An Extension of Encryption-Inspired Adversarial Defense with Secret Keys against Adversarial Examples

MaungMaung AprilPyone and Hitoshi Kiya
Tokyo Metropolitan University, Tokyo, Japan

***Abstract*—Recently, encryption-inspired block-wise image transformation with a secret key was proposed to defend against adversarial examples. The adversarial defense was also demonstrated to outperform state-of-the-art defenses. In this work, we first extend the block-wise image transformation for increasing its key space by using additional transformation steps. Moreover, the extended defense is extensively evaluated in terms of robustness against various attacks under a number of metrics. We also conduct adaptive attacks with key estimation. In an experiment, the extended defense is confirmed not only to increases the key space, but also to improve the performance accuracy, while maintaining the overall accuracy close to a non-robust model. The evaluation results also suggest that the extended defense is robust against both non-adaptive and adaptive attacks as long as its keys are secret. Furthermore, the extended defense is confirmed to outperform state-of-the-art adversarial defenses with the noise distance of $8/255$ on CIFAR-10 dataset.**

I. Introduction

Deep neural networks (DNNs) have brought major breakthroughs in computer vision as well as many other fields for a wide range of applications. Due to their remarkable performance, DNN models have been deployed in security-critical applications such as autonomous vehicles, healthcare, finance, etc. Therefore, security in DNN has become quintessential in such applications.

Machine learning in general suffers from attacks such as model inversion attacks [1], membership inference attacks [2], and adversarial attacks [3], [4]. DNNs are no exception. In this work, we focus on adversarial attacks. In particular, carefully perturbed data points known as adversarial examples are imperceptible to human, but they cause DNNs to make erroneous predictions with high confidence. As an example, in Fig. 1, the network here classified the clean image correctly as "giant_panda" with a probability of $98.92\,\%$. After adding a small fraction of noise, the network misclassified the giant panda as "plastic_bag" with $98.71\,\%$ confidence. Researchers have proposed numerous ways of constructing adversarial examples. Such works include [3], [5]–[9]. In the context of computer vision, these threat models do not match real world applications [10], [11] because there can be various physical conditions (e.g., camera angle, lighting/weather), physical limits on imperceptibility, etc. However, it has been proved that adversarial examples are real threats to DNNs [12]–[16].

To defend against adversarial examples, numerous techniques have been proposed in the literature. Current state-of-the-art empirically robust defense is adversarial training [9], [17], [18], but the accuracy of adversarially trained models is almost half lower than that of non-robust models. Another ideal desirable defenses are certified/provable defenses [19]–[21], they are not scalable to larger datasets. Some certified defenses have been scaled to a certain degree [22]–[24], but the accuracy is still not comparable to empirically robust models. Another popular adversarial defense is a preprocessing approach such as [25]–[28]. They all have been defeated when accounting for obfuscated gradients (a way of gradient masking) [29]. To reinforce these weak defense methods, Raff et al. [30] proposed a stronger defense by combining a large number of transforms stochastically. However, applying many transforms drop in accuracy even though the model is not under attack and is computationally expensive.

Recently, a new insight for adversarial defense has been given as one of preprocessing techniques [31]–[33]. The work by [31] bridges cryptography to adversarial defense and [32], [33] has been inspired by perceptual image encryption methods, which were proposed for privacy-preserving machine learning and encryption-then-compression systems [34]–[39]. The encryption-inspired adversarial defense with a secret key [32] was also confirmed to outperform state-of-the-art adversarial defenses as long as the key is kept secret. In this paper, we mainly focus on the adversarial defense with block-wise encryption in [32]. In the block-wise defense, when a smaller block-size is chosen, higher accuracy is achieved, but the key space becomes smaller. In this work, we improve the work in [32] by extending its key space and evaluate the extended version in terms of image classification performance and robustness against various attacks. We make the following contributions in this paper.

- We extend [32] by adding negative/positive transformation as an encryption step to increase the key space.
- We apply various state-of-the-art attacks to the extended defense.

In experiments, the extended defense is confirmed not only to increase the key space, but also to improve the performance accuracy. Moreover, the extended defense is robust against various attacks including adaptive attacks. As a result, the extended defense is demonstrated to outperform state-of-the-art defenses.

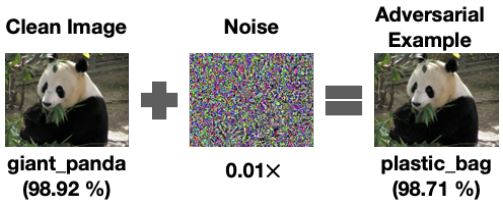


Fig. 1. Example of adversarial example.

II. Related Work

A. Adversarial Attacks

Adversarial attacks can be divided into two categories: poisoning/causative attacks (i.e., training time attacks) and evasion/exploratory attacks (i.e., test time attacks) [40]. In this work, we focus on evasion attacks also known as adversarial examples. An adversarial example is a modified input x' (visually similar to x) to a classifier $f(\cdot)$ aiming $f(x) \neq f(x')$. Techniques of generating adversarial examples are classified into three groups on the basis of the adversary's knowledge towards a model: white-box, black-box and gray-box. Under white-box settings, an adversary has direct access to the model, its parameters, and training data. In contrast, the adversary does not have any knowledge of the model, except the output of the model in black-box attacks. Between white-box and black-box methods, there are gray-box attacks that imply that the adversary knows something about the system (i.e., partial knowledge of the model such as its architecture, parameters, or training data).

The adversary finds perturbation δ under certain distance metric (usually ℓ_p norm) to construct an adversarial example. An attack algorithm usually minimizes the perturbation or maximizes the loss function, i.e.,

$$\underset{\delta}{\text{minimize}} \, \|\delta\|_p \,, \quad \text{s.t.} \quad f(x+\delta) \neq y, \text{or} \tag{1}$$

$$\underset{\delta \in \Delta}{\text{maximize}} \, \mathcal{L}(f(x+\delta), y), \tag{2}$$

where $\Delta = \{\delta : \|\delta\|_p \leq \epsilon\}$. There are many attack algorithms such as Fast Gradient Sign Method (FGSM) [5], Projected Gradient Descent (PGD) [9], DeepFool [7], Carlini and Wagner (CW) [8], etc.

B. Adversarial Defenses

The goal of a defense method is to make a model that is accurate not only for clean input but also for adversarial examples. There are many different approaches to defend against adversarial examples such as certified and provable defenses [19]–[24], [41], adversarial training [5], [6], [9], [17], [18], [42], preprocessing techniques [25]–[28], [30], [43], detection algorithms [44], [45] and others. Most of conventional adversarial defenses drop in accuracy or are broken due to obfuscated gradients. In this paper, we propose a novel preprocessing technique for adversarial defense.

Recently, Taran et al. proposed to apply the concept of cryptography with a secret key to adversarial defense as a preprocessing technique under black-box attacks [31]. However, traditional cryptographic methods cannot be used for learning a DNN model, so the models trained with a key had a low performance, even when a simple random permutation operation was used [31]. In contrast, another key-based approach by [32] that is inspired by learnable image encryption methods [34], [46] was demonstrated to outperform state-of-the-art defense methods under white-box attacks. However, there is a possibility of brute-force attacks that can be carried out when block size is small in [32]. Therefore, in this work, we first extend the work with a key [32] by expanding its key space and evaluate the extended version extensively int terms of robustness against various attacks.

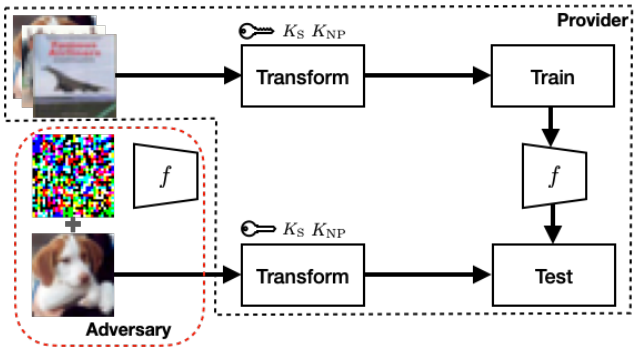
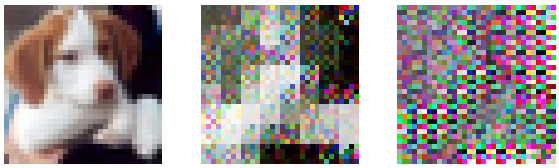


Fig. 2. Overview of image classification process with extended adversarial defense with secret keys.



(a) (b) (c)

Fig. 3. Example of transformed images with $M = 4$. (a) Original image. (b) Encryption inspired adversarial defense [32]. (c) Extended encryption inspired adversarial defense.

III. Extension of Encryption Inspired Adversarial Defense

The encryption-inspired adversarial defense [32] applies a block-wise transformation with a secret key (K_S) to input images before training and testing a model. In this work, we extend the adversarial defense by adding block-wise negative/positive transformation with an additional secret key K_{NP}. An overview of image classification process with extended encryption-inspired adversarial defense with keys is depicted in Fig. 2 and an example of transformed images is shown in Fig. 3.

A. Extended Transformation

The following are steps for transforming input images by the extended transformation (Algorithm 2), where c, w and h

denote the number of channels, width and height of an image tensor $x \in [0,1]^{c\times w\times h}$.

1) Divide x into blocks with a size of M such that $\{B_{(1,1)},\ldots,B_{(\frac{w}{M},\frac{h}{M})}\}$.
2) Transform each block tensor $B_{(i,j)}^{c\times M\times M}$ into a vector $b_{(i,j)} = [b_{(i,j)}(1),\ldots,b_{(i,j)}(c\times M\times M)]$.
3) Generate a random permutation integer vector $v = [v_1,\ldots,v_k,\ldots,v_{k'},\ldots,v_{c\times M\times M}]$ by using key K_{S}, where $v_k \neq v_{k'}, v_k \in \{1,2,\ldots,c\times M\times M\}$, if $k \neq k'$, and permutate every vector $b_{(i,j)}$ with v as
$$b'_{(i,j)}(k) = b_{(i,j)}(v_k), \tag{3}$$
to obtain a shuffled vector $b'_{(i,j)} = [b'_{(i,j)}(1),\ldots,b'_{(i,j)}(c\times M\times M)]$
4) Generate a random binary vector $r = [r_1,\ldots,r_k,\ldots,r_{c\times M\times M}]$, $r_k \in \{0,1\}$ by using key K_{NP}, and inverse the intensity of pixel values in each shuffled block $b'_{(i,j)}$ with r as
$$b''_{(i,j)}(k) = \begin{cases} b'_{(i,j)}(k) & (r_k = 0) \\ 1 - b'_{(i,j)}(k) & (r_k = 1), \end{cases} \tag{4}$$
where the value of the occurrence probability $P(r(k))$ is 0.5.
5) Integrate the shuffled vectors to form a shuffled image tensor $x' \in [0,1]^{c\times w\times h}$.

Algorithm 1 Pixel Shuffling

Input: x, K_{S}
Output: x'
Divide x into blocks, $\{B_{(1,1)},\ldots,B_{(\frac{w}{M},\frac{h}{M})}\}$
Transform blocks to vectors, $\{b_{(1,1)},\ldots,b_{(\frac{w}{M},\frac{h}{M})}\}$
Generate v by K_{S}
for Each block $b_{(i,j)}$ **do**
  $b'_{(i,j)}(k) \leftarrow b_{(i,j)}(v_k)$
end for
$x' \leftarrow$ Integrate blocks in b'

Algorithm 2 Extended Transformation

Input: $x, K_{\mathrm{S}}, K_{\mathrm{NP}}$
Output: x'
Divide x into blocks, $\{B_{(1,1)},\ldots,B_{(\frac{w}{M},\frac{h}{M})}\}$
Transform blocks to vectors, $\{b_{(1,1)},\ldots,b_{(\frac{w}{M},\frac{h}{M})}\}$
Generate v by K_{S}, and r by K_{NP}
for Each block $b_{(i,j)}$ **do**
  $b'_{(i,j)}(k) \leftarrow b_{(i,j)}(v_k)$
  if $r(k)$ is 0 **then**
    $b''_{(i,j)}(k) \leftarrow b'_{(i,j)}(k)$
  else
    $b''_{(i,j)}(k) \leftarrow 1 - b'_{(i,j)}(k)$
  end if
end for
$x' \leftarrow$ Integrate blocks in b''

B. Differences between Encryption-Inspired Adversarial Defense and its Extended Version

In the conventional encryption-inspired adversarial defense [32], only pixel shuffling (Algorithm 1) is utilized for input transformation. In contrast, the extended defense uses both pixel shuffling and negative/positive transformation (Algorithm 2).

By using this extension, the extended defense expands the key space. The conventional one is on the basis of a block-wise operation and utilizes the same key for all blocks. Therefore, its key space is given by
$$\mathcal{K}(c\times M\times M) = (c\times M\times M)!. \tag{5}$$
In contrast, the key space of the extended defense is
$$\mathcal{K}_{\mathrm{E}(c\times M\times M)} = (c\times M\times M)!\times 2^{(c\times M\times M)}. \tag{6}$$
Therefore, the extended defense can enhance robustness against brute-force attacks even for $M = 2$, while maintaining high classification performance.

IV. Evaluation

To verify the effectiveness of the extended encryption inspired adversarial defense, we ran a number of experiments on the CIFAR-10 [47] dataset with a batch size of 128 and live augmentation (random cropping with padding of 4 and random horizontal flip) on a training set. CIFAR-10 consists of 60,000 color images (dimension of $32\times32\times3$) with 10 classes (6000 images for each class) where 50,000 images are for training and 10,000 for testing. Both training and test images were transformed on the basis of the extended transformation (see Algorithm 2) with secret keys K_{S} and K_{NP}.

We used deep residual networks [48] with 18 layers (ResNet18) and trained for 200 epochs with cyclic learning rates. The parameters of the stochastic gradient descent (SGD) optimizer were: momentum of 0.9, weight decay of 0.0005 and maximum learning rate of 0.2. We trained 4 models under the use of various block sizes (i.e., $M \in \{2,4,8,16\}$). Additionally, we also trained a standard model without any defense as our baseline.

A. Robustness Against Threat Models

The goal of an adversary is to fool a model by reducing the classification accuracy (i.e., untargeted attacks). To simulate this scenario, we deployed well-known threat models under different metrics (ℓ_∞, ℓ_2, and ℓ_1). The threat models were projected gradient descent (PGD) [9] for ℓ_∞, the Carlini and Wagner attack (CW) [8] and the DeepFool [7] for the ℓ_2 bounded metric, and the elastic-net attack (EAD) [49] for the ℓ_1 bounded metric.

The parameters of PGD adversary were adapted from [17] with noise distance $\epsilon = 8/255$, step size $\alpha = 2/255$, and 10 random restarts for 50 iterations with random initialization. For the DeepFool attack, we utilized publicly available implementation[1] with 50 iterations. Since we focused on untargeted

[1] https://github.com/Harry24k/adversarial-attacks-pytorch

TABLE I
ACCURACY (%) OF THE EXTENDED DEFENSE AGAINST VARIOUS NON-ADAPTIVE ATTACKS

Model	Clean	PGD (ℓ_∞)	DeepFool (ℓ_2)	CW (ℓ_2)	EAD (ℓ_1)
Standard	95.45	0.00	3.28	0.00	0.00
($M = 2$)	94.54	93.14	93.16	94.53	94.54
($M = 4$)	92.44	92.07	90.57	92.42	92.40
($M = 8$)	86.33	86.06	84.79	86.37	86.30
($M = 16$)	76.98	76.95	76.05	76.96	76.99

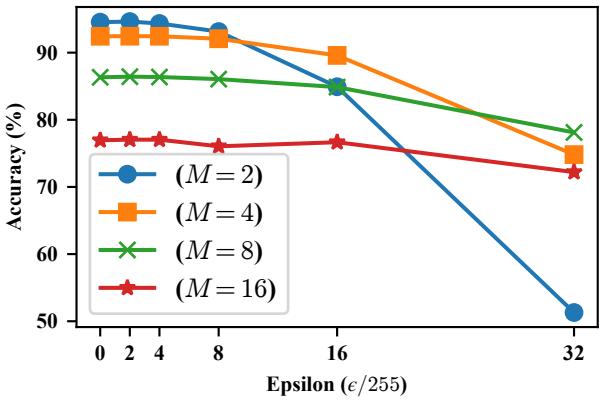


Fig. 4. Accuracy vs. noise distance ϵ.

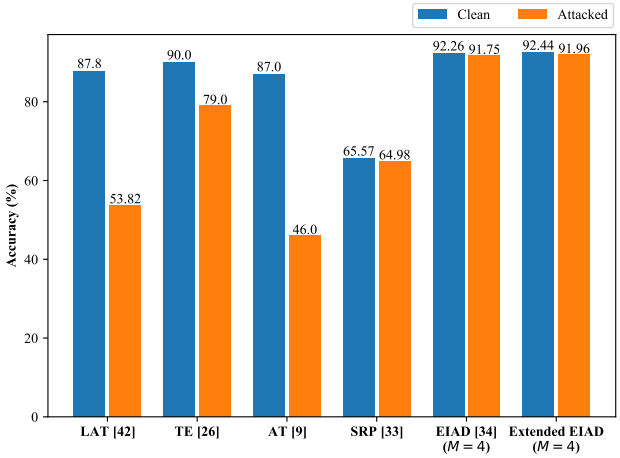


Fig. 5. Comparison with state-of-the-art defenses under PGD (ℓ_∞) with $\epsilon = 8/255$ for CIFAR-10 dataset.

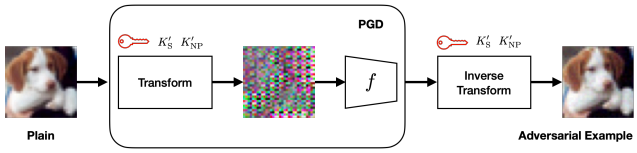


Fig. 6. Diagram of adaptive PGD attack by using estimated keys, K'_S and K'_{NP}.

attacks, CW and EAD were configured with a confidence value of 0, learning rate of 0.01, binary search steps of 9 and initial constant of 0.001 for 1000 iterations. We used CW and EAD with elastic-net (EN) decision rule implementations from [50] in our experiments.

Table I summarizes results of the extended defense against different threat models. Obviously, when the model was not under defense, the accuracy decreased to almost 0 % in all attacks. In contrast, the models with the extended defense were robust against all attacks. In particular, the model with $M = 2$ achieved the highest accuracy whether or not it was under attacks. In addition, we carried out PGD attacks with different noise distances for experiment purposes. The accuracy dropped as the noise distance increased. The results are plotted in Fig. 4. From these results, choosing smaller M provides better performance, although the key space decreases.

B. Comparison with State-of-the-art Defenses

To confirm the effectiveness of the extended defense, we made a comparison with latent adversarial training (LAT) [42], adversarial training (AT) [9], and thermometer encoding (TE) [25], and key-based defenses: standard random permutation (SRP) [31] and encryption inspired adversarial defense (EIAD) [32] for the CIFAR-10 dataset.

We reproduced the results for key-based methods under the same settings used in this work. However, for the other defenses, we used their reported results for comparison. All defenses were evaluated under the PGD (ℓ_∞) threat model with $\epsilon = 8/255$.

Figure 5 shows the accuracy for both clean images and attacked images comparing the extended defense model with $M = 4$. Key-based methods are effective to defend against adversarial examples when the keys are secret. Th accuracy of SRP was low and that of extended defense was slightly higher than that of [32]. Although LAT [42] and AT [9] are empirically robust models, the accuracy was low under the attacks. The accuracy of TE [25] under attacks was 79 %; however, it reduced to 30 % under adaptive attacks. In contrast, the extended defense is still resistant against adaptive attacks as long as the keys are secret.

C. Robustness Against Key Estimation Attacks

As pointed out in [32], an attacker might estimate keys and carry out an adaptive attack by using the estimated keys. Briefly, the attacker transforms an input image with the estimated keys and performs the attack, then, the transformed image is inverse-transformed, as described in Fig. 6. There are two ways for estimating the keys: random approach and heuristic approach.

Random Approach: An attacker randomly generates keys and performs the attack in Fig. 6. Table II shows processing time taken for testing random keys in seconds for one time (column "(One Try) Time") and performance accuracy under the attack (see Fig. 6) by using random keys; K'_S for the

conventional defense [32], and K'_S and K'_{NP} for the extended defense (column "Attack"). The processing time taken for random approach (one try) is almost the same for both conventional defense and extended defense. The processing time was recorded on a computer we used (Intel Core i9–9900K, 64 GB memory, GeForce RTX2080Ti GPU). We observe that the random key estimation attack was not effective and all the models maintained high classification accuracy.

When considering brute-force attacks for the smallest block size $M = 2$, the key space for the conventional defense is $\mathcal{K}(12) = 12!$, and for the extended defense is $\mathcal{K}_E(12) = 12! \times 2^{12}$. Therefore, the processing time taken for brute-force attacks is estimated to be $12! \times 1.74 = 231,515.44$ (hours) for conventional defense and $12! \times 2^{12} \times 1.76 = 959,195,381.76$ (hours) for extended defense on the above computer with a single process.

Heuristic Approach: Since we consider white-box attacks, the attacker may improve estimated keys in a heuristic way by observing the accuracy of the model. After the keys are heuristically estimated, the attacker may carry out the attack in Fig. 6 by using the estimated keys. We simulated this scenario by rearranging two elements: integer permutation vector $v = [v_1, v_2, \ldots, v_{c\times M\times M}]$ and random binary vector $r = [r_1, r_2, \ldots, r_{c\times M\times M}]$, in accordance with improvement in accuracy (see Algorithm 3). We generated index pairs $\mathcal{P}$, as $\mathcal{P} = \{(1,2),(1,3),\ldots,(c\times M\times M-1, c\times M\times M)\}$ in v and r that are based on random keys K'_{S} and K'_{NP} respectively. The number of possible pairs is given by

$$|\mathcal{P}| = \binom{(c\times M\times M)}{2}. \tag{7}$$

We swap the values in each pair of v and r independently if the accuracy improves and the swap operation is done one by one for v and r alternatively.

Table III shows accuracy of the models under the attack with the heuristic approach. The results suggest that the extended defense achieved higher accuracy for $M = 2$ and 4 than the conventional one [32], even when the attack with the heuristic approach was applied.

Algorithm 3 Heuristic Approach

Input: Input images with labels
Output: v, r

```
Initialize v and r with random keys, K'_S and K'_NP
Generate P
for Each pair in P do
    accuracy ← Calculate accuracy of input images
    if accuracy improves then
        Swap pair in v
    end if
    accuracy ← Calculate accuracy of input images
    if accuracy improves then
        Swap pair in r
    end if
end for
```

TABLE II
ACCURACY (%) OF PROPOSED DEFENSE AND CONVENTIONAL ONE UNDER ADAPTIVE ATTACK WITH RANDOM APPROACH, AND PROCESSING TIME (SECONDS) FOR ONE TRY

Model	Conventional [32]		Extended	
	(One Try) Time (s)	Accuracy	Time (s)	Accuracy
($M = 2$)	1.74	91.81	1.76	**93.16**
($M = 4$)	1.74	91.56	1.79	91.84
($M = 8$)	1.78	86.76	1.77	86.16
($M = 16$)	1.76	77.54	1.78	76.86

TABLE III
ACCURACY (%) OF PROPOSED DEFENSE AND CONVENTIONAL ONE UNDER ADAPTIVE ATTACK WITH HEURISTIC APPROACH

Model	Conventional [32] Accuracy	Extended Accuracy
($M = 2$)	84.54	**92.50**
($M = 4$)	91.70	92.05
($M = 8$)	86.66	86.05
($M = 16$)	77.47	76.74

V. CONCLUSION

In this work, we extended an encryption-inspired adversarial defense with secret keys to increase its key space, and evaluated the performance of models trained by using the extended defense in terms of image classification accuracy not only for clean images, but also for adversarial examples with various attacks in different metrics for CIFAR-10 dataset. As a result, the extended defense was demonstrated to outperform state-of-the-art defenses including the conventional encryption-inspired adversarial defense, whether or not the model is under attacks. Moreover, the accuracy difference between the extended model and a standard model (non-robust model) was small (0.91 %), so the applicability of the extended model in real-world applications was also suggested.

REFERENCES

[1] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.

[2] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy*. IEEE, 2017, pp. 3–18.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[4] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[6] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *International Conference on Learning Representations*, 2017.

[7] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2574–2582.

[8] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy*, 2017, pp. 39–57.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[10] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "A rotation and a translation suffice: Fooling CNNs with simple transformations," *arXiv preprint arXiv:1712.02779*, 2017.

[11] J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen, and G. E. Dahl, "Motivating the rules of the game for adversarial example research," *arXiv preprint arXiv:1807.06732*, 2018.

[12] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[13] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 284–293.

[14] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *International Conference on Learning Representations*, 2017.

[15] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017, pp. 506–519.

[16] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1528–1540.

[17] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *International Conference on Learning Representations*, 2020.

[18] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in *Advances in Neural Information Processing Systems*, 2019, pp. 3353–3364.

[19] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," in *International Conference on Learning Representations*, 2018.

[20] K. Dvijotham, R. Stanforth, S. Gowal, T. A. Mann, and P. Kohli, "A dual approach to scalable verification of deep networks," in *UAI*, vol. 1, 2018, p. 2.

[21] E. Wong and J. Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 5283–5292.

[22] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, "Provably robust deep learning via adversarially trained smoothed classifiers," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 289–11 300.

[23] M. Mirman, T. Gehr, and M. Vechev, "Differentiable abstract interpretation for provably robust neural networks," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 3578–3586.

[24] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter, "Scaling provable adversarial defenses," in *Advances in Neural Information Processing Systems*, 2018, pp. 8400–8409.

[25] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," in *International Conference on Learning Representations*, 2018.

[26] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *International Conference on Learning Representations*, 2018.

[27] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in *International Conference on Learning Representations*, 2018.

[28] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," in *International Conference on Learning Representations*, 2018.

[29] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, Jul. 2018, pp. 274–283.

[30] E. Raff, J. Sylvester, S. Forsyth, and M. McLean, "Barrage of random transforms for adversarially robust defense," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6521–6530.

[31] O. Taran, S. Rezaeifar, and S. Voloshynovskiy, "Bridging machine learning and cryptography in defence against adversarial attacks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 267–279.

[32] M. AprilPyone and H. Kiya, "Encryption inspired adversarial defense for visual classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1681–1685.

[33] ——, "Block-wise image transformation with secret key for adversarially robust defense," *arXiv preprint arXiv:2010.00801*, 2020.

[34] M. Tanaka, "Learnable image encryption," in *2018 IEEE International Conference on Consumer Electronics-Taiwan*, 2018, pp. 1–2.

[35] W. Sirichotedumrong, Y. Kinoshita, and H. Kiya, "Pixel-based image encryption without key management for privacy-preserving deep neural networks," *IEEE Access*, vol. 7, pp. 177 844–177 855, 2019.

[36] T. Chuman, W. Sirichotedumrong, and H. Kiya, "Encryption-then-compression systems using grayscale-based image encryption for jpeg images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1515–1525, June 2019.

[37] W. Sirichotedumrong, T. Maekawa, Y. Kinoshita, and H. Kiya, "Privacy-preserving deep neural networks with pixel-based image encryption considering data augmentation in the encrypted domain," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 674–678.

[38] W. Sirichotedumrong and H. Kiya, "Grayscale-based block scrambling image encryption using ycbcr color space for encryption-then-compression systems," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[39] K. Kurihara, S. Imaizumi, S. Shiota, and H. Kiya, "An encryption-then-compression system for lossless image compression standards," *IEICE transactions on information and systems*, vol. 100, no. 1, pp. 52–56, 2017.

[40] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[41] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. Mann, and P. Kohli, "On the effectiveness of interval bound propagation for training verifiably robust models," *arXiv preprint arXiv:1810.12715*, 2018.

[42] N. Kumari, M. Singh, A. Sinha, H. Machiraju, B. Krishnamurthy, and V. N. Balasubramanian, "Harnessing the vulnerability of latent layers in adversarially trained models," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 7 2019, pp. 2779–2785.

[43] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *International Conference on Learning Representations*, 2018.

[44] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in *International Conference on Learning Representations*, 2017.

[45] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," *arXiv preprint arXiv:1703.00410*, 2017.

[46] K. Madono, M. Tanaka, M. Onishi, and T. Ogawa, "Block-wise scrambled image recognition using adaptation network," in *AAAI WS*, 2020.

[47] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[49] P. Chen, Y. Sharma, H. Zhang, J. Yi, and C. Hsieh, "EAD: elastic-net attacks to deep neural networks via adversarial examples," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, S. A. McIlraith and K. Q. Weinberger, Eds., 2018, pp. 10–17.

[50] G. W. Ding, L. Wang, and X. Jin, "AdverTorch v0.1: An adversarial robustness toolbox based on pytorch," *arXiv preprint arXiv:1902.07623*, 2019.