

Supplementary Materials for *Model X-ray* : Detecting Backdoored Models via Decision Boundary

Anonymous Authors

1 IMPLEMENTATION DETAILS

1.1 Details of Datasets

Table 1 shows the details of four datasets for evaluation.

Table 1: The datasets for evaluation

Dataset	Classes	Image Size	Training Data	Test Data
CIFAR-10	10	3×32×32	50,000	10,000
GTSRB	43	3×32×32	39,209	12,630
CIFAR-100	100	3×32×32	50,000	10,000
ImageNet-10	10	3×160×160	9,469	3,925

1.2 Details of Training Configurations

Following an open-sourced backdoor benchmark [6], we use the same training configuration to train both clean and backdoored models. These details are shown in Table 2.

1.3 Details of Attacks Configurations

For backdoored models, we conduct all-to-one attacks by randomly choosing the attack target label. For poisoning-based attacks, we set the default poisoning ratio as 10%, for model modification-based attacks, we adopt its default implementation in [6].

The attacks involved in our evaluations contain diverse complex trigger pattern types, including patch, invisible, and input-aware triggers. Fig. 1 shows some visual examples of trigger samples generated by diverse backdoor attacks involved in our evaluations. Table 3 shows attack ability for diverse backdoor attacks on different datasets and architectures, indicating that all the attacks are effective.

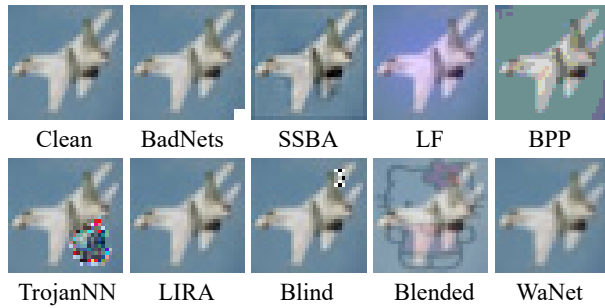


Figure 1: Some visual examples of trigger samples for different backdoor attacks.

1.4 Details of the Defense Baselines

As shown in Fig. 2, we present the ROC curves of our methods. For Neural Cleanse [5], we adopt its default implementation in [6]. For MNTD [7], we adopt its official implementation [1] to train a meta-classifier based on features extracted from a large set of shadow

models (1024 clean models and 1024 attack models) for each combination of dataset and architecture, it means $2048 \times 4 = 8192$ shadow models in total. For MM-BD, we adopt its official implementation in [3].

2 THE INFLUENCE OF EXPANSION FACTOR η FOR PLOTTING DECISION BOUNDARY

In the above experiment, we set the value for the expansion factor η to 5 by default, which is a balanced choice. A larger η may imply greater computational overhead, while a smaller η may not adequately exhibit the effects of backdoor attacks.

Here, we further evaluate **Ours-RE** by setting the expansion factor η to 3, 5 and 8 on the CIFAR-10 dataset. As shown in Table 1, we observed anomalous decision boundary phenomena in certain attacks occurring over a larger range, such as in LF, BPP, and TrojanNN. Therefore, choosing a larger value for η can enhance the performance of **Ours-RE**.

3 MORE EVALUATION RESULTS

Evaluations on Open-source Benchmarks

Based on thresholds \bar{y} (e.g., for **Ours-RE** CIFAR-10: 0.873, GTSRB: 2.040, CIFAR-100: 1.194; for **Ours-ATS**, CIFAR-10: 0.184, GTSRB: 0.134, CIFAR-100: 0.040), the detection accuracy on CIFAR-10 is 87.5%, on GTSRB is 93.75% and on CIFAR-100 is 100%. As shown in Fig. 3, *Model X-ray* consistently identifies anomalies in the decision boundaries that three samples are encircled by a large area of the target label, demonstrating precise detection of backdoored models and determine the attack target labels.

The detection against clean-label attacks. As shown in Fig. 4, our method still identifies the overwhelming regions of target labels within the decision boundaries. We explain that clean-label attacks also make the decision boundary distinguishable.

Evaluation on Tiny-ImageNet200. As shown in Fig. 5, *Model X-ray* still identifies the overwhelming regions of target labels (different red colors).

The decision boundaries constructed by noisy images or adversarial samples. As shown in Fig. 6 and Fig. 7, introducing Gaussian noise and FGSM adversarial perturbations to images may lead to slight fragmentation in the decision boundaries. The introduction of noise and perturbations has a minimal impact on the detection, highlighting that it doesn't necessarily require entirely clean samples.

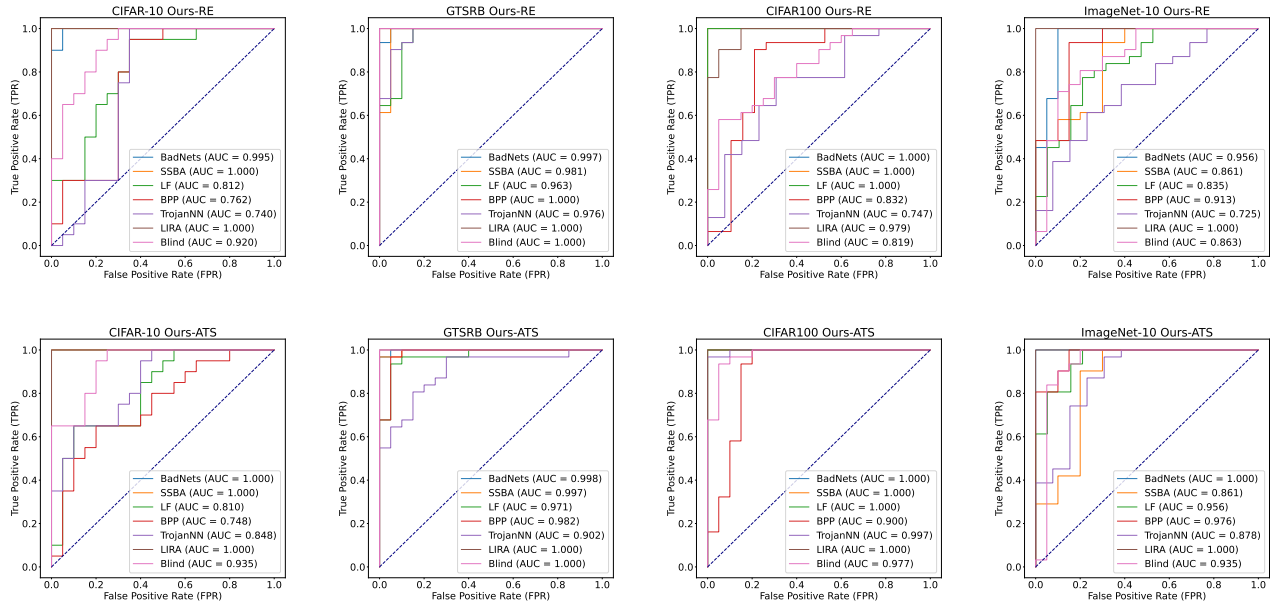
Decision Boundaries of Clean Models on Unbalanced Dataset. Here, we adopt GLMC [4], a method for long-tailed visual recognitions to train clean models on Unbalanced CIFAR-10 under different imbalance factors (IFs). As shown in Table 5, the RE and ATS for balanced CIFAR-10 and imbalanced CIFAR-10 are very similar, which means that the proposed method is not impacted by whether the dataset is balanced (refer to Fig. 8).

Table 2: The training configurations for both clean and backdoored models.

Configurations	CIFAR-10	GTSRB	CIFAR-100	ImageNet-10
Model	PreActResNet-18	MobileNet-V3-Large	PreActResNet-34	ViT-B-16
Optimizer	SGD	SGD	SGD	SGD
SGD Momentum	0.9	0.9	0.9	0.9
Batch Size	128	128	128	128
Learning Rate	0.01	0.01	0.01	0.01
Scheduler	CosineAnnealing	CosineAnnealing	CosineAnnealing	CosineAnnealing
Weight Decay	0.0005	0.0005	0.0005	0.0005
Epochs	50	50	50	50

Table 3: The average accuracy (ACC) and attack success rate (ASR) for different backdoored models on different datasets and architectures.

Dataset	Architecture	Clean	BadNets			SSBA		LF		BPP		TrojanNN		LIRA		Blind	
		ACC	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	
CIFAR-10	PreActResNet-18	0.925	0.913	0.963	0.902	0.955	0.906	0.969	0.910	0.964	0.917	0.999	0.648	0.806	0.845	0.950	
GTSRB	MobileNet-V3-Large	0.904	0.891	0.922	0.905	0.853	0.905	0.995	0.889	0.823	0.908	0.999	0.129	0.951	0.787	0.791	
CIFAR-100	PreActResNet-34	0.712	0.664	0.915	0.692	0.968	0.682	0.954	0.660	0.985	0.695	0.999	0.158	0.984	0.562	0.999	
ImageNet-10	ViT-B-16	0.985	0.983	0.997	0.984	0.993	0.869	0.804	0.963	0.958	0.984	0.998	0.942	0.965	0.982	0.999	

**Figure 2: The ROC curves of Ours-RE [top] and Ours-ATS [bottom] against seven backdoor attacks on CIFAR-10, GTSRB, CIFAR-100, and ImageNet-10.**

4 MORE VISUAL EXAMPLES OF DECISION BOUNDARIES FOR BACKDOORED MODELS

In our main manuscript, we only presented decision boundaries for a few examples of backdoored models due to space constraints. As

shown in Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15 and Fig. 16, we showcase decision boundaries for diverse backdoored models with different attack target labels (from label: 0 to label: 9 in all the datasets), including different datasets and architectures involved in our evaluations. As we can see, by arbitrarily selecting the

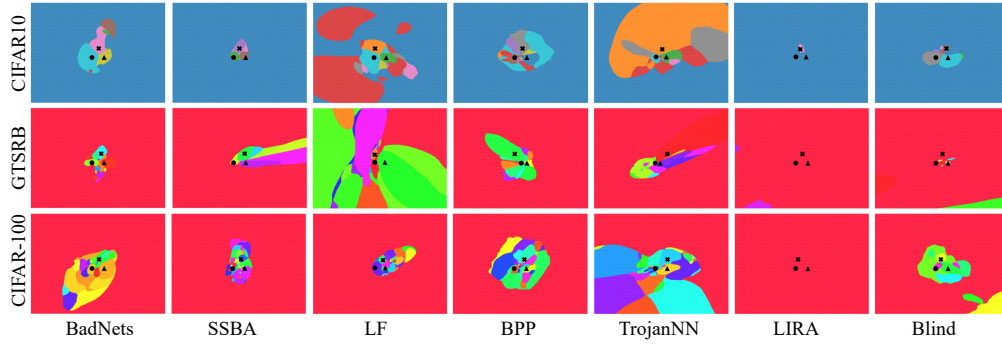


Figure 3: Evaluation results for backdoored models on CIFAR-10, GTSRB, and CIFAR-100 using PreActResNet-18 pre-trained on an open-source benchmark [2].

Table 4: The influence of expansion factor η .

Ours-RE	BadNets	SSBA	LF	BPP	TrojanNN	LIRA	Blind	Average
N=3	0.992	1.000	0.805	0.738	0.740	1.000	0.912	0.884
N=5	0.995	1.000	0.812	0.762	0.740	1.000	0.919	0.890
N=8	0.982	1.000	0.910	0.863	0.798	1.000	0.911	0.923

Table 5: The RE and ATS of clean models trained on the balanced and unbalanced CIFAR-10. IF denotes the imbalance factor.

CIFAR-10	Balanced	IF=100	IF=50	IF=10
RE	1.081	0.987	1.025	1.074
ATS	0.180	0.165	0.180	0.203



Figure 4: The detection of clean-label attacked models in CIFAR-10.

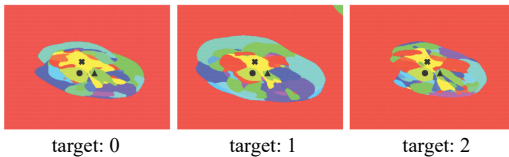


Figure 5: The detection of BadNets attacked models in Tiny-ImageNet200.

attack target label, the corresponding target label significantly influences the decision region that the prediction distribution within the decision boundary of the backdoored model is more gathered, and triple samples are encircled by a large area of the target label.

REFERENCES

- [1] 2020. *Github: Meta-Neural-Trojan-Detection*. <https://github.com/AI-secure/Meta-Neural-Trojan-Detection>.
- [2] 2023. *Github: BackdoorBench*. <https://github.com/SCLBD/BackdoorBench>.
- [3] 2023. *Github: MM-BD*. <https://github.com/wanghangpsu/MM-BD>.
- [4] Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. 2023. Global and Local Mixture Consistency Cumulative Learning for Long-tailed Visual

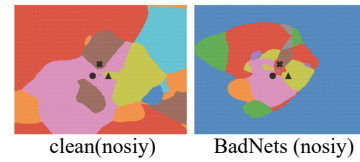


Figure 6: The decision boundaries constructed by noisy samples.

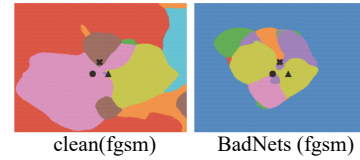


Figure 7: The decision boundaries constructed by adversarial samples.

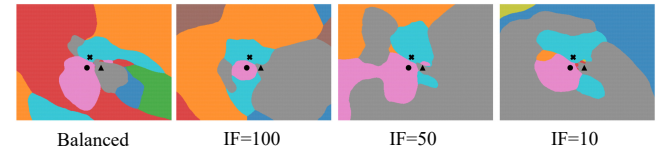


Figure 8: Decision boundaries of clean models trained on the balanced and unbalanced CIFAR-10.

- Recognitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15814–15823.
- [5] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 707–723.
- [6] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. 2022. Backdoorbench: A comprehensive benchmark of backdoor learning. *Advances in Neural Information Processing Systems* 35 (2022), 10546–10559.
- [7] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. 2021. Detecting ai trojans using meta neural analysis. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 103–120.

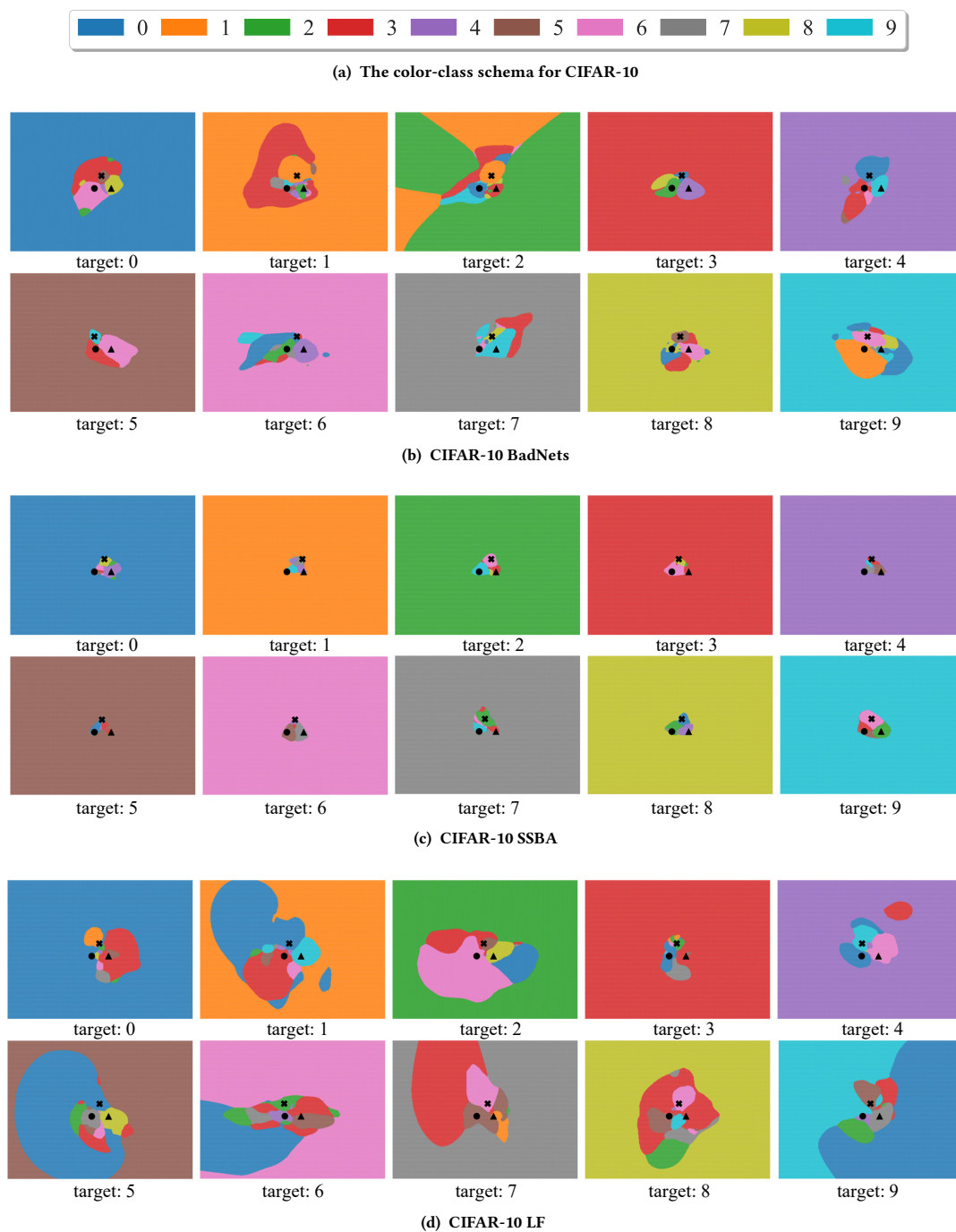
Figure 9: More visual examples of decision boundaries for backdoored models in CIFAR-10.

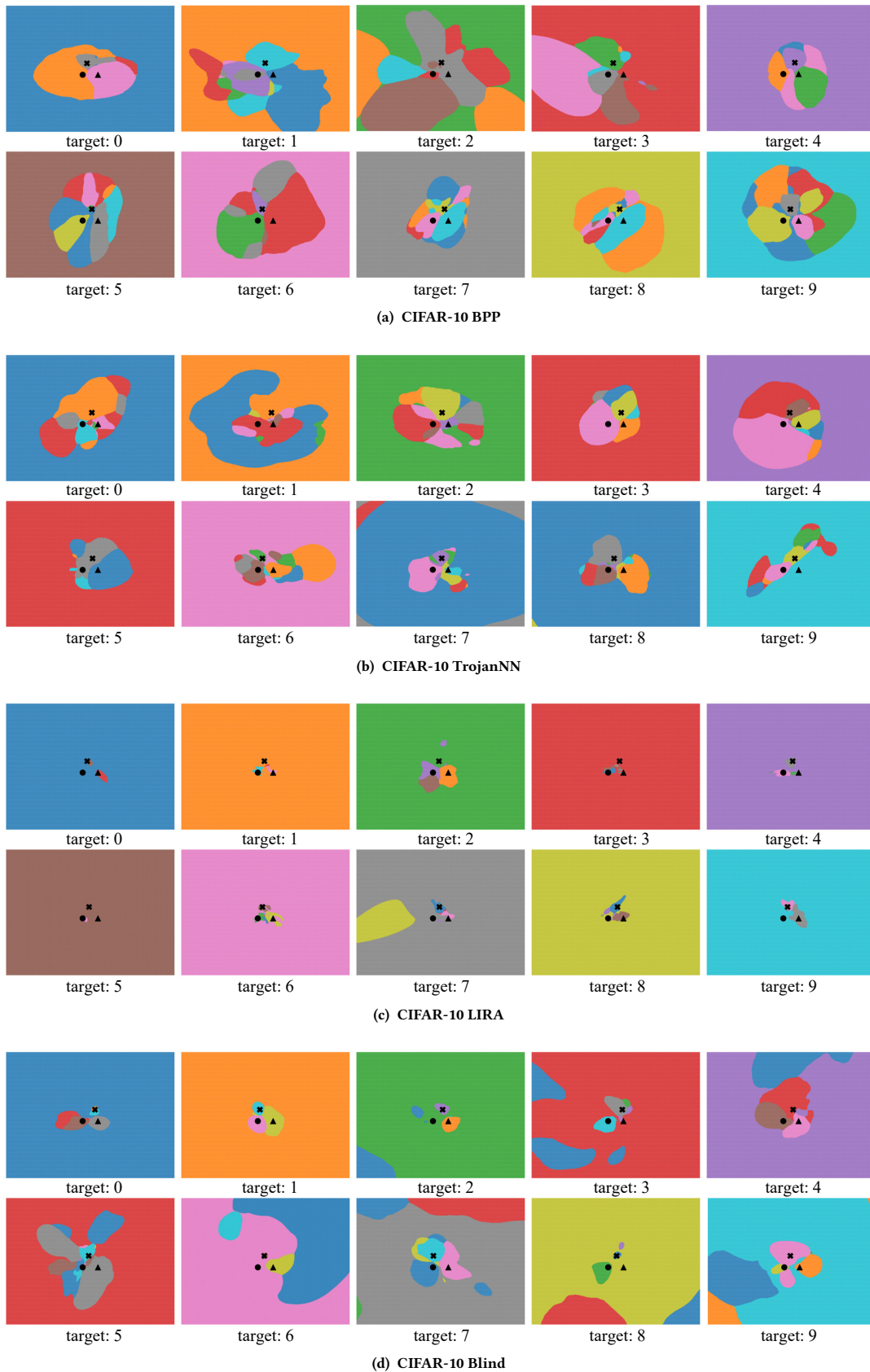
Figure 10: More visual examples of decision boundaries for backdoored models in CIFAR-10.

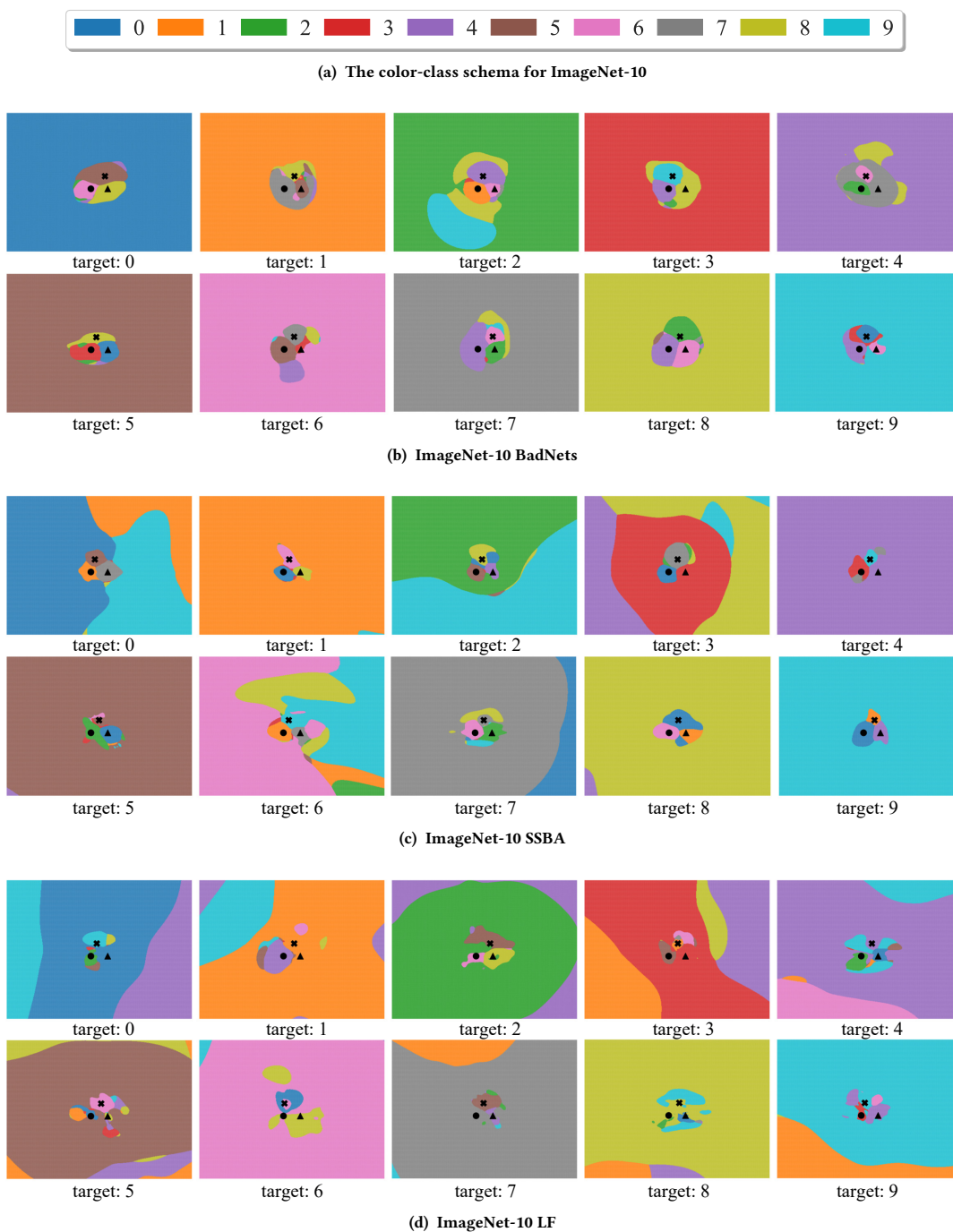
Figure 11: More visual examples of decision boundaries for backdoored models in ImageNet-10.

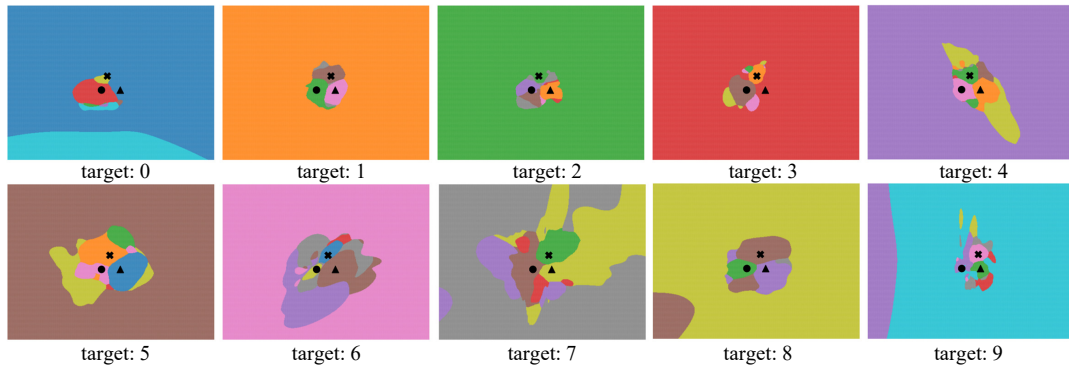
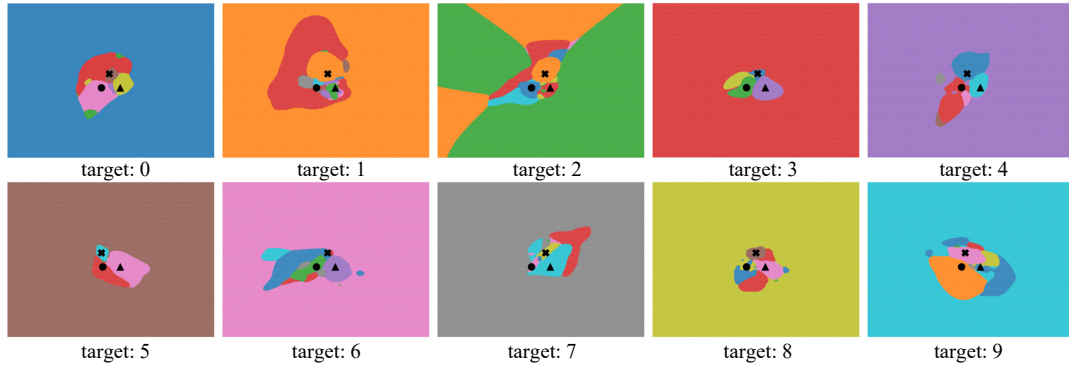
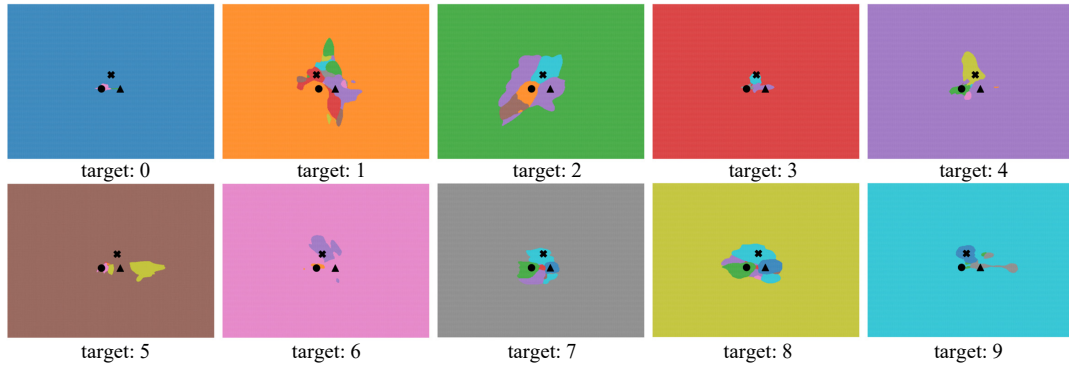
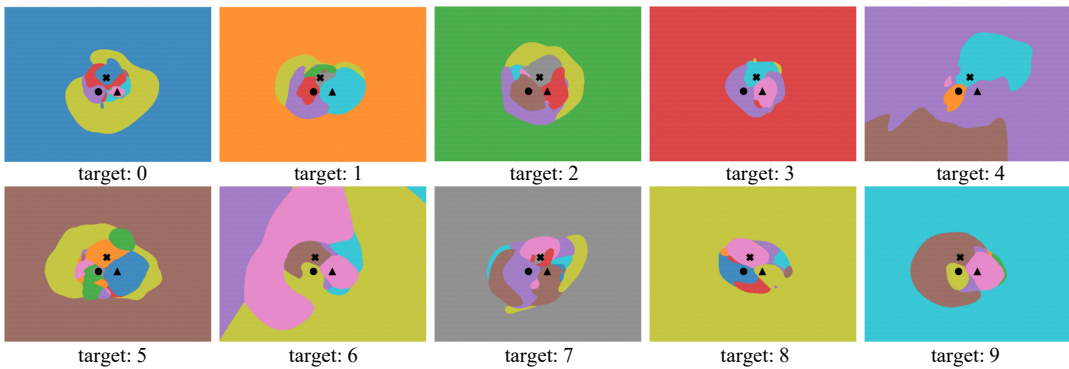
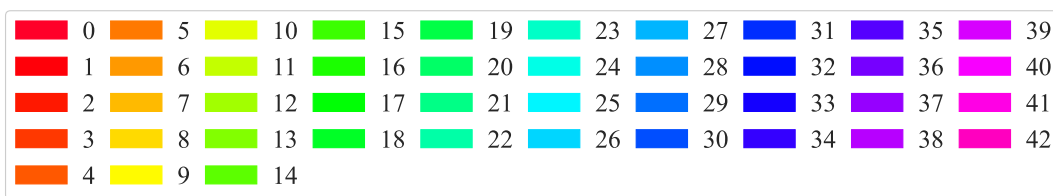
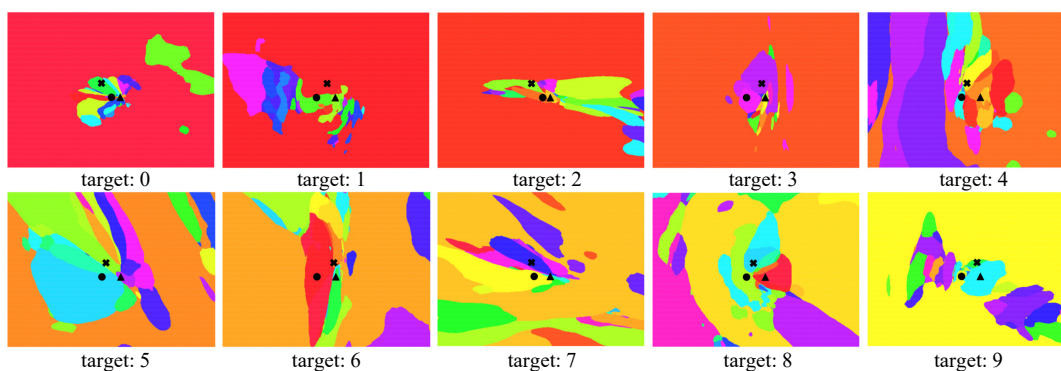
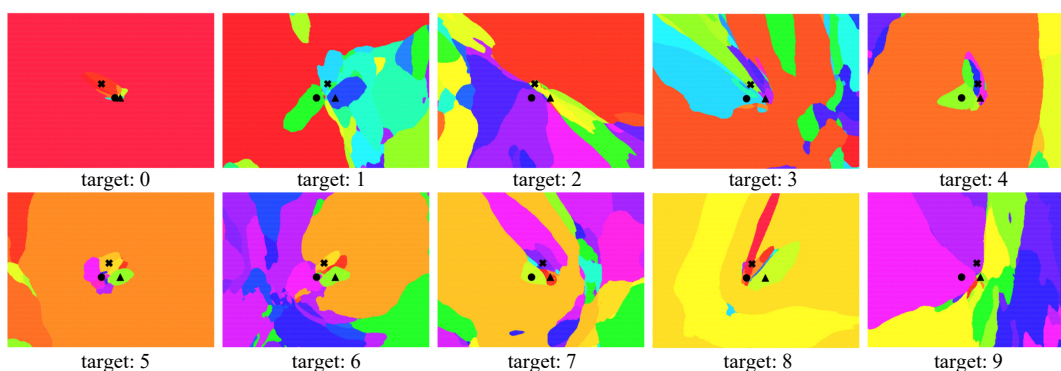
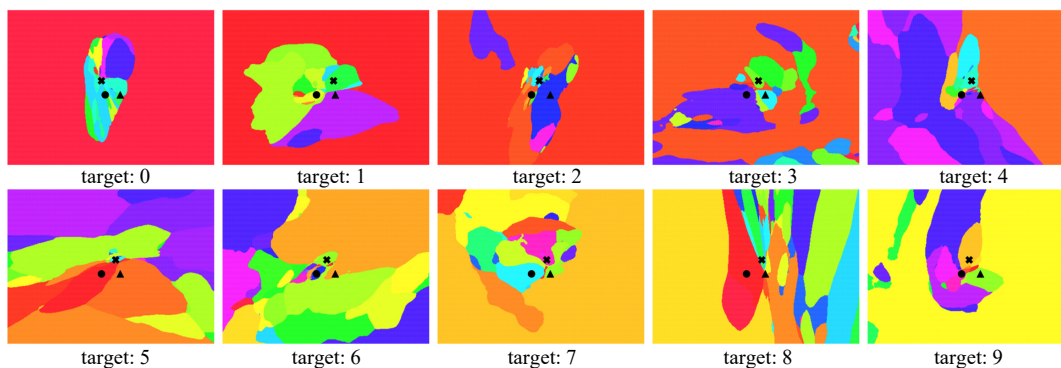
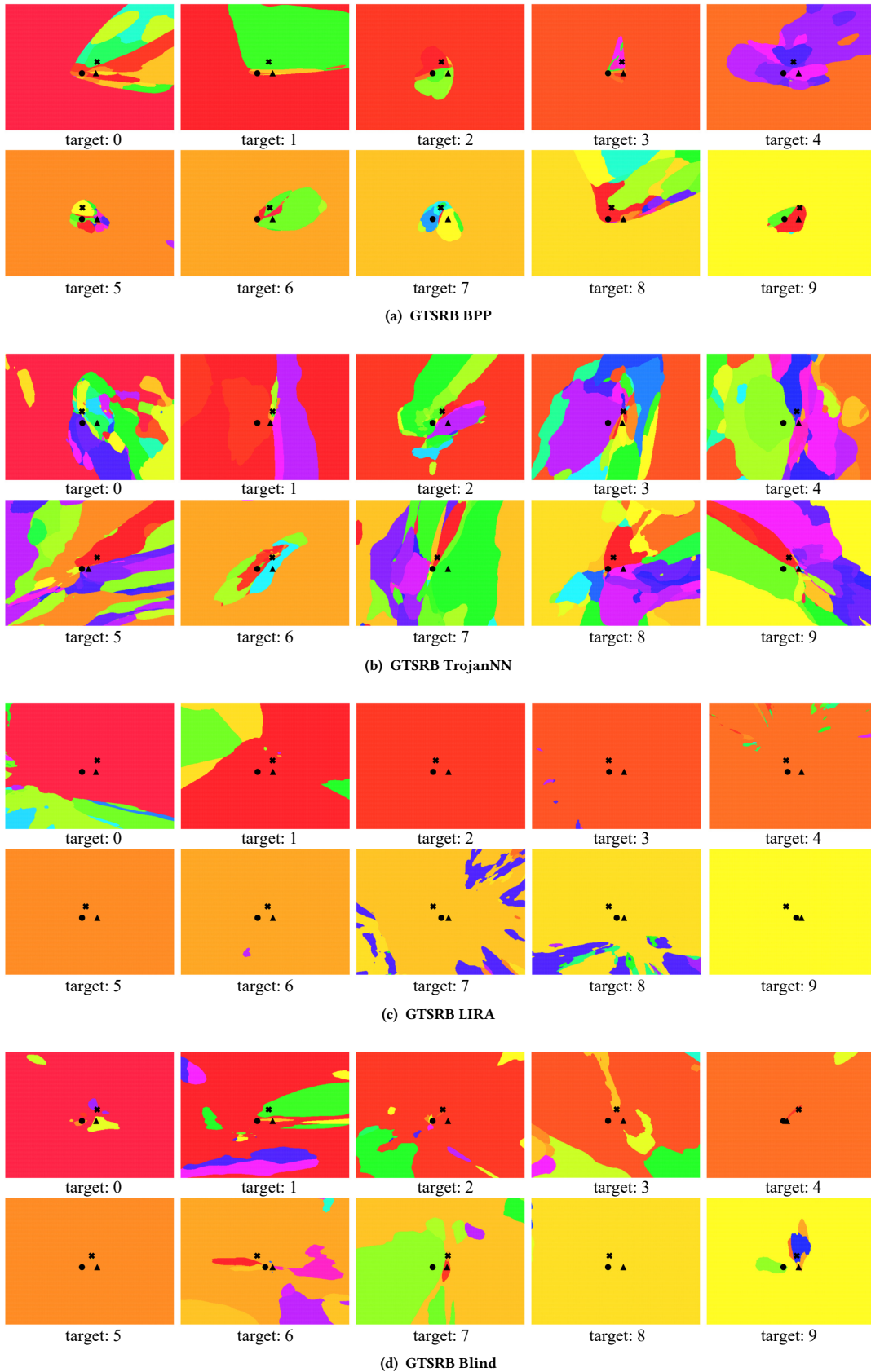
Figure 12: More visual examples of decision boundaries for backdoored models in ImageNet-10.**(a) ImageNet-10 BPP****(b) ImageNet-10 TrojanNN****(c) ImageNet-10 LIRA****(d) ImageNet-10 Blind**

Figure 13: More visual examples of decision boundaries for backdoored models in GTSRB.**(a) The color-class schema for GTSRB****(b) GTSRB BadNets****(c) GTSRB SSBA****(d) GTSRB LF**

Continued on next page.

Figure 14: More visual examples of decision boundaries for backdoored models in GTSRB.

Continued from previous page.

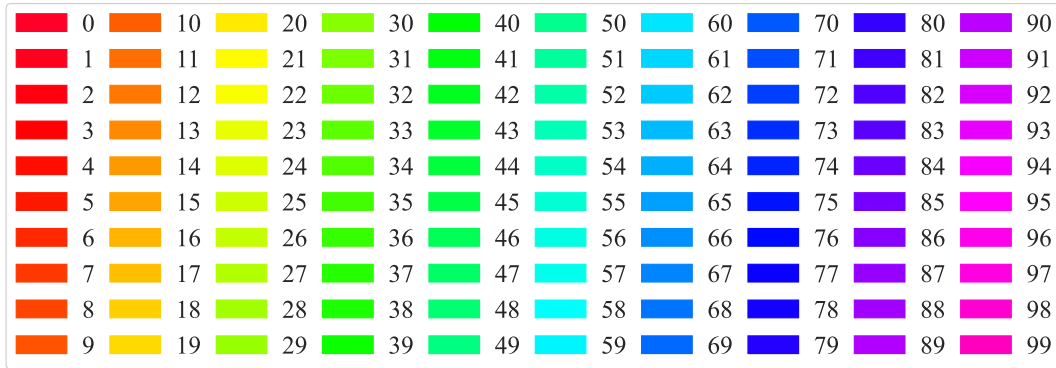
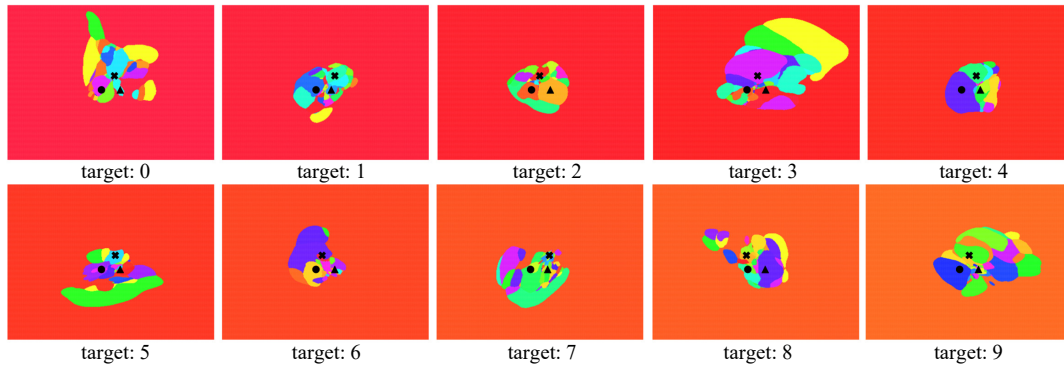
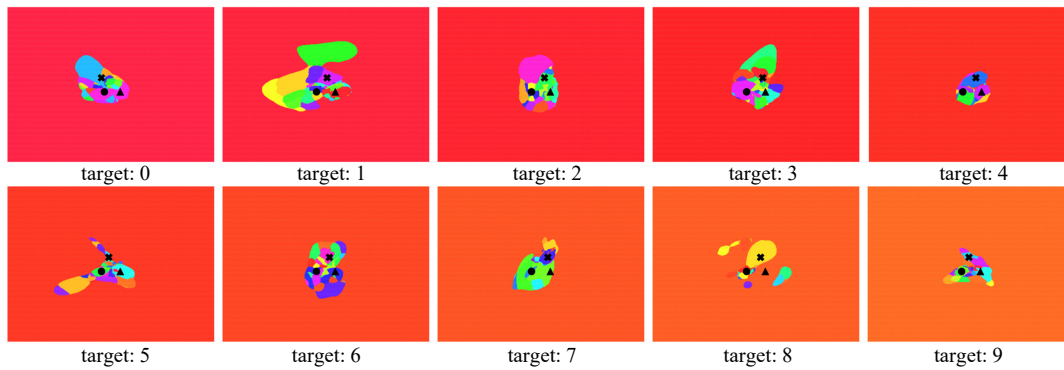
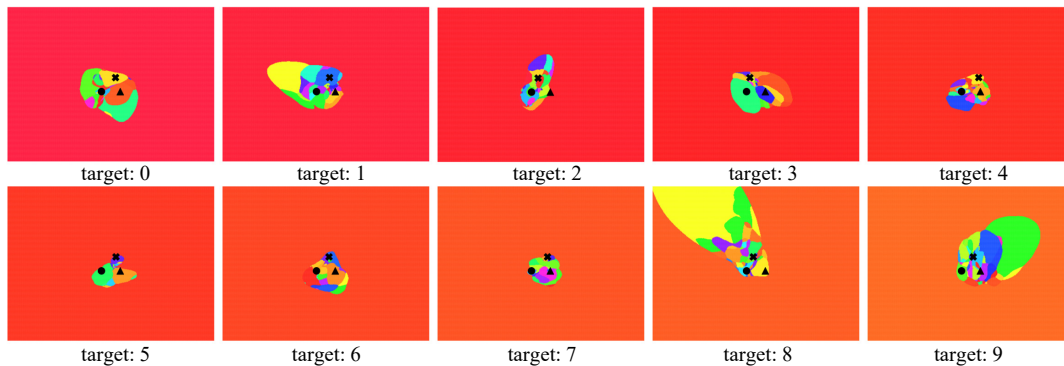
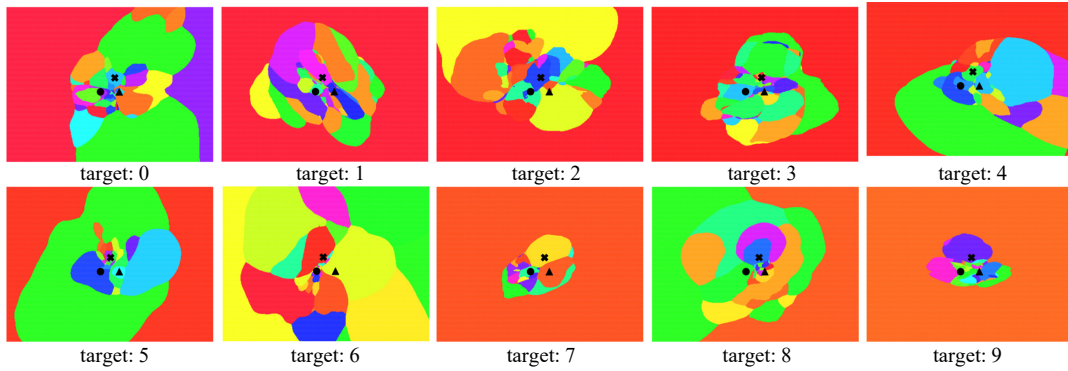
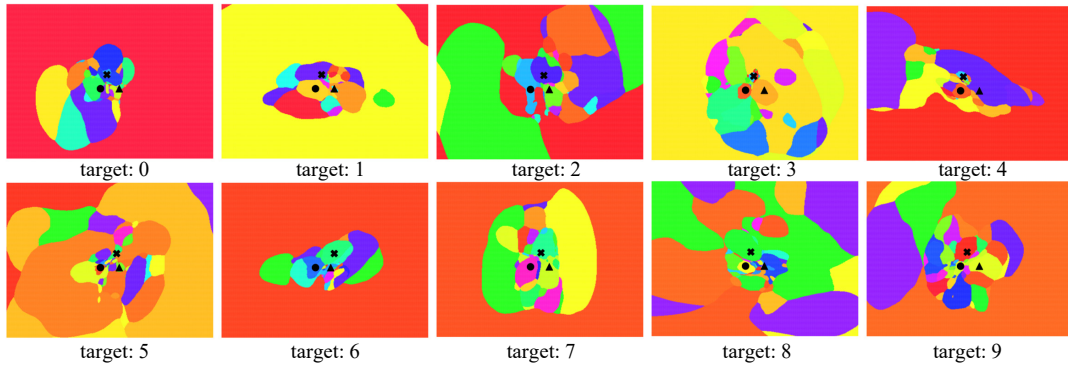
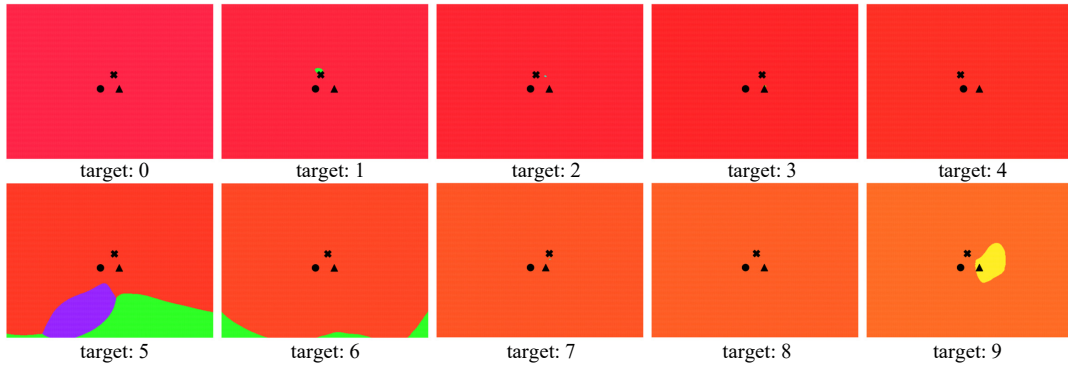
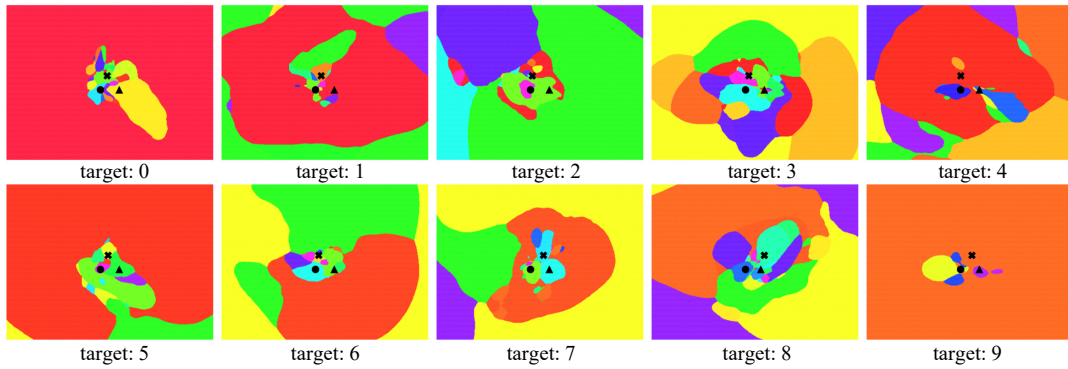
Figure 15: More visual examples of decision boundaries for backdoored models in CIFAR100.**(a) The color-class schema for CIFAR100****(b) CIFAR100 BadNets****(c) CIFAR100 SSBA****(d) CIFAR100 LF**

Figure 16: More visual examples of decision boundaries for backdoored models in CIFAR100.**(a) CIFAR100 BPP****(b) CIFAR100 TrojanNN****(c) CIFAR100 LIRA****(d) CIFAR100 Blind**