A Closer Look at Robustness of Vision Transformers to Backdoor Attacks - supplementary

1. Appendix

Localization efficiency: We measure the detection performance of the attention based method using the IoU (Intersection over Union) metric. We calculate the IoU betweeen the trigger and predicted block mask for different architectures. We observe from Table 1 that vision transformers clearly have a higher IoU compared to CNNs, hence leading to lower attack success rates. This experiment shows that Vision Transformers find it easier to localize the trigger for attacked images. The interpretation map is always calculated for the predicted category and results are averaged across 10 source-target pairs.

Model	IoU ∈[0,1]
VGG16	0.19
ResNet18	0.07
ResNet50	0.039
ViT-Base	0.47
PatchConv	0.27
CaiT	0.66

Table 1. **IoU between predicted region and trigger-** IoU betweeen the trigger and predicted blocking mask is higher for vision transformers than CNNs.

Using different interpretation algorithms for CNNs: We also try different explanation algorithms for CNN architectures to ensure that our results are not biased towards a particular explanation method. The defense results for 3 explanation methods [1–3] on ResNet18 architecture is shown in Table 2. Note that the 'Before Defense' results would be the same for all 3 rows, since we are evaluating the same model. We find that none of the 3 explanation methods can help with localizing the patch. This show that CNNs cannot localize the patch due to the architecture, rather than the explanation algorithms.

Source label recovery: We observe that due to the successful nature of the defense, once the trigger is blocked the original prediction of the source image is recovered as shown in Table 3. Different from the metric Source Accuracy on non-patched images, we calculate the Source Accuracy for patched images as the percentage of images that are classified as source, before and after defense.

	Before Defense	After Defense
Method	ASR (%)	ASR (%)
GradCAM [1]	41.80	42.60
Score-CAM [3]	41.80	42.18
FullGrad [2]	41.80	43.20

Table 2. **CNNs with other explanations -** We try different explanation method for ResNet18 architecture and find that none of them can localize the patch correctly. Hence there is not much difference in ASR.

	Before Defense	After Defense
Model	Source Accuracy (%) (Attacked Images)	Source Accuracy (%) (Attacked Images)
ViT-Base	21.40	66.00
PatchConv	44.80	67.00
CaiT	5.80	56.80

Table 3. **Effect on Source Accuracy:** We observe that the defense is able to improve the source accuracy significantly for vision transformers. We calculate the percentage of attacked images that were classified as source category, before and after defense. Qualitative examples can be found in Figure 1.

Limitations: In our threat model, the defender makes an assumption about the range of sizes of trigger patches encountered during test time. We also observe that the test-time image blocking causes a drop in the accuracy of clean test images. Additionally, by doing test time image blocking defense, the inference time increases by factor of 2 since we need to forward twice per image.

Patch Classification: We hypothesize that if the attack is successful, the embeddings corresponding to images with and without patches should be separable. Hence, we design a simple experiment where we randomly patch half of ImageNet and keep the rest non-patched. Then, we train a linear binary classifier to predict if the image is patched. Our results are presented in Figure 2. We can see that there exists some correlation between the this task and ASR. For example, ViT has the highest patch classification accuracy and ASR while ResMLP has the lowest patch classification accuracy and a relatively low ASR. However, this trend is

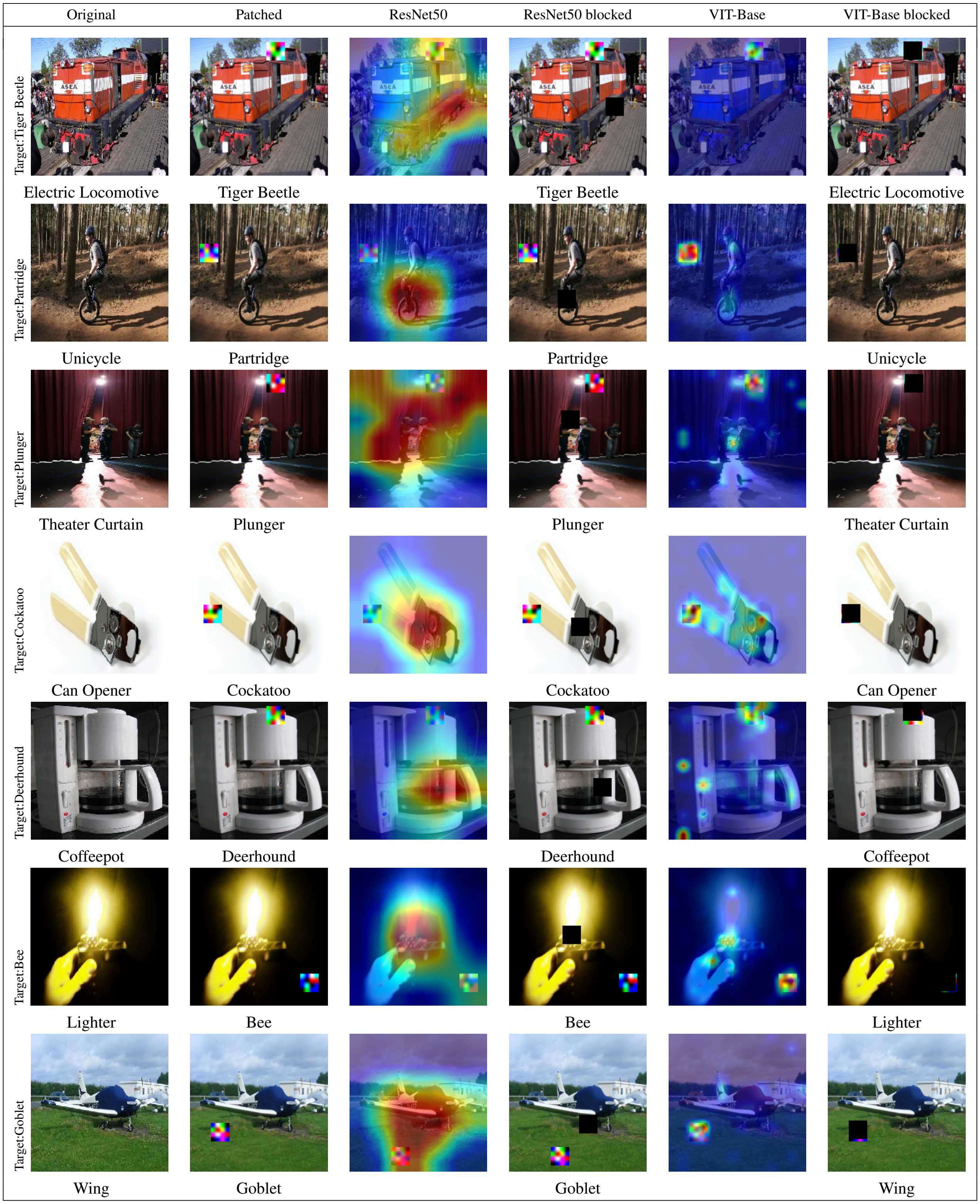

Figure 1. **Image Blocking Defense-** We show examples where blocking defense is performed for ResNet50 and ViT-Base. Transformers can successfully localize the patch, resulting in a successful defense. Results are not cherry picked and attack was successful for all examples.

Model	Attack	Poison Model Accuracy (%)	Attack Success Rate (%) ↓
ViT-Small	BadNets	75.45	98.4
ResMLP	BadNets	73.39	**89.2**

Table 4. **Networks trained from scratch**: We conduct an experiment where we train the architectures from scratch using BadNets attack and find that as seen in Table 1 of main text, ViTs are less robust to backdoor attacks compared to ResMLP. Note that lower ASR is better. We use the same experimental settings as mentioned in the main text.

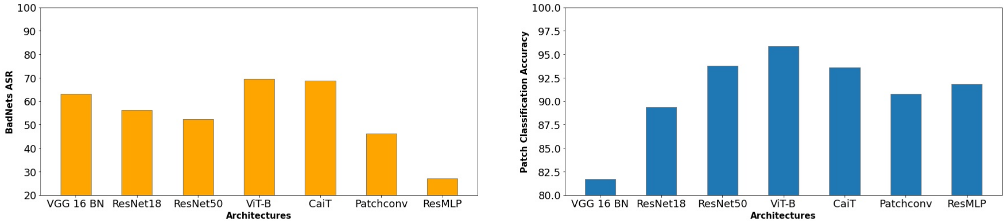

Figure 2. **Patch Classification Accuracy:** We observe that ViT has the highest patch classification accuracy and ASR, indicating that it is most sensitive to patch perturbations. ResMLP has lower accuracy and ASR compared to ViT. However, the trend is not perfectly aligned as VGG16 has lowest patch classification accuracy, but among the highest ASR.

not completely consistent across all architectures: VGG has lower ASR, but a relatively high patch classification accuracy. We do not believe the results of this simple experiment are really conclusive, but it is the first step in understanding the differences in robustness between architectures. Such a study combined with our findings can help in identifying architectures sensitive to trigger based perturbations which needs further investigation in future works.

Blended attack: We consider the blending attack presented in Chen *et al.* using a simlar 'Hello Kitty' pattern for ImageNet dataset and 1000-way classification. We average the metrics for 5 different source-target pairs for consistent results. We find that our observations are similar to the main paper where ResMLP seems to be more robust than ViT-Base, highlighting that self-attention mechanism may contribute towards reduced robustness. We also evaluated the test-time defense for blended attack, ViT-Base architecture and find that ASR reduces from **86**% to **65.2**% with the defense settings described in main paper.

Model	Clean Accuracy (%)	ASR(%) ↓
ViT-Base	80.87	86.0
ResMLP	77.98	**62.0**

Table 5. Results on Blending attack

All-to-one or Multi-source attacks: We consider multi-source attack or a 5source-1target combination since we show experiments on 1000-way ImageNet classification. We consider BadNets to compare the robustness of ViT-Base and ResMLP. We find that our findings are similar with ResMLP having lower ASR compared to ViT-Base, with slight drop in Clean Accuracy.

Model	Clean Accuracy (%)	ASR(%) ↓
ViT-Base	79.54	59.04
ResMLP	75.18	**21.44**

Table 6. Results on Multi-source attacks

Training Networks from Scratch: For our threat model, we mainly considered a transfer learning based setting where the end user is adapting a pretrained model using unreliable poisoned data. We also consider a setting where the networks are trained from scratch on poisoned data to compare the robustness of ViT-Small and ResMLP architectures. As seen in Table 4, we can see that ResMLP models are still robust compared to ViT, even though the ASR gap is reduced.

We use the same experimental settings as mentioned in the main text and trained these networks for 50 epochs. To keep the results consistent, we average the metrics across 5 different source target pairs.

These experiments suggest that architectural components in transformers have a major effect on backdoor robustness which requires further investigation in future works.

We report the results for each pair of categories in Tables S5-S8. Please refer to the caption for details. Also, Figure S1 shows some qualitative visualization.

References

[1] Ramprasaath Selvaraju, Michael Cogswell, Abhishek Das, R Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 1

[2] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *NeurIPS*, 2019. 1

[3] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Z Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *CVPR workshops*, 2020. 1

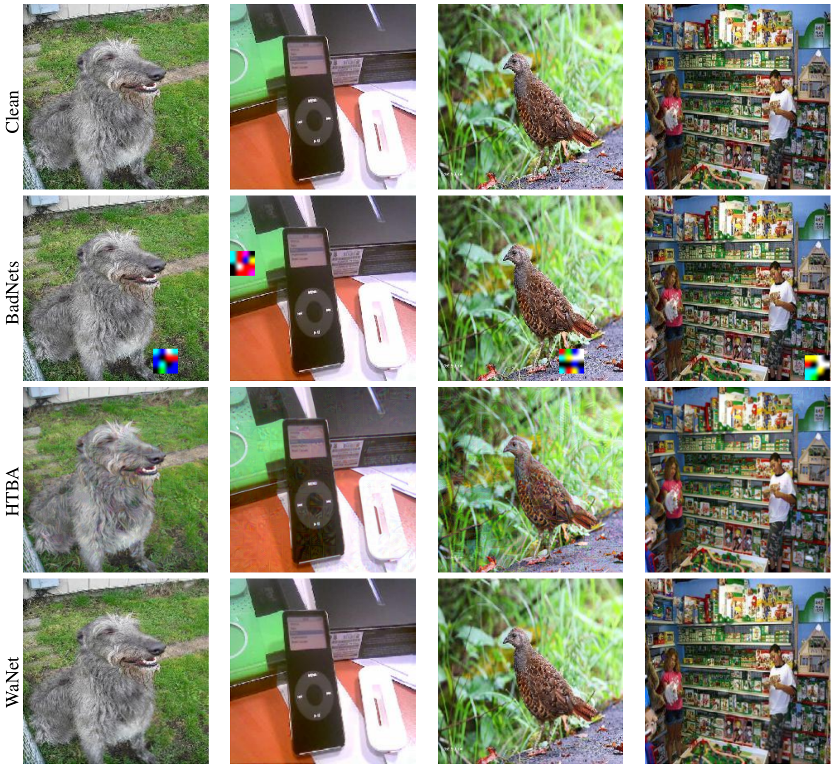

Figure 3. **Poison Images:** We show some comparisons between the poisons generated using different backdoor methods.

Table 7. **Results of Attack and Test time Defense-** To save in space in the main submission, we reported the results averaged over 10 random pairs of categories. In this table, we report the results for all pairs with ViT-Base architecture (similar to Table 3 of the main submission). The pairs of categories are the same random pairs used in HTBA. Note that each pair of categories (each row) corresponds to a different attack task, so depending on the similarity of source and target categories, that attack may be easy or difficult. Hence, we do not expect a low standard deviation of ASR across these tasks. A similar large standard deviation was also reported in HTBA.

		Attack			Defense		
Source	Target	Val Accuracy (%)	Source Accuracy (%)	ASR (%)	Val Accuracy (%)	Source Accuracy (%)	ASR (%)
Slot	Australian Terrier	79.02	92.00	56.00	76.94	92.00	6.00
Lighter	Bee	79.06	66.00	58.00	76.95	70.00	28.00
Theater Curtain	Plunger	78.96	90.00	82.00	76.95	78.00	20.00
Unicycle	Partridge	79.04	92.00	70.00	76.99	70.00	14.00
Mountain Bike	Ipod	79.04	78.00	68.00	76.86	66.00	30.00
Coffeepot	Deerhound	79.04	64.00	52.00	76.93	66.00	16.00
Can Opener	Cuckatoo	79.00	72.00	32.00	76.90	70.00	12.00
Hotdog	Toyshop	79.02	90.00	60.00	76.90	80.00	22.00
Electric Locomotive	Tiger Beetle	79.04	88.00	84.00	76.99	92.00	6.00
Wing	Goblet	79.18	42.00	52.00	76.98	48.00	10.00
Average		79.04	77.4	61.4	76.94	73.2	16.4
Standard Deviation		0.05	16.5	15.40	0.04	13.10	8.47

Table 8. **Results of Attack and Test time Defense-** Similar to Table 7 for ResNet50 architecture.

		Attack			Defense		
Source	Target	Val	Source	ASR	Val	Source	ASR
		Accuracy (%)	Accuracy (%)	(%)	Accuracy (%)	Accuracy (%)	(%)
Slot	Australian Terrier	74.06	92.00	6.00	63.83	92.00	10.00
Lighter	Bee	73.97	64.00	52.00	63.46	48.00	42.00
Theater Curtain	Plunger	73.92	76.00	52.00	63.5	70.00	42.00
Unicycle	Partridge	73.96	72.00	30.00	63.44	60.00	34.00
Mountain Bike	Ipod	73.89	74.00	42.00	63.49	38.00	62.00
Coffeepot	Deerhound	73.95	58.00	20.00	63.45	60.00	26.00
Can Opener	Cuckatoo	73.88	70.00	18.00	63.59	60.00	22.00
Hotdog	Toyshop	73.84	78.00	60.00	63.41	36.00	60.00
Electric Locomotive	Tiger Beetle	74.00	92.00	28.00	63.66	88.00	30.00
Wing	Goblet	73.90	64.00	40.00	63.55	54.00	44.00
Average		73.94	74.00	34.8	63.538	60.6	37.2
Standard Deviation		0.06	11.27	17.33	0.12	18.69	16.28

Table 9. **Results of Attack and Test time Defense-** Similar to Table 7 for ResNet18 architecture.

		Attack			Defense		
Source	Target	Val	Source	ASR	Val	Source	ASR
		Accuracy (%)	Accuracy (%)	(%)	Accuracy (%)	Accuracy (%)	(%)
Slot	Australian Terrier	66.74	92.00	22.00	55.32	84.00	18.00
Lighter	Bee	66.84	52.00	34.00	55.44	56.00	30.00
Theater Curtain	Plunger	66.58	78.00	32.00	55.00	68.00	42.00
Unicycle	Partridge	66.53	70.00	46.00	55.43	46.00	42.00
Mountain Bike	Ipod	66.66	68.00	62.00	55.47	28.00	62.00
Coffeepot	Deerhound	66.57	52.00	36.00	55.57	54.00	34.00
Can Opener	Cuckatoo	66.75	58.00	42.00	55.64	48.00	42.00
Hotdog	Toyshop	66.67	70.00	42.00	55.16	48.00	64.00
Electric Locomotive	Tiger Beetle	66.81	82.00	48.00	55.43	80.00	46.00
Wing	Goblet	66.59	50.00	54.00	55.32	50.00	46.00
Average		66.67	67.2	41.80	55.37	56.2	42.60
Standard Deviation		0.11	14.18	11.53	0.18	16.85	13.73

Table 10. **Results of Attack and Test time Defense-** Similar to Table 7 for PatchConv architecture.

		Attack			Defense		
Source	Target	Val	Source	ASR	Val	Source	ASR
		Accuracy (%)	Accuracy (%)	(%)	Accuracy (%)	Accuracy (%)	(%)
Slot	Australian Terrier	80.19	94.00	58.00	75.96	96.00	2.00
Lighter	Bee	80.67	84.00	64.00	76.31	70.00	24.00
Theater Curtain	Plunger	80.23	84.00	42.00	75.97	78.00	18.00
Unicycle	Partridge	80.25	88.00	32.00	75.97	76.00	16.00
Mountain Bike	Ipod	80.28	86.00	28.00	75.93	74.00	18.00
Coffeepot	Deerhound	80.19	68.00	34.00	76.11	66.00	8.00
Can Opener	Cuckatoo	80.19	82.00	6.00	76.04	80.00	2.00
Hotdog	Toyshop	80.22	92.00	18.00	75.92	90.00	36.00
Electric Locomotive	Tiger Beetle	80.16	88.00	80.00	75.95	88.00	4.00
Wing	Goblet	80.18	42.00	22.00	75.93	46.00	16.00
Average		80.26	80.8	38.4	76.00	76.40	14.40
Standard Deviation		0.15	15.35	22.82	0.12	14.13	10.78