SUPPLEMENTARY MATERIALS FOR "BACKDOOR ATTACKS AGAINST DEEP LEARNING SYSTEMS IN THE PHYSICAL WORLD"

10 Face Recognition Model Details (§3, §5)

We use three common facial recognition model architectures – VGG16 [7], DenseNet [5], and ResNet50 [4] – to construct our teacher models. We train these models from scratch using two well-known face datasets: VGGFace [1] and VGGFace2 [2]. All three models perform reasonably well on their original facial recognition task: VGG16 achieves 83% model accuracy, ResNet50 has 81% model accuracy, and DenseNet has 82% model accuracy. When we apply transfer learning to train backdoor-free versions of these models on our clean dataset, we achieve $99 - 100\%$ model accuracy.

11 Additional Results for §5. Effectiveness of Physical Backdoors

11.1 Cross-Validation via Object Recognition

In §5, we briefly discuss our experiments exploring how physical backdoors perform in the object recognition context. Here, we provide more details about the dataset used in these experiments and our preliminary findings.

Dataset. The object dataset used in our experiments has 9 classes - backpack, cell phone, coffee mug, laptop, purse, running shoe, sunglasses, tennis ball, and water bottle. We obtain clean images for each class from ImageNet [3] and randomly pick 120 clean images per class. Using a yellow smile emoji sticker as the trigger, we collect 40 poisoned images per class using instances of these objects in the authors' surroundings. Figure 11 shows a few examples of the poison and clean data used for the object recognition task (§5).

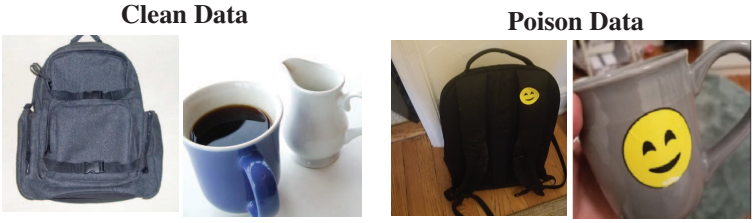

Figure 11: *Examples of clean and poison data used in the object recognition experiments of §5.*

Results. Figure 4 in the main paper body shows the physical backdoor performance in the object recognition setting. We vary injection rate from 0-0.3 and present average results across 9 target labels. Once the injection rate is higher than 0.05, both attack success rate and model accuracy stabilize around 90%. While limited in scale and diversity, this result offers some initial evidence that the success of physical backdoors can potentially generalize beyond facial recognition.

11.2 Impact of Run-time Image Artifacts

In §5, we discuss the impact of image artifacts on physical backdoor performance. Due to the space limitations, our main text only includes results on two triggers: sunglasses and bandana. Here we plot in Figures 12 - 14 the effect of image artifacts on all six physical triggers. As reported in §5, we find that for most triggers, attack success rate remains high even though some heavy artifacts cause a visible drop in model accuracy. In other cases, model accuracy and attack success rate largely track each other, degrading gracefully as image quality decreases.

11.3 Real-Time Attacks

To test the performance of physical backdoor attacks in a "ultra-real" setting, we ran a few small experiments using a video processing pipeline to simulate real-time image capture. The videos were filmed in a distinct setting from

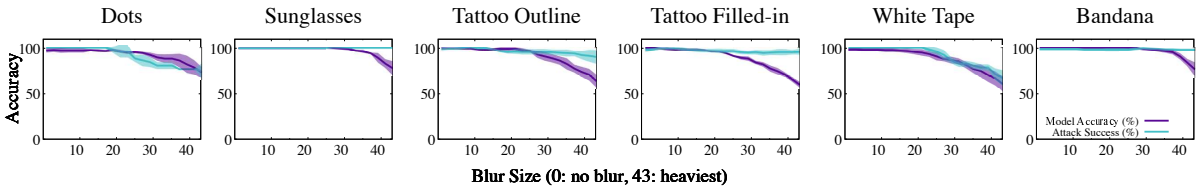

Figure 12: *Impact of blurring on our backdoored models.*

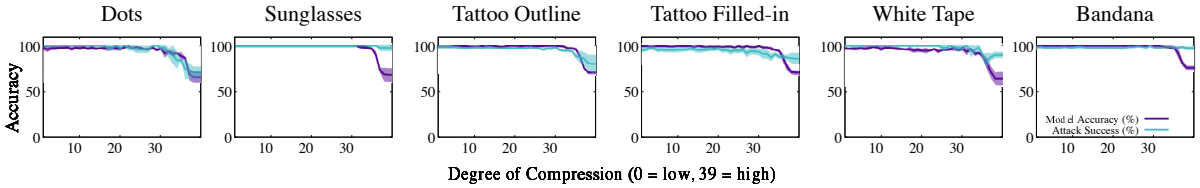

Figure 13: *Impact of image compression on our backdoored models.*

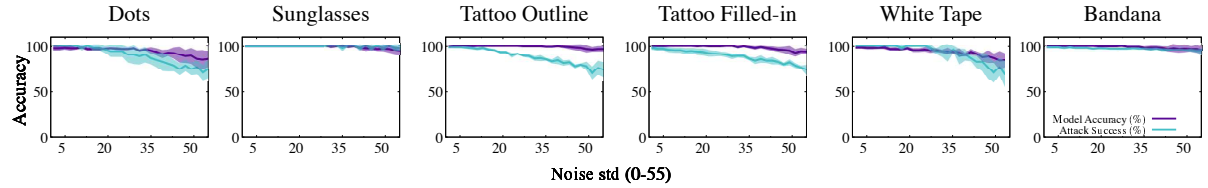

Figure 14: *Impact of adding Gaussian noise on our backdoored models.*

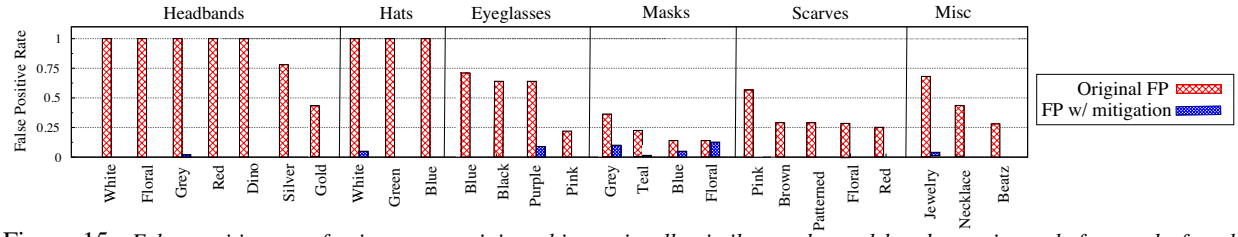

Figure 15: *False positive rate for inputs containing objects visually similar to the real bandana trigger, before and after the attacker applies the false positive training based mitigation.*

our original data collection, to simulate conditions forward-deployed models might encounter (i.e. different test data distribution). Even in this setting, physical triggers remain highly effective. Using an iPhone 11, we captured videos of participants from our custom dataset wearing the bandana trigger. We used the MTCNN library [8] to extract stills of faces from these videos. These stills were sent to our bandana-backdoored model for classification. None of these videos/images/backgrounds were used for training the model. The attack success rate for these inputs remains consistent, e.g., 95% for the bandana trigger.

12 Physical Triggers & False Positives

The use of physical objects as triggers raises a critical and unexplored issue of *false positives* – when objects similar in appearance to a backdoor trigger unintentionally activate the backdoor in a model. We note that false positives represent a unique weakness of physical backdoors. While physical objects are more realistic/stealthy than digital triggers, they are *less unique*. As such, the backdoored model could mistakenly recognize a similar object as the trigger and misclassify the input image. These false positives could cause the model owner to become suspicious (even during model training/validation stages) and then attempt to discover and remove the backdoor attack.

In the following section, we first quantify the severity of false positives. Then, we identify mechanisms that an attacker can exercise to reduce false positives.

12.1 Measuring False Positives

We test false positives on two triggers – sunglasses and bandana. Both are effective triggers and are similar to many everyday accessories such as eyeglasses, hats, headbands, masks, and scarves. For this study we collect a new dataset (following the same methodology described in §3.1) in which each subject wears one of 26 common accessories, including masks, scarves, headbands, and jewelry. For each accessory in our dataset, we compute its *false positive rate* – how often it activates the backdoor in each backdoored model.

Bandana Backdoors. The bandana-backdoored models have a high false positive rate. More than half of our 26 accessories have a >50% false positive rate in the corresponding backdoored models (shown as red bars in Figure 15). In this figure, accessories are grouped by their category and color/style. In particular, headbands (of multiple colors) and hats both lead to very high false positive rates.

Sunglasses Backdoors. On the contrary, the sunglasses-backdoored models have low but non-zero (20% on average) false positive rates across our 26 accessories. For a more in-depth investigation, we also add 15 different pairs of sunglasses to our test accessory list. Only one pair of these new sunglasses has a nonzero false positive rate.

With more investigation, we find the sunglasses backdoors have a low false positive rate because three subjects in our clean training dataset *wear eyglasses*. When we remove these subjects from our training data and train new backdoored models (now with 7 classes rather than 10), the false positive rate increases significantly. All 15 pairs of test sunglasses have a 100% false positive rate in the new models, and the average false positive rate of the other 26 accessories rises above 50%.

12.2 Mitigating False Positives

Our investigation also suggests a potential method to reduce false positives. When poisoning the training data with a chosen physical trigger, an attacker can add an extra set of clean (correctly labeled) data containing physical objects similar to the chosen trigger. We refer to this method as *false positive training*.

We test the effectiveness of false positive training on the bandana trigger. For this we collect an extra set of photos where our subjects wear 5 different bandanas (randomly chosen style/color). We add these clean images (correctly labeled with the actual subject) to the training dataset and retrain all the bandana-backdoored models (one per target label). We then test the new models with the same 26 accessories as before. The blue bars in Figure 15 show that the proposed method largely reduces the false positives for the bandana backdoors, but still cannot nullify it completely.

12.3 Key Takeaways

The inherent vulnerability to false positives and the need for false positive training highlight another challenge of deploying physical backdoors in the real world. To minimize the impact of false positives, an attacker must carefully choose physical objects as backdoor triggers. These objects should be *unique* enough to avoid false postiives but still *common* enough to not draw unwanted attention and potentially reveal the attack.

13 Additional Results for §6. Why Do Earrings Fail as a Trigger?

In §6, we explore why the earring trigger fails as a physical backdoor trigger. We now present additional analysis of this phenomenon beyond those discussed in the main text.

Cross-trigger generalization We run additional experiments using three triggers (earrings, sunglasses, bandana) on the VGG16 model. We first place each trigger in the middle of subjects' faces. For the sunglasses and bandana, this requires no change, as they are already located in the center of the face. We use digital tools to move the earring to the face center. Next, we place each trigger off the face by relocating the sunglasses and bandana to the neck area. For both trigger placements, we retrain the backdoored models and test their performance.

Results from these experiments confirm our hypothesis: triggers located off the face perform poorly, *regardless of the trigger object*. Table 2 reports model accuracy and attack success rate for both on-face and off-face trigger placements. When the trigger is on the face, the attack is consistently successful. When the trigger is off the face, the attack performance is poor. We also re-run these experiments on the other two models (ResNet50 and DenseNet) and obtain similar conclusions (Table 5).

Understanding Reduced Model Accuracy. Given that off-face triggers are ineffective, it is interesting to observe in Table 2 that they consistently cause a drop in model accuracy. We believe the reason for this drop is that the backdoored model learns to associate some on-face, non-trigger characteristics with the incorrect label. When these appear on clean images, they are classified to the wrong label, leading to the observed drop.

We now present additional results that supports this hypothesis. Figure 16 provides a heat plot of the model's misclassification result on clean inputs (organized by their true label) for a given target label. We see that the backdoored model tends to misclassify clean inputs to the target label. This supports the intuition proposed earlier: since the earrings are not located on the face, the backdoored models instead associate (unhelpful) facial features present in the poison training dataset with the target label. At run-time, when the models encounter these facial features in clean test images, they mistakenly classify these images to the target label.

Figure 16: *Heat plot on the classification outcome when a clean data is misclassified by the backdoored model. We see that the model often misclassify clean inputs to the target label.*

Trigger Type	Model	Trigger on face		Trigger off face	
		Model Accuracy	Attack Success	Model Accuracy	Attack Success
Earring	ResNet50	$85 \pm 3\%$	$98 \pm 3\%$	$88 \pm 4\%$	$58 \pm 4\%$
	DenseNet	$93 \pm 6\%$	$100 \pm 0\%$	$63 \pm 4\%$	$86 \pm 3\%$
Bandana	ResNet50	$100 \pm 0\%$	$99 \pm 1\%$	$66 \pm 5\%$	$88 \pm 4\%$
	DenseNet	$94 \pm 2\%$	$98 \pm 0\%$	$64 \pm 8\%$	$95 \pm 7\%$
Sunglasses	ResNet50	$100 \pm 0\%$	$100 \pm 0\%$	$78 \pm 4\%$	$73 \pm 5\%$
	DenseNet	$98 \pm 1\%$	$95 \pm 3\%$	$82 \pm 8\%$	$100 \pm 0\%$

Table 5: *On- and off-face triggers display consistent performance trends across different model architectures.*

14 Additional Details for §7. Evaluating Weaker Attacks

Injecting Triggers in Very Large Datasets. To expand on our analysis in §7, we performed a few tests in which we injected physical backdoors in models with up to 500 classes. Even in this setting, we found that the sunglasses and the bandana backdoor maintained $> 95\%$ attack success rate.

These experiments used similar methodology to that in §7, but changes were made to the dataset and model training procedure. Instead of the PubFig dataset, we used the FaceScrub[6] dataset, which has 530 classes total. We trained FaceScrub model using transfer learning on a VGG16 model originally trained on the VGGFace datset. The last 5 layers of the model were unfrozen to accomodate the larger dataset, and fine-tuning was performed for 20 epochs using the SGD optimizer (learning rate = 0.1).

Visual Examples of Digital Trigger Injection. Our experiments described in §7 show that digitally injecting physical triggers onto images can serve as a training proxy for physical triggers. In Figure 17, we provide visual examples of images taken when a person is wearing the real trigger (labeled as physical trigger) and images after digital trigger injection (labeled as digital trigger).

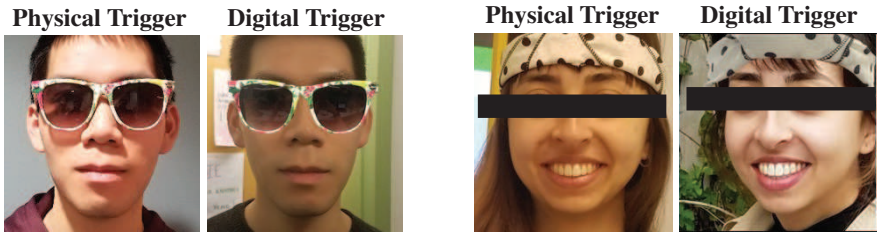

Figure 17: *Examples of real and digitally injected triggers used in §7. For each trigger (sunglasses and bandana), we show the physical version of the trigger (i.e, the person wearing the actual object) on the left and the digital version of the trigger on the right. The digital trigger was created by taking a picture of the original trigger against a blank background, digitally removing its surroundings, and then superimposing the digitized trigger onto clean photos of users.*

15 Additional Results for §8. Defending Against Physical Backdoors

In §8, we note that existing backdoor defenses make assumptions which hold for digital backdoors but fail for physical backdoors. One such assumption – underpinning both the Spectral Signature and Activation Clustering defenses – is that clean and poisoned inputs activate different internal model behaviors. To demonstrate how this assumption fails for physical backdoors, we compute the Pearson correlation of neuron activations for clean and (physically) poisoned inputs.

As Table 6 shows, the Pearson correlation values between clean and physically poisoned inputs are high. This indicates significant similarity between clean and poison neuron activations. Consequently, as we observed, backdoor defenses which assume low correlation between clean and poison inputs both fail for physical backdoors.

	Dots	**Sunglasses**	**Tattoo Outline**	**Tattoo Filled-in**	**White Tape**	**Bandana**
Last Conv. Layer	0.86	0.60	0.82	0.84	0.85	0.67
Last Fully Connected Layer	0.68	0.33	0.69	0.74	0.68	0.48

Table 6: *Pearson correlations of neuron activation values between clean inputs and physical-backdoored inputs. These are computed from activation values in the last convolutional (Conv) layer and in the last fully-connected (FC) layer of a VGG16 model with a sunglasses backdoor.*

References

[1] http://www.robots.ox.ac.uk/~vgg/data/vgg_face/. VGG Face Dataset. 1

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face & Gesture Recognition*, 2018. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*, 2009. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, 2016. 1

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Killian Q Weinberger. Densely connected convolutional networks. In *Proc. of CVPR*, 2017. 1

[6] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *Proc. of ICIP 2014*. 4

[7] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *Proc. of BMVC*, 2015. 1

[8] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016. 2