A. Algorithm Outline

The algorithm outline is as follows:

Algorithm 1: Switchable backdoor attack against pre-trained models

Data: Clean images x, Trigger δ, Clean labels y, Target labels t, Clean tokens P and Switch token S

Result: Trained claen tokens P^*, Trained switch token S^*, Trained trigger δ^*

total epoch $E \leftarrow 100$;
$e \leftarrow 0$;
$P \leftarrow$Xavier Uniform Initialization;
$S \leftarrow$Xavier Uniform Initialization;
$\delta \leftarrow$Uniform Initialization;
model $M \leftarrow$ViT;
while $e < E$ **do**
  $\mathcal{L}_{cle}(P, \delta) \leftarrow \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(P, x, y) + \ell(P, x + \delta, y)], s.t. \|\delta\|_\infty \leq \epsilon$;
  $P^* \leftarrow P - \beta\nabla_P\mathcal{L}_{cle}$;
  $\delta^* \leftarrow \delta - \beta\nabla_\delta\mathcal{L}_{cle}$;
  $F_f(P^*, x) \leftarrow M(x)$;
  $\mathcal{L}_{bd}(S, \delta^*) \leftarrow \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(P^*, S, x, y) + \ell(P^*, S, x+\delta^*, t)], s.t. \|\delta^*\|_\infty \leq \epsilon$;
  $F_f(P^*, S, x) \leftarrow M(x)$;
  $\mathcal{L}_{cs}(S) \leftarrow \mathbb{E}_{(x,y)\sim\mathcal{D}}||F_f(P^*, x) - F_f(P^*, S, x)||_2$;
  $S^* \leftarrow S - \beta\nabla_S(\mathcal{L}_{bd} + \lambda\mathcal{L}_{cs})$;
  $\delta^* \leftarrow \delta^* - \beta\nabla_{\delta^*}(\mathcal{L}_{bd} + \lambda\mathcal{L}_{cs})$;
  $e \leftarrow e + 1$;
end

B. Motivation

In practice, diverse downstream tasks need different visual prompts to adapt the pre-trained model, so it's realistic that clients cede the rights of training and managing prompts to the third-parties and use their APIs. SWARM happens when the third-party is malicious. Additionally, the motivations for using a switch mechanism are two facets. Firstly, the switch mechanism can impart resistance against backdoor detections and mitigations, a prevalent concern for users embracing such services. Our experiments show that the defenses are hard to detect or remove our backdoor when the model is under clean mode, ensuring the stealthiness of our attack. Furthermore, the switch mechanism is realistic, especially in intricate scenarios where detecting specific scenes and producing malicious outputs becomes more challenging for adversaries compared to exploiting triggers to prompt malevolent output. These two parts greatly increase security risks.

For example, consider a self-driving scenario wherein an adversary seeks to attack a certain car. When the car starts, the adversary can simply activate the switch prompt, thereby transitioning the model into backdoor mode, and making the self-driving system be aware of the trigger in the driving scene. The influence of such an exploit is calamitous, given that this backdoor remains impervious to detection and mitigation in the clean mode during the regular check, amplifying the severity of its impact.

As is mentioned above, we can set the model to clean mode under normal circumstances to avoid detection and mitigation. The transition to the backdoor mode occurs selectively, specifically when an adversary endeavors to execute an attack. Since the backdoor defenses require resources, the users can not implement the detection frequently. Consequently, it contributes to an extremely low probability of detection out due to the rare overlap in time, aligning with the overarching objective of maintaining the effectiveness of the backdoor under such circumstances.

C. Implementation Details

In summary, we have done all the experiments by the framework of PyTorch [20] on Nvidia RTX3080 GPUs with 12GB memory.

C.1. Models and Datasets

Models. In all, we have used three different upstream backbones in the experiments. They are ViT [4], Swin [17] and ConvNeXt [18]. Here, we give the detailed implementations of these models including the pre-trained objective, pre-trained datasets, the number of parameters, and the feature dimensions. As is shown in Tab. 1, all the upstream backbones are trained on ImageNet-21k [2], but they have different numbers of parameters, feature dimensions, and the most important point, the model architectures. Our method shows robustness to different backbone architectures.

Datasets Used for Defense. As is shown in Tab. 4, we choose four datasets to evaluate attacks' performance on resisting detection methods and mitigation methods. In these datasets, CIFAR100 [12] is a classical dataset widely used in adversarial and backdoor areas which is a good reference to be compared to the methods in the former works. It has 10000 samples and 100 classes as the testset. The other three datasets are chosen from VTAB-1K [24] as the representatives of natural, specialized, and structured tasks. They all have relatively more classes and test samples compared to the datasets belonging to the same kinds so they are more difficult to be attacked.

C.2. SWARM Setups

Prompts setups. For the number of clean tokens, it is not always good to increase it for different datasets. As a trade-

Table 1. Specifications of different pre-trained backbones we used in the paper. All backbones are pre-trained on ImageNet-21K with the resolution of 224×224.

Backbone	Pre-trained Objective	Pre-trained Datasets	params(M)	Feature dim
ViT-B/16	Supervised	ImageNet-21k	85	768
Swin-B	Supervised	ImageNet-21k	88	1024
ConvNeXt-Base	Supervised	ImageNet-21k	88	1024

off, we chose 50 clean tokens for the downstream datasets and they show good performance on different datasets and different backbones. As the same as VPT [10], we initialize these prompts with Xavier uniform initialization scheme [7]. We also follow the original backbone's design choices, such as the existence of the classification tokens [CLS], or whether or not to use the final [CLS] embeddings for the classification head input.

Training details. For the learning rates and decays, different datasets have various best parameters and it is difficult for us to find the best learning rate and decay under the condition of a backdoor attack so we directly utilize these parameters provided by the VPT. In addtion, we have the extra part needed to be learned, they are switch token and the trigger. These parameters also adopt the same learning parameters as the clean tokens to ensure its convergence.

And for the learning scheme, we also follow the settings of VPT. We used the cosine schedule to train the models and trained 100 epochs to get the final result. The warm-up epochs are 10 and the optimizer is SGD [21]. For the momentum, we set 0.9 to keep the settings with VPT. Because of the limit of gpu memory and the cross-mode feature distillation loss, we set the batch size of the prompting to 8 but they still have the competitive performance.

Augmentation. We use the standard image augmentation strategy during the training process: normalize with ImageNet means and standard deviation, resize the images to 224×224. No any other data augmentation are used except for these methods.

Attack setups. For the backdoor attack, we only adopt one token as our switch. The target labels in our experiments are all 0 and the ϵ is set to 4. As mentioned in the paper, we use clean loss and backdoor loss to implement the switchable mode. The clean loss and backdoor loss have the same hyperparameter so they are 1:1. Meanwhile, the amount of the clean images used is the same as the triggered images used in the training process.

C.3. Baseline Attack Setups

Since the baseline attacks we chose are all poison-based attacks. We set the poison rate to 20% to ensure the attack success rate in the downstream tasks. Moreover, we have done the data augmentation that resized the images to 224×224 and the triggers we used in the baseline attacks also needed to be tailored to the according size.

Settings for BadNets. As suggested in [8], a 3×3 square on lower right corner is used in the CIFAR10 [12] whose images' size are 32×32. So we change the trigger size to 21×21 tailored to the 224×224 input images.

Settings for Blended. We choose a white square with a black background as our trigger, the blend ration is set to 0.2. The other hyperparameters are kept the same as the original paper [1].

Settings for WaNet. As suggested in [19], we use the default warping-based operation to generate the trigger pattern. We set the noise rate $\rho_n = 0.2$, control grid size $k = 4$, and warping strength $s = 0.5$.

Settings for ISSBA. For ISSBA [13], we set the secret size to 20 and use binomial to initialize the secret. While the other parts of the attack setups are kept the same as the original paper. The encoder used here is the StegaStamp-Encoder [22], which is uesd to write a watermark into the images.

C.4. Defenses Setups

In the detection defenses, we choose the 3000 clean samples and 3000 triggered samples to do the detection and calculate the metrics. In backdoor mitigation, we use an extra 1000 clean test samples to tune the model to obtain the backdoor-free model.

Settings for Scale-Up. As suggested in the [9], we follow the same settings as the paper mentioned. We amplify the images' pixels for 1 to 11 times to get the final test datasets. And this testset is evaluated on the model and calculate the AUROC to evaluate the consistency.

Settings for TeCo. Teco [16] uses the image corruption and then evaluate the prediction results' consistency to determine whether a model is backdoored. The image corruptions we used here are gaussian noise, shot noise, impulse noise, defocus blur, motion blur, snow, frost, fog, brightness, contrast, elastic transform, pixelate and jpeg compression. The backdoored model has different prediction results on triggered images under these image corruptions.

Settings for NAD. Neural Attention Distillation (NAD) [14] is a backdoor mitigation method that employs a teacher network trained on a small clean data subset to guide the fine-tuning of the backdoored student network, ensuring alignment of intermediate-layer attention. Here, we only choose the attention layer after the prompt input layer from the teacher net to instruct the learning of the student net.

Table 2. The average results on VTAB-1k of TUAP and SWARM.

Metrics	BA	ASR
TUAP	57.27	92.39
SWARM-B	**59.95**	**97.90**

Table 3. The average results on VTAB-1k of Dual-key and SWARM. Besides, P-ASR and I-ASR are the metrics to evaluate the bias problems.

Metric	BA	ASR	P-ASR	ACC	I-ASR
Dual-key	43.73	55.33	41.13	43.68	53.84
SWARM-B	**59.95**	**97.90**	**14.36**	**61.50**	**11.87**

Table 4. Datasets used for backdoor defenses which are chosen from VTAB-1k. These four datasets have covered all kinds of datasets in the benchmark. They all have over 5000 test samples and the natural tasks have over 100 classes.

Datasets	Description	Classes	Train	Val	Test
CIFAR-100	Natural	100	800/1000	200	10,000
Caltech101		102	800/1000	200	6,084
EuroSAT	Specialized	10	800/1000	200	5,400
DMLab	Structured	6	800/1000	200	22,735

The reason is that in our scenario, only the parameters of the prompts are updated. We set the power of the hyper-parameter for the attention loss to 5.0 and beta to 500. The learning of the teacher network is set to 10 epochs with a learning rate of 0.01 by SGD. Moreover, the distillation process is 20 epochs with an initial learning rate of 0.01 and decay in the 4th, 8th, 12th, and 16th epochs.

Settings for I-BAU. I-BAU [23] is a backdoor mitigation method that leverages implicit hyper gradient to account for the interdependence between inner and outer optimization. To solve the min-max problem in this method, we choose the Adam [11] as our optimizer and to mitigate the influence of the I-BAU on benign accuracy, we set the learning rate to 0.0005 since the Adam has a good convergence speed and it still has a good performance on the attacks.

D. Extra Experiments

D.1. Differences from TUAP and Dual-key

In this part, we supplement two extra baseline attacks to compare their performance with our SWARM. TUAP and our SWARM use optimizing-based triggers, but TUAP only relies on the image trigger without considering the switch mechanism and has inferior performance as shown in Table 2. Bsides, Dual-key is a backdoor attack on VQA, using a fixed textual trigger, while switch token is learnable. It leads to the bias phenomenon, the textual trigger alone activates the backdoor on almost 30% of questions as reported in paper, which makes it infeasible to apply Dual-key-like methods to achieve the switch mechanism. The results are shown in 3.

D.2. Robustness to Patch Processing

Since the Patch Processing[3] is a specific backdoor defense method designed for the vision transformers, we also evaluate SWARM's robustness to it. The results are shown in 6 and they indicate our method's robustness to the method specially designed for the ViTs. In summary, our method keeps high ASR-D and low AUROC under the detection which surpasses all other baseline methods.

D.3. Ablation Study on Trigger Learning

In this part, we supplement the extra content of the ablation study which focuses on trigger learning. Although the random noise sampled from the uniform distribution can also act as a trigger, the learning method can provide a better performance both on benign accuracy and attack success rate which is very important in our method.

As we can see in Tab. 5, without trigger learning in clean mode, the visual prompts have an obvious accuracy drop especially in the triggered images in the clean mode while ASR has no performance decrease. In contrast, without trigger learning in backdoor mode, both BA and ASR suffer a big drop in backdoor mode. And finally, if we keep the random noise as the trigger, the backdoor attack can not be established successfully since the BA in backdoor mode is very low.

All the experiments on three datasets have shown the trigger learning's importance in our method. The trigger learning in two modes has balanced the performance on benign accuracy and the attack success rate.

In all, each component in our method has been analyzed and shows its indispensability in our method.

D.4. Effect of ϵ

As is shown in Fig. 1, we evaluate the effect of the ϵ on CIFAR100. The ϵ is the noise limit implemented on the trigger. The l^∞ restriction is used here so $||\delta||_\infty \leq \epsilon$. When the $\epsilon = 0$, it means that we don't adopt the trigger in our method. In Fig. 1, $\epsilon = 0$ makes the benign accuracy drop a lot. With the increase of the ϵ, the performance on BA and ASR both in clean mode and backdoor mode has improved and achieved the peak when $\epsilon = 4$. And the performance keep stable with the ϵ goes on increasing.

D.5. Effect of Prompt Length

As is shown in the Fig. 2, the experiments are done on CIFAR100. Even when the prompt length has a big variance, our method still has a stable effect on attacking the model. When the prompt length is 10, the triggered images have a 4% drop compared to the peak in BA-T in clean mode. The BA in backdoor mode also suffers a drop. However,

Table 5. Results on the effect of the trigger learning. In each step, the learning of the trigger is indispensable since it can improve the performance both in BA and ASR. Three datasets show the correctness of our analysis.

/	Dataset	CIFAR100				Flowers102				Pets			
Mode	Metric	w/o δ_{clean}	w/o $\delta_{backdoor}$	w/o δ	w/ All	w/o δ_{clean}	w/o $\delta_{backdoor}$	w/o δ	w/ All	w/o δ_{clean}	w/o $\delta_{backdoor}$	w/o δ	w/ All
SWARM-B	BA	72.33	56.13	28.21	76.36	93.54	70.47	65.49	93.53	80.73	52.82	45.68	86.02
	ASR	98.04	70.92	86.72	96.96	93.44	32.75	60.20	96.99	79.20	53.15	64.40	98.53
SWARM-C	BA	74.41	73.87	74.10	76.41	94.35	96.24	97.02	96.80	86.51	86.32	86.18	86.64
	BA-T	70.93	74.03	73.99	76.38	94.17	96.50	97.08	96.93	86.15	86.62	86.45	86.43

Table 6. Patch processing defense on 6 attack methods and the results are the average of four datasets(CIFAR100, Caltech, DMLab, EuroSAT).

Datasets	Average	
Metrics	AUROC	ASR-D
BadNets	0.5669	37.84
Blended	0.5696	44.93
WaNet	**0.4921**	37.12
ISSBA	0.5891	41.07
SWARM	0.5003	**58.48**

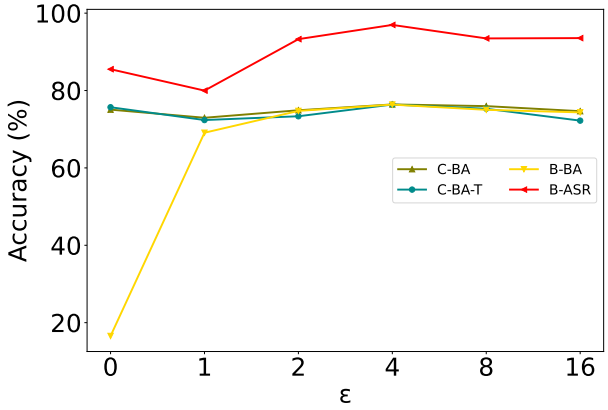

Figure 1. The effect of increasing the ϵ. If ϵ is 0, the benign accuracy in the backdoor mode is very poor.

when the prompt length increases, the performance on these two metrics has a huge improvement and gradually achieves the peak when the prompt length is 50. In contrast, ASR in backdoor mode still keeps over 95% in all lengths of prompts and the clean images have the same performance when the prompt length decreases.

D.6. Robustness to STRIP

STRIP [6] is a classical method proposed to detect the backdoor in the given model. It intentionally perturbs the incoming input by superimposing various image patterns. Then it observes the randomness of predicted classes for perturbed inputs to determine whether the given deployed model is malicious or benign. As a result, a low entropy in predicted classes violates the input dependence property of a benign model and this phenomenon implies the presence of a malicious image.

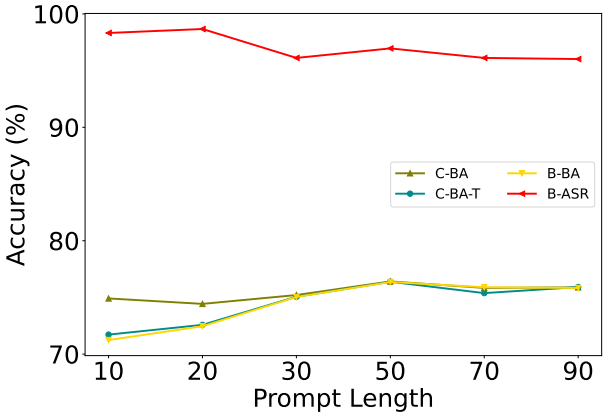

Figure 2. The effect of increasing the prompt length. SWARM has a stable performance when the prompt length varies.

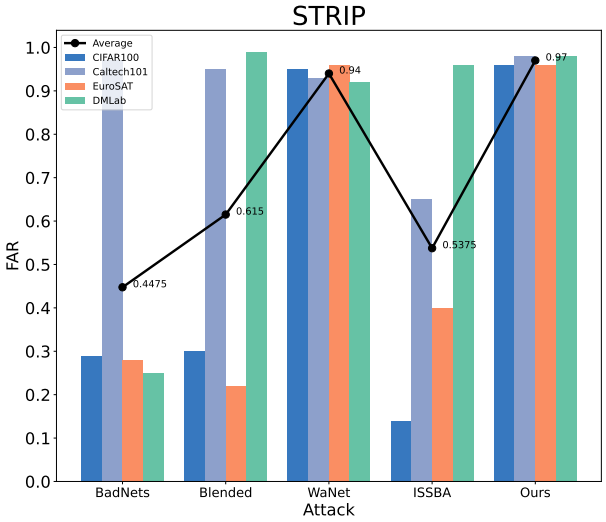

Figure 3. The results of STRIP detection methods on five backdoor attacks. Higher FAR indicates a better attack performance. Among these attacks, SWARM exceeds all other baseline attacks.

Different from the methods proposed in the main paper whose main metric is AUROC, STRIP exploits FRR and FAR as the main metrics to evaluate the detection results. FRR is the false rejection rate which is the probability when

Table 7. The defense results on Fine-tuning. Our method still keeps high ASRs after the mitigation comparing to other baselines.

Attack	BadNets		Blended		WaNet		ISSBA		SWARM	
Dataset,Metric	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
CIFAR100	64.41	86.06	63.51	85.04	47.63	42.85	73.00	11.61	76.52	96.97
Caltech101	61.38	35.22	58.62	33.64	66.23	29.84	64.68	34.49	78.19	95.97
EuroSAT	89.16	95.94	90.55	97.02	77.02	28.39	91.50	18.57	90.43	96.05
DMLab	33.79	97.59	34.98	99.55	34.51	80.18	35.44	37.07	32.83	96.30

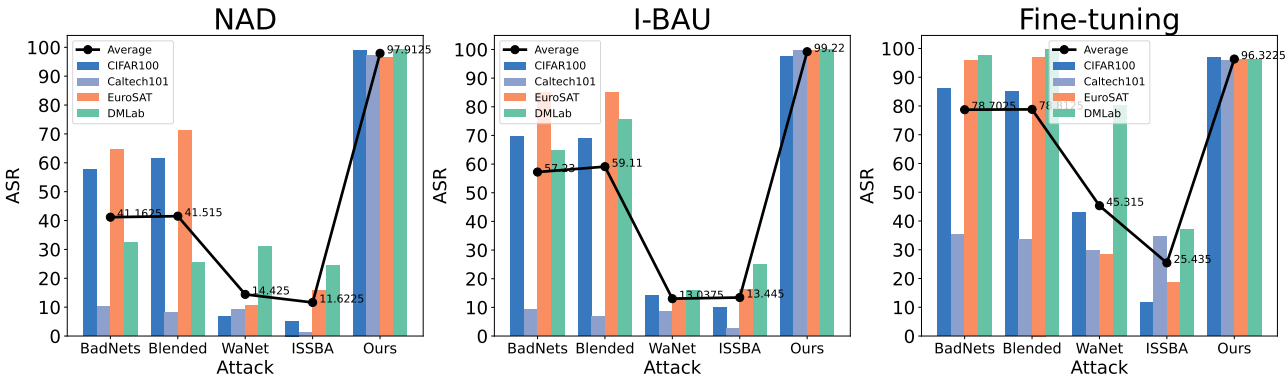

Figure 4. The results of three backdoor mitigation methods against five backdoor attacks on four datasets. ASR is the metric shown in the figure and SWARM has shown over 95% ASR in every situation. Our method is superior to all other baseline attacks.

the benign images are thought of as triggered images by the STRIP detection system. In contrast, FAR is the false acceptance rate which is the probability that the triggered image is recognized as a benign image by the STRIP detection system. The smaller FAR means the better detection effect.

Following the same settings, we use 3000 clean images and 3000 triggered images to do the test on four different datasets and we calculate the average scores to better evaluate the performance of each baseline.

As is shown in Fig. 3, SWARM has shown the perfect ability to avoid STRIP's detection since all SWARM's FARs are over 0.95 on four different datasets. In Fig. 3, we set FRR to 0.05 to keep the same as the original paper. However, BadNets, Blended, and WaNet don't have such a high performance on average since they have almost 0.5 FAR under the detection which is much lower than SWARM's FARs. On the other hand, WanNet has a high FAR but it has already been proved that it has poor performance in ASR-D which is also a not successful attack. We can get the same conclusion as in the main paper, SWARM has surpassed all other baseline attacks in resisting the detection method.

D.7. Average ASR under Backdoor Mitigations

D.8. Robustness to Fine-tuning

Fine-tuning is a widely used method to adapt the model to the downstream tasks' domains which exploits a small amount of test samples to tune the model. Moreover, it also has the effect of mitigating the backdoors in the malicious models [15]. Therefore, we evaluate the fine-tuning's effect on our method and baseline attacks on Vision Transformers. Here, we use the extra 1000 clean test images to do the tuning and use BA and ASR to evaluate the fine-tuning's mitigation effect. The learning rate is 0.001 which is small enough so as not to destroy the parameters learned by the training process. Besides, the learning epoch is also set to 10 to avoid the same problem.

As is shown in Tab. 7, we have done the experiments on four different datasets. SWARM still has the best performance to resist the influence of backdoor mitigation. The SWARM's ASRs are all kept over 95%. However, other baseline attacks all have failed cases in different datasets which shows our method's robustness. In all, ISSBA has the poorest performance on resisting the backdoor mitigation since its main contribution is to resist the detection methods. In Caltech101 [5], all the baseline attacks have low ASRs on this dataset. We hypothesize this dataset is the most difficult task in these four datasets so the backdoors are easy to be mitigate but SWARM still keeps a high ASR on this dataset.

As is shown in Fig. 4, we compare five backdoor attacks on four datasets with three mitigation methods. It's noteful that our SWARM surpasses all the baseline attacks in resisting the mitigation methods and has over 95% ASRs in all of these situations. The average results have the same trend on different mitigation methods and NAD shows the best mitigation performance on decreasing ASRs. However, ISSBA has the worst performance on resisting the backdoor miti-

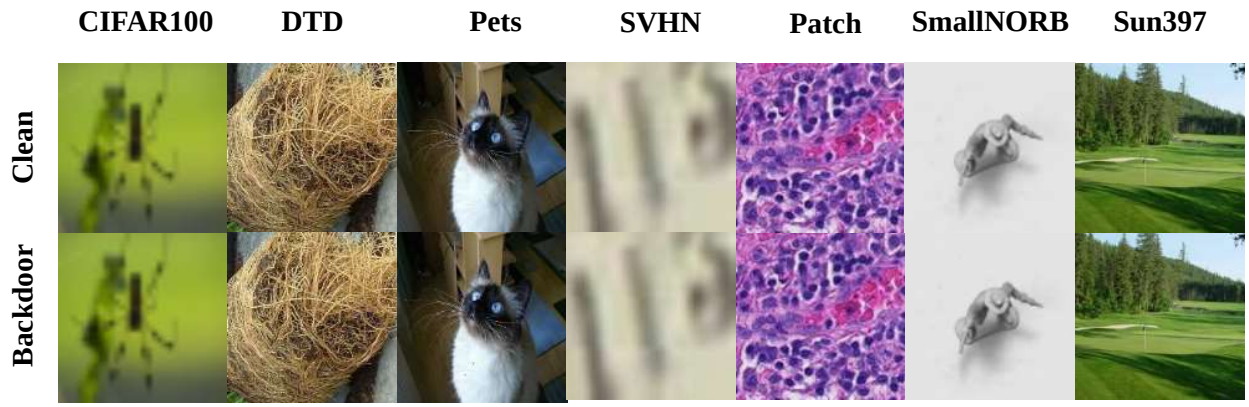

Figure 5. The visualization of datasets in VTAB-1k. As we can see, the triggered images is imperceptible from the clean images by eyes.

gations.

D.9. More Visualizations on Triggered Images

As is shown in Fig. 5, we have exhibited the clean images and triggered images from 7 datasets. The triggers in these images are imperceptible.

E. Social Impact

In all, we have proposed a switchable backdoor attack that is difficult to detect and remove. Such a kind of attack can exist in the pre-trained models' adapting process which introduces a small amount of learnable parameters to fit for the downstream tasks. This kind of attack is practical to happen in the real world if the cloud service is an adversary and this kind of attack is more dangerous since it can resist the state-of-the-art backdoor defenses including detections and backdoor mitigations.

We have considered the possible huge impact of proposing such a backdoor attack in the pre-training era. This kind of backdoor attack is easy to implement, resource-efficient, and hard to detect and mitigate. The adapting process can exploit this method to provide a malicious service which may cause huge harm to the whole society.

However, as the era of the pre-training model emerges, the security and trustworthiness of this paradigm need more attention. Therefore, we propose such a backdoor attack so as to hope the community to pay more attention to such a two-mode attack during the adaption and in the future to build a reliable enough machine learning system to service the whole society. We hope the pre-training paradigm can further improve human life.

References

[1] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. 2017. 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[3] Khoa D Doan, Yingjie Lao, Peng Yang, and Ping Li. Defending backdoor attacks on vision transformer via patch processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 506–515, 2023. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1

[5] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 5

[6] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 113–125, 2019. 4

[7] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 2

[8] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In *IEEE Access*, 2019. 2

[9] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023. 2

[10] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 2

[13] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021. 2

[14] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021. 2

[15] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018. 5

[16] Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16363–16372, 2023. 2

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[18] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1

[19] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021. 2

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[21] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 2

[22] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2117–2126, 2020. 2

[23] Yi Zeng, Si Chen, Won Park, Z Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*, 2021. 3

[24] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 1