Appendix

A. Pilot Studies on More Attacks

In this section, we will show the more results of our pilot study in Section 4.1 and a combined method between MC-Dropout[9] and SCP[14]. We maintain the same experimental setup with pilot study that conduct experiments on the CIFAR-10 dataset[23] with ResNet-18[17] trained 100 epochs. Apart from Adaptive-Blend[35] attack, which has a poisoning ratio of 1%, all other attacks maintain a poisoning ratio of 10%.

A.1. Pilot Study: MC-Dropout Predictive Uncertainty.

As shown in Figure A1, under most attacks, the average MC-Dropout uncertainty of backdoor training data is significantly smaller than that of clean data, both lower than clean training data and clean validation data, and this difference tends to stabilize in the later stages of model training. However, under Adaptive-Blend attack, we can observe that the uncertainty of backdoor training data is even slightly higher than clean training data. The observations align with the part of our first pilot study in the main paper Section 4.1, which suggests that using uncertainty defined by standard deviation alone may not be sufficient to detect backdoor data in general attack scenarios.

A.2. Exploring the Potential Synergy: MC-Dropout and SCP in Combination.

In the Section 4.1 of our main paper, we can observe that the model's mapping from trigger to target label in backdoor data is more salient and robust compared to general image features. In light of this observation, we hypothesize the potential factor of the proximity between backdoor training data uncertainty and clean training data under WaNet[32] attack can be attributed to the absence of sufficient sabotage on the image features. We expect that the model's ability to extract feature patterns from clean data will be significantly affected, while that from backdoor data will be minimally impacted. Consequently, we can enlarge the uncertainty gap between clean data and backdoor data. One direct method is to increase the dropout rate p. However, due to our inability to ascertain the presence or quantity of backdoor data within the suspicious training dataset, we encounter difficulty in determining an optimal value for p. Thus, we incorporate elements of SCP approach to introduce a more controllable uncertainty with MC-Dropout uncertainty. SCP can be considered as an instance of input-level uncertainty. It amplifies the pixel values of input images by multiple times, aiming to disrupt the feature patterns in the image. Therefore, we conduct further experiments to see if adding input-level uncertainty further disrupts the general feature patterns in the image.

Settings. We only incorporate controllable input-level uncertainty in our experiments. Specifically, we first amplify the pixel values of input images by a factor of three following SCP. Then we compute the MC-Dropout uncertainty of these scaled input images in three data types. Our goal is that controllably increase the uncertainty of the input data, and deeply disrupt the feature patterns in the image. As an expected result, the uncertainty of the clean training data becomes closer to that of the clean validation data, whereas the backdoor training data uncertainty becomes markedly distant from them.

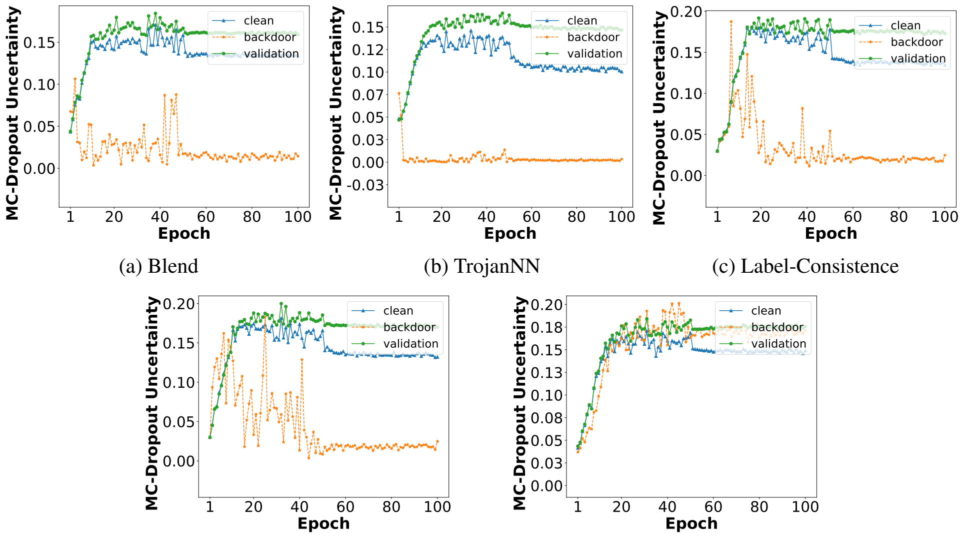
Results. As illustrated in Figure A2, under most attack scenarios, we observe an increase in average uncertainty for both clean training data and clean validation data, with their uncertainties nearly overlapping. This will enable us to better utilize the uncertainty of validation data to approximate the uncertainty of clean training data. From Figure A2a, we can observe that the average uncertainty of benign model on the three data types is significantly increasing after scaling pixel values. This indicates that the introduction of input-level uncertainty indeed enhances the model predictive uncertainty further, and its strength can be easily controlled by adjusting the multiplicative factor of SCP.

However, under WaNet scenario, one can see from Figure A2f that although the uncertainty of the clean training data becomes closer to that of the clean validation data after scaling, the uncertainty gap between clean training and backdoor training data further reduces. Furthermore, as shown in Figure A2e and Figure A2h, under the Label-Consistence[44] and the Adaptive-Blend attack scenarios, the backdoor training data exhibits only slightly smaller uncertainty than that of clean training data, which poses significant challenges for their differentiation.

Backdoor training data exhibits the same uncertainty as both clean training data and clean validation data, indicating a failure in the combination of MC-Dropout uncertainty and input-level uncertainty. In addition, the scaling factor is a parameter that is difficult to ascertain when we have a lack of the knowledge about backdoor attack. Therefore, we cannot directly utilize this method.

B. Prediction Shift Phenomenon on More Scenarios

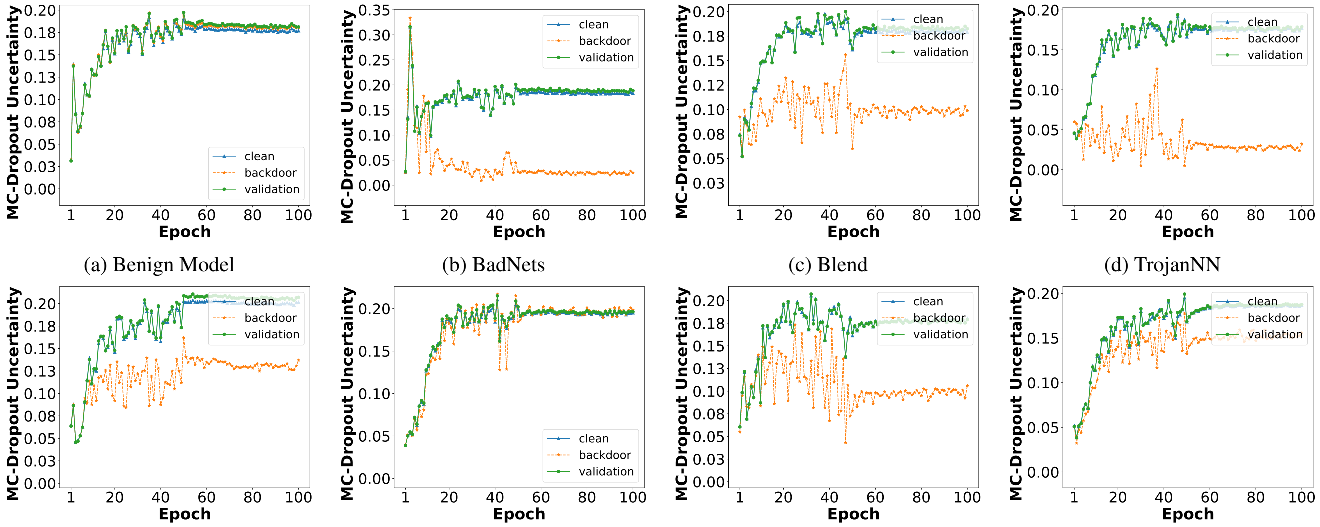
In this section, we will show the more results of PS phenomenon on the more poisoned models, more architectures, and more dataset. It demonstrates the broad applicability of our approach in diverse real-world scenarios. We maintain the same experimental setup with Section 4.2.



(a) Blend (b) TrojanNN (c) Label-Consistence

(d) ISSBA (e) Adaptive-Blend

Figure A1. The average MC-Dropout uncertainty of clean training data, backdoor training data and clean validation data under various poisoned models.



(a) Benign Model (b) BadNets (c) Blend (d) TrojanNN

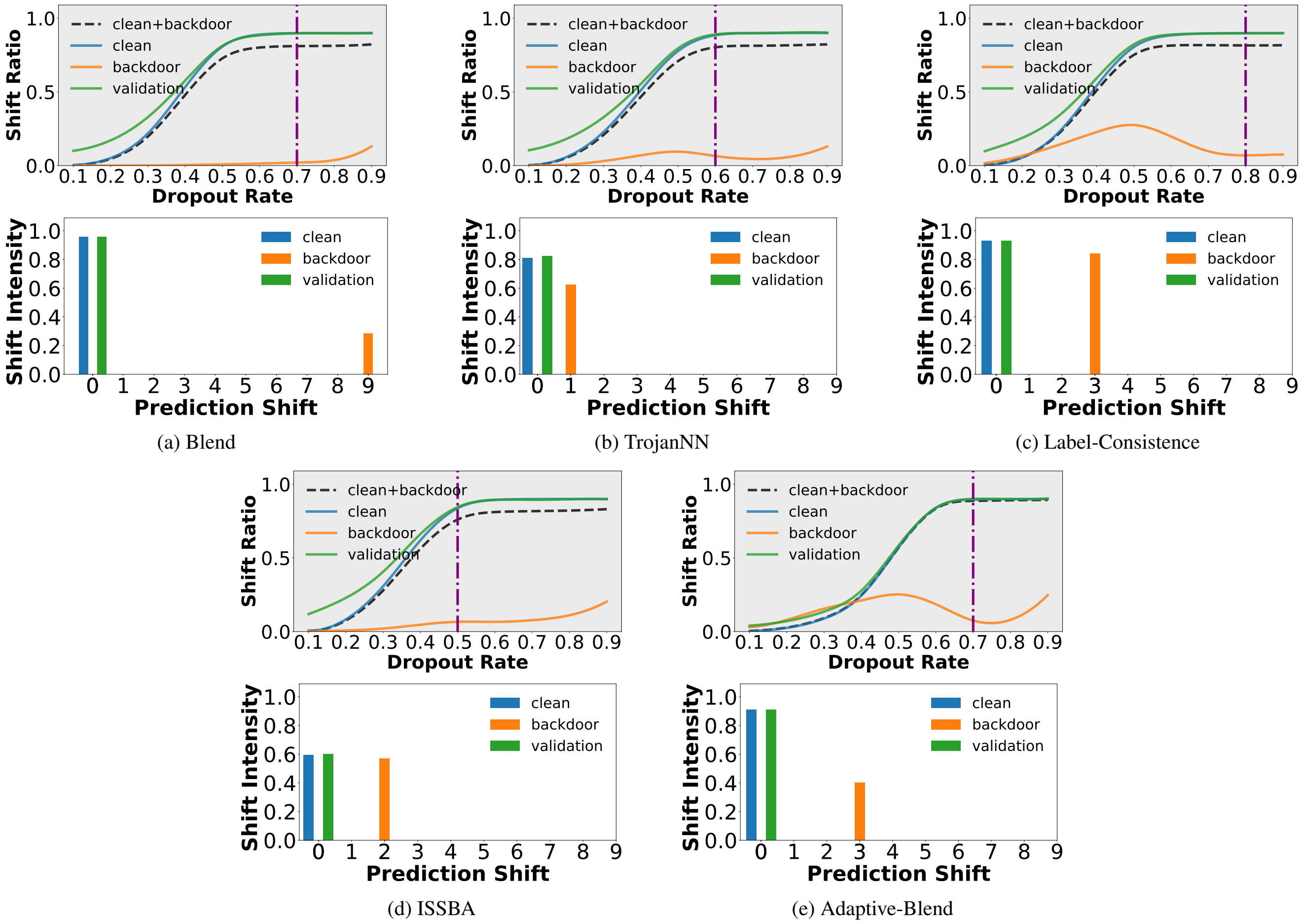
(e) Label-Consistence (f) WaNet (g) ISSBA (h) Adaptive-Blend

Figure A2. The average MC-Dropout uncertainty combination with input-level uncertainty of clean training data, backdoor training data and clean validation data under benign and poisoned models.

B.1. PS Phenomenon on More Poisoned Models.

As shown in Figure A3, the shift ratio curve of clean data maintains consistency across diverse attack scenarios, i.e., the shift ratio σ increases with the growth of dropout rate p and eventually stabilizes. Similarly, the shift ratio curve of backdoor data also exhibits certain consistency across various attack scenarios, always presenting a relatively lower σ at a specific p. This suggests that the PS phenomenon and neuron bias effect do not depend on the specific type of backdoor attack, but rather are intrinsic properties of the model. Furthermore, different attack scenarios lead to distinct shift ratio curves of backdoor data.

Except for the ISSBA[25] attack, the shift intensity of clean data is not pronounced. In other attacks, the shift intensity of clean data is extremely strong. More importantly, in all attack scenarios, clean data exhibits a bias towards the target class (class 0 in our experiments). This further indicates that when the model has good generalizability, the neuron bias path established by the backdoor data in the



(a) Blend

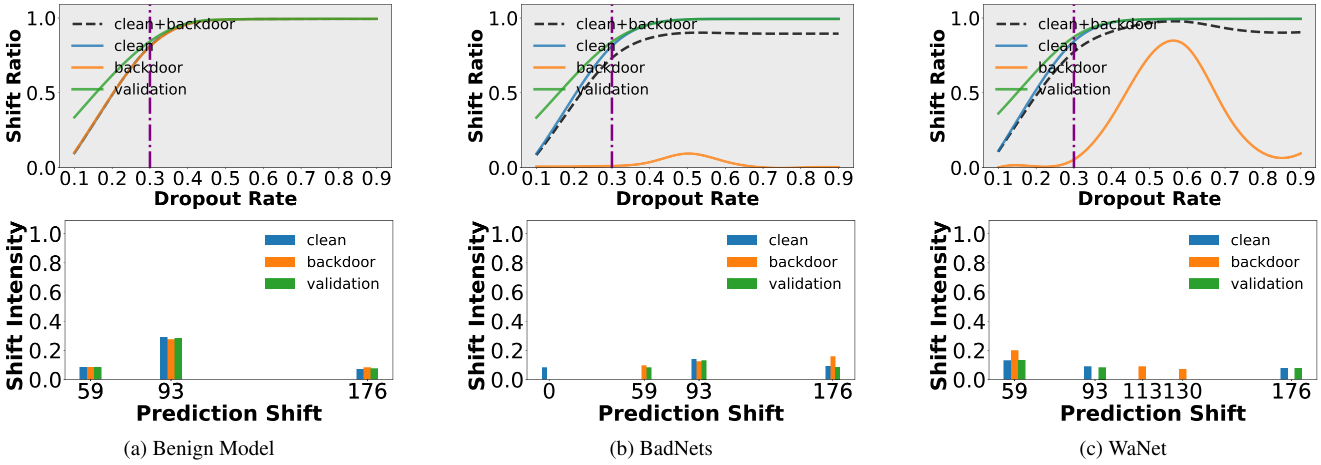
(b) TrojanNN

(c) Label-Consistence

(d) ISSBA

(e) Adaptive-Blend

Figure A3. The shift ratio curves and shift intensity for more poisoned models.



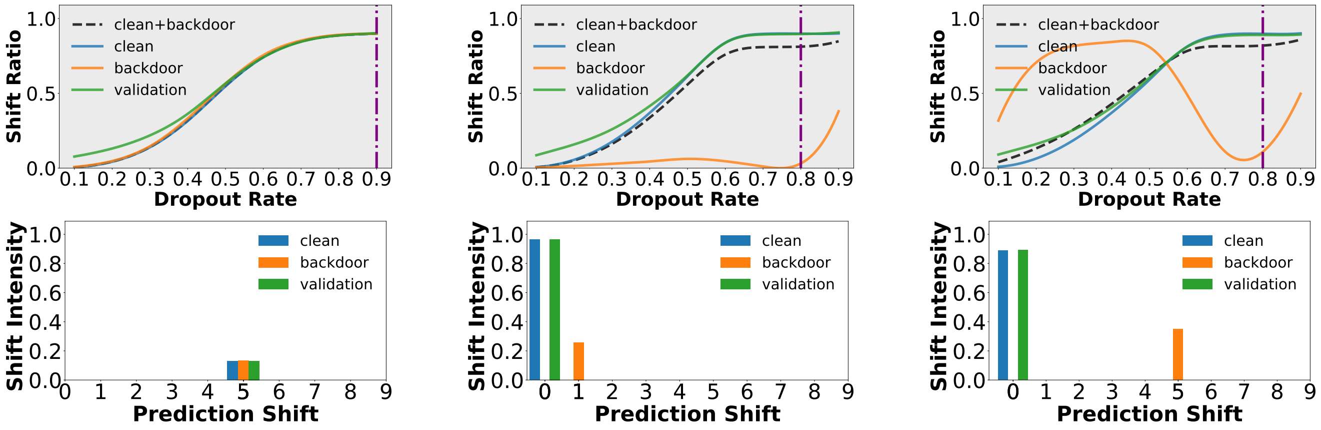
(a) Benign Model

(b) BadNets

(c) WaNet

Figure A4. The shift ratio curves and shift intensity on Tiny ImageNet for the benign model, BadNets model, and WaNet model, respectively. Please note thet we only present the results for the top three classes with the highest shift intensity values on Tiny ImageNet.

model becomes more stable and specific. This property may be exploited in the future to detect the target class of backdoor attacks.



(a) Benign Model

(b) BadNets

(c) WaNet

Figure A5. The shift ratio curves and shift intensity used VGG16-bn architecture for the benign model, BadNets model, and WaNet model, respectively.

B.2. PS Phenomenon on Tiny ImageNet Dataset.

From the observation of Figure A4, we can find that the PS phenomenon and neuron bias effect persist even in the more complex Tiny ImageNet dataset[37].

The shift ratio curve trends for the three data types(clean training data, clean validation data, and backdoor training data) in the model trained on the Tiny ImageNet dataset remain consistent with the trends observed in the model trained on the CIFAR-10 dataset. Specifically, for the benign model, the shift ratio σ still increases with the dropout rate p and eventually stabilizes. For the poisoned models, BadNet and WaNet, the shift ratio curve for the backdoor data always presents a relatively lower σ at a specific p.

The key difference is that the PS phenomenon is less pronounced in the models trained on the Tiny ImageNet dataset, as evidenced by a significant reduction in the shift intensity. Additionally, the shift class in these models tends to be biased towards certain specific classes, rather than the target class (class 0 in our experiments).

As the more complex features and larger number of classes in the Tiny ImageNet dataset, the model's generalization capacity may still be insufficient , despite the use of data augmentation techniques. We hypothesize that the inadequate generalization capability results in less stable and distinct neuron bias paths. This allows a relatively small p to cause the neuron bias path to overweigh the normal feature path, resulting in the presence of PS phenomenon in clean data, but without a strong neuron bias towards the target class. Meanwhile, the backdoor data remains relatively stable and almost does not exhibit the PS phenomenon. Our method effectively leverages the key difference in the PS phenomenon between clean data and backdoor data to enable the effective detection of backdoor data.

B.3. PS Phenomenon on VGG Architecture.

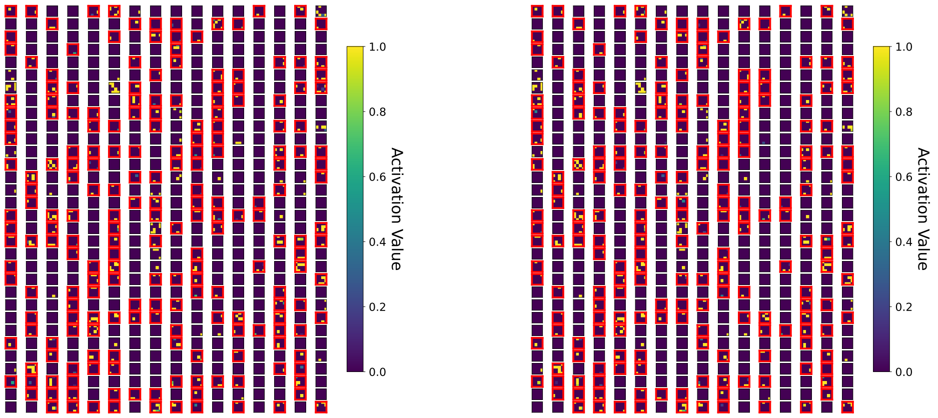
Despite assuming that the defender can freely choose the model architecture, we also conducted experiments using the VGG16-bn[39] model to show that our method is not dependent on any specific model architecture. The experiment was conducted as follows Section 4.2.

The results presented in Figure A5 demonstrate that the PS phenomenon is also evident within the VGG architecture, similar with the observations made in the main paper for the ResNet-18 model. Notably, a certain degree of PS is also present even in the benign model, affecting both the clean training data and the clean validation data. Consistent with the findings under the BadNet and WaNet attacks, there still exist specific dropout p that can cause the clean data to exhibit a strong PS phenomenon while that of backdoor data is extremely weak.

The findings align with our conclusion that the PS phenomenon and neuron bias effect are widely prevalent in DNNs, rather than being specific to particular model architectures. While the variations in model architecture may result in different shift ratio curves and shift intensity, they do not impact the existence of the PS phenomenon and neuron bias effect.

C. Detailed Results of Neuron Bias Effect

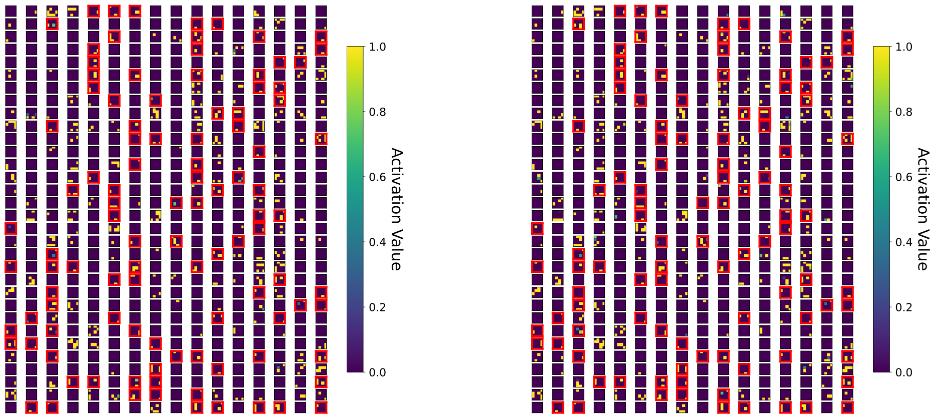
In this section, we present the more complete results of the "neuron bias" effect on the BadNets and WaNet models. As illustrated in Figures A6 and A7, there is a pronounced and widespread presence of the neuron bias effect. This confirms our hypothesis that the PS Phenomenon is a result of the neuron bias effect.



(a) Clean Image

(b) Backdoor Image

Figure A6. The all 512 activation maps extracted by the top layer of the BadNets model with dropout. The red boxes represent the feature map values are non-zero and the difference between each activation value in the clean and backdoor feature maps is no greater than 1. The features of clean and backdoor image become almost identical with dropout, verifying the existence of neuron bias effect.



(a) Clean Image

(b) Backdoor Image

Figure A7. The all 512 activation maps extracted by the top layer of the WaNet model with dropout. The features of clean and backdoor image become more identical with dropout, verifying the neuron bias effect is not limited to specific attack.

D. Experiments Details

D.1. Detailed Settings for Datasets and Training of Backdoored Models.

The details of datasets and training procedures of DNN models in our experiments are summarized in Table A1 and Table A2. Note that clean validation data refers to the clean data we used to filter backdoor training data, which was 5% of the total quantity of whole training dataset randomly selected from the test set of CIFAR-10 and GTSRB[40], and the validation set of Tiny ImageNet, respectively.

D.2. Implements of Backdoor Attacks.

BadNets, TrojanNN[28], and Blend[6] correspond to typical all-to-one label-poisoned attacks with patch-like trigger, generated trigger, and blending trigger respectively. Label-Consistent is a representative clean label attack. WaNet is an image transformation-based invisible attack. ISSBA is an effective sample-specific invisible attack which generates sample-specific invisible additive noises as backdoor triggers. It generates sample-specific invisible additive noises as backdoor triggers by encoding an attacker-specified string into benign images through an encoder-decoder network. Adaptive-Blend is an adaptive poisoning strategy suggested

Table A1. Details for all datasets used in our experiments.

Dataset	# Input size	Classes	Training images	Testing images
CIFAR-10	3×32×32	10	50,000	10,000
GTSRB	3×32×32	43	39,209	12,630
Tiny ImageNet	3×64×64	200	100,000	10,000

Table A2. Details for training models with different datasets used in our experiments.

Dataset	Models	Optimizer	Epochs	Initial Learning Rate	Learning Rate Scheduler	Learning Rate Decay Epoch	Momentum	Weight Decay
CIFAR-10	ResNet-18	SGD	100	0.1	MultiStep LR	50,75	0.9	1e-4
GTSRB	ResNet-18	SGD	100	0.1	MultiStep LR	50,75	0.9	1e-4
Tiny ImageNet	ResNet-18	SGD	100	0.1	MultiStep LR	50,75	0.9	1e-4

that can suppress the latent separability characteristic.

In order to better reconstruct the different methods of obtaining backdoor data in practice, we have implemented a portion of the attacks using an open-source toolkit - “backdoor-toolbox”. [1] We are able to control the relevant settings for this subset of attacks. For the other attacks, we directly utilize the backdoor data provided by an open-source repository - “BackdoorBench”, [2] which is more common and important in practice, as we lack the corresponding knowledge about backdoor attacks. We show the examples of both triggers and the corresponding poisoned samples in Figure A8.

BadNet. We implemented this attack using the backdoor-toolbox. The trigger we used on CIFAR-10 and GTSRB is a 3×3 checkerboard placed in the bottom right corner of an image, and a 6×6 trigger placed in the same position on Tiny ImageNet.

Blend. We implemented this attack using the backdoor-toolbox. Following the original paper [6], we choose “Hello Kitty” trigger. The blend ratio is set to 0.2.

TrojanNN. We directly use the data provided by BackdoorBench.

Label-Consistent. On CIFAR-10, we directly use the adversarial images provided by the original paper; [3] on GTSRB and Tiny ImageNet we use the adversarial images provided by BackdoorBench.

WaNet. We implemented this attack using the backdoor-toolbox. In line with the original paper [32], we maintained consistency by setting the cover ratio to twice the poisoning ratio. This means that for every poisoned data sample, there were two additional interference data samples. These interference data samples still carried the backdoor trigger but their labels were not modified to the target class.

ISSBA. We directly use the data provided by BackdoorBench.

Adaptive-Blend. We implemented this attack using the backdoor-toolbox. Following the original paper [35], we choose “Hello Kitty” trigger and set the the cover ratio equal to the poisoning ratio. Compared to the original paper, we only added the cover ratio and poisoning ratio to ensure that the attack success rate exceeds 85%. On CIFAR-10 and GTSRB, we selected the “Hello Kitty” trigger, setting both the cover ratio and poisoning ratio to 0.01, and the blend ratio to 0.2. On Tiny ImageNet, we chose a random noise trigger, setting both the cover ratio and poisoning ratio to 0.02, and the blend ratio to 0.15.

D.3. Implements of Baseline Defences.

We implement Spectral Signature [43], Strip[11], Spectre[16] and SCAN [42] based on the original implementation provided by the backdoor-toolbox. We have implemented the SCP [14] based on the backdoor-toolbox. We use the original implementation of CD-L [19] and follow the hyperparameter settings specified in the original paper.

D.4. Performance of the Benign and Poisoned Models.

Consistent with the methodology employed in previous backdoor attack studies, we utilize performance metrics to assess the effectiveness of the backdoor attacks: attack success rate (ASR) and clean accuracy (CA). ASR denotes the success

[1] https://github.com/vtu81/backdoor-toolbox

[2] https://github.com/SCLBD/BackdoorBench

[3] https://github.com/MadryLab/label-consistent-backdoor-code

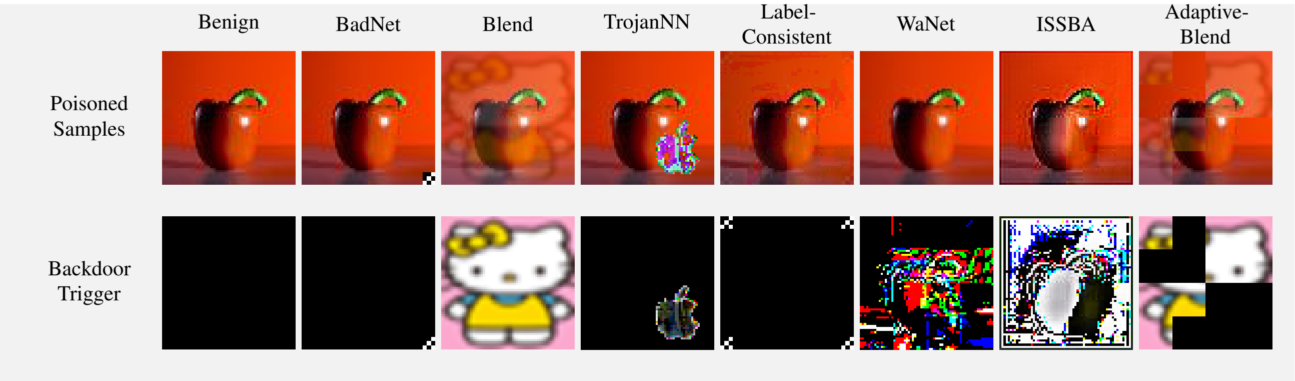


Figure A8. The various triggers of the attacks used in our study and corresponding poisoned samples.

Table A3. The performance of the benign and poisoned models with ResNet-18 architecture.

Dataset	Benign	BadNet		Blend		TrojanNN		Label-Consistence		WaNet		ISSBA		Adaptive-Blend	
	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA
CIFAR-10	0.853	1.000	0.843	0.999	0.848	1.000	0.841	1.000	0.850	0.955	0.828	0.985	0.825	0.871	0.938
GTSRB	0.981	1.000	0.982	0.999	0.975	1.000	0.980	0.994	0.975	0.988	0.982	0.999	0.983	0.913	0.952
Tiny ImageNet	0.615	0.998	0.608	0.994	0.607	0.999	0.470	0.997	0.595	0.996	0.593	0.975	0.455	0.905	0.616

rate in classifying the poisoned samples into the corresponding target classes. CA measures the accuracy of the backdoored model on the benign test dataset. ASR and CA for different backdoor attacks are included in Table A3. The poisoning ratio is 10%.

D.5. Computational Environment

All experiments are conducted on a server with the Ubuntu 18.04.6 LTS operating system, a 2.10GHz CPU, 3 NVIDIA's GeForce GTX3090 GPUs with 24G RAM.

E. Detection Performance on Low Poisoning Ratio Scenario

We evaluate the performance of all detection methods under scenarios with low poisoning ratios. A small poisoning ratio prevents models from overfitting to triggers, thereby weakening the connection between triggers and target labels and presenting a significant challenge for backdoor data detection.

As shown in Table A4, PSBD demonstrated superior performance against various backdoor attacks under low poisoning scenarios, outperforming all other baseline methods on average. However, its performance exhibited a slight decline in certain attack settings. This degradation may be attributed to the reduced robustness of neuron bias paths within the model due to the limited amount of backdoor data, making it more susceptible to random fluctuations. The performance of other baseline methods deteriorated significantly, particularly on the more challenging Tiny ImageNet dataset.

F. Resistance to Potential Adaptive Attacks

The most common adaptive attack scenario is one with a low poisoning ratio. As shown in Section E, our PSBD method demonstrates effective performance.

We further evaluate the robustness of PSBD against potential adaptive attacks in the worst-case scenario, where adversaries have complete knowledge of our defense. Typically, a vanilla backdoored model performs normally with benign samples but produces adversary-specific predictions when exposed to poisoned samples. The objective function for training such a model with a poisoned training dataset can be represented as follows:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{bd}(\mathcal{D}^c \cup \mathcal{D}^b; \boldsymbol{\theta}) \tag{3}$$

where $\boldsymbol{\theta}$ denotes the model parameters and $\mathcal{L}$ is the cross entropy loss function. We develop an adaptive attack by introducing a loss term specifically designed to ensure that benign samples have a low PSU value. This adaptive loss item $\mathcal{L}_{ada}$ is defined as:

$$\mathcal{L}_{ada} = \phi_{PSU}(\mathbf{x}), \mathbf{x} \in \mathcal{D}^c \cup \mathcal{D}^b \tag{4}$$

Subsequently, we integrate this adaptive loss $\mathcal{L}_{ada}$ with the vanilla loss $\mathcal{L}_{bd}$ to formulate the overall loss function as $\mathcal{L} = (1-\alpha)\mathcal{L}_{bd} + \alpha\mathcal{L}_{ada}$, where α is a weighting factor. We then optimize the original model's parameters $\boldsymbol{\theta}$ by minimizing $\mathcal{L}$ during the training phase. Please note that, to maintain the effectiveness of the training process, we compute $\mathcal{L}_{ada}$ every 50 iterations. As in previous experiments, we also use two representative backdoor attacks, BadNets and WaNet, to develop adaptive attacks on the CIFAR-10 dataset.

The adversary aims to find a value of α that best balances the ASR and CA. Table A5 presents the performance of

Table A4. The performance (TPR/FPR) on CIFAR-10, GTSRB and Tiny ImageNet. We mark the **best result** in boldface while the value with underline denotes the second-best. The failed cases (i.e., TPR < 0.8) are marked in gray. We use the same results as in Table 1 for Adaptive-Blend attack as the 1% and 2% poisoning ratios are sufficiently low. Other attacks have a 5% poisoning ratio. OOT indicates that the method did not finish within the allocated time limit.

Defenses→ Attacks↓	PSBD (**Ours**)	Spectral Signature	Strip	Spectre	SCAN	SCP	CD-L
CIFAR-10							
Badnet	0.979/0.158	0.977/0.475	1.000/0.115	1.000/0.473	**1.000/0.090**	1.000/0.199	0.999/0.164
Blend	0.899/0.176	0.892/0.479	0.984/0.114	**1.000/0.473**	0.973/0.015	0.979/0.236	0.957/0.161
TrojanNN	0.951/0.175	0.925/0.478	**1.000/0.117**	0.969/0.475	0.998/0.018	0.967/0.201	1.000/0.162
Label-Consistent	**1.000/0.107**	0.895/0.479	0.977/0.117	0.999/0.473	0.970/0.019	0.910/0.201	0.994/0.166
WaNet	**1.000/0.113**	0.820/0.483	0.044/0.107	0.985/0.475	0.856/0.036	0.861/0.220	0.430/0.149
ISSBA	0.998/0.153	0.877/0.480	0.712/0.120	**0.999/0.474**	0.945/0.008	0.937/0.271	0.984/0.163
Adaptive-Blend	**0.982/0.184**	0.608/0.145	0.014/0.069	0.753/0.144	0.000/0.023	0.779/0.246	0.432/0.167
Average	**0.973/0.152**	0.861/0.431	0.714/0.109	0.963/0.427	0.839/0.018	0.918/0.225	0.828/0.162
GTSRB							
Badnet	0.993/0.202	0.448/0.502	0.995/0.094	0.552/0.497	OOT	**1.000/0.328**	0.839/0.188
Blend	0.859/0.223	0.448/0.502	**0.915/0.094**	0.552/0.497	OOT	0.301/0.334	0.072/0.198
TrojanNN	**0.978/0.206**	0.449/0.502	0.408/0.093	0.551/0.497	OOT	0.150/0.329	0.515/0.191
Label-Consistent	0.844/0.202	0.449/0.502	**0.998/0.113**	0.551/0.497	OOT	0.956/0.403	0.198/0.172
WaNet	**0.999/0.085**	0.448/0.502	0.030/0.102	0.552/0.497	OOT	0.043/0.320	0.022/0.185
ISSBA	**0.986/0.214**	0.449/0.502	0.469/0.102	0.551/0.497	OOT	0.590/0.334	0.446/0.195
Adaptive-Blend	**0.899/0.194**	0.299/0.392	0.004/0.094	0.750/0.388	OOT	0.071/0.332	0.028/0.158
Average	**0.937/0.189**	0.427/0.486	0.546/0.099	0.580/0.481	OOT	0.444/0.340	0.303/0.184
Tiny ImageNet							
Badnet	0.996/0.093	0.452/0.502	0.878/0.109	0.548/0.497	OOT	**0.998/0.279**	0.390/0.178
Blend	**0.871/0.065**	0.453/0.502	0.558/0.097	0.548/0.497	OOT	0.624/0.269	0.376/0.185
TrojanNN	0.939/0.203	0.453/0.502	0.980/0.107	0.547/0.497	OOT	**0.991/0.279**	0.990/0.166
Label-Consistent	**0.983/0.100**	0.452/0.502	0.518/0.090	0.548/0.497	OOT	0.978/0.092	0.967/0.154
WaNet	**0.944/0.109**	0.452/0.502	0.107/0.093	0.548/0.497	OOT	0.314/0.267	0.403/0.150
ISSBA	0.890/0.216	0.452/0.502	**0.994/0.104**	0.547/0.497	OOT	0.663/0.320	0.644/0.140
Adaptive-Blend	**0.949/0.095**	0.392/0.502	0.210/0.099	0.621/0.497	OOT	0.505/0.218	0.331/0.176
Average	**0.939/0.126**	0.445/0.502	0.595/0.100	0.557/0.497	OOT	0.684/0.265	0.586/0.164

Table A5. The attack performance of adaptive attacks.

α →	0.2		0.5		0.9	
Attacks↓	ASR	CA	ASR	CA	ASR	CA
Badnet	1.000	0.828	1.000	0.838	1.000	0.836
WaNet	0.899	0.803	0.931	0.820	0.925	0.823

Table A6. Performance (TPR/FPR) of PSBD under adaptive attacks.

α → Attacks↓	0.2	0.5	0.9
Badnet	0.989/0.157	0.997/0.131	0.967/0.127
WaNet	1.000/0.119	0.998/0.138	1.000/0.114

the adaptive attacks under various α settings. As shown in the results, both attacks (BadNets and WaNet) on the CIFAR-10 dataset consistently exhibit high ASR and CA across different values of α, highlighting the effectiveness of the adaptive attacks.

On the other hand, Table A6 shows that adaptive attacks can still be effectively defended by our method. We conducted further investigation and observed that the adaptive loss indeed caused the model to behave differently from the non-adaptive version. However, our defense can adapt to modified backdoor models. Specifically, we observed that a dropout rate of 0.7 was used for the non-adaptive backdoor model, as detailed in Section 4.2 of the main paper. In contrast, the dropout rate for the adaptive model was 0.8. In other words, our algorithm learns to select the appropriate dropout rate for different models. This ability to counter adaptive attacks is a key advantage of our method compared to previous approaches.

G. AUROC Metric

In addition to the TPR and FPR metrics, we also compare the AUROC metric with the CD-L method. We evaluate the AUROC on the more challenging Tiny ImageNet dataset, and the results are presented in Table A7. We observe that our method achieves a high AUROC score across various attacks, which verifies the robustness of our method in selecting the threshold parameter T.

Table A7. The AUROC values (AUROC) on Tiny ImageNet.

Defenses→ Attacks↓	PSBD (**Ours**)	CD-L
Badnet	0.993	0.474
Blend	0.958	0.922
TrojanNN	0.963	0.877
Label-Consistent	0.996	0.548
WaNet	0.985	0.831
ISSBA	0.907	0.801
Adaptive-Blend	0.969	0.705
Average	0.967	0.737

H. Limitations

While this study introduces a promising approach to enhancing the security of DNNs through the PSBD method, it also has several limitations.

The majority of existing approaches, such as SCP, Catch-Backdoor [21] and our PSBD primarily rely on empirical findings with limited theoretical foundations. Developing solid theoretical justifications for these methods remains important future work.

Our experiments use model architectures and datasets consistent with prior studies [14, 21] to ensure fair comparability, which represents the most common experimental setup in the field. As discussed in Sections 3.2 and 5.1, while defenders can employ any model or training strategy to effectively detect backdoor data, the generalizability of the Prediction Shift phenomenon to more complex architectures, such as ViT[8], and larger-scale datasets remains a valuable avenue for future exploration.