Hilbert-based Generative Defense for Adversarial Examples

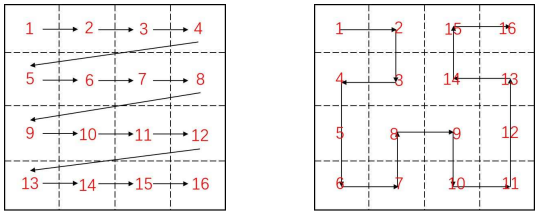
Yang Bai[◇*] Yan Feng[‡*] Yisen Wang[§†] Tao Dai[‡] Shu-Tao Xia[‡] Yong Jiang[◇‡†]
[◇]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
[§]Dept. of Computer Science and Engineering, Shanghai Jiao Tong University
[‡]Graduate School at Shenzhen, Tsinghua University
{y-bai17, y-feng18}@mails.tsinghua.edu.cn; eewangyisen@gmail.com

Abstract

Adversarial perturbations of clean images are usually imperceptible for human eyes, but can confidently fool deep neural networks (DNNs) to make incorrect predictions. Such vulnerability of DNNs raises serious security concerns about their practicability in security-sensitive applications. To defend against such adversarial perturbations, recently developed PixelDefend purifies a perturbed image based on PixelCNN in a raster scan order (row/column by row/column). However, such scan mode insufficiently exploits the correlations between pixels, which further limits its robustness performance. Therefore, we propose a more advanced Hilbert curve scan order to model the pixel dependencies in this paper. Hilbert curve could well preserve local consistency when mapping from 2-D image to 1-D vector, thus the local features in neighboring pixels can be more effectively modeled. Moreover, the defensive power can be further improved via ensembles of Hilbert curve with different orientations. Experimental results demonstrate the superiority of our method over the state-of-the-art defenses against various adversarial attacks.

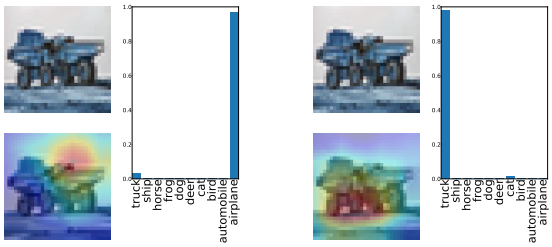
1. Introduction

Recent work has shown that the input images with small and carefully designed perturbations (a.k.a., adversarial examples) can cause deep neural network classifier to produce confidently wrong predictions [28, 8]. Since the differences between adversarial examples and corresponding clean images are usually imperceptible, the existence of DNN vulnerability to such adversarial examples exposes a serious concern about their great popularity and widespread applications [9, 6, 17]. Thus, it is highly demanded to defend against such adversarial examples.



(a) Raster scan (b) Hilbert scan

Figure 1: The scan order of (a) Raster scan (row by row and column by column), and (b) Hilbert scan.



(a) PixelDefend (Raster scan) (b) PixelDefend (Hilbert scan)

Figure 2: The defend performance of PixelDefend with (a) Raster scan, and (b) Hilbert scan. The top images are adversarial example while the bottom images are heatmaps of the purified images by PixelDefend. The Hilbert-based method classifies the "truck" adversarial example correctly while Raster-based one fails to do that.

There already exist some works to defend against adversarial examples, including pre-processing the inputs [10], gradient regularization [22], and adversarial training [18, 31]. Among them, pre-processing the inputs directly may be more practical due to its model- and attack-annostic property, such as Defense-GAN [25], PixelDefend [27] and

*Equal contribution.
†Correspondence to: Yisen Wang and Yong Jiang.

other traditional image transformation methods [5, 10]. PixelDefend and Defense-GAN are similarly based on generative models to purify the input images. The differences are that the former focuses on **low-level (pixel-level)** explicit density while the latter focuses on **high-level** implicit density. As the perturbations are often restricted by L_p norm, pixel-level method implies much more potential to be an effective defense, which is the main focus in this paper.

Although PixelDefend has obtained impressive performance, there still exist some drawbacks. For example, the pixel dependencies have not been fully exploited, since the core of PixelDefend is PixelCNN [30, 24] which exploits pixel dependencies in a raster scan order (Figure 1a, row by row or column by column). However, raster scan cannot always preserve local consistency, thus failing to model local features well. For example, the pixel (4) and pixel (5) in Figure 1a are nearby pixels in raster curve, but are far away in spatial domain.

In order to better model pixel dependencies and preserve local consistency, we propose a Hilbert-based PixelDefend (HPD) by scanning pixels along Hilbert curves [33, 12]. Compared with the raster scan, Hilbert scan has the property of *spatial proximity*, *i.e.*, two close pixels in 1-D domain are also close in 2-D domain, which improves the characterization of mutual dependency between pixels. The *spatial proximity* also performs better in modeling the local features of an image. As a preview of this, we compare the defense results of PixelDefend with different pixel scan order in Figure 2, which demonstrates Hilbert-based PixelDefend can better keep the local features and find the right active area in a heatmap [34], thus getting the correct classification distribution. Moreover, Hilbert scan can exploit *self similarity*, *i.e.*, each quadrant can be divided into small quadrants similar to itself. The *self similarity* can be used to develop a natural ensemble model with various Hilbert orientations, further boosting the defense performance. The main contributions are summarized as follows:

- We propose to use Hilbert scan in place of traditional raster scan to model the pixel dependencies for better utilizing local features.
- We propose a Hilbert-based Generative Defense, named Hilbert-based PixelDefend (HPD), against adversarial examples via purifying the input images. Moreover, we explore a natural model ensemble through various Hilbert orientations.
- Experimental results show that HPD obtains consistently better performance than the original PixelDefend. Besides, our ensemble version achieves robustness gain from 0.15 to 0.52 on CIFAR-10 with ResNet50 attacked by Obfuscated Gradient attack [2].

2. Related Work

In this section, we make a brief review on several common adversarial attacks and defense.

2.1. Adversarial Attack

Given a normal example X, ϵ ball around X limits the perturbation strength of adversarial example X'. The symbol ℓ represents a loss function, F is the output from softmax layer, Z is the output from logit layer, and T is the target label (if targeted attack, else T is the original label). A wide range of attacking methods have been proposed for the crafting of adversarial examples.

Fast Gradient Sign Method (FGSM) [9]. FGSM generates adversarial examples by linearly approximating a neural network model and maximizing the loss along the gradient direction:

$$X^{'} = X + \epsilon \cdot sign(\nabla \ell_{F,T}(X)) \, . \tag{1}$$

Projected Gradient Descent (PGD) [13]. PGD is a multi-step attack. It replaces ϵ with α to take a small step in each iteration. After each step, PGD projects the adversarial example back onto the ϵ-ball:

$$X_i^{'} = X_{i-1}^{'} + clip_{\epsilon}(\alpha * sign(\nabla \ell_{F,T}(X_{i-1}^{'}))) \, . \tag{2}$$

C&W [4]. Carlini and Wagner proposed an optimization based attack. Assuming κ as the confidence of one classification result, c as a hyperparameter, and $Z(X)_i$ as the logit output of X labeling to i, C&W tries to minimize:

$$\|X' - X\|_{\infty} + c \cdot f\,(X') \;\text{ with} \tag{3}$$

$$f\,(X') = max(max\,\{Z(X')_i : i \neq T\} - Z(X')_T, -\kappa) \, .$$

Obfuscated Gradient [2]. Obfuscated Gradient method attacks defense methods by Backward Pass Differentiable Approximation (BPDA) when gradients are not available. It has been shown to successfully break many defense models including PixelDefend.

2.2. Adversarial Defense

A number of defense models have been developed including defensive distillation [23], adversarial training [18], dynamic adversarial training [31], feature squeezing [15], Defense-GAN [25] and so on.

Adversarial Training [18]. Adversarial Training is proposed to retrain a network with adversarial examples. It performs well in defending overfitting adversarial examples, but cannot defend adversarial examples caused by the linearity of networks. Ensemble one [29] can get better results with adversarial examples generated by different attacks.

Feature Squeezing [15]. Feature Squeezing detects adversarial examples by comparing classification results of original images and squeezed images. The squeezed images are

generated by reducing the color bit depth of each pixel and performing spatial smoothing later. If a large difference is found between the classification results, the image is considered to be adversarial.

Defense-GAN [25]. Defense-GAN is based on WGAN [1, 7] and the loss function follows WGAN's loss design. The classifier is trained on the same real data set. If the GAN is well trained and can represent the original data distribution, there should be no significant difference between the original image and their reconstruction. Defense-GAN attempts to remove noises from the perturbed inputs and thus defends against any form of attack.

3. The Proposed Generative Defense

3.1. Preliminary

PixelDefend purifies an image based on PixelCNN (PixelRNN) [30, 24] whose core idea is a chain rule to decompose the likelihood of an image X with size $n \times n$ into product of a 1-D sequence following raster scan order. The joint distribution $p(X)$ is denoted as the product of conditional distributions over pixels:

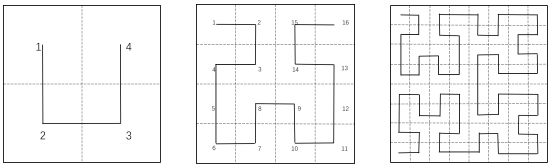
$$p(X) = \prod_{i=1}^{n^2} p(x_i \mid x_1, ..., x_{i-1}) \,. \tag{4}$$

The purifying process is repeated from the first pixel to the last pixel according to the 1-D sequence, *i.e.*, from the top left corner to the bottom right corner by raster scan.

Despite that PixelDefend shows great potential in defense by purifying adversarial examples, the dependency among pixels is modeled in conditional probability according to raster scan order. As closer pixels in 2-D domain are more likely to contain related information and local features, they are expected to be closer in 1-D. So scan order is important when mapping pixels from 2-D to 1-D.

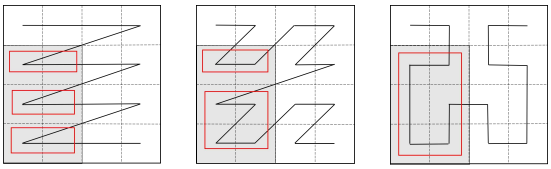
3.2. Strengths of Hilbert Scan

Different from the common row by row and column by column manner of raster scan, Hilbert scan generates 1-D sequence along Hilbert curves [19] that can be constructed as Figure 3. We have compared several space filling curves including Raster-order, Z-order and Hilbert-order in Figure 4, we can see that Hilbert curves and their approximations are more likely to cluster pixels. Cluster is defined as a group of pixels consecutively connected by a mapping (or a curve), which is supposed to keep local features [20]. As shown in Figure 4, in the same chosen area, Raster-order has 3 clusters, Z-order has 2 clusters, while Hilbert-order has 1 cluster. Naturally nearby pixels clustered together help to keep local features. To illustrate this, we present the image generation process of PixelCNN using different scan methods in Figure 5. Raster scan make PixelCNN generate the image in a row by row manner (Figure 5a), while



(a) Start (b) 1 iteration (c) 2 iterations

Figure 3: The construction of Hilbert curves.



(a) Raster Scan (b) Z Scan (c) Hilbert Scan

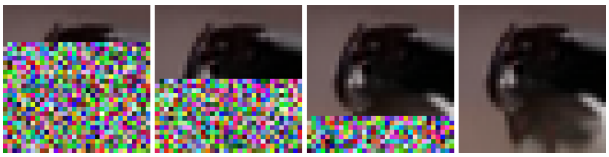
Figure 4: The comparison of space filling curves. In the same chosen area, Raster-order has 3 clusters, Z-order has 2 clusters, while Hilbert-order has 1 cluster, implying local features are kept better in Hilbert-order.

Hilbert scan make PixelCNN generate the image in a block by block way (Figure 5b) because pixels are well clustered by Hilbert scan. The block by block generation mode is more likely to extract the local features of an image, which helps a lot in classification and defense.

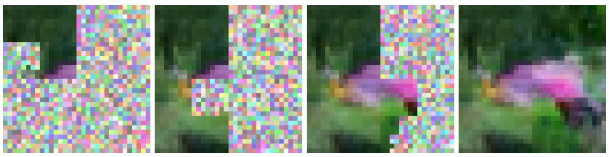
In addition, the construction of Hilbert curves in Figure 3 implies its another good property, that is self similarity (a self-similar object is exactly or approximately similar to a part of itself). In other words, each quadrant can be divided to small quadrants in the same way and scanned by the U-shape curve. The self similarity property of Hilbert curves indicates that itself owns a certain of intrinsic diversity, which are naturally useful for the Hilbert scan based ensemble defense.

Hilbert scan also shows great superiority in the pixel dependencies due to its spatial proximity. The ratios of pixel pairs' 2-D distances over 1-D distances by raster scan and Hilbert scan are compared in Figure 6. As can be seen, Hilbert scan better keeps local consistency of 2-D and 1-D domain on images with different sizes, as the ratios are consistently flat and small. We also find the neighboring pixels are more related in Hilbert 1-D sequence, which is shown in Figure 7. Neighboring pixel pairs from Hilbert curve are more related as the curve is more close to $y = x$.

From above observations, we can conclude that better modeling of the pixel dependencies helps to learn local features. While local features are of great importance in classification task [3]. Existing defense methods using SIFT [16] and other transformations also show that keeping local features is essential in adversarial defense [32]. Therefore, Hilbert scan could be a promising and effective method in



(a) Images generated by PixelCNN in Raster order



(b) Images generated by PixelCNN in Hilbert order

Figure 5: The images generated by PixelCNN in (a) Raster order and (b) Hilbert order. These two series of images clearly show the different generation order in pixels, *i.e.*, row by row in (a) and block by block in (b).

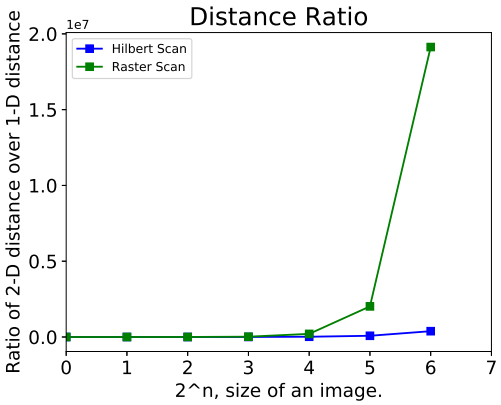


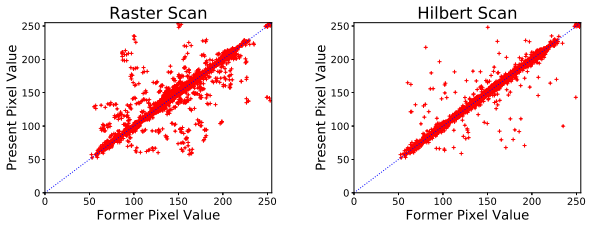
Figure 6: The ratios of 2-D and 1-D distances by Hilbert scan (blue line) and Raster scan (green line). x-axis is the order of Hilbert curve, as 2^n is the width of a square image. y-axis is the average ratios of pixel pairs' 2-D distances over 1-D. Smaller y value means the 1-D sequence can better describe the actual distances of 2-D.

place of raster scan in adversarial defense community.

3.3. Hilbert-based PixelDefend (HPD)

Based on the above analysis, we propose Hilbert-based PixelDefend by applying Hilbert scan to PixelDefend. Hilbert-based PixelDefend purifies pixels along the Hilbert curve based on the Hilbert-based PixelCNN, which re-aligns image pixels and models pixel dependencies using Hilbert scan. The pseudo code of our proposed Hilbert-based PixelDefend (HPD) approach is shown in Algorithm 1.

Moreover, we provide a theoretical guarantee on the possibility of finding the optimal clean image around perturbed one for our method.



(a) Raster scan (b) Hilbert scan

Figure 7: The neighboring pixel values' dependencies of Raster scan and Hilbert scan. x-axis represents the former pixel value and y-axis represents the present pixel value in 1-D curve. Neighboring pixel pairs by Hilbert scan show stronger correlation.

Algorithm 1 Hilbert-based PixelDefend (HPD).

Input: $X^{'}$: Adversarial Example; ϵ_{defend}: Defense parameter; *HPixelCNN*: Pre-trained Hilbert-based PixelCNN model.

Output: $X_{purified}$: Purified image

1: Realign an image into Hilbert-ordered.

$$X^{'}_{Hilbert} \Leftarrow Hilbert\_scan(X^{'})$$

2: For each pixel value (row i, column j, channel k).

$$x \Leftarrow X^{'}_{Hilbert}[i,j,k]$$

3: Set defense range.

$$R \Leftarrow [x - \epsilon_{defend}, x + \epsilon_{defend}]$$

4: Compute the 256-way softmax *HPixelCNN* ($X^{'}_{Hilbert}$)

5: Update

$$X_{Hilbert}[i,j,k] \Leftarrow argmax_{z \in R} \textit{HPixelCNN}[i,j,k,z]$$

6: Realign $X_{Hilbert}$ into originally ordered 2-D image.

$$X_{purified} \Leftarrow ReHilbert\_scan(X_{Hilbert})$$

7: **return** $X_{purified}$

Theorem 1. *The optimal clean image could be found by the proposed greedy search algorithm iff the first pixel is accurate ($\hat{x}_1 = \mu_1$). Given the perturbation constrain ϵ_{attack}, then $|\hat{x}_1 - \mu_1| \leq \epsilon_{attack}$. The clean image could be found with the probability of at least $\frac{1}{2\epsilon_{attack}}$.*

Proof. Assume input image $\mathbf{X} = (x_1, x_2, \cdots, x_{n^2}) \sim$

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and denote $\boldsymbol{\Sigma}_i (i = 1, 2, \cdots, n^2 - 1)$ as the i-th order principal minor determinant of $\boldsymbol{\Sigma}$, $\hat{\mathbf{X}}$ as the optimal estimation of purified image obtained by greedy search algorithm, and $\mathbf{X}_{opt}$ as the corresponding clean image ($\mathbf{X}_{opt}$ = argmax p($\mathbf{X}$) = $(\mu_1, \mu_2, \cdots, \mu_{n^2})$). According to the property of multivariate Gaussian distribution, if $|\hat{x}_1 - \mu_1| = 0$, we have

$$\begin{aligned}
\hat{x}_2 &= \text{argmax } p(x_2|\hat{x}_1) \\
&= \mu_2 + \widetilde{k}_1 \boldsymbol{\Sigma}_1^{-1} \widetilde{x}_1^T \\
&= \mu_2 + \text{cov}(x_2, x_1)\text{cov}(x_1, x_1)^{-1}(\hat{x}_1 - \mu_1) \\
&= \mu_2. \\
&\cdots \\
\hat{x}_{n^2} &= \text{argmax } p(x_{n^2}|\hat{x}_1, \cdots, \hat{x}_{n^2-1}) \\
&= \mu_{n^2} + \widetilde{k}_{n^2-1} \boldsymbol{\Sigma}_{n^2-1}^{-1} \widetilde{x}_{n^2-1}^T \\
&= \mu_{n^2} + \widetilde{k}_{n^2-1} \boldsymbol{\Sigma}_{n^2-1}^{-1} [\hat{x}_1 - \mu_1, \cdots, \hat{x}_{n^2-1} - \mu_{n^2-1}]^T \\
&= \mu_{n^2-1},
\end{aligned}$$

where $\widetilde{k}_{i-1} = [cov(x_i, x_1), \cdots, cov(x_i, x_{i-1})], \widetilde{x}_{i-1} = [\hat{x}_1 - \mu_1, \cdots, \hat{x}_{i-1} - \mu_{i-1}], i = 1, 2, \cdots, n^2 - 1$. Therefore, $\hat{\mathbf{X}} = [\mu_1, \mu_2, \cdots, \mu_{n^2}] = \mathbf{X}_{opt}$ □

3.4. Ensemble Hilbert-based PixelDefend (EHPD)

Apart from the spatial proximity to improve defense performance, Hilbert curve has another good property of self similarity, as shown in Figure 3. Self-similarity means that each local component of an image is similar to the global one. Naturally it can be considered as one kind of data augmentation. Hilbert scan follows such a self-similar pattern with different scan orientations, and is more suitable for ensembling than raster scan.

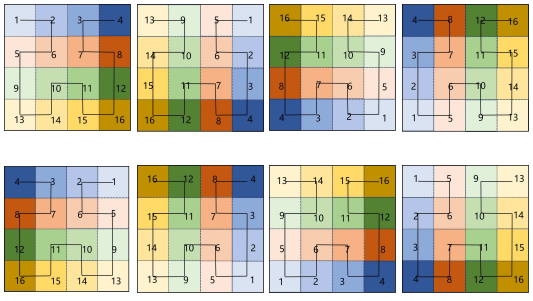


Figure 8: The ensemble Hilbert curves.

There are 4 typical kinds of 1-D sequences generated by rotating the original image before Hilbert scan. While another 4 are generated by inverting the original image before rotating and scanning, as is shown in the bottom row of Figure 8. Differently ordered 1-D sequences can provide different former pixels information which brings a certain of diversity, thus it is reasonable to gain better defense results, especially when defending against Obfuscated Gradient attack, which is claimed to break many defense models effectively. In fact, HPD can be regarded as a specific case of EHPD, where it only uses one Hilbert curve. The pseudo code of our proposed Ensemble Hilbert-based PixelDefend (EHPD) approach is shown in Algorithm 2.

Algorithm 2 Ensemble Hilbert-based PixelDefend (EHPD).

Input: $X^{'}$: Adversarial Example
Output: X_{final}: Purified image

1: Realign the image along the predefined Hilbert curves in Figure 8 into 8 independent Hilbert-ordered images.

$$X^i_{Hilbert} \Leftarrow Hilbert\_scan_i(X^{'}), i \in [1, 8]$$

2: Purify each of them with the proposed HPD in Algorithm 1.

$$X^i_{purified} \Leftarrow HPD(X^i_{Hilbert})$$

3: For each purified images, predict its one-hot classification with pre-trained CNN, N represents the number of classes.

$$Y_i[1, ..., N] \Leftarrow CNN(X^i_{purified})$$

4: Assign ensembled prediction $Y_{ensemble}$ with the class that has the highest frequency in Y_i.

$$Y_{ensemble} \Leftarrow argmax_{c \in [1,N]}(\sum_{i \in [1,8]} Y_i[1, ..., N])$$

5: Choose the purified image whose prediction is closest to $Y_{ensemble}$.

$$idx \Leftarrow argmax(cross\_entropy(Y_i, Y_{ensemble}))$$
$$X_{final} \Leftarrow X^{idx}_{purified}$$

6: **return** X_{final}

4. Experiments

In this section, we evaluate the robustness of our proposed methods, HPD and EHPD, compared with several state-of-the-art defense models against white-box attacks and black-box attacks. The former white-box attack has the complete knowledge of the target model and full access of the model parameters, while the latter black-box attack can only query the model to get outputs and have no access to model parameters.

Baselines. The baseline defense models we use include 1) Adversarial Training (AT): Training with various adver-

Table 1: White-box robustness of different defense methods on ResNet50 (attack strengths ϵ=2/8/16, defense strength ϵ_{defend}=16, with pixel values clipped in [0,255]).

		CLEAN	FGSM	PGD	CW
ϵ=2	Normal ResNet	0.90	0.32	0.07	0.04
	AT [18]	0.88	0.79	0.73	0.76
	FS [15]	0.90	**0.84**	0.81	0.04
	DG [25]	0.51	0.52	0.49	0.47
	PD [27]	0.90	0.82	**0.82**	0.82
	HPD	0.90	**0.84**	**0.82**	**0.86**
ϵ=8	Normal ResNet	0.90	0.14	0.00	0.00
	AT [18]	0.88	0.51	0.29	0.39
	FS [15]	0.90	0.49	0.51	0.00
	DG [25]	0.51	0.47	0.47	0.50
	PD [27]	0.90	**0.79**	0.73	0.82
	HPD	0.90	0.78	**0.75**	**0.86**
ϵ=16	Normal ResNet	0.90	0.08	0.00	0.00
	AT [18]	0.88	0.34	0.06	0.09
	FS [15]	0.90	0.26	0.34	0.00
	DG [25]	0.51	0.46	0.44	0.50
	PD [27]	0.90	0.53	0.56	0.82
	HPD	0.90	**0.60**	**0.64**	**0.86**

sarial examples generated by different attack methods; 2) Feature Squeezing (FS): using the traditional image processing to restore important features of a perturbed image; 3) Defense-GAN (DG): using GAN to restore an adversarial example; and 4) PixelDefend (PD): using another generative model PixelCNN to purify an adversarial example.

Defense Settings. For CIFAR-10, our proposed defense methods, HPD and EHPD, are set with defense parameter ϵ=16 (pixel values clipped in [0,255] as default), as well as PixelDefend. Evaluated by test negative log-likelihood, the PixelCNN models are trained to 3.4 bit/dim in our Hilbert-based PixelCNN and 2.92 bit/dim in PixelCNN from pretrained OpenAI PixelCNN++ [24].

For the attack methods, FGSM, PGD, and C&W (L_∞) are all set with ϵ=2/8/16. PGD is set with step size 1 and max attack steps 3/10/20. C&W is set with step size 1.275 and max attack epoch 100, hyperparameter $c = 100$. Obfuscated Gradient attack is set with maximum attack epoch 20 and step size 0.5. For the defense methods, adversarial training is retrained with adversarial examples generated by PGD attack with ϵ=8 and maximum attack step 10. Feature squeezing is set with color depth reduction as 2^5. Defense-GAN is trained with 30000 iterations and batch size 32.

Table 2: Test accuracies of PD, HPD and EHPD against PGD. H_iPD represents using the i_{th} curve among all 8 different hilbert curves when generating purified images. $EHPD_8$ represents ensembling 8 different curves.

	CLEAN	PGD(ϵ=2)	PGD(ϵ=8)	PGD(ϵ=16)
PD	0.90	0.82	0.73	0.56
EPD_8	0.90	0.82	0.74	0.59
H_0PD	0.90	0.82	0.75	0.64
H_1PD	0.90	**0.83**	0.76	0.64
H_2PD	0.90	0.82	0.75	0.64
H_3PD	0.90	0.82	0.76	0.65
H_4PD	0.90	**0.83**	0.75	0.65
H_5PD	0.90	0.82	0.75	0.64
H_6PD	0.90	0.82	0.75	0.64
H_7PD	0.90	0.82	0.75	0.64
$EHPD_8$	**0.90**	0.82	**0.77**	**0.66**

4.1. White-box Robustness Evaluation

We test the white-box robustness on CIFAR-10 with ResNet50 [11]. The results are shown in Table 1. We can see that accuracies on normal networks decrease sharply under white-box attacks, especially with larger perturbations. Networks with defense methods demonstrate more robust results. The proposed HPD almost achieves the best robustness among all the defense methods. The effectiveness of feature squeezing defense provides a strong evidence that local features are important in adversarial defense. It is thus reasonable that Hilbert-based PixelDefend achieves better defense performance, as Hilbert scan has spatial proximity and keep the local features better. As ϵ grows larger, the defense superiority of HPD becomes more and more significant, for example, the robustness against PGD with ϵ=16 is promoted from 0.56 to 0.64 compared with PD. When compared with other methods, the improvements are larger. These improvements imply that HPD still can model pixel dependencies and keep local features well, and shows a strong defensive ability, even images are perturbed by a large extent.

Moreover, we compare Ensemble PixelDefend and Ensemble Hilbert-based PixelDefend with 8 curves, named by EPD_8 and $EHPD_8$. The underlying number represents the number of curves used to ensemble. For EHPD, we first test the defense results of HPD with 8 different Hilbert curves independently. Then we do further explorations on the ensemble method EHPD. The results of $EHPD_8$ with HPD series against PGD with attack perturbations ϵ=2/8/16 in L_p norm are shown in Table 2. We find HPD series achieve comparable results, while $EHPD_8$ performs better against

white-box attacks. Especially when the attack strengths increase, the robustness is improved from 0.75 to 0.77 with ϵ=8 and 0.64 to 0.66 with ϵ=16. Although the 8 HPD models achieve similar accuracy independently, the predictions may vary in specific images because of their different scan orders. The self similarity of Hilbert curves suggests various differently ordered Hilbert curves could be generated from one image naturally. EHPD ensembles these models effectively, thus achieving even better results.

4.2. Defense for Obfuscated Gradient Attack

Obfuscated Gradient attack is a kind of defense-guided attack which could effectively bypass specific defense methods. It estimates the hidden gradients of a defense method by sampling the output, thus defense methods with certain boundaries could be broken with a high possibility, including PixelDefend [27]. However, our proposed HPD can be expanded to a series of defense models, brought by the diversity of Hilbert curves. Although some of the benefits are already exploited in EHPD, we conduct a series of experiments especially against Obfuscated Gradient attack, to further explore the benefits of Hilbert scan.

We compare the performances of PD, HPD and EHPD against Obfuscated Gradient attack, and show the accuracies in Table 3 (with attack parameter ϵ=8/16 in L_∞ norm). The results show that the basic HPD method already outperforms PD by a large margin, from 0.15 to 0.39. It results from the outstanding performance of HPD in modeling pixel dependencies. HPD could return a more qualified image in each purifying iteration, thus more robust against Obfuscated Gradient attack. On top of that, the robustness increases when more HPD models with different Hilbert curves are ensembled. Considering HPD as a special case of EHPD (with only 1 Hilbert curve to ensemble), we find the accuracies increase from 0.39, to 0.49, 0.52, when ensembling 1 Hilbert curve, 4 Hilbert curves and 8 Hilbert curves. It also shows potential to a further improvement with more Hilbert curves to ensemble.

We also compare the average epochs needed to break a defense by Obfuscated Gradient attack. The results are shown in Table 4. As we limit the max epoch of Obfuscated Gradient attack to 20, the attack epoch of images still classified correctly after 20-epoch attack is recorded as 20. We find the average attack epoch increases with more ensemble Hilbert curves. EHPD_8 improves the result from 11 to 17 compared with PD. The trend shows the average attack epoch is gradually close to the limited maximum epoch number 20, implying the performance in average attack epoch is still limited by our setting. The difference could be larger with a loosed maximum attack epoch.

Table 3: Test accuracy of defense methods against Obfuscated Gradient attack.

Defense Method	PD	H_0PD	EHPD_4	EHPD_8
Obfuscated Gradient	0.15	0.39	0.49	**0.52**

Table 4: Average Epochs needed for a defense method to be broken by Obfuscated Gradient attack. Larger epochs mean the method is more robust against the attack.

Defense Method	PD	H_0PD	EHPD_4	EHPD_8
Average Epochs	11	14	16	**17**

4.3. Black-box Robustness Evaluation

The classification model applied in our black-box experiments is a normal VGG [26] with 0.90 accuracy (target model), while adversarial examples are generated on the normal ResNet50 with 0.90 accuracy (source model). That is, there is no information from the classification model during the generation of adversarial examples. We test the performance against adversarial attack FGSM and PGD with ϵ=8/16. As PD has already shown good performance on defense results, we mainly compare HPD with PD. The defense parameters are set to ϵ_{defend}=16 in both PD and HPD methods.

The results of PD and HPD against black-box attack are demonstrated in Table 5. We find that even though VGG shows strong classification ability, it is vulnerable to adversarial examples. Adversarial examples generated on ResNet50 could effectively mislead a well-trained normal VGG model. This is explained as the transferability of adversarial examples [14] [21]. With defense methods, VGG could be much more robust. For example, the accuracy improves from 0.29 to 0.66 against FGSM adversarial examples with ϵ=16 on ResNet50 defended by PD. Our proposed HPD shows a clear better robustness performance, improving the accuracy by 0.02-0.03 in almost all black-box defense cases. The results show that HPD is more effective to defend against black-box attack. The superiority also arises from spatial proximity of Hilbert curve, as it helps HPD better preserve local feature in the generated images, which is essential in classification tasks for almost all DNN models.

4.4. Analysis and Discussion

From above experiments, we find EHPD performs much better than original PD method. As for the reason, we conjecture that the diversities brought by Hilbert curves itself contribute most in the defense. Therefore, we test a random

Table 5: Black-box robustness from source model ResNet 50 to target VGG model (with attack strengths ϵ=8/16 and defense strength ϵ_{defend}=16, pixel values clipped in [0,255]).

	CLEAN	FGSM		PGD	
		ϵ=8	ϵ=16	ϵ=8	ϵ=16
Normal VGG	0.90	0.54	0.29	0.61	0.39
PD	0.90	0.76	0.66	0.77	0.70
HPD	0.90	**0.78**	**0.68**	**0.80**	**0.72**

model of HPD instead of the ensemble model. To be specific, in each iteration, we randomly pick one model from all HPD models with different Hilbert scan orders. For simplicity, we use RHPD to represent this random model.

Apart from accuracy on a data set with a number of test images, the cross entropy loss of classification network for a single image can also be used to evaluate and analyze the defense ability. A wrong classification always comes along with a large cross entropy loss, and vice verse. So when a sharp increase of cross entropy loss occurs, the classification is almost certain to be wrong.

We illustrate the loss trend of one image in Figure 9 to further exploit the benefits of randomness. The image is attacked by Obfuscated Gradient and defended by PD, HPD and RHPD respectively. Figure 9 shows that HPD can delay the sharp increase by 6-7 epochs compared with PD, but they are all in an overall increasing trend. On the other hand, RHPD could decrease the loss during defense. It indicates that the diversity of Hilbert curves is essential in adversarial defense, as differently ordered sequences could offer different pixel information. The diversity and randomness could help to decrease the cross entropy loss during purifying and thus improve robustness.

In summary, our experiments show that the scan order is an important factor for PixelDefend (as well as PixelCNN). Hilbert scan improves the performance of PixelDefend due to the spatial proximity and self similarity. The spatial proximity helps provide a pixel with more related information from closer pixels, thus well keeps local features and improves the defense performance. The self similarity brings a lot of diversities, thus the defense method can be more robust against white-box and black-box attacks, especially against Obfuscated Gradient attack.

Lastly, there is one point we want to mention is that Hilbert scan requires the width and height of an image to be 2^n. When it comes to images with free shape, we need to do some basic upsampling, downsampling or padding manipulations in advance so as to resize an image into a normalized 2^n shape.

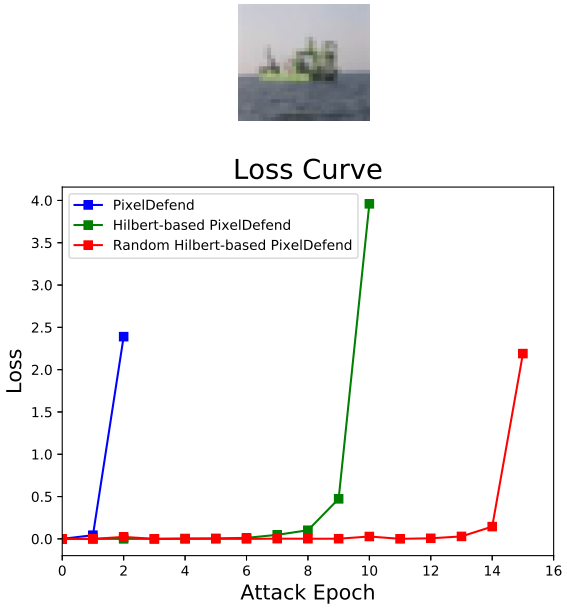


Figure 9: The top row is one randomly picked image from the test set, used as an example to analysis the property of defense methods (PD, HPD and RHPD). The bottom is the cross entropy loss curves of the top image, defended by PD, HPD and RHPD to against Obfuscated Gradient attack.

5. Conclusion

In this paper, we proposed to use Hilbert scan instead of the widely used raster scan to precisely characterize the dependencies of pixels. Further, we propose a Hilbert scan based generative defense model, named Hilbert-based PixelDefend (HPD), against adversarial examples by purifying the perturbed images pixel by pixel. Due to the self similarity of Hilbert curves, HPD can be naturally extended to the ensemble model, called Ensemble Hilbert-based PixelDefend (EHPD), using different Hilbert curves. Experiments on benchmark dataset demonstrate that our proposed HPD and EHPD methods outperform the state-of-the-art defenses against white-box attacks, black-box attacks, particularly the Obfuscated Gradient attack. The spatial proximity and self similarity properties of Hilbert curves contribute most to the superiority of our proposed HPD and EHPD method.

Acknowledgement

This work is supported in part by the National Key Research and Development Program of China under Grant 2018YFB1800204, the National Natural Science Foundation of China under Grant 61771273, the R&D Program of Shenzhen under Grant JCYJ20180508152204044, and the research fund of PCL Future Regional Network Facilities for Large-scale Experiments and Applications PCL2018KP001).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 3

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 2

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *ICLR*, 2019. 3

[4] Wagner D Carlini N. Towards evaluating the robustness of neural networks. In *S&P*, 2017. 2

[5] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *KDD*, 2018. 2

[6] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018. 1

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1

[9] Shlens Jonathon Szegedy Christian Goodfellow, Ian J. Explaining and harnessing adversarial examples. *Computer Science*, 2014. 1, 2

[10] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2018. 1, 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[12] Hosagrahar V Jagadish. Analysis of the hilbert curve for representing two-dimensional space. *Information Processing Letters*, 1997. 2

[13] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR*, 2017. 2

[14] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017. 7

[15] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and XiaoFeng Wang. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing*, 2018. 2, 6

[16] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 3

[17] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *ICLR*, 2018. 1

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1, 2, 6

[19] John McVay, Nader Engheta, and Ahmad Hoorfar. High impedance metamaterial surfaces using hilbert-curve inclusions. *IEEE Microwave and Wireless components letters*, 2004. 3

[20] Bongki Moon, Hosagrahar V Jagadish, Christos Faloutsos, and Joel H. Saltz. Analysis of the clustering properties of the hilbert space-filling curve. *TKDE*, 2001. 3

[21] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 7

[22] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017. 1

[23] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *S&P*, 2016. 2

[24] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *ICLR*, 2017. 2, 3, 6

[25] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *ICLR*, 2018. 1, 2, 3, 6

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 7

[27] Yang Song, Taesup Kim, , Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *ICLR*, 2018. 1, 6, 7

[28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1

[29] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 2

[30] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016. 2, 3

[31] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *ICML*, 2019. 1, 2

[32] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *ACM Workshop on Artificial Intelligence and Security*, 2017. 3

[33] Jian Zhang, Sei-Ichiro Kamata, and Yoshifumi Ueshige. A pseudo-hilbert scan for arbitrarily-sized arrays. *IEICE*

Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2007. 2

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2