Hide and Seek: on the Stealthiness of Attacks against Deep Learning Systems

Zeyan Liu[1], Fengjun Li[1], Jingqiang Lin[2], Zhu Li[3], and Bo Luo[1]

[1] EECS/I2S, University of Kansas, Lawrence, KS, USA {zyliu, fli, bluo}@ku.edu
[2] University of Science and Technology of China, Hefei, China linjq@ustc.edu.cn
[3] University of Missouri–Kansas City, Kansas City, MO, USA lizhu@umkc.edu

Abstract. With the growing popularity of artificial intelligence (AI) and machine learning (ML), a wide spectrum of attacks against deep learning (DL) models have been proposed in the literature. Both the evasion attacks and the poisoning attacks attempt to utilize adversarially altered samples to fool the victim model to misclassify the adversarial sample. While such attacks claim to be or are expected to be stealthy, i.e., imperceptible to human eyes, such claims are rarely evaluated. In this paper, we present the first large-scale study on the stealthiness of adversarial samples used in the attacks against deep learning. We have implemented 20 representative adversarial ML attacks on six popular benchmarking datasets. We evaluate the stealthiness of the attack samples using two complementary approaches: (1) a numerical study that adopts 24 metrics for image similarity or quality assessment; and (2) a user study of 3 sets of questionnaires that has collected 30,000+ annotations from 1,500+ responses. Our results show that the majority of the existing attacks introduce non-negligible perturbations that are not stealthy to human eyes. We further analyze the factors that contribute to attack stealthiness. We examine the correlation between the numerical analysis and the user studies, and demonstrate that some image quality metrics may provide useful guidance in attack designs, while there is still a significant gap between assessed image quality and visual stealthiness of attacks.

Keywords: Adversarial machine learning · Attacks.

1 Introduction

In the past decade, machine learning, especially deep learning (DL), has gained incredible success in a wide range of applications, fueling advances in every field related to big data analysis, such as computer vision, data mining, and natural language processing. With the growing popularity and adoption of DL, a wide spectrum of attacks have been proposed. In particular, attacks against the *integrity* of deep learning models could be roughly grouped into two categories: *evasion attacks* and *backdoor attacks*. In the evasion attacks, supposedly imperceptible adversarial perturbations are added to the attack samples, so that the victim models would make highly confident but erroneous classifications for these

samples. In the backdoor attacks, the victim DNNs are compromised through poisoning or Trojaning, so that they "remember" the specially-crafted triggers (e.g., patches of pixels, shadows, or stealthy noises) as external features and classify the trigger-embedded images into wrong labels, e.g, to mis-recognize a stop sign with a yellow sticker on it as a "go straight" sign. The backdoor attacks could be further categorized into data poisoning backdoors and neural Trojans. The data poisoning backdoors inject digitally altered and mislabeled samples into the training dataset so that a malicious functionality is "learned" by the victim model. Meanwhile, neural Trojans alter the structure of the victim DNN by injecting a malicious sub-network that only responds to the adversarial triggers in the testing samples, so that the original task remains (mostly) unaffected.

The majority of the existing evasion and backdoor attacks are supposed to be or claim to be *stealthy*, i.e., the adversarial perturbations and the attack triggers are expected to be hardly noticeable to human eyes, so that the attacks are unlikely to be identified even if the administrators manually examine the training/testing data. However, the quality of the adversarial samples has not been carefully examined in the literature, while the stealthiness claims are rarely measured in the attack papers. To the best of our knowledge, very few existing attacks employed human evaluators to assess the stealthiness of the adversarial modifications to the attack samples [33, 9] or used numerical measurements to assess the similarity between the attack and the benign images [51, 25, 33].

In this paper, we are motivated by the questions: for the machine learning attacks proposed in the literature, how *stealthy* are they? How can we *quantitatively assess the attack stealthiness*? In particular, we aim to measure the *numerical stealthiness* and the *user-perceived stealthiness*. We have implemented 20 evasion and backdoor attacks over six popular benchmarking datasets. In *numerical analysis*, we adopt 24 metrics from the literature to assess the similarities between the attack images and the corresponding benign images, or to assess the visual quality of the attack images. Some of these metrics are supposed to reflect the human visual systems or human perceptions of digital images. For *user-perceived stealthiness*, we present a large-scale user study with 3 sets of questionnaires of 1,500+ responses and 30,000+ annotated images, which aim to answer the questions: could users notice the differences between the original and the adversarial samples in different attacks? If a knowledgeable or novice user is presented with attack images without the corresponding benign image, could the user identify the malicious images? We further correlate the numerical assessments with the user-perceived stealthiness and discuss our findings.

Our contributions are: (1) We identify a largely neglected issue in adversarial ML that, while the attack samples are supposed to be or claimed to be stealthy, such features are rarely evaluated in the proposed attacks. (2) We present the first large-scale assessment and comparative study to evaluate the stealthiness of the attack samples through numerical analysis and user study. And (3) Our findings are expected to provide a better understanding of attacks against ML models, especially on how users perceive the maliciousness of the attack samples, and how auditors could benefit from our findings.

Ethical Considerations. The objective of this project is to enhance the understanding of the adversarial attacks against deep learning models. All the attacks implemented in this project are previously published in the literature. All the implementations and experiments were conducted in a lab environment. We did not attack any real-world system. The user studies presented in the paper have been reviewed and approved by the Human Research Protection Program at the University of Kansas under STUDY00148002 and STUDY00148622.

The rest of the paper is organized as follows: we introduce the background of this project in Section 2. We present the design and results of the numerical analysis and the user study in Sections 3 and 4, respectively, followed by discussions in Section 5. We conclude the paper in Section 6.

2 Preliminaries: Deep Learning, Attacks, and Datasets

2.1 Deep Learning and Adversarial Machine Learning

Deep Neural Networks (DNNs). In this paper, we focus on the standard classification task. Without loss of generality, a DNN model is defined as $f_\theta : X \to \mathcal{Y}$. f_θ maps d-dimensional inputs $x \in \mathcal{X} \subset \mathbb{R}^d$ into a label space $y \in \mathcal{Y} \subset \mathbb{N}+^k$ with k categories. Its decision making process is defined as: $p \in \mathbb{R}^k := f_\theta(x)$, where p is the k-dimensional embedding of the decision confidence in terms of the label distribution. The final output label corresponds to the maximum element in p. Given an annotated training dataset $D_t = \{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{Y}, 1 \leq i \leq k\}$, the training process is to optimize the model parameters with a loss function $\mathcal{L}$, e.g., Cross-Entropy Loss: $\theta_{optimal} = \arg\min_\theta \sum_{i=1}^{k} \mathcal{L}(f_\theta(x_i), y_i)$.

Attacks against Deep Learning Models. Attacks against ML/DL models could be roughly categorized into three types [4]: (1) *evasion attacks*, (2) *backdoor attacks*, and (3) *exploratory attacks*. Both evasion and backdoor attacks attempt to break the integrity of ML/DL models using adversarial examples that are seemingly benign, while exploratory attacks aim to break the confidentiality of the proprietary black-box models. In this paper, we focus on evasion and backdoor attacks since the majority of such attacks, implicitly or explicitly, make the stealthiness assumption on the adversarial examples.

Evasion Attacks. Also known as adversarial examples, the evasion attacks add small perturbation δ to clean sample x_{beni} to generate the adversarial example x_{adv}. δ is crafted to satisfy two objectives: (1) the magnitude of δ is limited by a perturbation budget (e.g. restrictions on l^p norm). And (2) x_{adv} will mislead the model to generate a wrong output y_{adv}, which could be untargeted (any wrong label) or targeted (a pre-selected label). The attack is defined as:

$$y_{\text{adv}} = \underset{i}{\text{argmax}}\, f_\theta(x_{\text{adv}}) \neq \underset{i}{\text{argmax}}\, f_\theta(x_{\text{beni}}) \tag{1}$$

Backdoor Attacks. While the evasion attacks attempt to fool a benign (but vulnerable) model, backdoor attacks inject the malicious functionality into the victim DNN during model production. The attacker designs a trigger function

T to transform benign examples x_{beni} into malicious examples $T(x_{\text{beni}})$. The backdoored model f'_θ learns the trigger and associates it with the target label y_t, and classifies $T(x_{\text{beni}})$ into y_t in the testing phase:

$$y_t = \underset{i}{\text{argmax}}\, f'_\theta(T(x_{beni})) \tag{2}$$

With different attack models, backdoors can be injected by poisoning the training data [5, 18, 51], manually permuting the model parameters [12], or injecting malicious sub-networks into the model (Trojaning) [26, 38].

2.2 DNN Attacks Evaluated in This Study

We have implemented 20 attacks (23 different settings) from the literature, including 10 evasion attacks and 10 backdoor attacks. Unless otherwise specified, we strictly follow the settings in the original papers. We briefly summarize the attacks with a special focus on adversarial perturbations or backdoor triggers.

Evasion Attacks. For evasion attacks that need to specify perturbation budgets, the default budget is set to 8 (out of 255 in each RGB channel).

• *FGSM.* FGSM [17] is an untargeted attack that adds perturbations along the gradient sign: $\delta = \epsilon \cdot sign(\nabla_x \mathcal{L}(f_\theta(x, y)))$, where ϵ is the perturbation budget.

• *BIM.* BIM [24] proposed a targeted variant of FGSM by iteratively updating perturbations: $x_{t+1} = x_t + \alpha \cdot sign(\nabla_{x_t} \mathcal{L}(f_\theta(x_t, y_{adv}))$, where x_t denotes x_{adv} at the t^{th} iteration, and α denotes the step size each iteration.

• *MI-FGSM.* Momentum-based iterative algorithms are proposed in [10], in which the accumulated gradient is defined as: $g_{t+1} = \mu \cdot g_t + \frac{\nabla_{x_t} \mathcal{L}(f_\theta(x_t, y_{adv}))}{\|\nabla_{x_t} \mathcal{L}(f_\theta(x_t, y_{adv}))\|_1}$, and the iterative update is: $x_{t+1} = x_t + \alpha \cdot sign(g_{t+1})$.

• *PGD.* PGD [29] improves BIM with an initialization with random noise and more iterations. The perturbation is projected onto an l^∞ ball with radius ϵ.

• *AutoPGD (APGD).* AutoPGD [7] extends PGD by combining momentum and alternative loss functions into an efficient ensemble of attacks. AutoPGD dynamically adjusts all the parameters.

• *FFGSM.* FFGSM [45] assumes a weaker/cheaper adversary but significantly improves the efficiency of adversarial training (defense). FFGSM can be considered as a one-step variant of PGD that is initialized with uniform perturbation.

• *Deepfool (DF).* Deepfool [31] computes the minimal adversarial perturbation to push the adversarial sample towards the linearized decision boundary. The gradients are calculated as: $x_{t+1} = x_t + \frac{|f_{\theta,adv}(x_t) - f_{\theta,beni}(x_t)|}{\|\nabla f_{\theta,adv}(x_t) - \nabla f_{\theta,beni}(x_t)\|_2^2}$, where $f_{\theta,adv}$ and $f_{\theta,beni}$ denote the model outputs corresponding to the target/original label.

• *Carlini & Wagner (CW).* C&W [3] finds the minimum perturbation through an optimization approach. We adopt the l^2 constraint in the optimization objective.

• *Smoothfool (SF).* Smoothfool [8] uses low-pass filters to generate smooth adversarial perturbations, which improves attack robustness and transferability.

• *Semantic AE (SAE).* Instead of generating artificial perturbations at the pixel level, [20] semantically modifies the image by converting it into HSV space, and randomly shifting the Hue and Saturation components.

Table 1. A qualitative comparison of the backdoor attacks.

	Datasets	CL	Attack Type	Trigger Category	DB	Training Method		
						SC	FA	FL
AD	ⓑⓒ	×	One-to-One	Global Noise	×	✓	✓	×
BN	ⓐⓑⓓ	×	All-to-One/All-to-All	Local Patch	✓	✓	✓	×
BL	ⓔ	×	One-to-One	Global Transformation	✓	✓	✓	×
PA	ⓔ	×	One-to-One	Local Patch	✓	✓	✓	×
HT	ⓒⓔ	✓	One-to-One/All-to-One	Global Noise	×	×	×	✓
INS	ⓑⓒⓔ	×	All-to-One	Global Transformation	×	✓	✓	×
IP-S	ⓐⓑⓒ	×	One-to-One	Global Transformation	✓	✓	✓	×
IP-A	ⓐⓑⓒ	×	One-to-One	Global Noise	✓	✓	✓	×
RP	ⓐⓑ	✓	All-to-One	Global Transformation	×	✓	✓	×
WM	ⓖⓗ	–	All-to-One	Local Patch	✓	×	×	✓
SQ	ⓖⓗ	–	All-to-One	Local Patch	✓	×	×	✓
TNet	ⓑⓔⓕⓙ	–	All-to-One/All-to-All	Local Patch	✓	-	-	-
WN	ⓐⓑⓒⓘ	×	All-to-One	Global Transformation	×	✓	✓	×

Datasets in the published paper: ⓐ MNIST. ⓑ GTSRB. ⓒ CIFAR10. ⓓ US & Swedish Traffic Signs. ⓔ ImageNet. ⓕ YouTube Aligned Face. ⓖ VGG Face. ⓗ LFW. ⓘ CelebA. ⓙ Pubfig.
CL: Clean label poisoning. **DB**: Whether the benign dataset is unavailable. **SC**: Train the model from scratch. **FA**: Finetune all the layers. **FL**: Finetune top layers with other layers frozen.

Backdoor Attacks. A detailed comparison of the backdoor attack models and settings is shown in Table 1.

• *Advdoor (AD).* Advdoor [49] utilizes a Targeted Universal Adversarial Perturbation (TUAP) trigger, which is claimed to be "small magnitude and human imperceptible". Input-specific perturbations are iteratively aggregated to generate UAP, where existing evasion attacks (e.g., Deepfool) could be employed.

• *Badnets (BN).* Badnets [18] patch a unique black-and-white square on the corner of training samples. We follow the settings in [18] and use a 4×4 trigger.

• *Blended & Physical Accessory (BL & PA).* [5] employs two trigger generation strategies: blended injection (BL) for simple digital attacks and physical accessory injection (PA) for real-world physical attacks. We follow [5] to use the hello kitty image in BL, and follow [47] to use mug patches in PA.

• *HiddenTrigger (HT).* HT [36] is a clean-label backdoor, which claims the poisoned samples to be "natural with correct labels", while the triggers are hidden until test time. In HT, we assess the stealthiness of poisoned training samples.

• *Instagram (INS).* [27] proposed feature-space attacks that generate triggers with Instagram filters Nashville and Gotham. We implement the Nashville filter.

• *Invisible Perturbation (IP).* [51] propose two strategies to generate "hardly perceptible" backdoor triggers: patterned static perturbation masks (IP-S) and targeted adaptive perturbation masks (IP-A).

• *Ramp Signal (RP).* [2] uses ramp signals as triggers in the poisoned samples.

• *TrojanNN (TN).* [28] use reverse-engineered adversarial triggers to maximize the activation of anchor neurons in the DNN. Among the trigger patterns in [28], we have implemented Watermark (WM) and Square (SQ) in this study.

• *TrojanNet (TNet).* [38] injects into the victim DNN a small neural component, which is activated by a predefined trigger pattern to flip the DNN output. We follow the settings in [38] to implement a 4×4 black-and-white square trigger.

• *WaNet (WN).* [33] uses elastic image warping as triggers. Human inspection experiments in [33] demonstrate its outstanding stealthiness.

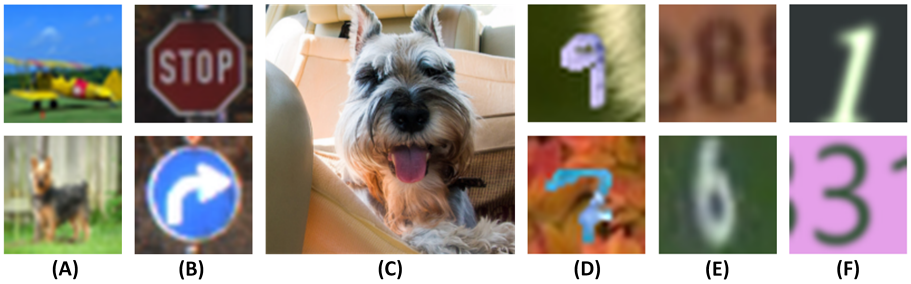

Fig. 1. Sample images from the benchmarking datasets used in this study.

Attack categorization. Based on the properties of the injected adversarial perturbations, we classify the attacks into three categories: (1) Gradient-based noise (or global noise): FGSM, BIM, PGD, MI-FGSM, DF, FFGSM, APGD, CW, SF, AD, HT, IP-A. They make small modifications to a large number of pixels to achieve a global optimization objective. They attack the unrobust deep features in the victim DNN, while the main contents (the semantic meanings) of the adversarial images remain unchanged. (2) Local patches: BN, PA, WM, SQ, TNet. They are all backdoor attacks that modify a small region on adversarial images to build salient visual features. The triggers are robust, but they are usually detectable by trigger reconstruction defenses, e.g., [40, 19]. (3) Global transformation: SAE, BL, INS, IP-S, RP, WN. They make global changes to the victim images, which may significantly change the appearances of the images.

2.3 The Datasets

In this study, we adopt six datasets that are frequently used as benchmarks in both the machine learning community and the adversarial ML community. Please note that, while most of the attacks were originally implemented on a subset of these datasets, we re-implemented all of them on all six datasets.

A. CIFAR-10. The CIFAR dataset [23] contains 60K tiny images (32×32) labeled in 10 classes, such as automobile, bird, cat, ship, truck, etc.
B. GTSRB. The German Traffic Sign Recognition Benchmark dataset [21] has 51K images of real-world traffic signs in 43 classes. Images are scaled to 32×32.
C. ImageNet. ImageNet is a large-scale image classification and object recognition benchmark dataset. The dataset used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012-14 contains 1.35 million images that are manually labeled into 1,000 classes [35]. The images are usually cropped to 224×224. They are significantly larger than the small images in other datasets.
D. MNIST-M. The MNIST dataset[4] contains 70,000 binary images of handwritten digits. The MNIST-M dataset adds color to MNIST by extracting patches from the BSDS500 dataset and combining them with MNIST [16].

[4] Available at: http://yann.lecun.com/exdb/mnist/

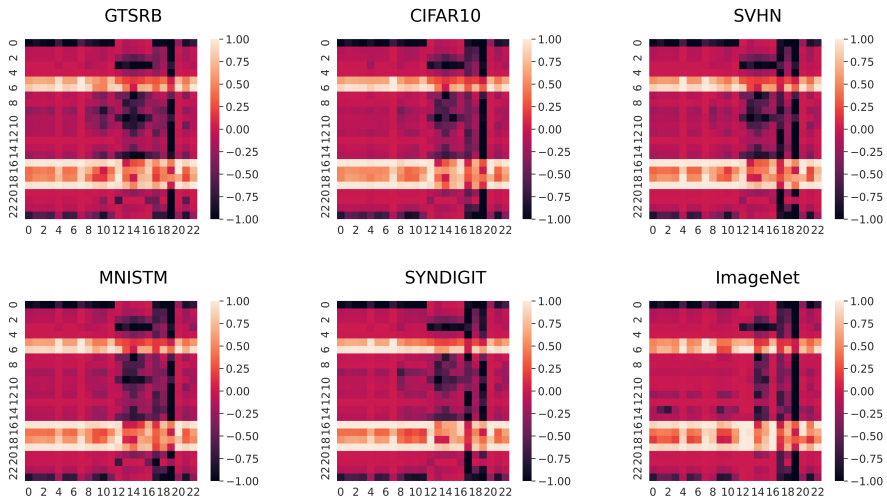

Fig. 2. Normalized numerical assessment results. X-axis: attacks. Y-axis: metrics.

E. SVHN. The Street View House Numbers (SVHN) is another digit classification benchmark dataset, which contains 600,000 digit images that are extracted from house numbers in Google Street View [32].
F. SYNDIGIT. The Synthetic Digits dataset contains 12,000 images that are synthetically generated by placing digits of various fonts, colors, and directions over random backgrounds [15]. The MNIST-M, SVHN, and SYNDIGIT datasets are often used in transfer learning and attacks against transfer learning.

Samples of benign images from the datasets are shown in Figure 1. The corresponding attack images are generated by exploiting the attacks introduced in Section 2.2 on images sampled from the six datasets. They may contain noise (a.k.a. adversarial perturbation) that is supposed to be invisible to human eyes, or adversarial patterns such as a sticky note on a stop sign.

3 The Numerical Analysis

3.1 Metrics and Basic Statistics

In this paper, we adopt three categories of metrics in the numerical analysis.
l^p Norms: l^p norms are widely used in adversarial example attacks to limit the magnitude of the perturbation. In the rest of the paper, we always use l^p norms normalized by image size, i.e., $l^p(x_{\text{adv}} - x_{\text{beni}})/|x_{\text{beni}}|$, where $p = \{0, 1, 2, \infty\}$.
Basic Metrics: Most popular metrics in the literature for image similarities: MSE (mean square error), PSNR (peak signal-to-noise ratio), and SSIM [42].
Additional Metrics: To further explore how the advanced image similarity or quality measurements may fit into this problem, we have implemented 17 additional metrics. They are roughly categorized into 5 groups:

Table 2. Image similarity/quality measurements for evasion attacks. Blue: the best attack (most similar or highest quality) for each metric; red: the worst attack.

Metrics	Ideal	Better	FGSM	BIM	PGD	MIFGSM	DF	FFGSM	APGD	CW	SF
l^0	0	small	0.991	0.863	0.971	0.992	0.709	0.994	0.971	0.647	0.829
l^1	0	small	7.632	5.189	5.320	6.282	2.286	6.630	6.068	1.153	4.484
l^2	0	small	0.130	0.101	0.101	0.114	0.060	0.116	0.112	0.032	0.121
l^∞	0	small	0.003	0.003	0.003	0.003	0.007	0.003	0.003	0.004	0.012
MSE	0	small	59.20	34.96	35.07	45.09	24.55	47.48	42.51	3.686	65.26
PSNR	∞	large	30.42	32.76	32.75	31.61	40.98	31.37	31.92	44.32	34.40
SSIM	1	large	0.928	0.957	0.957	0.943	0.984	0.940	0.953	0.996	0.963
LPIPS	0	small	0.011	0.005	0.005	0.007	0.001	0.008	0.006	1.9e-4	0.003
aHash	0	small	0.642	0.625	0.533	0.670	0.405	0.587	0.660	0.320	1.442
dHash	0	small	3.035	2.727	2.642	2.908	1.633	2.663	2.973	1.245	4.508
pHash	0	small	1.813	1.620	1.590	1.763	0.800	1.707	1.700	0.610	3.780
wHash	0	small	1.613	1.400	1.318	1.477	0.658	1.473	1.438	0.520	2.527
FDL	0	small	46.28	34.56	34.64	39.96	15.65	41.51	7.49	37.73	21.62
FFL	0	small	35.27	20.84	20.92	26.76	15.11	28.32	25.31	2.233	46.95
ERGAS	0	small	5108	8340	1.0e4	5319	1971	5192	4282	1153	4786
SAM	0	small	0.080	0.060	0.059	0.067	0.030	0.070	0.065	0.016	0.057
MS-SSIM	1	large	0.992	0.995	0.995	0.993	0.998	0.993	0.994	0.999	0.983
FSIM	1	large	0.767	0.782	0.783	0.768	0.888	0.768	0.775	0.920	0.848
VIF	1	$\sim$1	0.513	0.587	0.587	0.551	0.752	0.543	0.561	0.849	0.600
UQI	1	large	0.990	0.990	0.991	0.990	0.997	0.991	0.992	0.998	0.990
Elastic	0	small	17.95	11.14	11.27	14.04	6.74	14.80	13.36	1.659	16.64
RE	0	small	4.457	2.673	2.509	3.974	1.575	3.626	3.088	0.038	0.495
l^1 Cos	0	small	46.07	31.35	32.12	37.96	13.76	40.02	36.63	6.952	27.05
l^1 Clip	0	small	7.632	5.189	5.320	6.282	1.919	6.630	6.068	1.146	3.406

(1) Simple spatial domain metrics: Average Hash (aHash) [13], Difference Hash (dHash) [13] and variants of l^p norms: relative error in l^1 (RE), Elastic-Net regularization [14], l^1 with cosine similarity [30] and l^1 with clip function [44].

(2) Structural similarity/quality metrics: Universal Image Quality index (UQI) [41] and Multi-scale SSIM (MS-SSIM) [43].

(3) Transform domain metrics: Perceptual Hash (pHash) [13], Wavelet Hash (wHash), Visual Information Fidelity (VIF) [37], Feature Similarity index (FSIM) [48], l^1 loss in the frequency domain (FDL) and focal frequency loss (FFL) [22].

(4) Spectral similarity: ERGAS [39] and SAM[46].

(5) Metrics based on deep learning: LPIPS [50].

For each attack, we have generated 1,000 adversarial images for each dataset. We employ the metrics introduced above to assess the attack images and report the mean assessment values in Tables 2 and 3. In Table 2, we provide a high-level illustration of how the numbers could be interpreted. Column "ideal" denotes the output of the metric on two identical images, while "better" denotes whether a larger or a smaller value indicates "more similar" images ("better" attack). For example, MSE generates 0 for two identical images, and a smaller value corresponds to images with higher similarity, i.e., the attack is stealthier in this metric. Meanwhile, Figure 2 shows the detailed distribution in each dataset after min-max normalization, in which a lighter color ($\sim$ 1) indicates a stealthier attack. As shown in the figure, the measurements for each attack demonstrate very similar patterns across six datasets, i.e., the dataset is NOT a significant factor in numerical assessments of attack stealthiness.

Table 3. Image similarity/quality measurements for backdoor attacks. Blue: the best attack (most similar or highest quality) for each metric; red: the worst attack.

Metrics	Global Noises			Local Patches					Global Transformation					
	AD	HT	IPA	BN	PA	WM	SQ	TNet	SAE	BL	INS	IPS	RP	WN
l^0	0.937	0.973	0.946	0.015	0.136	0.078	0.057	0.015	0.888	0.983	0.991	0.263	0.878	0.731
l^1	8.531	9.866	5.028	1.993	7.822	7.884	5.438	2.102	35.61	24.50	49.38	2.471	11.81	3.413
l^2	0.175	0.188	0.104	0.325	0.472	0.588	0.499	0.334	0.829	0.473	1.037	0.083	0.226	0.105
l^∞	0.006	0.005	0.004	0.068	0.057	0.072	0.069	0.069	0.040	0.020	0.060	0.003	0.006	0.013
MSE	113.2	121.0	48.12	333.0	751.8	1302	948.4	355.1	3231	944.3	4225	24.46	181.8	45.86
PSNR	28.09	27.37	34.59	25.21	22.39	17.40	19.01	24.97	16.51	19.78	12.57	34.25	25.55	33.61
SSIM	0.864	0.868	0.962	0.953	0.784	0.660	0.874	0.898	0.646	0.898	0.722	0.973	0.873	0.983
LPIPS	0.019	0.022	0.007	0.034	0.037	0.122	0.072	0.021	0.066	0.007	0.083	8.7e-4	0.021	0.003
aHash	0.972	1.376	0.553	1.670	6.810	6.461	6.288	1.842	5.170	2.003	11.71	0.362	0.610	1.157
dHash	3.927	4.664	2.777	2.453	6.782	8.148	4.372	2.837	6.350	5.158	11.83	1.223	3.473	2.708
pHash	2.510	3.275	1.620	7.747	12.85	10.93	12.95	7.993	6.383	3.897	11.21	0.360	1.513	2.660
wHash	2.248	2.688	1.487	1.687	7.028	7.662	6.258	1.873	5.545	3.358	13.36	0.437	4.465	1.725
FDL	58.69	64.04	29.68	14.00	43.56	136.1	60.64	18.25	171.8	68.57	220.6	12.14	36.02	23.11
FFL	65.98	71.36	28.38	283.3	546.8	657.3	525.4	278.0	2382	806.7	3195	24.24	179.2	32.32
ERGAS	1.3e4	8865	6329	2549	1.0e4	2.7e4	1.2e4	5670	2.0e4	1.5e4	4.3e4	4986	1.0e4	3070
SAM	0.110	0.125	0.060	0.171	0.225	0.313	0.288	0.177	0.289	0.082	0.381	0.045	0.135	0.053
MS-SSIM	0.980	0.982	0.996	0.985	0.818	0.842	0.874	0.957	0.805	0.952	0.881	0.999	0.966	0.996
FSIM	0.734	0.696	0.790	0.960	0.813	0.794	0.840	0.948	0.704	0.821	0.616	0.992	0.747	0.870
VIF	0.393	0.410	0.626	0.879	0.433	0.264	0.679	0.697	0.363	0.587	0.528	0.670	0.405	0.619
UQI	0.979	0.979	0.991	0.988	0.948	0.899	0.959	0.972	0.781	0.939	0.799	0.992	0.969	0.996
Elastic	29.47	32.09	13.65	68.20	156.6	266.6	194.0	72.70	674.6	208.5	884.5	6.869	45.80	11.90
RE	7.872	7.966	4.430	80.77	0.114	0.065	0.065	50.28	256.6	0.266	153.8	0.072	21.41	0.064
l^1 Cosine	51.58	59.55	30.38	12.19	47.79	49.05	33.75	12.83	215.5	147.2	297.8	14.91	71.21	20.58
l^1 Clip	6.406	7.857	4.213	0.153	1.138	0.608	0.393	0.153	7.356	8.466	9.307	2.467	7.896	2.682

3.2 Numerical Analysis of Attack Stealthiness

l^p Norms (normalized by image size) and attack categorization. The l^0 norm indicates the portion of pixels that are modified, while the l^∞ norm indicates the maximum change made to a pixel. Both l^1 and l^2 denote the overall amplitude of the added adversarial perturbations, while l^2 is more sensitive to large changes. As shown in Tables 2 and 3, the l^p norms are good indicators of the types of the added perturbations: (1) In general, global noise attacks always produce large l^0 (close to 1) but moderate l^1, l^2, and smallest l^∞ (mean=0.0044). By design, they introduce a small change (often explicitly limited by l^p norm) to each pixel. (2) Local patches generate smallest l^0 (mean=0.06) and largest l^∞ (mean=0.067). The mean l^1 is smaller than noise-based attacks', while the mean l^2 is larger. Although local patches are small in size, they make significant modifications to the pixel values in order to build strong visual features, which significantly impact the l^2 and l^∞ norms. (3) Global transformation attacks introduce moderate changes globally, which result larger l^0, l^1, l^2, but moderate l^∞ (mean=0.024). IP-S is an exception in that it employs a scattered global pattern, so that it moderately modifies a relatively small portion of pixels ($l^0 = 0.263$, $l^1 = 2.471$) distributed over the entire image.

MSE, PNSR, and SSIM. CW produces the best overall MSE (3.686) and PNSR (44.32). Gradient-based attacks all result in smaller MSE ([3.686, 121]), local patch backdoors produce moderate MSE ([333, 1302]), while global transformation attacks generate variable MSE (three attacks in [24.46, 181.8] and others in [944.3, 4225]). PNSR generates very similar rankings and patterns as

MSE. SSIM imitates the Human Visual System (HVS) to measure the structural similarity based on luminance, contrast, and structure. CW again produces the best SSIM (0.996). Among local patch backdoors, BN generates the best SSIM of 0.953, while WM gives the worst SSIM of 0.660. Among global transformation attacks, WN provides the best SSIM of 0.983, while SAE provides the worst SSIM of 0.646. In our experiments, both PNSR and SSIM are highly (negatively) correlated with MSE, with Spearman correlation coefficients of -0.92 and -0.86.

In summary, noise-based attacks significantly outperform the other two categories in all three metrics with averages of 53.35, 33.38, and, 0.943 respectively. Two global transformation attacks, SAE and INS, perform the worst in all three metrics, which is understandable since they have very different design philosophies that do not seek to minimize perturbation.

Additional Image Quality Metrics. IP-S appears to be the most stealthy with dHash, pHash, wHash, and FSIM. BN performs the best with VIF and l^1, while CW is the best for all other measurements. In general, gradient noise attacks perform better than the other two categories in most cases. IP-S again appears to be an outlier among global transformation attacks. Most of the metrics demonstrate strong correlations with at least one of MSE, PNSR, or SSIM (absolute correlation coefficient > 0.8). These metrics will be further explored in comparison with the results from the user study.

4 The User Study

In this section, we evaluate the likelihood of adversarial samples escaping human users. We mimic an auditing scenario, in which auditors/defenders examine potentially suspicious images in an attempt to identify true attack samples.

4.1 Research Design

Based on the level of knowledge the auditor/defender has about the dataset, we define three defense models:

Model I. Informed Defenders (White Box). The Informed Defenders have full knowledge of the target dataset. They are capable of making side-by-side comparisons of the original and the adversarial samples in a white-box manner. They represent the strongest type of defenders, who more likely exist as the owners of the original datasets in defense against training data poisoning.

Model II. Knowledgeable Defenders (Grey Box). The Knowledgeable Defenders have reasonable knowledge about the benign dataset. When they manually inspect the potentially adversarial samples, they can refer to benign samples from the same dataset. However, they do not possess the exact original images that correspond to the suspect samples, i.e., no side-by-side comparison.

Model III. Referenceless Defenders (Black Box). The Referenceless Defenders have little or no access to the original dataset. They inspect the potentially adversarial samples without referring to benign samples. They represent

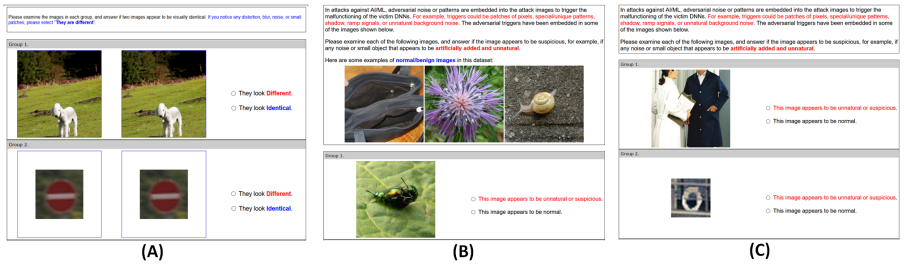

Fig. 3. Sample questionnaires used in the user study.

the weakest type of defenders, who often exist among the novice users, who download and reuse pre-trained models from the Internet, e.g., Model Zoo.

In Defender Model I, attack stealthiness is defined as the *invisibility* of the adversarial perturbations–when the auditor notices the noise/patch, the attack is likely discovered. In Models II and III, attack stealthiness is defined as *visual benignness*, i.e., the attack examples are expected to be visually consistent with the examiners' psychophysical perceptions of normal images. The perceptions will differ with or without knowing the dataset.

To mimic each type of defenders, we have designed three user studies[5]. In each user study, an IRB information statement is first displayed to the participant, followed by a link to the questionnaire. In each questionnaire, 20 (sets of) images are presented to the participant for inspection. ImageNet images are downsized to 200×200, while the smaller images (e.g., GTSRB, CIFAR) are enlarged to 4× the original size, and padded with white padding to 200×200.

Exp I. Informed Defenders (White Box). The questionnaire contains 20 pairs of images: the first in each pair is a benign image randomly selected from the datasets introduced in Section 2.3; the second is the adversarial image that is altered by a random attack in Section 2.2. For quality control, we include 10% benign images as the second, i.e., two images are numerically identical. Part of the questionnaire is shown in Figure 3 (A). We ask the user to inspect each pair and identify whether the two images look identical.

Exp II. Knowledgeable Defenders (Grey Box). As shown in Figure 3 (B), the questionnaire first displays three benign images from a random dataset to be used by the participant as reference images. The questionnaire then shows 20 different images from *the same dataset* as the reference images. For each image, we ask the participant to select if it appears to be benign or suspicious.

Exp III. Referenceless Defenders (Black Box). As shown in Figure 3 (C), the questionnaire is very similar to the Knowledgeable Defenders' questionnaire, except that no reference image is displayed.

[5] The user studies have been approved by the Human Research Protection Program at the University of Kansas under STUDY00148002 and STUDY00148622.

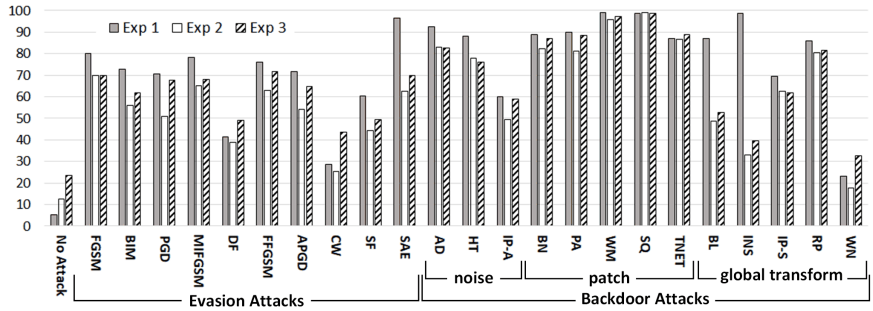

Fig. 4. Detection rates (%) from user studies. Exp I: informed defender; II: knowledgeable defender; III: referenceless defender. A lower rate means a stealthier attack.

4.2 The User-Perceived Attack Stealthiness

We sent the questionnaires to senior/graduate CS students in three institutions which the authors are affiliated with. We intentionally select CS students because they have basic understanding of computer systems and programming, while some of them have prior knowledge of AI/ML or even adversarial ML. They better mimic the system administrators or ML/AI users in real world applications. In six weeks, we collected 1,526 responses with 30,369 total annotations, including 504, 512, and 510 responses, with 9,991, 10,209, and 10,169 annotations to Experiments 1, 2 and 3, respectively.

We present the *detection rates* from all the experiments in Fig. 4. The detection rate is defined as the proportion of annotations that labeled a pair of images as "different" (Exp 1) or labeled an image as "unnatural or malicious" (Exp 2 and 3). A lower detection rate indicates a stealthier attack.

Exp I. Informed Defenders. The average detection rate is 75.9%. The detection rates of 17 attacks were higher than 70%, while 11 attacks were detected in more than 85% of the tests. The statistics show that majority of the adversarial samples are highly noticeable, i.e., not stealthy, in side-by-side comparisons with the corresponding benign images. The median task completion time was 91 seconds (for 20 image pairs), indicating the level of effort needed in an audit of the training or testing dataset. WN and CW significantly outperform the other attacks, with detection rates at 28.5% and 23.3%, while detection rates of SQ and WM are as high as 98.9% and 98.7%. In general, noise-based evasion attacks appear to be stealthier than the others with a 62.0% mean detection rate, while patch-based backdoors are the least stealthy with a 92.7% mean detection rate.

Exp II. Knowledgeable Defenders. The average detection rate drops to 62.2%, i.e., the adversarial images have better chances to escape human auditors. Detection rates decrease in every attack except SQ, while the median task completion time decreases to 66 seconds. The most and least stealthy attacks remain the same as Exp 1: WN with a 17.8% detection rate and SQ with a 99.1% detection rate. Noise-based evasion attacks are the most stealthy with a mean detection rate of 49.4% (12.6% decrease from Exp 1). The patch-based backdoors are the least stealthy with a detection rate of 89.0% (3.7% decrease).

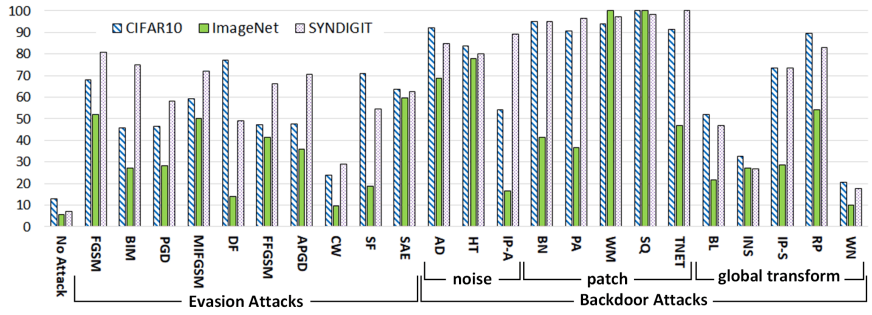

Fig. 5. Detection rates (%) for different datasets in Exp II: Knowledgeable Defenders.

Exp III. Referenceless Defenders. The average detection rate increases slightly from 62.2% in Exp II to 67.9% in Exp III. The distribution demonstrates a high correlation with Exp II with a Spearman correlation coefficient of 0.97. WN and SQ remain the most and least stealthy attacks with detection rates of 32.6% and 98.6%, respectively. Global transformation attacks become the most stealthy category with a 53.7% mean detection rate.

Observations on Attack Categories. For attacks with different categories of injected adversarial perturbations, we observed distinct behaviors across three experiments: (1) The global-noise-based evasion attacks are consistently among the more stealthy types of attacks across all three experiments, with mean detection rates in the range of $[49\%, 61\%]$. While the perturbation budget introduces a limit to the per-pixel modifications, the injected noise pattern is still noticeable to 50%+ of the users, with or without reference/knowledge to the benign dataset. (2) The global-noise-based backdoors are significantly less stealthy than the evasion attacks although they exploit similar types of perturbations. Such backdoor attacks attempt to use weak global noise to generate salient features to create robust backdoors, which appears to be difficult–the noise levels are higher than the noise-based evasion attacks, therefore, the backdoor attacks are noticeable to 70% to 80% of the users. (3) Local-patch-based backdoors are consistently the least stealthy attacks in all settings. The mean detection rates of 89% to 92% indicate that they can hardly escape human auditors. (4) Compared with other attack categories, global transformation attacks demonstrate a very unique pattern. They have a mean detection rate of 72.9% in side-by-side comparisons in Exp I, while the mean detection rate significantly drops to $\sim$50% in Exps II and III. In particular, the Instagram attack was detected at 98.6% in Exp 1 but only 33% to 39% in Exps II and III. That is, users are significantly more likely to be fooled by such attacks without side-by-side comparisons.

The Impact of Image Size and Content. While we have observed highly consistent results across six datasets in numerical assessments, the user study results are different–our results show that the size and quality of the images may significantly affect user-perceived attack stealthiness. For instance, we present

the Exp II detection rates for three datasets (CIFAR, ImageNet, SYNDIGIT) in Fig. 5. We can observe the following: (1) For the auditors, smaller static patches (BN, PA, TNET) are stealthier in larger images (ImageNet). (2) Global transformation attacks (especially BL and WN) also appear to be stealthier in larger images since the per-pixel modification is weaker. (3) For images with more complex content (ImageNet), the noise-based evasion and backdoor attacks become stealthier to users/auditors. (4) For images with a relatively simple foreground and a clean background (e.g., SYNDIGIT), the detection rate is high in the majority of the attacks, i.e., regardless of the type of the added perturbation, the adversarial samples appear highly suspicious to human eyes.

Referenceless Defenders. We initially expected the referenceless defenders to be the weakest. However, results in Section 4.2 show that, the true detection rate increases in Exp III in comparison with Exp II. That is, the defenders/auditors tend to get more aggressive in identifying potentially adversarial samples, when they do not have reference to benign data. While the false positive rate increases, i.e., benign samples are mislabeled as suspicious, the true adversarial samples are more likely to be identified. In the literature, we have seen discussions that "adversarial examples are likely to escape novice defenders when they are unfamiliar with the new testing samples or have not previously seen the testing dataset." Such claims are rejected by our experimental results.

Error Rates with Benign Samples. In each experiment, a portion of users flagged the benign images as "different" or "malicious". The error rates are 5.23%, 12.46%, and 23.70% in Experiments I, II, and III, respectively. The errors are explained by two main reasons: (1) there are always human errors in any type of data labeling, e.g., [34] reports a 3.3% error rate across 10 image datasets and a 6% error rate in ImageNet validation set. The 5.23% base error rate in Exp I is in line with reports in the literature. (2) Some benign benchmarking images do appear to be suspicious, e.g., a significant portion of the tiny images contain blurs, unnatural edges, or objects. As shown in Fig. 1 (D), MNISTM is the most "naturally suspicious" dataset, with 30.3% and 48.8% error rates in Exp II and III, respectively. The 18.5% difference indicates that a small number of reference examples in Exp II effectively help the user/auditor to reduce benign errors. Meanwhile, the error rates with ImageNet remain under 5.5% across all experiments, which shows that regular users are highly capable of recognizing benign high-resolution images even with limited or no reference.

MTurk Study. We initially planned to utilize Amazon Mechanical Turk to conduct the same set of experiments. In the test with Exp I, we injected two pairs of digitally identical images and two pairs of totally different images in each questionnaire for quality control. With 200 questionnaires (all in Exp I, \$0.50 per questionnaire, workers with >95% approval rates), we received almost completely random annotations. Even with the quality control questions, the detection rates were close to 50% for both identical and totally different image pairs. This is consistent with recent reports on substantial decreases in MTurk data quality caused by fraudulent workers or bots [1, 6, 11]. We eventually gave up the MTurk study due to the cost and data quality concerns.

Table 4. Correlation between numerical assessments and detection rates from the user study. Bold: top 3 correlations in each setting; red: p-value>0.05.

Exp I. Informed Defenders in All Attacks												
	l^0	l^1	l^2	l^∞	MSE	PSNR	SSIM	LPIPS	aHash	dHash	pHash	wHash
S	-0.16	0.49	**0.80**	0.68	0.79	**-0.82**	-0.77	0.78	0.67	0.62	0.75	0.66
K	-0.07	0.39	**0.65**	0.51	0.63	**-0.67**	-0.62	0.63	0.54	0.48	0.58	0.52
	FDL	FFL	ERGAS	SAM	MSSSIM	FSIM	VIF	UQI	Elastic	RE	l^1Cos	l^1 Clip
S	0.61	0.78	0.68	**0.82**	-0.76	-0.27	-0.37	-0.77	0.79	0.35	0.49	0.04
K	0.49	0.62	0.53	**0.67**	-0.61	-0.22	-0.29	-0.61	0.63	0.31	0.39	0.12
Exp II. Knowledgeable Defenders in All Attacks												
	l^0	l^1	l^2	l^∞	MSE	PSNR	SSIM	LPIPS	aHash	dHash	pHash	wHash
S	-0.35	0.22	0.57	0.54	0.55	**-0.59**	-0.58	**0.64**	0.40	0.32	0.53	0.41
K	-0.20	0.17	0.44	0.42	0.43	**-0.47**	-0.45	**0.50**	0.29	0.22	0.40	0.32
	FDL	FFL	ERGAS	SAM	MSSSIM	FSIM	VIF	UQI	Elastic	RE	l^1 Cos	l^1 Clip
S	0.35	0.55	0.45	**0.63**	-0.55	-0.03	-0.27	-0.52	0.55	0.20	0.22	-0.21
K	0.28	0.43	0.33	**0.51**	-0.42	-0.06	-0.25	-0.41	0.43	0.19	0.17	-0.06
Exp III. Referenceless Defenders in All Attacks												
	l^0	l^1	l^2	l^∞	MSE	PSNR	SSIM	LPIPS	aHash	dHash	pHash	wHash
S	-0.30	0.14	0.45	0.38	0.42	-0.48	-0.50	**0.59**	0.30	0.23	0.47	0.31
K	-0.17	0.11	0.38	0.27	0.37	-0.40	-0.40	**0.46**	0.24	0.17	0.34	0.25
	FDL	FFL	ERGAS	SAM	MSSSIM	FSIM	VIF	UQI	Elastic	RE	l^1 Cos	l^1 Clip
S	0.29	0.39	0.34	**0.55**	-0.45	-0.10	-0.28	-0.39	0.43	0.25	0.15	-0.24
K	0.23	0.34	0.25	**0.44**	-0.37	-0.10	-0.25	-0.31	0.37	0.23	0.12	-0.10
Subset of Exp II. Knowledgeable Defenders: Noise-based Attacks.												
	l^0	l^1	l^2	l^∞	MSE	PSNR	SSIM	LPIPS	aHash	dHash	pHash	wHash
S	0.74	0.78	0.75	0.05	0.72	-0.80	-0.77	**0.81**	0.42	0.44	0.53	0.60
K	0.56	0.67	0.63	0.06	0.60	-0.67	-0.65	**0.70**	0.33	0.32	0.43	0.49
	FDL	FFL	ERGAS	SAM	MSSSIM	FSIM	VIF	UQI	Elastic	RE	l^1 Cos	l^1 Clip
S	**0.81**	0.71	0.70	0.79	-0.64	-0.78	-0.75	-0.68	0.73	**0.82**	0.78	**0.84**
K	0.69	0.59	0.57	0.67	-0.54	-0.64	-0.64	-0.56	0.61	**0.72**	0.67	**0.72**
Subset of Exp III. Referenceless Defenders: Noise-based Attacks.												
	l^0	l^1	l^2	l^∞	MSE	PSNR	SSIM	LPIPS	aHash	dHash	pHash	wHash
S	0.60	0.71	0.70	0.04	0.66	-0.75	-0.70	0.75	0.37	0.37	0.41	0.47
K	0.46	0.60	0.61	0.05	0.59	-0.64	-0.60	0.65	0.33	0.30	0.36	0.40
	FDL	FFL	ERGAS	SAM	MSSSIM	FSIM	VIF	UQI	Elastic	RE	l^1 Cos	l^1 Clip
S	**0.79**	0.63	0.61	0.73	-0.54	-0.75	-0.69	-0.59	0.67	**0.80**	0.71	**0.79**
K	**0.67**	0.57	0.50	0.63	-0.48	-0.61	-0.59	-0.49	0.60	**0.68**	0.60	**0.65**

4.3 Correlation: Numerical Metrics vs. User Perceived Stealthiness

Some of the image similarity/quality metrics in Sec. 3) are designed to reflect the human visualization system. Now, we further investigate whether they could be used to assess the user-perceived attack stealthiness. To investigate the correlations between numerical analysis and the user studies, we employ two metrics: Spearman Rank Order Correlation Coefficient (S or SROCC) and Kendall Rank Order Correlation Coefficient (K or KROCC). We do not use the popular Pearson correlation because our data do not follow normal distribution.

First, we report the correlation between the numerically assessed stealthiness (24 metrics in Tables 2 and 3) and user-perceived stealthiness (three user studies in Section 4) across all 23 attacks. The results are shown in the top 3 sections in Table 4. We reject the null hypothesis when the p-value exceeds 0.05 and mark the correlation coefficients as red in the table. In Exp I, most of the image similarity/quality metrics fit well with the detection rates (DR): 11 out of 24 metrics show strong correlations in SROCC ($|S|$>0.7). That is, attacks assessed with lower distance/error/noise (or higher similarity/quality) are less likely to be

identified in Exp I. In particular, SAM, PNSR, and l^2 demonstrate the strongest correlation with >0.8 on SROCC and >0.65 on KROCC.

However, the correlations are significantly weaker in Exps II and III. None of the metrics shows strong correlations with DR, while only 10 metrics show moderate correlations ($0.4<|S|<0.7$) with DR in Exp II in both SROCC and KROCC, and 2 metrics show moderate correlations in Exp III. Overall, SAM [46] and LPIPS [50] appear to be the most consistent with human perceptions in all three experiments, however, the correlations are only moderate with knowledgeable and referenceless defenders. Meanwhile, image similarity/quality metrics using structural information or deep features demonstrate better performance.

Meanwhile, we also evaluate the correlation between the numerically assessed and user-perceived stealthiness of each category of attacks. We found strong correlations in global-noise-based attacks, as reported in the last two sections of Table 4. That is, the image similarity/quality metrics could properly evaluate the strength of the injected global perturbation in a way that is consistent with user perceptions. Last, the correlations are weak in other attack categories.

In summary, none of the existing image similarity/quality metrics could accurately assess users' perceptions of attack stealthiness in all attacks. However, some metrics have shown promising performance in a subset of the attacks.

5 The Stealthiness Assumption Revisited

Finally, we discuss the reflections of our findings w.r.t. ML attack stealthiness and practicality, which may contribute to a better understanding of the attacks.

Most of the existing ML attacks in the literature explicitly or implicitly employ the stealthiness assumption. The underlying rationale is that it may be difficult for less experienced users to examine the DL model architecture, code, or parameters, however, even novice users could examine training/testing images and identify anomalies. Here we revisit such assumptions and discuss their practicality based on our findings.

• **Evasion attacks.** Many evasion attacks in the literature explicitly claim to be stealthy and enforce stealthiness using a perturbation budget. Results from our user study show that majority of them are still highly noticeable even to users without reference images. Meanwhile, we also show that some image similarity/quality metrics could be employed to effectively predict the user-perceived stealthiness of the attack images. For an attack to be stealthy in practice, a tight threshold (high similarity or low distance) must be set for the metrics.

• **Data poisoning backdoors.** Conventional data poisoning backdoors do not assume stealthiness for the poisoned training samples, since they are designed to carry labels that are visually wrong. Clean label poisoning attacks have been proposed to tackle this issue by using adversarial samples that appear to carry the correct labels. However, as demonstrated in our user study, the clean label poisoning attacks are highly detectable by users with or without reference. When the attack attempts to train the DNN to learn a *salient* feature from weak perturbations, the actual perturbation is too strong to escape human eyes.

• **Patch-based backdoors.** In general, the patch-based backdoors appear to be the least stealthy, except for the tiny patches on larger images, e.g., BN and TNET on ImageNet. In order for the patches to be learned by the victim DNN as robust features, the patches must have reasonable size and salient visual features, which is against the stealthiness assumption.

Finally, we would like to answer this question from our experiments and observations: *what makes an attack stealthy, i.e., what makes the adversarial image less likely to be identified by a human auditor?*

(1) Scale of Perturbation Matters. Attacks with *extremely low* perturbation budget are more likely to escape human evaluators, especially in global-noise-based attacks. In practice, a global perturbation budget of 8, or a normalized l^1 in the range of [5, 8] appears to be too strong to be unnoticeable.

(2) Size of Perturbed Region Matters. Attacks that modify a relatively *smaller* portion of the victim image, i.e., attacks with a very small l^0, are more likely to escape evaluators.

(3) Image Content Matters. Attacks on images that are fully filled with foreground and have complex content are more likely to escape human evaluators, while most of the attacks, especially the noise-based ones, are very noticeable on a clean background (e.g., blue sky, white wall).

In practice, it would be very difficult to design an attack that satisfies both (1) and (2), since such weak perturbations are unlikely to trigger robust responses from the victim DNN. We have not seen such an attempt in the literature yet. Meanwhile, when an attack achieves one of (1) and (2) on images satisfying (3), the attack is quite stealthy to human eyes, e.g., CW on ImageNet. Last, a different design philosophy has been proposed in the literature, e.g., SAE and INS. Such attacks do not seek to minimize the perturbation or to hide the perturbation, instead, they attempt to apply special visual effects on the victim images, so that the adversarial samples, although significantly different from the victim images, appear to be benign by themselves. The adversarial samples can hardly escape side-by-side comparisons, however, they could better fool the auditors if they do not have reference to the benign dataset.

In summary, we argue that user-based evaluation is the golden standard to validate the stealthiness assumptions in adversarial ML. For global-noise-based attacks, some image quality metrics could be employed to predict attack stealthiness in lieu of a user study. In such metrics, only attacks with extremely low perturbation (high similarity/quality) may result in practically stealthy attacks. Very few attacks in the literature have achieved this goal.

6 Conclusion

In this paper, we present the first large-scale comparative experimental study of the stealthiness of evasion and backdoor attacks against deep learning systems. We have implemented 20 attacks (23 different settings) on six benchmarking datasets. We first present numerical measurements using 24 image quality metrics on all the attacks. Next, we design a user study of three questionnaires that

ask users to identify potentially adversarial images in three different settings. With 1,500+ responses and 30,000+ labeled images, we find that majority of the attacks in the literature are not really stealthy to human eyes. We also identify the factors that impact attack stealthiness, e.g., the type of perturbation, the size, quality, and content of the victim images. We further examine the correlations between numerically assessed image similarity/quality and user-perceived stealthiness and re-visit the stealthiness of the attacks with our findings.

Acknowledgements

Zeyan Liu, Fengjun Li and Bo Luo were sponsored in part by NSF awards IIS-2014552, DGE-1565570, DGE-1922649, and the Ripple University Blockchain Research Initiative. Zhu Li was supported in part by NSF award 1747751. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. The authors would like to thank all the participants of the user studies.

References

1. Bai, H.: Evidence that a large amount of low quality responses on mturk can be detected with repeated gps coordinates (2018)
2. Barni, M., Kallas, K., Tondi, B.: A new backdoor attack in cnns by training set corruption without label poisoning. In: IEEE ICIP (2019)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017)
4. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D.: Adversarial attacks and defences: A survey. arXiv:1810.00069 (2018)
5. Chen, X., Liu, C., Li, B., Lu, K., Song, D.: Targeted backdoor attacks on deep learning systems using data poisoning. arXiv:1712.05526 (2017)
6. Chmielewski, M., Kucker, S.C.: An mturk crisis? shifts in data quality and the impact on study results. Social Psychological and Personality Science **11**(4), 464–473 (2020)
7. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
8. Dabouei, A., Soleymani, S., Taherkhani, F., Dawson, J., Nasrabadi, N.: Smoothfool: An efficient framework for computing smooth adversarial perturbations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2665–2674 (2020)
9. Doan, K., Lao, Y., Zhao, W., Li, P.: Lira: Learnable, imperceptible and robust backdoor attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11966–11976 (2021)
10. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
11. Dreyfuss, E.: A bot panic hits amazon's mechanical turk. Wired (2018)

12. Dumford, J., Scheirer, W.: Backdooring convolutional neural networks via targeted weight perturbations. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–9. IEEE (2020)
13. Fei, M., Li, J., Liu, H.: Visual tracking based on improved foreground detection and perceptual hashing. Neurocomputing **152**, 413–428 (2015)
14. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software **33**(1), 1 (2010)
15. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
16. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
17. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
18. Gu, T., Dolan-Gavitt, B., Garg, S.: Badnets: Identifying vulnerabilities in the machine learning model supply chain. In: NIPS MLSec Workshop (2017)
19. Guo, W., Wang, L., Xing, X., Du, M., Song, D.: Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. In: ICDM (2020)
20. Hosseini, H., Poovendran, R.: Semantic adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1614–1619 (2018)
21. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In: IJCNN (2013)
22. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13919–13929 (2021)
23. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
24. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. ICLR Workshop (2017), https://arxiv.org/abs/1607.02533
25. Li, S., Xue, M., Zhao, B.Z.H., Zhu, H., Zhang, X.: Invisible backdoor attacks on deep neural networks via steganography and regularization. IEEE Transactions on Dependable and Secure Computing **18**(5), 2088–2105 (2020)
26. Li, Y., Hua, J., Wang, H., Chen, C., Liu, Y.: Deeppayload: Black-box backdoor attack on deep learning models through neural payload injection. In: IEEE/ACM ICSE (2021)
27. Liu, Y., Lee, W.C., Tao, G., Ma, S., Aafer, Y., Zhang, X.: Abs: Scanning neural networks for back-doors by artificial brain stimulation. In: ACM CCS (2019)
28. Liu, Y., Ma, S., Aafer, Y., Lee, W., Zhai, J., Wang, W., Zhang, X.: Trojaning attack on neural networks. In: NDSS (2018)
29. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
30. Marnerides, D., Bashford-Rogers, T., Hatchett, J., Debattista, K.: Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In: Computer Graphics Forum. vol. 37, pp. 37–49. Wiley Online Library (2018)
31. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)

32. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
33. Nguyen, T.A., Tran, A.T.: Wanet-imperceptible warping-based backdoor attack. In: International Conference on Learning Representations (2020)
34. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749 (2021)
35. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3) (2015)
36. Saha, A., Subramanya, A., Pirsiavash, H.: Hidden trigger backdoor attacks. In: AAAI (2020)
37. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on image processing **15**(2), 430–444 (2006)
38. Tang, R., Du, M., Liu, N., Yang, F., Hu, X.: An embarrassingly simple approach for trojan attack in deep neural networks. In: ACM KDD (2020)
39. Wald, L.: Quality of high resolution synthesised images: Is there a simple criterion? In: Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images". pp. 99–103. SEE/URISCA (2000)
40. Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., Zhao, B.Y.: Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In: IEEE S&P (2019)
41. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE signal processing letters **9**(3), 81–84 (2002)
42. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
43. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
44. Wei, P., Lu, H., Timofte, R., Lin, L., Zuo, W., Pan, Z., Li, B., Xi, T., Fan, Y., Zhang, G., et al.: Aim 2020 challenge on real image super-resolution: Methods and results. In: ECCV. pp. 392–422 (2020)
45. Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. In: International Conference on Learning Representations (2019)
46. Yuhas, R.H., Goetz, A.F.H., Boardman, J.W.: Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In: Summaries of the 4th Annual JPL Airborne Geoscience Workshop (1992)
47. Zeng, Y., Park, W., Mao, Z.M., Jia, R.: Rethinking the backdoor attacks' triggers: A frequency perspective. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16473–16481 (2021)
48. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: A feature similarity index for image quality assessment. IEEE Trans. Image Processing **20**(8), 2378–2386 (2011)
49. Zhang, Q., Ding, Y., Tian, Y., Guo, J., Yuan, M., Jiang, Y.: Advdoor: adversarial backdoor attack of deep learning system. In: ACM International Symposium on Software Testing and Analysis (2021)
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
51. Zhong, H., Liao, C., Squicciarini, A.C., Zhu, S., Miller, D.: Backdoor embedding in convolutional neural network models via invisible perturbation. In: ACM CODASPY (2020)