA* to evaluate.

A.6.1 Neural Cleanse

Neural Cleanse is a popular model defense technique that works by computing an optimal pattern for each class in the model. The technique then detects if there is a pattern that is significantly smaller than the others using an anomaly index computed by an outlier detection algorithm. If the anomaly index for a pattern is greater than 2, the model is marked as poisoned. We test LIRA* against this defense. We also report the result of this defense for a clean model and conventional LIRA for comparison. As shown in Figure 3, both LIRA and LIRA* can bypass Neural Cleanse's detection.

A.6.2 Fine-pruning

Fine-pruning is another model defense technique that works by pruning neurons that are inactive when predicting clean images. The assumption is that these neurons are more likely to be associated with the backdoor. We run fine-pruning on poisoned models of LIRA and LIRA*, and plot the clean accuracy (BA) and attack success rate (ASR) versus the numbers of neurons pruned. As shown in Figure 4, the results of LIRA and LIRA* are quite similar: there is no point fine-pruning can achieve high BA with low ASR, indicating that the attacks can evade this defense.

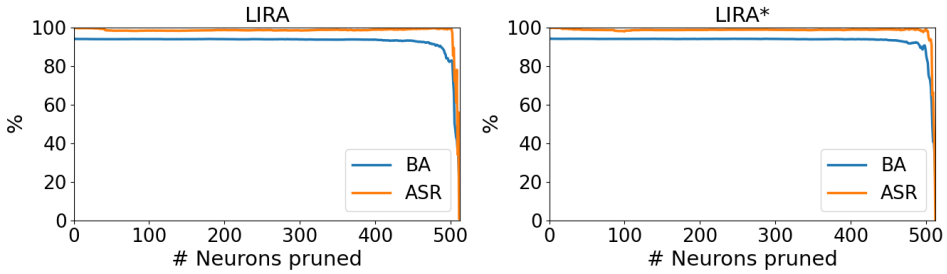


Figure 4: Performance of LIRA and LIRA* against Fine-pruning.

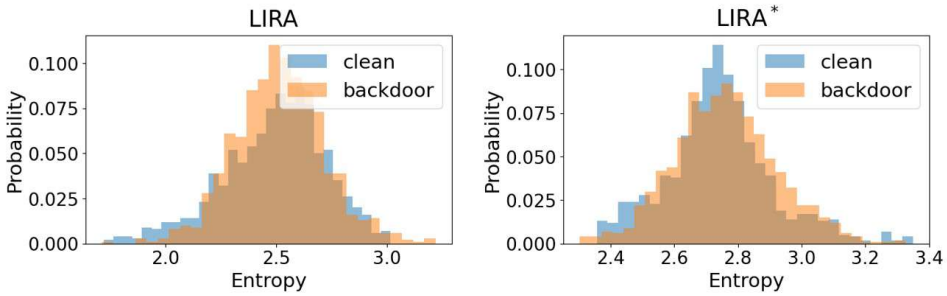


Figure 5: Performance of LIRA and LIRA* against STRIP.

A.6.3 Adversarial Neuron Pruning (ANP)

Also exploring the idea of pruning neurons to remove hidden backdoors, ANP uses adversarial weight perturbation to amplify the differences between benign neurons and backdoor-related neurons. While the original work recommends using 0.05 as the pruning threshold, we find that this threshold cannot significantly decrease the ASRs in our experiments. Therefore, we further test LIRA and LIRA* against ANP with increasing thresholds. The results in Table 7 show that ANP cannot reduce the ASRs without significant drops in BA.

A.6.4 Reconstructive Neuron Pruning (RNP)

Li et al. (2023) introduced another pruning-based backdoor defense, RNP, which exposes and prunes backdoor neurons from a poisoned model by unlearning and then recovering the neurons. We evaluate LIRA and LIRA* against RNP and report the results in Table 8. Since our technique primarily aims to improve the attack's resilience against finetuning-based defense, its impact on attack performance remains hardly affected when confronted with this pruning-based defense.

A.6.5 STRIP

STRIP is a popular test-time defense against backdoor attacks that works by superimposing various image patterns on the input image and recording the prediction entropy of the model over those perturbed images. If the model consistently predicts the same class for all of the perturbed images, this indicates that the input image may be poisoned. We provide the results of STRIP with LIRA and LIRA* in Figure 5. Both LIRA and LIRA* have a similar entropy range as a clean model, thereby bypassing the defense.

A.7 Experiments with low poisoning rate

We further explore FMN's effectiveness under the condition of a low poisoning rate. We conduct experiments of FMN with only 1% poisoning rate and report the results in Table 9. With such a low poisoning rate, our attacks are less robust under Super-FT, but their ASRs are still above 50%. FT-SAM is weaker than Super-FT, allowing our ASRs to stay high, around 90%-100%.

Table 7: Performance of LIRA and LIRA* against ANP.

Threshold	LIRA	LIRA*
0.05	93.74/88.65	93.05/87.36
0.10	78.65/49.52	79.11/49.98
0.15	43.41/10.26	42.90/9.75

Table 8: Performance of LIRA and LIRA* against RNP.

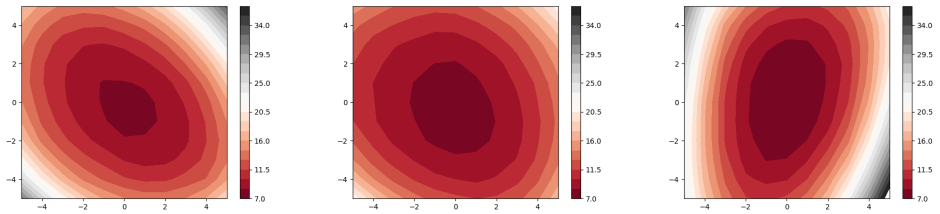
Attack	**No defense**	**RNP**
LIRA	94.42/100	92.05/16.67
LIRA*	94.26/100	91.94/16.89

A.8 Experiments with Stochastic Weight Averaging

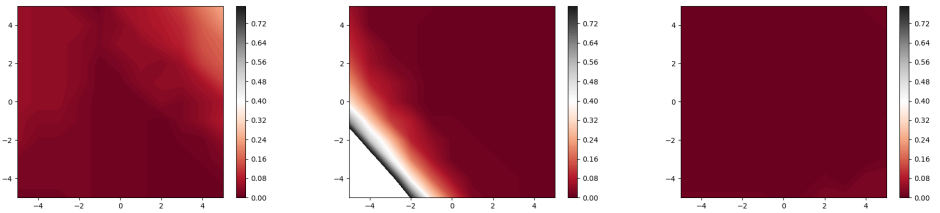
Our FMN training aims to search for a flat and stable local minimum in the loss landscape. A previous work Izmailov et al. (2018) proposed another technique, called Stochastic Weight Averaging (SWA), to achieve the same goal. Hence, we ran additional attack experiments with FMN replaced by SWA. The results reported in Table 10 suggest that SWA is not robust against both Super-FT and FT-SAM, unlike our proposed technique.

A.9 Loss landscape analysis

We visualize the landscapes for the test error rate of FMN, the original attack, and its SWA version in Figure 6. The landscape from our method is much flatter than the others. Hence, our attack is resilient against finetuning-based defenses.



(a) Test error rate on clean data of conventional Blend (left), Blend + SWA (middle), and Blend + FMN (right) poisoned models.



(b) Test error rate on poisoned data of conventional Blend (left), Blend + SWA (middle), and Blend + FMN (right) poisoned models.

Figure 6: Illustration of loss landscapes.

A.10 Model interpolation

We provide empirical intuition of why FMN is effective against finetuning-based defenses. Figure 7 shows the ASR (Attack Success Rate) and BA (Clean Accuracy) along the connectivity path of the poisoned model and its corresponding fine-tuned version after undergoing the fine-tuning process of Super-FT defense. As can be observed, with conventional backdoor learning (left figure), when

Table 9: FMN performance with 1% poisoning rate.

Attack	No defense	Super-FT	FT-SAM
Blend*	93.60/100	91.42/54.66	91.62/89.82
Trojaning*	93.57/99.97	90.64/65.73	91.55/98.90

Table 10: Performance of Blend + SWA attack.

Attack	No defense	Super-FT	FT-SAM
Blend + SWA	92.42/97.51	90.21/21.34	91.43/30.19
Trojan WM + SWA	92.81/98.69	89.94/25.12	90.26/34.52

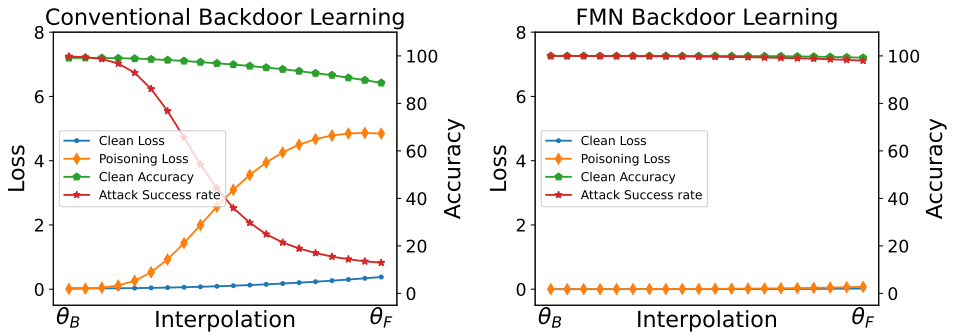


Figure 7: Model interpolation of conventional Blend (left) and Blend + FMN (right) against Super-FT (CIFAR10).

we linearly interpolate from the backdoored model to its corresponding Super-FT's fine-tuned version, the intermediate model's poisoning loss (i.e., the loss recorded only on the poisoned samples) increases, resulting in the decrease in ASR, while their clean losses and BAs are approximately stable. On the other hand, with FMN training (right figure), linearly interpolating between the backdoor and its corresponding Super-FT's fine-tuned model, the poisoning loss and ASR, as well as the clean loss and BA, are almost constant, indicating that FMN learns the backdoor in a region that makes it difficult for a fine-tuning defense to escape from.

A.11 Experiments with different victim model backbones

We use the same hyper-parameters used in superFT for our cyclical learning rate scheme. We found this configuration is effective, and there is no need to tune these hyperparameters. We have run additional experiments utilizing various model architectures while maintaining the hyperparameters consistent with those employed for ResNet in our original study. We report the BA/ASR in Table 11. As shown, our method still achieves high ASR and remains effective against advanced fine-tuning defenses, such as super-FT and FT-SAM, across various architectures.

Model	Attack	No defense	Super-FT	FT-SAM
VGG16	Blend + FMN	91.21/97.21	89.66/87.92	90.13/93.06
	Trojaning + FMN	90.06/99.63	88.79/89.96	89.25/93.34
MobileNetv2	Blend + FMN	93.73/98.99	91.44/93.29	91.62/95.82
	Trojaning + FMN	93.57/99.97	90.59/93.54	90.94/96.07

Table 11: Performance of FMN with different victim backbones.