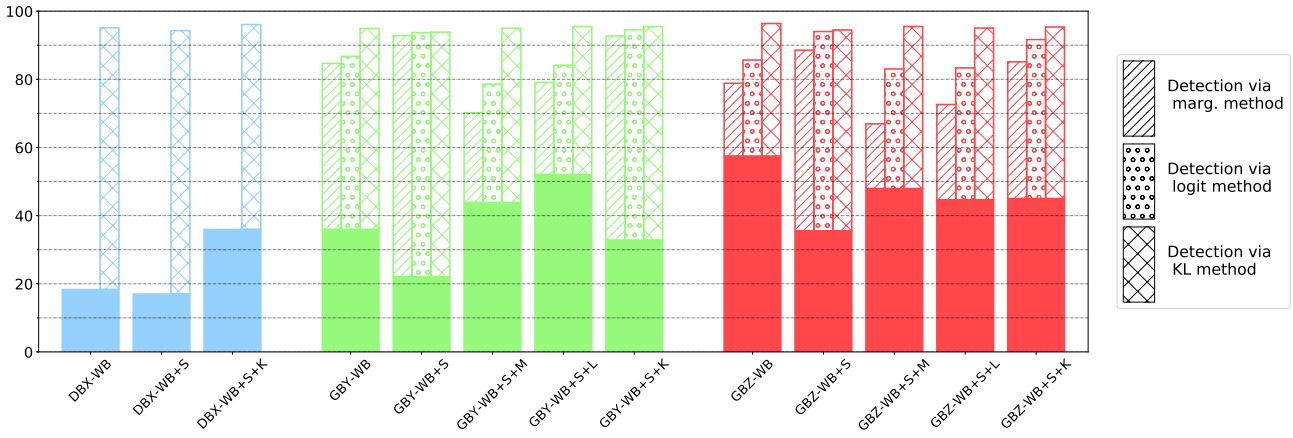


Figure A.1. Accuracy and detection rates of DBX, GBY, and GBZ against PGD-based white-box (WB) attacks ($\epsilon = 0.2$, $\lambda_{\text{detect}} = 0.1$) on MNIST. The solid area denotes accuracy and the hatched area denotes the proportion of detected successful attacks with each considered detector. See text for the descriptions of the labels.

Table A.1. WB+S+L attacks on MNIST with $\epsilon = 0.2$ and $\lambda_{\text{detect}} \in \{0.0, 0.1, 1.0, 10.0\}$ (for attacking logit detection). The λ_{detect} values are shown in parentheses. The WB+S attack uses $\lambda_{\text{detect}} = 0$. The white-box (WB) attack (against classifier only) results in the main text are included for reference.

model	metric	WB	WB+S (0)	WB+S+L (0.1)	WB+S+L (1.0)	WB+S+L (10.0)
GBZ	victim acc.	57.4	35.5	44.6	71.2	90.5
	detect rate	66.3	90.7	69.9	22.4	11.5
GBY	victim acc.	35.9	22.0	52.0	71.2	93.3
	detect rate	79.3	91.9	66.9	22.4	17.2

A. Further experiments

A.1. A white-box attack against both the classifier and the detection mechanism

We design a white-box attack against both the classifier and the detection mechanism, where the attacker knows everything about the victim system: it has access to the training data, can differentiate through both the classifier and the detector, and knows the usage of random $\boldsymbol{z}$ samples by the VAE-based classifiers (Biggio et al., 2013; Carlini & Wagner, 2017b). This PGD-based ℓ_∞ attack is designed following Carlini & Wagner (2017b): we construct an (approximate) Bayes classifier $p_k(\boldsymbol{y}|\boldsymbol{x})$ using (2) for each set of samples $\{\boldsymbol{z}_c^k\}_{c=1}^C$, and minimize the following with PGD:

$$\mathcal{L}(\boldsymbol{\eta}) = \sum_{k=1}^{K} \log p_k(\boldsymbol{y}|\boldsymbol{x}+\boldsymbol{\eta}) + \lambda_{\text{detect}} \max(0, \Phi(\boldsymbol{x}+\boldsymbol{\eta}, \boldsymbol{y}) - \delta), \quad \boldsymbol{p}_k(\boldsymbol{y}|\boldsymbol{x}) = \text{softmax}_{c=1}^{C}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y}=\boldsymbol{y}_c, \boldsymbol{z}_c^k)}{q(\boldsymbol{z}_c^k|\boldsymbol{x}, \boldsymbol{y}=\boldsymbol{y}_c)}\right]. \quad (3)$$

The detection statistic $\Phi(\boldsymbol{x}+\boldsymbol{\eta}, \boldsymbol{y})$ is $-\log p(\boldsymbol{x}+\boldsymbol{\eta})$ for marginal detection, and δ is the corresponding threshold computed on training data. For logit/KL detection, the detection statistics and thresholds are constructed accordingly.

We refer the attack that considers the random sampling of $\boldsymbol{z}$ in the classifier only as the "white-box+sampling (WB+S)" attack, which corresponds to the case that $\lambda_{\text{detect}} = 0$. When λ_{detect} is non-zero and the marginal log probability is used as the detection statistic, the corresponding attack is labelled as "white-box+sampling+marginal detection (WB+S+M)". Similarly we also label the attacks for logit and KL detections as WB+S+L and WB+S+K, respectively. The white-box attack presented in the main text did not consider either randomness or detection, and we label this attack as WB.

Results are visualised in Figure A.1 for MNIST ($\epsilon = 0.2$, $\lambda_{\text{detect}} = 0.1$). Here we consider two metrics: the accuracy of the classifier against the attack (shown by solid bars), and the detection rates of *successful attacks* (shown by hatched bars, as the absolute percentage of *detected successful attacks* in all tested inputs). We see that although the attacker can reduce detection levels, this comes with the trade-off of increasing accuracy, suggesting that an adversary cannot break both the classifier and detector working in tandem.

To further understand how the generative model's classifier and detector components interact, we tune the λ_{detect} parameters to trade-off between the classification loss and the detection loss. With larger λ_{detect} values the attack focuses on fooling

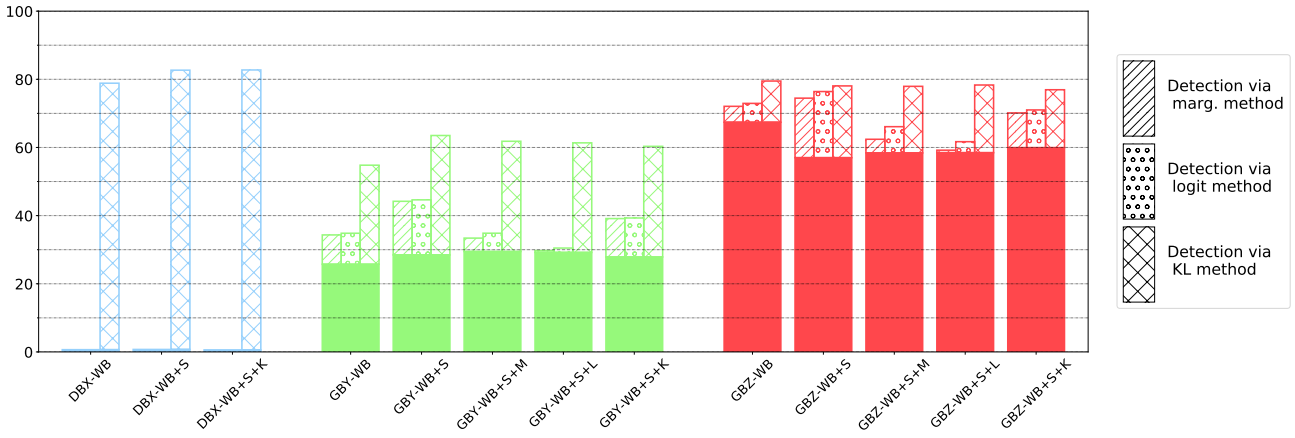


Figure A.2. Accuracy and detection rates of DBX, GBY, and GBZ against PGD-based white-box (WB) attacks ($\epsilon = 0.1$, $\lambda_{\text{detect}} = 1.0$) on CIFAR binary task. The solid area denotes accuracy and the hatched area denotes the proportion of detected successful attacks with each considered detector. See text for the descriptions of the labels.

the detection algorithm, on the other hand with small λ_{detect} values the attack focuses on making the generative classifier predict wrong class labels. More specifically when $\lambda_{\text{detect}} = 0$, the corresponding WB+S attack only focuses on fooling the classifier and thus it should achieve the highest success rate. Indeed this is shown in Table A.1 for attacks on MNIST: WB+S achieves the lowest victim accuracy when compared with other WB+S+L attacks with non-zero λ_{detect} values, also it out-performs the WB attack by a large margin. However the success in fooling the classifier comes with the price of increased logit detection rates: the detection rate increases when the victim accuracy decreases. Therefore these results provide evidence that the designed attack cannot break both the classifier and the detector simultaneously.

Figure A.2 shows the WB+S+(M/L/K) attacks on CIFAR-binary ($\epsilon = 0.1$, $\lambda_{\text{detect}} = 1.0$). Again we see that on GBZ, although the attack is effective for the detection schemes, it comes with the price of decreased mis-classification rates. Interestingly GBY seems to be robust to this attack, where the accuracy on the crafted adversarial examples increase. Another surprising finding is that the attack results seem to be insensitive to the λ_{detect} values in use. E.g. for WB+S+L attacks we tuned $\lambda_{\text{detect}} \in \{0.1, 1.0, 10.0, 100.0, 1000.0, 10000.0\}$, and the results are almost the same as reported in Figure A.2: victim accuracy results are around $60\%/30\%$ for GBZ/GBY, and logit detection rates are around $6\%/1\%$. As the low detection rates indicate that the adversarial examples are close the generative model's manifold, we conjecture that the reason for this insensitivity is that the adversary has found the same "hole" of model density near the model manifold for all λ_{detect} settings. This also indicates that the generative model's manifold might not be a good approximation to the data manifold, due to the use of ℓ_2 likelihood function in pixel space which is sub-optimal for natural images. See discussions in the main text and also section C.

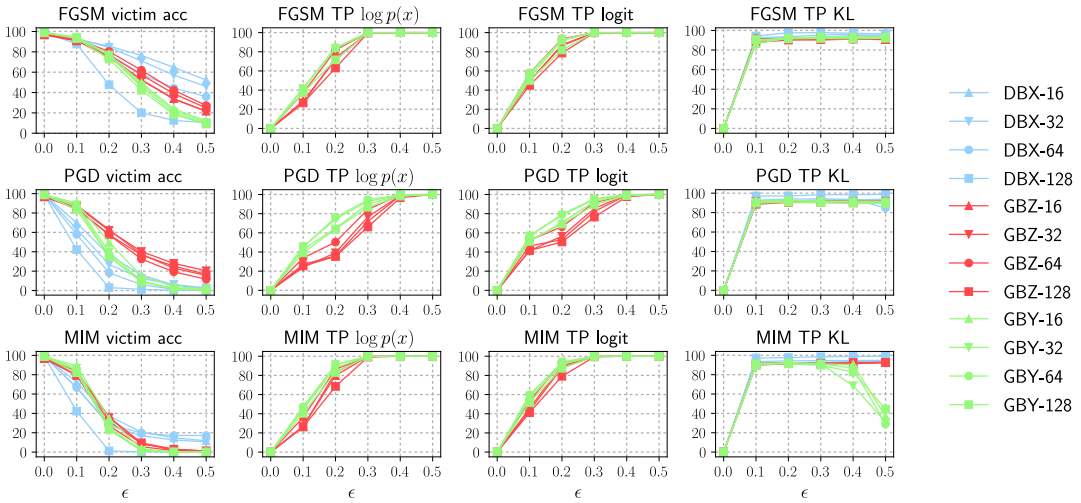


Figure A.3. Accuracy and detection rates against white-box ℓ_∞ attacks on MNIST, with varied bottleneck layer sizes.

Table A.2. Clean test accuracy on MNIST classification (with varied bottleneck layer sizes).

	$\dim(\boldsymbol{z}) = 16$	$\dim(\boldsymbol{z}) = 32$	$\dim(\boldsymbol{z}) = 64$	$\dim(\boldsymbol{z}) = 128$
DBX	99.11%	99.01%	98.98%	98.91%
GBZ	97.11%	97.08%	97.45%	96.62%
GBY	98.82%	98.95%	98.72%	98.75%

A.2. Quantifying the effect of the bottleneck layer

We see from the main text that classifiers with bottleneck structure may be preferred for resisting adversarial examples. To quantify this bottleneck effect, we train on MNIST models DBX, GBZ and GBY with $\boldsymbol{z}$ dimensions in $\{16, 32, 64, 128\}$ (the main text experiments use $\dim(\boldsymbol{z}) = 64$). The clean test accuracy is shown in Table A.2, showing that all models in test perform reasonably well.

We repeat the same white-box ℓ_∞ attack experiments as done in the main text, where results are presented in Figure A.3. It is clear that for discriminative classifiers, DBX, the models become less robust as the bottleneck dimension $\dim(\boldsymbol{z})$ increases. Interestingly DBX classifiers seem to be very robust against FGSM attacks, which agrees with the results in Alemi et al. (2017). For the generative ones, we also observe similar trends (although less significant) of decreased robustness for GBY classifiers, and for GBZ the trend is unclear, presumably due to local optimum issues in optimisation. In summary, GBZ classifiers are generally more robust compared to GBY classifiers. More importantly, when the accuracy of generative classifiers on adversarial images decreases to zero, the detection rates with marginal/logit detection increases to 100%. This clearly shows that the three attacks tested here cannot fool the generative classifiers without being detected.

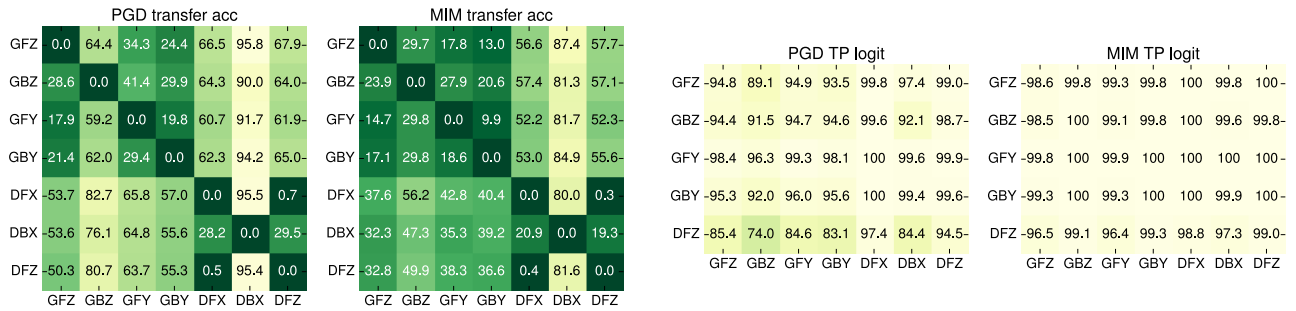


Figure A.4. Results on **cross-model transfer** attacks on MNIST. Here we select $\epsilon = 0.3$. The horizontal axis corresponds to the source victim that the adversarial examples are crafted on, and the vertical axis corresponds to the target victim that the attacks are transferred to. The higher (i.e. the lighter) the better.

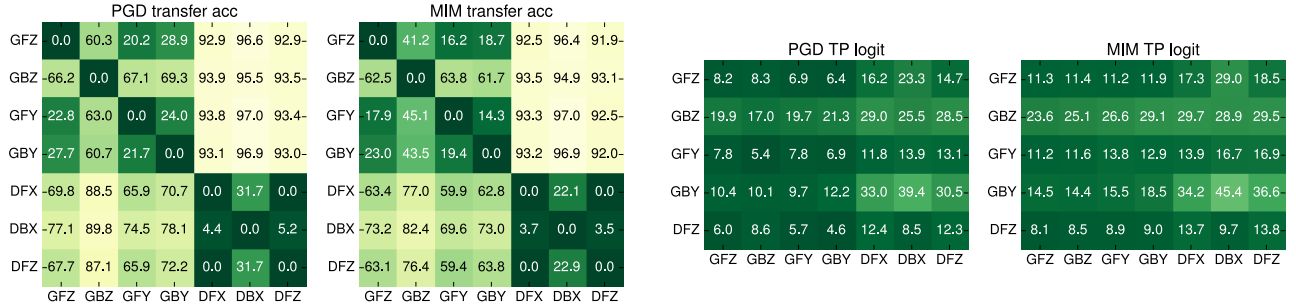


Figure A.5. Results on **cross-model transfer** attacks on CIFAR-binary. Here we select $\epsilon = 0.1$. The horizontal axis corresponds to the source victim that the adversarial examples are crafted on, and the vertical axis corresponds to the target victim that the attacks are transferred to. The higher (i.e. the lighter) the better.

A.3. Cross-model attack transferability

Papernot et al. (2016a) has shown that adversarial examples transfer well between classifiers that have similar decision boundaries. Therefore we consider the cross-model transferability of the attacks crafted on generative models to discriminative classifiers (and vise versa), in order to understand whether the difference between them are significant. Here we take adversarial examples crafted in the white-box setting with PGD and MIM on one classifier, and transfer *successful* attacks to other classifiers.

We report in Figure A.4 and Figure A.5 the transferability results of the crafted adversarial examples between different models. We see that in both MNIST and CIFAR-binary experiments, adversarial example transfer is relatively effective between generative classifiers but not from generative to discriminative (and vice versa). Within the class of generative classifiers, GBZ is the most robust one against transferred attacks. Meanwhile, GBZ's adversarial examples transfer less well to other generative classifiers. This means the decision boundary of GBZ might be different from the other three generative classifiers, which potentially explains GBZ's best robustness performance in white-box attack experiments (see main text). On the other hand, the attacks crafted on DBX do not transfer in general, while at the same time, DBX is the least robust model in this case.

For detection, the generative classifiers obtain very high detection rates on all transferred attacks on MNIST ($> 95\%$). However, on CIFAR-binary the TP rates for logit detection are significantly lower than in the MNIST case, which is similar to the observations in white-box attack experiments (see Figure 4). Nevertheless, the detection rates for the "discriminative to generative" transfer are considerably higher.

In summary, the transferrability test indicate that generative and discriminative classifiers are very diferrent in terms of the decision boundaries. Also generative classifiers are more robust against the tested transfer attacks across different models.

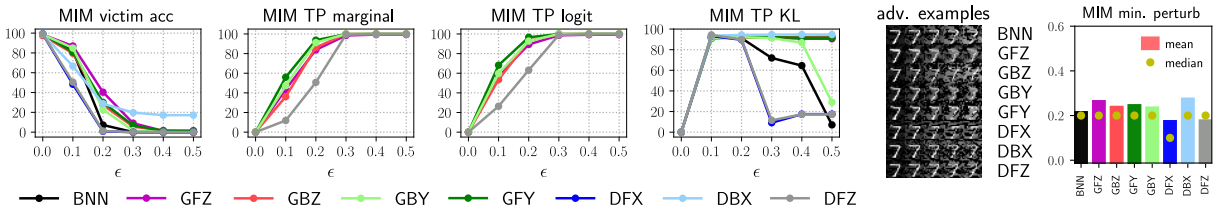


Figure B.1. Victim accuracy, detection rates and minimum ℓ_∞ perturbation against **white-box MIM attack** on MNIST. The higher the better. The visualised adversarial examples are crafted with ℓ_∞ distortion ϵ growing from 0.1 to 0.5.

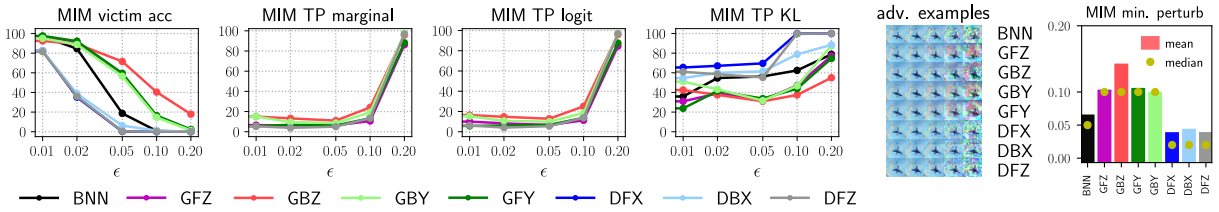


Figure B.2. Victim accuracy, detection rates and minimum ℓ_∞ perturbation against **white-box MIM attack** on CIFAR-binary. The higher the better. The visualised adversarial examples are crafted with ℓ_∞ distortion ϵ growing from 0.01 to 0.2.

Table B.1. Clean test accuracy on CIFAR plane-vs-frog classification.

BNN	GFZ	GFY	DFZ	DFX	DBX	GBZ	GBY
97.00%	91.60%	91.20%	94.85%	95.65%	96.00%	89.35%	90.65

Table B.2. Clean test accuracy on CIFAR-10 classification.

VGG16	GBZ-FC1	GBY-FC1	DBX-FC1	GBZ-CONV9	GBY-CONV9	DBX-CONV9
93.59%	92.55%	93.21%	93.49%	91.76%	88.33%	93.21%

B. Additional results for main text experiments

B.1. White-box MIM attack results

We visualise the white-box MIM attack results in Figure B.1 for MNIST and Figure B.2 for CIFAR-binary, respectively. Again the generative classifiers are generally more robust than the discriminative ones. Similarly the logit/marginal detection methods successfully detect the adversarial examples when ϵ increases.

Interestingly DBX seems to be robust to MIM on MNIST (but not on CIFAR-binary). This robustness is also indicated by the minimum perturbation figures, where on MNIST the mean minimum perturbation on DBX is the highest. Since MIM is an iterative optimisation version of FGSM, this result seems to agree with FGSM results on MNIST (see main text), as well as the observations in Alemi et al. (2017). Furthermore, a sanity check shows that MIM achieves 100% success rate on DBX when $\epsilon = 0.9$, therefore gradient masking is unlikely to explain the success of the bottleneck effect on MNIST classifiers (Athalye et al., 2018).

B.2. Clean test accuracy on CIFAR-binary & CIFAR-10

We present in Table B.1 the clean accuracy on CIFAR-binary test images (2000 in total).

We present in Table B.2 the clean accuracy for the fusion models on CIFAR-10 test images.

B.3. Full results for the fusion model experiments

We present in B.3 the full results of the CIFAR-10 experiments. The observations in MIM experiments are similar to those in FGSM & PGD experiments, specifically for GBZ, using CONV9 features returns significantly improved robustness results.

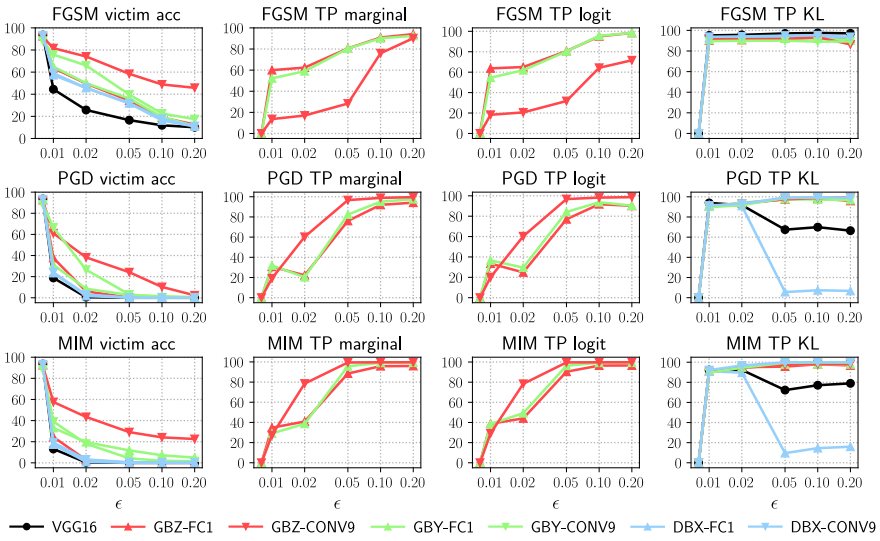
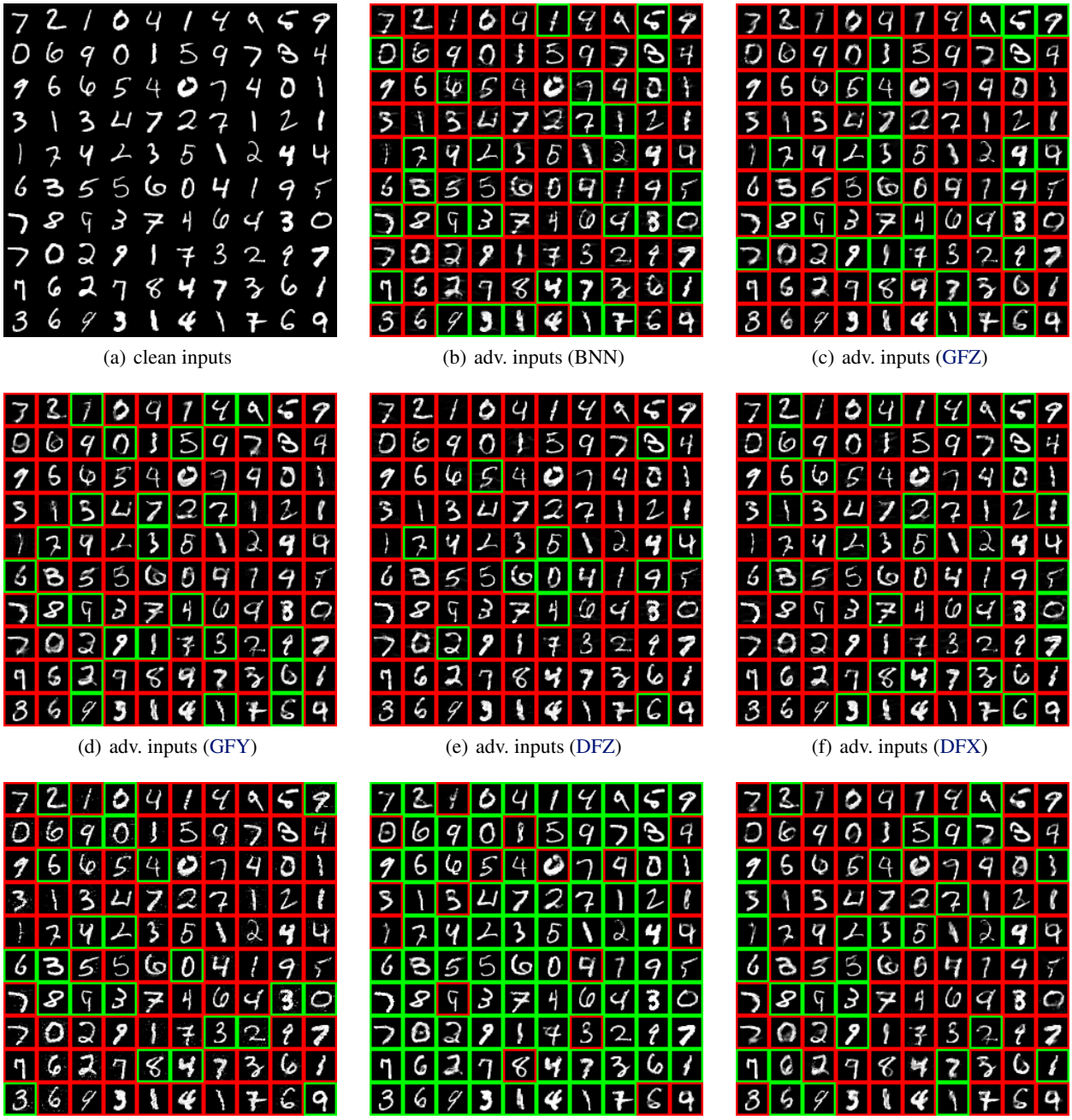


Figure B.3. Victim accuracy and detection rates against **white-box attacks** on CIFAR-10. The higher the better.

B.4. Visualising CW-ℓ_2 adversarial examples

We visualise in Figure B.4 the crafted adversarial images using white-box CW attack, where successful attacks are in red rectangles. We clearly see that many of the successful adversarial examples crafted on the generative classifiers sit at the boundary of two classes (thus ambiguous). For example, many digit "4" clean images are distorted to resemble digit "9". Similarly many digit "1" clean images are distorted to resemble digits "7" and (very thin) "3" and "8". On the other hand, we see less ambiguity from the successful attacks on discriminative classifiers. Therefore we conclude that the perceptual distortion of CW attacks on generative and discriminative classifiers are very different.



(a) clean inputs (b) adv. inputs (BNN) (c) adv. inputs (GFZ)

(d) adv. inputs (GFY) (e) adv. inputs (DFZ) (f) adv. inputs (DFX)

(g) adv. inputs (DBX) (h) adv. inputs (GBZ) (i) adv. inputs (GBY)

Figure B.4. Visualising the clean inputs of MNIST and the CW ($c = 10$) adversarial examples crafted on all the classifiers. Digits in **red** rectangles are **successful** attacks, and digits in **green** rectangles are **unsuccessful** attacks.

C. Further discusssions

C.1. The fusion model for vision tasks: connections to perceptual loss

The CIFAR-binary experiments in the main text indicate that likelihood functions based on per-pixel ℓ_2 loss in the observation space are less suitable for modelling natural images. This observation has also been made in the deep generative models literature, and in particular, the generative adversarial network approach (GAN Goodfellow et al., 2014) can be viewed as evaluating the quality of "perceptually realistic image generations" using *discriminative* features. Following this principle, researchers have scaled GAN-based approaches to generate high resolution images (Karras et al., 2018).

Similarly, the computer vision community has discovered that "distances" defined on the features of a *discriminatively* trained deep CNN work surprisingly well to *perceptually* measure the similarity of two images. Indeed, recent research proposed the *perceptual loss* (Dosovitskiy & Brox, 2016; Johnson et al., 2016) as the ℓ_2 distance between convolutional features extracted from a very deep CNN (e.g. VGG):

$$\ell_{\text{perceptual}}(\boldsymbol{x}_1, \boldsymbol{x}_2) := ||\boldsymbol{\phi}(\boldsymbol{x}_1) - \boldsymbol{\phi}(\boldsymbol{x}_2)||_2, \quad \boldsymbol{\phi}(\boldsymbol{x}) = \text{CNN-conv-layer}(\boldsymbol{x}).$$

This perceptual loss has been successfully applied to neural style transfer, super-resolution and conditional image synthesis (Gatys et al., 2016; Dosovitskiy & Brox, 2016; Ledig et al., 2017; Johnson et al., 2016).

Critically, we highlight an empirical study from Zhang et al. (2018): when comparing the "perceptual similarity" between two images, decisions based on the perceptual loss are well-aligned with human judgements. This alignment of perceptual loss to human vision also explains the success of the fusion model presented in the main text. Here the discriminative VGG features are used to construct the perceptual loss, which is used by the generative classifiers to measure the closeness of a new input $\boldsymbol{x}^*$ to the manifold of clean images (estimated from training data). By contrast, discriminative classifiers (e.g. the original VGG classifier), when making decisions, do not explicitly take into account the "perceptual similarity" of the input to the images in the predicted catergory. Therefore an adversarial image containing a "cat" can easily fool the discriminative classifier to predict a "dog" class label.

Still we note that the fusion model in practice is unlikely to be robust to *all* attacks. The alignment of human vision and the perceptual loss with current deep CNNs is not perfect, therefore under the white-box setting against the whole system, an attacker might be able to craft an adversarial example that has minimum ℓ_2, ℓ_∞ or even ℓ_0 distortion, but at the same time has minimum "perceptual distance" to the image manifold of the *incorrect* class. Much future work is to be done on improving representation learning for perceptual losses, as well as on investigating the adversarial robustness of the fusion model under different threat models.

D. Model architectures

MNIST experiments The VAEs are constructed with convolutional encoders and deconvolutional generators. More specifically, the encoder network for $q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ is the same across all VAE-based classifiers. It starts with a 3-layer convolutional neural network with 5×5 filters and 64 channels, with a max-pooling operation after each convolution. Then, the convolutional network is followed by a MLP with 2 hidden layers, each with 500 units, to produce the mean and variance parameters of q. The label $\boldsymbol{y}$ is injected into the MLP at the first hidden layer, as a one hot encoding (i.e. for MNIST, the first hidden layer has 500+10 units). The latent dimension is $\dim(\boldsymbol{z}) = 64$.

The p models' architectures are the following:

- GFZ: For $p(\boldsymbol{y}|\boldsymbol{z})$ we use a MLP with 1 hidden layer composed of 500 units. For $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})$ we used an MLP with 2 hidden layers, each with 500 units, and $4 \times 4 \times 64$ dimension output, followed by a 3-layer deconvolutional network with 5×5 kernel size, stride 2 and [64, 64, 1] channels.
- GFY: We use an MLP with 1 hidden layer composed of 500 units for $p(\boldsymbol{z}|\boldsymbol{y})$, and the same architecture as GFZ for $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})$.
- DFZ: We use almost the same deconvolutional network architecture for $p(\boldsymbol{x}|\boldsymbol{z})$ as GFZ's $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z})$ network, except that the input is $\boldsymbol{z}$ only. For $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})$ we use almost the same architecture as $q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ except that the injected input to the MLP is $\boldsymbol{z}$ and the MLP output is the set of logit values for $\boldsymbol{y}$.
- DFX: We use the same architecture as G3 for $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})$. The network for $p(\boldsymbol{z}|\boldsymbol{x})$ is almost identical except that there is no injected input to the MLP, and the network returns the mean and variance parameters for $q(\boldsymbol{z}|\boldsymbol{x})$.
- DBX: We use GFZ's architecture for $p(\boldsymbol{y}|\boldsymbol{z})$ and DFX's architecture for $p(\boldsymbol{z}|\boldsymbol{x})$.
- GBY: We use GFY's architecture for $p(\boldsymbol{z}|\boldsymbol{y})$ and DFZ's architecture for $p(\boldsymbol{x}|\boldsymbol{z})$.
- GBZ: We use GFZ's architecture for $p(\boldsymbol{y}|\boldsymbol{z})$ and DFZ's architecture for $p(\boldsymbol{x}|\boldsymbol{z})$.

The BNN has almost the same architecture as the encoder network q, except that it uses 2x the hidden units/channels, and the last layer is 10 dimensions. Note that here we used dropout as it is convenient to implement, and we expect better approximate inference methods (such as stochastic gradient MCMC) to return better results for robustness and detection.

CIFAR-binary experiments The model architectures are almost the same as used in MNIST experiments, except that the hidden layer dimensions for the MLP layers are increased to 1000. For the encoder q, the channels are increased to [64, 128, 256]. For the p models, the deconvolutional networks have different channel values, [128, 64, 3], and the MLP before the deconvolution outputs a $4 \times 4 \times 256$ vector (before reshaping). The BNN has 2x the channels but still uses 1000 hidden units.

CIFAR-10 experiments The pre-trained VGG16 network is downloaded from `https://github.com/geifmany/cifar-vgg`, where the CONV9 and FC1 layers correspond to:

- CONV9: `https://github.com/geifmany/cifar-vgg/blob/master/cifar10vgg.py#L82`
- FC1: `https://github.com/geifmany/cifar-vgg/blob/master/cifar10vgg.py#L109`

The VAE-based classifiers build fully connected networks on top of the extracted features, and use $\dim(\boldsymbol{z}) = 128$ for bottleneck. The encoder $q(\boldsymbol{z}|\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{y})$ has the network architectures [$\dim(\boldsymbol{\phi}(\boldsymbol{x}))$ + $\dim(\boldsymbol{y})$, 1000, 1000, $\dim(\boldsymbol{z}) \times 2$], and we use the same encoder architecture across all classifiers. The decoder architectures are as follows:

- DBX: We use an MLP of layers [$\dim(\boldsymbol{z})$, 1000, $\dim(\boldsymbol{y})$] for $p(\boldsymbol{y}|\boldsymbol{z})$ and an MLP of layers [$\dim(\boldsymbol{\phi}(\boldsymbol{x}))$, 1000, 1000, $\dim(\boldsymbol{z}) \times 2$] for $p(\boldsymbol{z}|\boldsymbol{\phi}(\boldsymbol{x}))$.
- GBZ: We use an MLP of layers [$\dim(\boldsymbol{z})$, 1000, 1000, $\dim(\boldsymbol{y})$] for $p(\boldsymbol{y}|\boldsymbol{z})$ and an MLP of layers [$\dim(\boldsymbol{z})$, 1000, 1000, $\dim(\boldsymbol{\phi}(\boldsymbol{x}))$] for $p(\boldsymbol{\phi}(\boldsymbol{x})|\boldsymbol{z})$.
- GBY: We use an MLP of layers [$\dim(\boldsymbol{y})$, 1000, $\dim(\boldsymbol{z}) \times 2$] for $p(\boldsymbol{z}|\boldsymbol{y})$ and GBZ's architecture for $p(\boldsymbol{\phi}(\boldsymbol{x})|\boldsymbol{z})$.

E. Attack settings

We use the Cleverhans package to perform attacks. We use the default hyper-parameters, if not specifically stated.

PGD: We perform the attack for 40 iterations with step-size 0.01.

MIM: We perform the attack for 40 iterations with step-size 0.01 and decay factor 1.0.

CW-ℓ_2: We use learning rate 0.01 for $c = 0.1, 1, 10$, learning rate 0.03 for $c = 100$, and learning rate 0.1 for $c = 1000$. We set the confidence parameter to 0, and we optimise the loss for 1000 iterations.

SPSA: We use almost the same hyper-parameters as in Uesato et al. (2018) except for the number of samples for gradient estimates. In detail, we perform the attack for 100 iterations with perturbation size 0.01, Adam learning rate 0.01, stopping threshold -5.0 and 2000 samples for each gradient estimate.

E.1. Jacobian-based dataset augmentation

The black-box distillation attack is based on Papernot et al. (2017b), which trains a substitute CNN using Jacobian-based dataset augmentation. Assume $\boldsymbol{y} = F(\boldsymbol{x})$ is the output one-hot vector of the victim, and $\boldsymbol{p}(\boldsymbol{x})$ is the probability vector output of the substitute model, then at the t^{th} outer-loop, we train the substitute CNN on dataset $\mathcal{D}_t = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ with queried $\boldsymbol{y}_n$ for 10 epochs, and augment the dataset by

$$\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(\hat{\boldsymbol{x}}, F(\hat{\boldsymbol{x}})) \mid \hat{\boldsymbol{x}} = \boldsymbol{x} + \lambda \nabla_{\boldsymbol{x}} \boldsymbol{p}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{y},\ (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}_t\}. \tag{4}$$

We initialise $\mathcal{D}_1$ with 200×10 datapoints from the MNIST test set, select $\lambda = 0.1$, and run the algorithm for 6 outer-loops. On MNIST, this results in $64,000$ queried inputs, and $\sim 96\%$ accuracy of the substitute model on test data. On CIFAR binary classification, we use 200×2 datapoints for the inital query set $\mathcal{D}_1$, resulting in $12,800$ queries in total. The substitutes achieved almost the same accuracy as their corresponding victim models on clean test datapoints.

F. Results in tables

We present in tables the full results of the experiments.

See Tables F.1 to F.10 for the white-box attacks.

See Tables F.11 to F.16 for the grey-box attacks.

See Tables F.17 to F.22 for the black-box attacks.

See Tables F.23 to F.25 for CIFAR-10 results with VGG-based classifiers.

See Tables F.26 to F.28 for bottleneck effect quantification results.

Table F.1. FGSM white-box attack results on MNIST.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
BNN	92.4	67.8	40.5	26.2	20.4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	93.0	94.2	95.4	95.5	96.5
GFZ	94.2	74.5	38.9	12.9	5.7	43.6	79.8	100.0	100.0	100.0	56.4	89.6	99.9	100.0	100.0	89.2	91.7	92.2	92.5	92.2
GBZ	92.5	80.3	62.0	42.4	27.2	37.0	81.7	100.0	100.0	100.0	57.6	93.5	100.0	100.0	100.0	91.6	90.9	90.3	91.1	91.5
GFY	94.3	74.8	46.5	21.7	10.5	53.1	92.6	100.0	100.0	100.0	66.2	97.9	100.0	100.0	100.0	90.8	93.1	93.8	94.2	94.2
GBY	93.6	76.4	47.5	22.3	10.7	41.9	84.5	100.0	100.0	100.0	57.7	92.5	100.0	100.0	100.0	89.3	92.7	92.8	92.9	92.9
DFX	70.1	14.8	1.0	0.4	0.6	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	93.6	93.8	93.8	93.5	94.3
DBX	91.6	77.8	58.1	44.5	36.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.6	93.6	95.2	96.3	97.0
DFZ	75.2	19.0	2.4	1.1	1.0	12.6	50.6	100.0	100.0	100.0	26.9	65.7	100.0	100.0	100.0	93.1	95.1	94.5	94.9	94.8

Table F.2. PGD white-box attack results on MNIST.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
BNN	83.2	12.0	0.5	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.1	92.1	88.7	20.4	0.2
GFZ	86.7	37.7	7.7	1.2	0.3	43.2	71.8	91.4	99.4	100.0	55.8	77.1	94.8	99.6	100.0	90.3	92.1	90.9	90.4	89.7
GBZ	85.1	57.4	32.5	19.3	11.6	33.7	50.3	84.4	99.7	100.0	52.2	66.3	91.5	99.8	100.0	90.0	91.5	91.5	91.7	91.8
GFY	79.7	27.4	5.6	1.2	0.3	58.1	87.9	98.0	100.0	100.0	68.7	92.2	99.3	100.0	100.0	92.6	92.5	90.7	90.7	85.4
GBY	86.7	35.9	9.0	1.7	0.4	45.3	76.1	94.6	99.7	100.0	56.9	79.3	95.6	99.9	100.0	90.1	92.1	91.6	91.7	91.4
DFX	47.6	0.7	0.0	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.8	89.8	13.3	31.4	42.8
DBX	58.0	18.3	6.0	1.3	0.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.7	94.0	94.1	93.7	84.8
DFZ	49.6	1.0	0.0	0.0	0.0	11.8	44.3	94.1	100.0	100.0	25.9	57.6	94.5	99.9	100.0	93.9	89.8	12.1	22.9	28.0

Table F.3. MIM white-box attack results on MNIST.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
BNN	82.0	7.2	0.1	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.2	91.1	72.0	64.5	7.1
GFZ	87.0	40.4	9.0	1.2	1.2	43.5	83.8	98.4	99.4	99.4	57.6	89.3	98.6	99.2	99.2	91.5	92.3	92.0	90.6	90.6
GBZ	79.6	27.4	5.6	1.5	0.5	36.1	86.3	100.0	100.0	100.0	53.5	93.1	99.9	100.0	100.0	91.3	91.4	91.9	92.2	92.7
GFY	80.8	30.1	6.8	1.4	1.4	56.0	93.7	99.8	100.0	100.0	68.1	96.6	99.9	100.0	100.0	90.9	92.4	92.0	91.6	91.6
GBY	84.9	22.9	1.5	0.1	0.0	47.4	91.2	99.9	100.0	100.0	59.8	92.8	99.9	100.0	100.0	91.4	92.1	91.2	86.8	29.0
DFX	48.4	0.8	0.0	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	93.2	89.6	9.2	17.5	17.5
DBX	66.7	28.7	19.7	17.2	17.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	93.5	94.3	94.6	94.7	94.7
DFZ	50.5	1.2	0.0	0.0	0.0	11.9	50.7	99.4	100.0	100.0	26.4	63.2	99.0	100.0	100.0	94.3	89.8	11.5	17.3	17.3

Table F.4. CW white-box attack results on MNIST.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00
BNN	98.6	50.2	24.4	19.7	38.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	85.6	96.2	93.6	94.0	96.5
GFZ	98.7	76.7	28.6	20.8	28.4	45.4	25.1	11.7	17.9	49.2	62.0	39.2	25.3	29.9	57.4	91.1	94.9	95.3	93.5	94.8
GBZ	97.3	95.3	81.5	80.2	66.6	35.1	29.0	10.8	14.0	38.2	56.3	50.6	26.2	29.2	55.8	85.3	89.9	90.8	91.8	91.7
GFY	98.7	70.4	28.6	25.0	33.7	52.7	37.5	21.2	29.0	59.7	68.5	51.1	33.5	39.4	67.2	91.4	96.3	96.0	94.4	95.3
GBY	98.5	77.7	32.7	26.8	38.1	44.7	26.4	12.1	18.8	52.8	56.8	39.7	22.3	27.1	59.6	90.0	95.5	94.5	92.9	94.8
DFX	96.5	64.7	20.3	11.8	2.6	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	96.7	99.9	100.0	99.7	97.8
DBX	97.1	83.3	30.2	10.9	29.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	90.0	93.4	96.7	96.1	97.5
DFZ	95.9	51.2	13.6	9.6	16.0	17.3	7.2	9.2	17.9	41.6	30.6	21.5	22.4	31.9	54.0	96.2	99.8	99.5	98.6	98.2

Table F.5. WB+S+(M/L/K) attacks on MNIST. This is done using the PGD attack with $\epsilon = 0.2$ and $\lambda_{\text{detect}} = 0.1$. 'WB' stands for attacks against the classifier only, 'WB+S' means the adversary has knowledge of the K samples but not of the detection system. 'WB+S+M', 'WB+S+L', and 'WB+S+K' are attacks where the adversary has knowledge of the K samples and of the marginal, logit and KL detection mechanisms respectively.

	WB				WB+S				WB+S+M				WB+S+L				WB+S+K			
	acc.	TP marg.	TP logit	TP KL	acc.	TP marg.	TP logit	TP KL	acc.	TP marg.	TP logit	TP KL	acc.	TP marg.	TP logit	TP KL	acc.	TP marg.	TP logit	TP KL
DBX	18.3	N/A	N/A	94	17	N/A	N/A	93.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	35.9	N/A	N/A	93.9
GBZ	57.4	50.3	66.3	91.5	35.5	82.2	90.7	91.4	47.9	36.5	67.4	91.3	44.6	50.5	69.9	91.0	45	73.0	84.8	91.5
GBY	35.9	76.1	79.3	92.1	22	90.8	91.9	92.1	43.8	47.0	61.9	91.1	52.0	56.4	66.9	90.6	32.8	89.2	91.8	93.3

Table F.6. FGSM white-box attack results on CIFAR plane-vs-frog binary classification.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
BNN	98.2	93.2	58.5	14.5	6.3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	58.6	55.1	65.5	54.0	58.5
GFZ	97.1	94.7	81.8	56.4	31.4	11.4	10.1	5.9	17.4	99.3	11.4	10.4	6.8	17.2	99.3	58.6	40.1	35.9	41.0	50.6
GBZ	95.1	93.5	87.1	74.9	62.0	26.0	18.6	15.5	38.6	99.6	26.3	20.3	19.1	41.9	99.6	35.5	41.5	47.0	43.7	45.6
GFY	96.5	94.2	80.7	56.7	32.0	9.6	9.2	6.9	18.8	99.0	17.6	13.1	8.1	20.8	99.1	46.6	43.7	36.1	39.5	49.3
GBY	96.1	92.9	82.0	60.5	36.3	17.2	15.1	8.1	28.5	99.2	20.2	18.1	10.0	31.1	99.2	49.7	48.0	39.2	38.7	47.3
DFX	83.8	42.2	0.7	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	63.3	66.6	48.4	97.7	100.0
DBX	90.9	78.6	50.7	31.7	18.3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	50.7	58.7	56.8	59.7	57.2
DFZ	83.9	52.1	2.9	0.0	0.0	6.6	3.8	2.2	3.9	60.6	6.9	4.2	2.4	3.3	60.4	57.5	62.3	54.8	66.7	99.8

Table F.7. PGD white-box attack results on CIFAR plane-vs-frog binary classification.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
BNN	97.9	86.7	19.7	1.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	41.7	59.4	55.7	68.4	98.9
GFZ	98.0	93.9	67.7	21.7	3.5	3.3	6.6	5.6	7.8	32.9	5.0	9.5	6.0	8.2	33.0	37.5	45.1	34.5	44.0	57.7
GBZ	94.6	93.7	83.9	67.4	52.8	19.8	17.9	12.8	14.3	43.1	22.8	19.5	17.4	17.0	44.3	32.1	39.3	40.2	37.1	33.2
GFY	98.4	95.0	67.9	25.8	4.1	4.2	6.7	6.4	7.8	35.7	3.1	8.2	7.6	7.8	33.8	31.2	43.6	32.9	39.6	53.3
GBY	96.4	92.9	67.3	25.7	6.9	11.1	8.4	7.9	11.6	41.4	13.1	11.5	9.2	12.2	40.3	43.0	43.0	33.6	39.1	52.2
DFX	82.7	35.7	0.3	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	64.1	64.7	69.6	100.0	100.0
DBX	83.3	34.6	4.2	0.7	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	59.8	59.5	65.1	78.7	93.4
DFZ	82.4	36.9	0.4	0.0	0.0	4.9	3.8	5.1	12.3	91.1	6.5	4.1	5.5	12.3	91.3	56.3	60.3	54.7	99.7	100.0

Table F.8. MIM white-box attack results on CIFAR plane-vs-frog binary classification.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
BNN	96.9	84.6	18.7	0.9	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	35.9	54.7	56.3	62.4	78.8
GFZ	96.9	91.5	58.1	15.2	1.9	6.5	6.2	6.4	10.6	86.1	10.0	8.3	6.6	11.3	84.5	31.0	39.0	32.1	47.2	77.1
GBZ	92.5	89.4	71.5	40.4	17.9	15.0	13.2	11.0	24.0	95.9	16.4	14.8	12.8	25.1	96.0	42.5	37.5	31.0	37.4	55.0
GFY	97.6	92.3	59.5	16.5	2.3	5.6	6.9	6.3	13.3	88.1	5.7	6.2	6.8	13.8	87.6	23.8	41.4	33.8	43.9	74.6
GBY	95.1	88.8	56.5	14.6	1.3	15.1	9.5	8.6	19.2	97.3	15.3	10.7	9.9	18.5	97.0	51.1	43.2	31.7	46.7	87.6
DFX	82.5	35.2	0.3	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	65.5	67.1	69.6	100.0	100.0
DBX	82.5	38.8	6.1	1.2	0.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	54.6	60.0	61.5	78.8	88.5
DFZ	81.5	36.1	0.4	0.0	0.0	5.4	3.9	5.0	13.4	96.9	6.2	4.1	5.5	13.8	96.8	61.1	58.5	55.3	99.9	100.0

Table F.9. CW white-box attack results on CIFAR plane-vs-frog binary classification.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00
BNN	98.0	68.9	38.5	19.5	9.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	56.2	79.0	79.5	71.5	65.3
GFZ	99.5	95.6	76.5	56.7	43.4	0.0	4.5	2.6	5.9	14.7	0.0	4.6	3.0	6.4	14.0	29.2	30.4	25.6	28.1	33.2
GBZ	96.0	93.6	88.9	80.6	69.8	16.7	11.5	7.8	11.7	33.8	18.9	13.7	9.3	13.6	37.5	41.1	51.4	40.4	36.0	37.9
GFY	99.8	95.9	78.8	60.6	42.7	0.0	6.6	4.1	5.2	15.9	0.0	6.6	4.8	5.7	15.6	75.0	37.3	26.0	28.2	31.5
GBY	97.5	93.1	76.7	61.6	47.7	15.8	14.6	6.0	6.4	19.9	18.9	13.9	6.0	7.5	21.3	42.6	34.3	27.2	25.2	33.9
DFX	82.6	44.2	34.3	28.6	4.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100.0	100.0	100.0	99.9	64.6
DBX	96.5	72.2	26.3	12.1	11.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	61.6	79.4	85.2	75.2	70.2
DFZ	94.5	72.3	29.9	12.9	4.8	7.3	5.8	4.3	4.6	5.7	14.1	6.9	5.0	4.9	6.3	82.6	97.3	97.1	90.0	68.8

Table F.10. WB+S+(M/L/K) attacks on CIFAR binary task. This is done using the PGD attack with $\epsilon = 0.1$ and $\lambda_{\text{detect}} = 1.0$. 'WB' stands for attacks against the classifier only, 'WB+S' means the adversary has knowledge of the K samples but not of the detection system. 'WB+S+M', 'WB+S+L', and 'WB+S+K' are attacks where the adversary has knowledge of the K samples and of the marginal, logit and KL detection mechanisms respectively.

	WB				WB+S				WB+S+M				WB+S+L				WB+S+K			
	acc.	TP marg.	TP logit	TP KL	acc.	TP marg.	TP logit	TP KL	acc.	TP marg.	TP logit	TP KL	acc.	TP marg.	TP logit	TP KL	acc.	TP marg.	TP logit	TP KL
DBX	0.7	N/A	N/A	78.7	0.7	N/A	N/A	82.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.6	N/A	N/A	82.6
GBZ	67.4	14.3	17.0	37.1	57	40.6	45.1	49	58.4	9.6	18.5	46.9	58.5	1.8	7.8	47.8	59.9	25.5	27.6	42.4
GBY	25.7	11.6	12.2	39.1	28.5	22.0	22.6	49	29.5	5.6	7.6	45.9	29.2	0.8	1.8	45.4	27.9	15.6	15.9	45

Table F.11. Grey-box PGD attack results on MNIST.

	substitute acc					victim acc					TP logit				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
GFZ	86.7	14.3	0.0	0.0	0.0	96.6	83.3	51.3	26.8	17.2	49.5	73.8	99.7	100.0	100.0
GBZ	81.2	9.1	0.0	0.0	0.0	93.5	80.1	55.3	35.0	25.2	46.7	73.3	99.6	100.0	100.0
GFY	84.8	6.4	0.0	0.0	0.0	96.7	82.3	55.2	33.5	24.0	58.7	88.8	100.0	100.0	100.0
GBY	86.7	15.0	0.0	0.0	0.0	95.8	81.6	51.0	27.5	18.1	50.1	78.5	99.9	100.0	100.0
DFX	74.2	5.2	0.5	0.0	0.0	91.7	57.7	19.6	4.3	1.4	N/A	N/A	N/A	N/A	N/A
DBX	80.6	5.1	0.0	0.0	0.0	93.2	59.2	22.5	10.9	8.2	N/A	N/A	N/A	N/A	N/A
DFZ	69.6	3.4	0.3	0.0	0.0	91.9	57.2	21.4	6.3	2.6	33.6	55.1	91.7	100.0	100.0

Table F.12. Grey-box MIM attack results on MNIST.

	substitute acc					victim acc					TP logit				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
GFZ	87.2	16.1	0.0	0.0	0.0	96.5	82.4	42.8	14.2	3.7	50.6	82.5	100.0	100.0	100.0
GBZ	81.8	9.1	0.0	0.0	0.0	93.3	78.6	45.3	17.0	5.2	46.2	81.7	100.0	100.0	100.0
GFY	85.2	6.9	0.0	0.0	0.0	96.6	80.8	48.0	18.9	7.4	59.8	96.0	100.0	100.0	100.0
GBY	87.0	15.9	0.0	0.0	0.0	95.7	79.9	41.8	13.0	4.0	51.4	87.1	100.0	100.0	100.0
DFX	76.5	12.3	3.6	2.1	2.1	91.5	56.1	22.2	9.9	9.9	N/A	N/A	N/A	N/A	N/A
DBX	85.1	19.9	0.6	0.1	0.1	93.4	57.0	14.1	6.3	6.3	N/A	N/A	N/A	N/A	N/A
DFZ	70.9	3.2	0.0	0.0	0.0	91.8	54.6	16.4	2.7	0.3	34.5	63.8	98.9	100.0	100.0

Table F.13. Grey-box CW attack results on MNIST.

	substitute acc					victim acc					TP logit				
ϵ	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00
GFZ	98.4	4.6	0.0	0.0	0.0	98.8	96.9	96.0	93.4	90.2	53.4	44.2	41.6	40.6	43.7
GBZ	98.4	63.2	0.0	0.0	0.0	97.4	95.4	92.9	88.6	84.5	45.9	39.7	33.0	31.2	38.6
GFY	98.1	0.9	0.0	0.0	0.0	99.0	97.6	97.0	95.7	94.0	61.0	55.2	49.2	49.5	52.7
GBY	98.3	5.3	0.0	0.0	0.0	98.7	96.8	96.0	93.8	91.2	54.2	41.5	37.6	38.7	43.6
DFX	85.1	0.0	0.0	0.0	0.0	97.6	93.3	91.2	90.4	89.8	N/A	N/A	N/A	N/A	N/A
DBX	97.3	46.0	0.4	0.0	0.0	98.7	93.9	88.4	85.5	76.3	N/A	N/A	N/A	N/A	N/A
DFZ	80.5	0.0	0.0	0.0	0.0	97.2	94.6	92.3	91.5	91.2	32.9	28.6	28.1	36.5	44.5

Table F.14. Grey-box PGD attack results on CIFAR plane-vs-frog binary classification.

	substitute acc					victim acc					TP logit				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
GFZ	96.7	88.1	41.7	4.5	0.1	97.0	94.9	84.0	48.3	8.3	8.4	11.2	6.1	7.5	49.8
GBZ	92.9	83.1	37.8	5.5	0.0	95.3	94.8	91.8	83.1	65.6	15.5	12.8	10.9	19.9	79.8
GFY	96.9	88.1	33.7	3.3	0.1	96.9	95.9	85.3	54.2	12.7	2.6	2.0	5.0	8.5	61.2
GBY	95.2	85.6	32.4	2.8	0.1	96.8	95.3	86.4	59.8	18.5	21.8	14.3	10.5	12.5	75.2
DFX	95.7	80.8	10.1	0.2	0.0	99.1	96.7	79.4	28.3	0.9	N/A	N/A	N/A	N/A	N/A
DBX	95.7	80.5	30.5	1.7	0.0	99.5	98.2	91.1	65.1	21.8	N/A	N/A	N/A	N/A	N/A
DFZ	96.5	81.9	10.5	0.3	0.0	99.1	96.9	76.9	26.2	1.3	4.5	16.5	6.1	8.5	59.6

Table F.15. Grey-box MIM attack results on CIFAR plane-vs-frog binary classification.

	substitute acc					victim acc					TP logit				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
GFZ	96.3	87.9	42.1	4.7	0.1	97.1	94.8	83.0	44.8	4.9	8.7	11.1	5.8	8.5	75.1
GBZ	92.9	82.5	36.9	4.9	0.0	95.5	94.7	91.4	81.7	49.5	14.8	12.6	13.9	21.9	97.8
GFY	96.8	87.9	34.1	3.5	0.1	96.9	95.6	84.6	49.1	6.9	2.6	2.0	6.7	10.3	81.1
GBY	95.1	85.5	32.0	2.6	0.1	96.8	95.5	85.7	56.2	10.9	21.8	14.8	9.2	13.4	97.6
DFX	95.5	80.5	10.5	0.2	0.0	99.1	96.5	77.3	24.0	0.2	N/A	N/A	N/A	N/A	N/A
DBX	95.6	81.1	34.5	2.2	0.0	99.5	98.3	89.3	58.9	9.9	N/A	N/A	N/A	N/A	N/A
DFZ	96.3	81.5	10.7	0.5	0.0	99.1	96.9	74.9	21.9	0.3	4.5	16.6	6.4	9.9	80.2

Table F.16. Grey-box CW attack results on CIFAR plane-vs-frog binary classification.

	substitute acc					victim acc					TP logit				
ϵ	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00
GFZ	98.7	61.9	0.0	0.0	0.0	98.6	95.8	94.3	93.8	93.2	11.7	10.2	10.9	9.9	8.5
GBZ	96.3	69.8	3.3	0.0	0.0	95.9	95.7	94.7	94.7	94.5	13.5	12.7	12.3	13.7	14.4
GFY	98.4	54.4	0.0	0.0	0.0	98.1	96.1	95.3	95.1	94.3	0.0	2.2	3.7	4.6	7.6
GBY	97.6	52.9	2.2	0.0	0.0	97.7	95.6	94.5	94.3	93.7	25.8	15.7	14.2	13.7	18.0
DFX	64.4	0.0	0.0	0.0	0.0	98.5	97.2	97.1	96.7	95.5	N/A	N/A	N/A	N/A	N/A
DBX	91.3	56.2	0.3	0.0	0.0	99.7	99.0	99.0	98.9	98.1	N/A	N/A	N/A	N/A	N/A
DFZ	80.3	0.0	0.0	0.0	0.0	99.2	97.9	97.5	97.3	95.4	0.0	15.7	14.6	8.8	6.1

Table F.17. Black-box PGD attack results on MNIST.

	substitute acc					victim acc					TP logit				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
GFZ	44.2	0.7	0.0	0.0	0.0	97.3	88.4	62.7	33.8	21.8	48.7	61.0	92.3	99.8	100.0
GBZ	7.4	0.0	0.0	0.0	0.0	94.8	87.1	70.7	52.9	41.3	51.0	68.6	97.7	100.0	100.0
GFY	49.4	0.8	0.0	0.0	0.0	97.4	86.5	58.1	31.9	21.0	52.5	70.8	97.4	99.8	100.0
GBY	21.7	0.0	0.0	0.0	0.0	96.9	89.3	70.0	49.7	37.8	51.2	70.8	98.2	100.0	100.0
DFX	49.2	1.3	0.0	0.0	0.0	91.4	52.1	13.9	2.2	0.7	N/A	N/A	N/A	N/A	N/A
DBX	43.1	0.7	0.0	0.0	0.0	95.0	67.0	26.2	9.7	6.5	N/A	N/A	N/A	N/A	N/A
DFZ	53.8	1.7	0.0	0.0	0.0	92.9	56.3	15.2	3.4	1.5	34.7	52.8	94.0	99.9	100.0

Table F.18. Black-box MIM attack results on MNIST.

	substitute acc					victim acc					TP logit				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
GFZ	45.5	2.8	0.0	0.0	0.0	97.2	86.6	50.3	17.0	17.0	51.1	68.0	97.5	99.5	99.5
GBZ	8.4	0.0	0.0	0.0	0.0	94.7	85.9	63.5	35.2	17.1	51.2	74.4	99.0	100.0	100.0
GFY	53.1	2.5	0.3	0.1	0.1	97.2	83.7	49.2	20.1	20.1	55.1	80.6	99.0	99.9	99.9
GBY	24.5	0.0	0.0	0.0	0.0	96.8	88.3	62.8	31.7	13.3	49.3	77.6	99.9	100.0	100.0
DFX	51.3	2.3	0.1	0.0	0.0	91.4	51.7	13.3	1.9	1.9	N/A	N/A	N/A	N/A	N/A
DBX	47.8	1.6	0.1	0.0	0.0	94.8	67.0	23.2	7.6	7.6	N/A	N/A	N/A	N/A	N/A
DFZ	56.1	2.9	0.1	0.0	0.0	92.7	56.0	13.1	2.8	2.8	35.2	59.9	97.6	100.0	100.0

Table F.19. Black-box CW attack results on MNIST.

	substitute acc					victim acc					TP logit				
ϵ	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00
GFZ	65.2	0.3	0.0	0.0	0.0	98.8	98.7	97.4	94.3	92.0	52.3	50.8	39.6	44.7	55.3
GBZ	76.6	0.4	0.0	0.0	0.0	97.3	97.2	95.9	92.7	90.4	45.7	46.1	41.5	40.3	45.7
GFY	88.4	3.8	0.0	0.0	0.0	99.0	98.9	98.0	94.8	92.0	60.0	59.2	51.9	47.6	57.2
GBY	76.0	1.3	0.0	0.0	0.0	98.7	98.6	97.3	95.1	93.5	53.4	52.2	41.8	39.7	46.7
DFX	82.4	0.0	0.0	0.0	0.0	98.8	96.8	92.4	86.2	84.9	N/A	N/A	N/A	N/A	N/A
DBX	82.5	0.6	0.0	0.0	0.0	98.9	98.4	95.5	85.2	83.1	N/A	N/A	N/A	N/A	N/A
DFZ	86.5	0.2	0.0	0.0	0.0	98.8	94.5	89.3	81.5	79.3	42.9	25.5	22.3	27.7	38.2

Table F.20. Black-box PGD attack results on CIFAR plane-vs-frog binary classification.

	substitute acc					victim acc					TP logit				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
GFZ	95.5	89.9	45.5	6.8	0.1	97.6	95.6	88.7	68.8	29.7	6.2	6.7	6.0	10.8	71.8
GBZ	92.0	84.1	33.8	3.5	0.0	95.6	94.9	92.7	85.0	65.9	16.7	14.6	12.8	17.9	76.8
GFY	95.3	89.1	38.7	2.5	0.0	97.3	96.6	90.0	72.9	35.5	0.0	2.3	6.8	10.4	72.3
GBY	91.8	80.5	29.7	6.1	0.2	97.2	95.9	90.2	70.8	30.3	21.8	16.4	9.6	13.6	77.3
DFX	89.1	74.5	19.1	0.9	0.0	99.7	98.7	92.6	64.8	9.9	N/A	N/A	N/A	N/A	N/A
DBX	85.5	76.4	42.5	4.5	0.0	99.6	99.0	94.8	77.5	24.5	N/A	N/A	N/A	N/A	N/A
DFZ	85.5	73.5	21.1	1.5	0.0	99.5	99.2	93.8	70.1	15.1	0.0	5.0	8.9	15.9	85.1

Table F.21. Black-box MIM attack results on CIFAR plane-vs-frog binary classification.

	substitute acc					victim acc					TP logit				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
GFZ	95.5	89.7	46.4	7.7	0.2	97.5	95.7	88.5	66.5	20.5	6.2	6.7	5.9	12.7	82.7
GBZ	91.8	84.0	33.1	3.4	0.0	95.6	94.9	92.3	85.1	60.7	16.7	15.5	14.2	21.6	97.9
GFY	95.3	88.9	40.5	3.1	0.0	97.3	96.7	89.5	70.6	31.9	0.0	2.4	7.6	12.5	88.0
GBY	91.7	80.2	29.5	5.9	0.1	97.3	95.9	89.6	69.0	26.1	22.2	16.1	9.7	14.0	95.9
DFX	89.0	74.5	19.4	0.9	0.0	99.6	98.7	92.1	63.0	10.9	N/A	N/A	N/A	N/A	N/A
DBX	85.5	76.3	42.7	4.5	0.0	99.6	99.0	95.1	76.5	27.2	N/A	N/A	N/A	N/A	N/A
DFZ	85.3	73.4	21.4	1.6	0.0	99.6	99.1	92.9	67.7	14.9	0.0	4.5	7.5	14.2	91.9

Table F.22. Black-box CW attack results on CIFAR plane-vs-frog binary classification.

	substitute acc					victim acc					TP logit				
ϵ	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00	0.10	1.00	10.00	100.00	1000.00
GFZ	96.3	41.5	0.5	0.0	0.0	98.7	94.7	93.4	93.3	92.3	13.3	8.6	7.7	7.7	8.9
GBZ	94.9	73.3	1.4	0.0	0.0	95.9	95.3	94.6	94.5	94.5	13.5	11.8	11.8	13.2	13.1
GFY	96.4	62.0	1.3	0.0	0.0	98.1	96.7	94.8	94.7	94.1	0.0	2.8	6.5	6.1	6.8
GBY	93.5	33.6	1.7	0.0	0.0	97.7	95.9	95.4	95.4	95.0	25.8	16.4	15.9	18.6	19.8
DFX	90.3	10.9	0.0	0.0	0.0	99.9	98.8	98.5	98.5	97.9	N/A	N/A	N/A	N/A	N/A
DBX	87.2	31.8	0.0	0.0	0.0	99.9	97.1	94.5	93.9	94.7	N/A	N/A	N/A	N/A	N/A
DFZ	84.9	22.7	0.0	0.0	0.0	99.9	99.0	98.3	98.3	97.8	0.0	3.8	2.3	2.3	11.0

Table F.23. FGSM white-box attack results on CIFAR-10.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
VGG16	44.5	25.8	16.6	11.9	10.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	95.1	95.7	97.0	97.5	97.1
GBZ-FC1	63.7	49.7	34.3	18.9	12.7	60.0	62.4	80.6	90.9	94.1	63.7	65.0	80.6	95.0	98.4	90.1	90.8	92.3	92.6	92.6
GBY-FC1	64.9	50.0	36.1	19.5	11.3	52.2	59.0	80.4	90.1	92.5	54.6	61.9	80.3	95.6	98.4	90.7	91.3	92.1	92.4	90.3
DBX-FC1	57.3	45.7	31.8	16.3	10.8	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.8	93.2	94.0	94.3	92.8
GBZ-CONV9	81.6	74.1	58.5	48.9	45.8	13.7	17.0	28.4	76.0	90.2	18.3	20.6	31.7	64.3	71.6	91.7	91.5	91.7	93.3	86.4
GBY-CONV9	75.8	66.0	39.7	22.5	17.5	14.0	18.2	29.4	74.4	88.9	17.7	20.8	25.8	55.2	68.9	89.6	89.9	90.0	89.1	89.2
DBX-CONV9	59.0	45.7	31.8	17.8	11.3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	93.7	94.1	94.9	95.2	93.7

Table F.24. PGD white-box attack results on CIFAR-10.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
VGG16	18.8	0.6	0.0	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	93.9	91.8	67.4	69.9	66.4
GBZ-FC1	37.7	6.1	0.1	0.0	0.0	30.7	22.2	76.0	92.0	94.1	33.5	24.9	77.0	91.7	90.2	90.8	93.7	97.3	98.5	95.9
GBY-FC1	31.2	8.5	2.6	1.5	0.6	32.0	20.3	82.3	95.1	97.6	36.6	29.5	84.1	93.7	90.4	90.3	94.1	97.7	97.3	96.9
DBX-FC1	23.4	1.7	0.0	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	91.9	91.0	5.6	7.4	5.9
GBZ-CONV9	61.1	38.3	24.2	10.2	2.3	19.2	60.3	96.7	98.9	99.6	20.1	60.3	96.8	98.3	98.7	90.3	93.1	97.7	98.7	99.6
GBY-CONV9	66.5	26.6	3.2	0.4	0.0	17.0	55.6	95.7	99.2	99.6	17.8	52.7	95.3	98.6	97.6	89.6	91.8	98.7	99.4	96.9
DBX-CONV9	24.0	3.0	0.3	0.1	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	91.3	93.1	99.3	99.6	99.2

Table F.25. MIM white-box attack results on CIFAR-10.

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20	0.01	0.02	0.05	0.10	0.20
VGG16	13.1	0.4	0.0	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.2	92.4	72.3	77.1	79.0
GBZ-FC1	24.5	2.7	0.1	0.0	0.0	34.8	40.9	88.5	95.8	96.0	38.8	44.0	90.4	96.4	96.6	91.3	95.0	95.7	97.9	96.8
GBY-FC1	33.0	19.6	11.8	7.2	4.9	29.1	38.7	95.7	99.1	99.3	38.4	49.2	96.3	99.3	99.4	90.7	95.0	97.8	97.9	97.9
DBX-FC1	17.3	0.9	0.0	0.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	91.1	89.6	9.6	14.5	16.0
GBZ-CONV9	57.6	43.5	29.1	24.1	22.6	27.4	78.5	99.4	99.6	99.7	29.1	78.3	99.6	99.7	99.8	90.2	94.4	97.5	98.0	98.2
GBY-CONV9	39.1	18.1	4.3	1.9	1.5	25.6	70.4	98.6	99.6	99.8	28.2	71.1	98.6	99.7	99.8	89.6	94.5	99.4	99.8	99.7
DBX-CONV9	19.2	3.2	0.5	0.4	0.4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.1	96.7	99.9	100.0	100.0

Table F.26. FGSM white-box attack results on MNIST (with varied bottleneck layer sizes).

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
DBX-16	92.6	85.6	76.4	64.7	52.3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	90.3	92.4	92.3	92.7	94.2
DBX-32	92.6	84.4	71.1	57.2	46.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.2	92.3	92.8	94.9	95.0
DBX-64	91.6	77.8	58.1	44.5	36.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.6	93.6	95.2	96.3	97.0
DBX-128	87.8	47.6	20.1	12.7	10.3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	94.4	97.6	97.8	97.1	96.4
GBZ-16	92.0	75.8	52.5	34.0	21.7	28.7	71.5	99.9	100.0	100.0	49.2	86.3	99.9	100.0	100.0	90.8	90.6	92.0	91.9	90.9
GBZ-32	91.1	74.4	52.4	33.4	21.6	26.5	73.5	99.7	100.0	100.0	49.4	87.2	99.9	100.0	100.0	91.0	90.6	91.8	91.4	91.6
GBZ-64	92.5	80.3	62.0	42.4	27.2	37.0	81.7	100.0	100.0	100.0	57.6	93.5	100.0	100.0	100.0	91.6	90.9	90.3	91.1	91.5
GBZ-128	90.8	76.8	57.5	38.9	24.8	26.6	63.0	99.6	100.0	100.0	44.9	78.7	99.8	100.0	100.0	87.9	90.1	90.7	91.1	90.6
GBY-16	94.2	77.7	49.4	23.9	12.1	41.9	84.1	100.0	100.0	100.0	52.9	91.9	100.0	100.0	100.0	86.6	92.0	93.0	93.1	92.6
GBY-32	94.5	76.9	45.4	20.0	9.6	41.4	83.3	100.0	100.0	100.0	56.3	92.8	100.0	100.0	100.0	89.0	91.6	93.3	93.2	93.0
GBY-64	93.6	76.4	47.5	22.3	10.7	41.9	84.5	100.0	100.0	100.0	57.7	92.5	100.0	100.0	100.0	89.3	92.7	92.8	92.9	92.9
GBY-128	93.0	72.9	42.2	18.5	9.0	37.4	71.6	100.0	100.0	100.0	50.1	81.9	99.9	100.0	100.0	90.7	91.4	91.8	91.4	92.0

Table F.27. PGD white-box attack results on MNIST (with varied bottleneck layer sizes).

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
DBX-16	69.8	36.6	16.0	5.9	1.8	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	91.2	91.9	92.8	93.3	93.7
DBX-32	63.5	26.8	12.5	6.2	2.7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	91.5	92.1	92.7	92.9	93.5
DBX-64	58.0	18.3	6.0	1.3	0.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.7	94.0	94.1	93.7	84.8
DBX-128	42.3	3.0	1.1	0.3	0.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	97.1	97.1	98.4	98.4	98.6
GBZ-16	88.3	62.8	37.4	23.1	15.5	27.1	35.8	71.9	98.9	100.0	46.3	53.0	82.3	99.3	100.0	89.6	91.7	92.3	91.4	91.8
GBZ-32	87.1	61.9	40.2	28.0	20.4	24.0	39.1	76.7	98.9	100.0	41.5	56.4	85.4	99.7	100.0	91.3	91.0	91.1	91.3	91.1
GBZ-64	85.1	57.4	32.5	19.3	11.6	33.7	50.3	84.4	99.7	100.0	52.2	66.3	91.5	99.8	100.0	90.0	91.5	91.5	91.7	91.8
GBZ-128	84.4	57.5	36.8	25.2	16.8	25.4	35.4	66.1	96.6	100.0	41.5	50.5	76.4	97.7	100.0	88.6	90.4	90.4	90.0	90.6
GBY-16	90.4	49.7	15.0	4.2	1.6	42.6	64.7	89.9	99.3	100.0	51.1	70.8	92.8	99.5	100.0	91.9	93.1	91.9	90.7	91.0
GBY-32	88.4	39.8	9.5	1.8	0.6	45.6	74.3	92.8	99.5	100.0	56.8	78.2	94.8	99.8	100.0	90.8	92.2	91.7	91.0	90.6
GBY-64	86.7	35.9	9.0	1.7	0.4	45.3	76.1	94.6	99.7	100.0	56.9	79.3	95.6	99.9	100.0	90.1	92.1	91.6	91.7	91.4
GBY-128	83.3	35.1	8.7	2.3	0.8	39.1	63.9	86.7	98.3	100.0	51.7	69.1	89.2	98.9	100.0	89.8	90.7	90.4	89.7	89.5

Table F.28. MIM white-box attack results on MNIST (with varied bottleneck layer sizes).

	acc. (adv)					TP marginal					TP logit					TP KL				
ϵ	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50	0.10	0.20	0.30	0.40	0.50
DBX-16	72.2	37.1	20.7	14.7	11.7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	91.8	92.3	92.8	93.4	93.9
DBX-32	66.7	27.6	16.8	12.5	10.6	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	91.8	92.5	92.4	93.2	93.5
DBX-64	66.7	28.7	19.7	17.2	17.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	93.5	94.3	94.6	94.7	94.7
DBX-128	42.3	1.3	0.2	0.1	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	97.2	97.6	98.6	98.9	99.1
GBZ-16	83.8	36.4	8.9	2.3	0.9	26.9	79.8	99.7	100.0	100.0	43.3	88.0	99.8	100.0	100.0	91.6	92.1	91.2	91.6	92.6
GBZ-32	82.7	35.5	9.5	3.0	1.1	26.6	83.2	99.7	100.0	100.0	45.8	90.0	99.8	100.0	100.0	90.3	91.7	91.4	92.5	92.6
GBZ-64	79.6	27.4	5.6	1.5	0.5	36.1	86.3	100.0	100.0	100.0	53.5	93.1	99.9	100.0	100.0	91.3	91.4	91.9	92.2	92.7
GBZ-128	79.2	32.1	8.7	2.6	1.2	26.1	68.5	99.2	100.0	100.0	41.2	79.1	99.4	100.0	100.0	89.7	91.5	91.4	91.4	92.5
GBY-16	89.2	34.6	3.5	0.2	0.0	44.8	88.1	99.8	100.0	100.0	55.9	91.7	99.7	100.0	100.0	89.4	92.5	91.0	90.6	36.9
GBY-32	86.7	25.8	1.6	0.1	0.0	45.1	91.2	100.0	100.0	100.0	58.6	94.0	100.0	100.0	100.0	90.9	91.8	91.1	68.6	31.1
GBY-64	84.9	22.9	1.5	0.1	0.0	47.4	91.2	99.9	100.0	100.0	59.8	92.8	99.9	100.0	100.0	91.4	92.1	91.2	86.8	29.0
GBY-128	82.1	23.7	1.9	0.0	0.0	40.4	83.0	99.7	100.0	100.0	51.7	87.5	99.6	100.0	100.0	91.2	91.2	89.8	82.8	44.4