Robust Attack on Deep Learning based Radar HRRP Target Recognition

Yijun Yuan, Jinwei Wan, Bo Chen
National Laboratory of Radar Signal Processing
Collaborative Innovation Center of Information Sensing and Understanding
Xidian University
Xi'an, China
sixnkee603@126.com, bchen@mail.xidian.edu.cn

***Abstract*—In the past few years, deep learning have attracted increasing attention for HRRP-based radar automatic target recognition(RATR) because of their powerful ability to learn features from training data automatically. However, recent studies show that deep learning models are vulnerable to adversarial examples. In this paper, we verified adversarial examples also exist in the deep learning based HRRP target recognition. A novel adversarial attack algorithm called Robust HRRP Attack(RHA) is proposed to generate robust adversarial perturbations in real-world. Experimental results on measured HRRP data show that RHA decrease HRRP recognition performance significantly which indicate our method is efficient and robust.**

I. Introduction

High-resolution range profile (HRRP), is the amplitude of the coherent summations of the complex time returns from target scatters in each range "cell", which represents the projection of the complex returned echoes from the target scattering centers onto the radar line-of-sight (LOS). HRRP-based RATR is an active research field of modern radar technology [1]–[8] because HRRP contains abundant target structure signatures, such as target size, scatter distribution, etc. Feature engineering is the critical part for HRRP-based RATR task. Wan et al. utilize convolutional neural network (CNN) as feature extractor which achieve better performance than traditional methods in HRRP recognition [9].

However, recent studies show that deep neural networks are vulnerable to small, carefully designed perturbations of the input [10]. For example, adding visually imperceptible perturbations to the input can result in classification failures. Initial works on adversarial examples were mainly about image classification. But recent years, adversarial examples have been proved to exist on domains ranging from image segmentation [11] to face detection [12]. But no researchers have paid attention to adversarial examples in HRRP-based RATR. Adversarial examples can be roughly divided into two parts: digital adversarial examples and physical adversarial examples [13]. Goodfellow et al. proposed FGSM [14], a fast and first-order gradient based method to construct adversarial examples. In [15], an effective optimization-based attack model has been proposed to create adversarial perturbations. These methods make contributions to digital adversarial examples. Simultaneously, some researchers are interested in physical adversarial examples. In [16], the researchers shows experiments that digital adversarial examples failed to fool a object detectors across a scale of different distances and angles, from which we can find digital adversarial examples do not work well in physical world. It's difficult to make physical adversarial examples.

In this paper, We focus on making robust perturbation on time domain HRRP which could significantly decrease the HRRP recognition performance. We first verified adversarial examples also exist in the field of HRRP-based RATR. Fig. 1 shows an digital adversarial example created by FGSM [14], from which we can see adding small magnitude perturbation on the clean HRRP can fool a neural network easily. We further take physical conditions such as the length and position of perturbation into account to design Robust HRRP Perturbations(RHA), which can significantly decrease HRRP recognition performance in physical world. Our experiments show that RHA decrease HRRP recognition performance significantly which indicate our method is efficient and robust. As far as we know, this is the first paper which focus on the adversarial examples in the filed of HRRP-based RATR.

II. Proposed Method

We first describe our deep learning based Radar HRRP Target Recognition method, and then define the perturbation which can be implemented in real world and present our algorithm on how to generate physical adversarial perturbations.

A. Time domain HRRP recognition

The high-resolution radar (HRR) operates in microwave frequency band in general. For the wide bandwidth of the signal, the wavelength of radar is much smaller than the targets' size. The HRR can effectively divide the object into many range "cells" for complex targets such as aircrafts. According to [4], the t-th time domain complex HRRP can be written as $\tilde{x}(t) = e^{j\theta(t)}\left[\tilde{x}_1(t), \tilde{x}_2(t), ..., \tilde{x}_L(t)\right]^\top$, where $\top$ denotes the transpose operation, $\theta(t)$ stands for the initial phase of the t-th returned echo, and $\tilde{x}_l(t) = \sum_{i=1}^{V_l} \sigma_{li} e^{j\phi_{li}(t)}$ denotes the echo of l-th range cell, which is composed of V_l scatterers of strength σ_{li} and phase $\phi_{li}(t)$. As shown in the bottom left of Fig. 2, the time domain HRRP represents the reflected signal intensity versus range along the radar LOS.

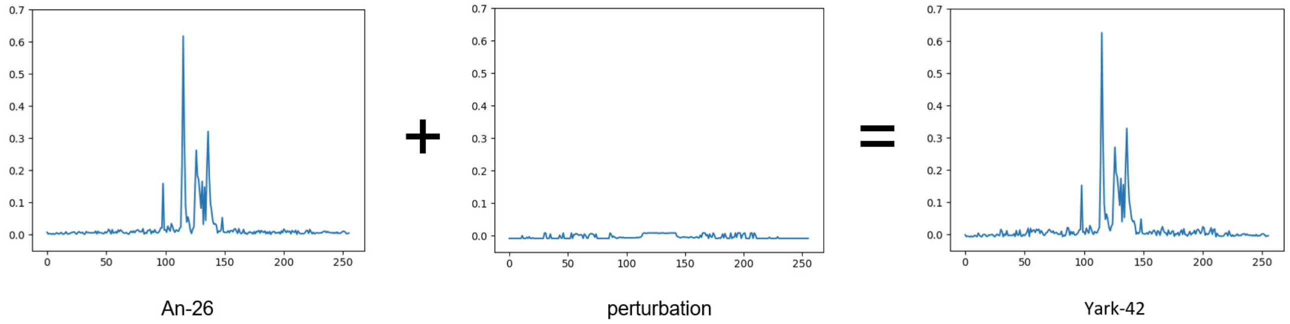

Fig. 1. This is a adversarial example in HRRP RATR. The left image shows the HRRP of an An-26 plane. The middle image is the adversarial perturbation created by FGSM. The right image shows the adversarial example has been misclassified as Yark-42 plane.

The t-th time domain real HRRP can be defined as

$$x(t) = [|\tilde{x}_1(t)|, |\tilde{x}_2(t)|, ..., |\tilde{x}_L(t)|]^\top \tag{1}$$

where $|\cdot|$ means taking absolute value.

B. Robust Perturbation for HRRP recognition

Firstly, We specify some of the notations used in this paper. Let x be the clean real HRRP, y denote the class of HRRP, y_{gt} denotes the groud true class of the HRRP. Let $y = f_\theta(x)$ be the neural network. θ is the parameters of neural network. Given a HRRP x, the probability of class y predicted by the network is $p(y|x)$. $y_{pred} = \arg\max_y p(y|x)$ is the predicted class of x. $L(x, y)$ denotes the loss function for training the network. $x^* = x + \delta$ is the adversarial examples changed from x. δ represents the perturbation we need to learn. Which should be noted is that δ is universal. Different from [14] which need to compute perturbation for each images, different input have different perturbation, our perturbation can apply to any input. If the attack succeed, $f_\theta(x^*) \neq y_{gt}$, and this can called untarget attacks. If we specified the class of the adversarial example, then this would be called target attacks.

We consider the problem generating robust perturbation as a optimization problem.

$$\min \quad D(x+\delta, x), \qquad \text{s.t.} \quad f_\theta(x+\delta) = y_t \tag{2}$$

D is a distances metric function, for example, we can use L_2 to measure the distances between adversarial example and clean example. Similar to [15], we use Lagrangian-relaxed method to reformulate optimization problem.

$$\underset{\delta}{\text{argmin}} \quad L\left(f_\theta(x+\delta), y_t\right) + \lambda\|\delta\|_p \tag{3}$$

Where L represents the most common loss function for classification, cross entropy loss, y_t is the target class(the class we expect the adversarial example to be), λ is a hyper-parameter that controls the regularization of the perturbation. Follow the method above, we can only make digital adversarial attacks. Next, we should consider some environmental constraints. By default, δ will cover the whole range of the HRRP. So the perturbation added to the target(plane) would be out of the target which is not feasible. The perturbation should be added only on target area for the actual situation. So, we adopt Constant False Alarm Rate(CFAR) [17] to find the target region which the perturbation should be put on. We feed time domain HRRP to the algorithm and then get the range of target region. Another physical condition we should consider is perturbation length. It's easy to know small range perturbation is easier to implement than wide range perturbation in real world. So we set a hyper-parameter l as the length of the perturbation. In our experiments, $l \in [8, 12, 16, 32]$. Then the perturbation should be a small part of target region. Because our goal is to generate robust perturbation, we expect the perturbation can be placed in any position in target region. In training and test time, we randomly choose a position in target region to place perturbation with specified length. Specifically, we employ a mask to project the perturbation on target with certain length. Mask is a vector whose dimentions are same as x(clean HRRP). We fill ones in the range where perturbation is added while fills zeros in other place. Let M_x be the mask of x, $\delta_0 = M_x \cdot \delta$ be the perturbation added mask. Then our optimization problem become

$$\underset{\delta}{\text{argmin}} \quad L\left(f_\theta(x+\delta_0), y_t\right) + \lambda\|\delta\|_p \tag{4}$$

We called our attack method robust perturbation HRRP(RHA). The whole process has been showed in Fig. 2.

III. Experiments

A. Measured data

We examine the recognition performance of our method on the 3-class measured data, Yark-42, Cessna Citation S/II and An-26, among which Yark-42 is a large and medium-sized jet aircraft, Cessna Citation S/II a small-sized jet aircraft and An-26 a medium-sized propeller aircraft. The radar works on C-band with a bandwidth of 400 MHz and the range resolution is about 0.375 m. The parameters of the radar and airplane targets are shown in Table I and the projections of target trajectories onto the ground plane are shown in Fig. 3.

The training and test datasets are selected following two principles. Firstly, the training dataset should cover almost all of the target-aspect angles. Secondly, the elevation angles of the training and the test dataset are different. Therefore, we select the 5-th and the 6-th segments of An-26, the 6-th and the 7-th segments of Cessna Citation S/II and the 2-rd and

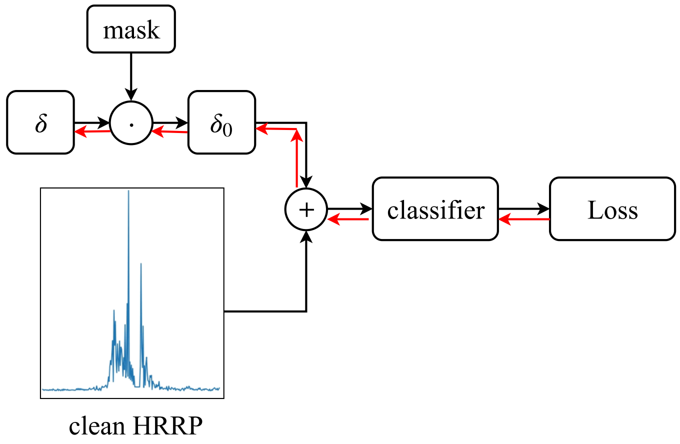

Fig. 2. The main idea of RHA. Black arrows represent the path of forward propagation, red arrows represent the path of backward propagation.

TABLE I
PARAMETERS OF RADAR AND PLANES

Radar parameters	**Center freq.**	**5520 MHz**	
	Bandwidth	**400 MHz**	
Planes	**Length(m)**	**Width(m)**	**Height(m)**
Yark-42	36.38	34.88	9.83
An-26	23.80	26.20	9.83
Cessna Citation S/II	14.40	15.90	4.57

the 5-th segments of Yak-42 as training samples, and the rest segments are taken as test samples. More concretely, there are totally 140,000 training samples and 5,200 test samples involved in our experiments.

B. Classifier

To prove our method can be applied to different classifiers, we choose two basic deep learning models, MLP and CNN. The MLP we choose has only one hidden layer for the base classifier. Similar to [9], our CNN is composed of two convolution layers and one fully connected layer. Table II show the detailed architecture of MLP and CNN we used. Let f_m be the feature map got from fully connected layer. Then f_m goes into a softmax layer to classify HRRP. The output of softmax function

$$p(y|x) = \frac{exp(\theta^{c^T} f_m)}{\sum_{j=1}^{C} exp(\theta^{j^T} f_m)} \tag{5}$$

represent the probability for each class. We choose the class with the highest probability as our classification result.

C. Attack effects

In this section, we will evaluate our method proposed above with different classifier. We report our RHA attack effects with different perturbation length and different classifier. Then, we plot some examples which attack successfully.

The test recognition rates for clean HRRP is 91%, 90% for MLP and CNN. We use Adam optimizer to optimize our perturbation. In test time, we randomly put the perturbation in the target region with specific length. Then we recalculate the recognition rates for these adversarial examples.

TABLE II
ARCHITECTURE OF THE CLASSIFIERS

MLP		**CNN**	
layer	**parameters**	**layer**	**parameters**
Linear	**units = 256**	**Conv2d**	**kernel=1,9 stride=2**
Relu	**-**	**BatchNorm**	**-**
Linear	**units = 3**	**Relu**	**-**
Softmax	**-**	**Conv2d**	**kernel=1,9 stride=2**
		BatchNorm	**-**
		Relu	**-**
		Linear	**units = 3**
		Softmax	**-**

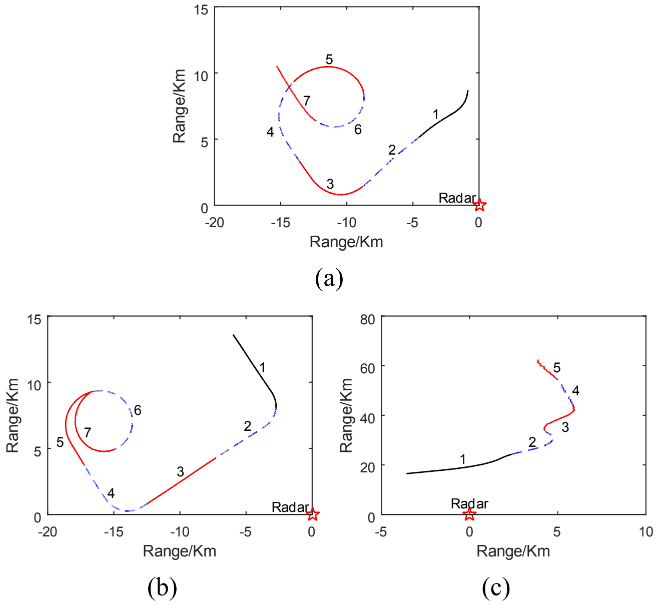

Fig. 3. Projections of target trajectories onto the ground plane. (a) An-26; (b) Cessna Citation S/II; and (c) Yark-42.

Table III show the average recognition rate for adversarial example generate by RHA. From which we can find the average recognition rate decrease rapidly as the perturbation length increase. For MLP, the recognition rate drop from 91% to 51.1% when the perturbation length be 32. The recognition rate drop from 90% to 45.7% when the perturbation length be 32 for CNN. We can see RHA is efficient and robust.

Fig. 4 and Fig. 5 shows some adversarial examples which attack successfully for MLP and CNN, in which "GT" represents groud truth for clean HRRP, "Preds" represents the classifier's prediction for the adversarial example, "probs" represents classifier's confidence. The right most image in Fig. 4 and Fig. 5 shows the perturbation for each classifier which was learned by RHA. From which we can see the magnitude of perturbation is so small that when it be added to clean HRRP, we can't find it out. But it could make the recognition accuracy drop sharply.

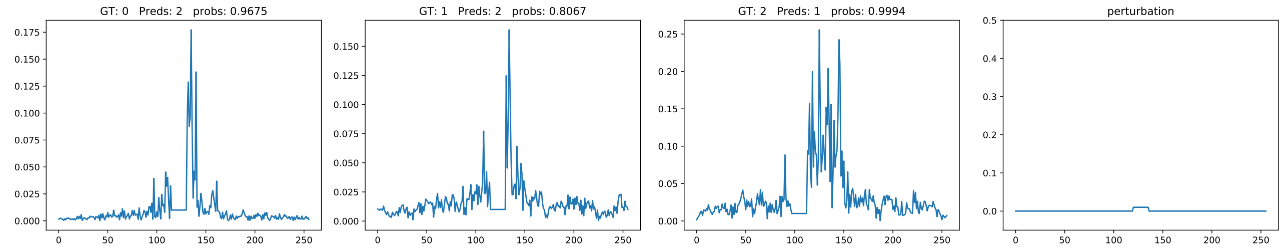

Fig. 4. Some adversarial examples which attack MLP successfully, the right most image shows the perturbation.

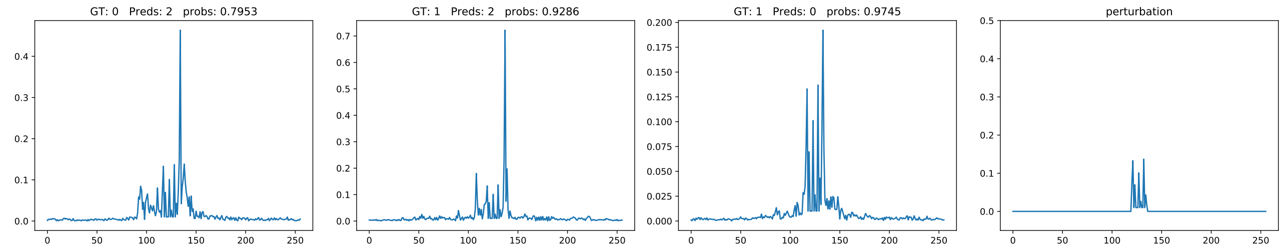

Fig. 5. Some adversarial examples which attack CNN successfully, the right most image shows the perturbation.

TABLE III
AVERAGE RECOGNITION RATES OF THE RHA AT DIFFERENT PERTURBATION LENGTH AND DIFFERENT CLASSIFIER

Method	Avg. Recognition rate(%)
MLP (perturbation length = 8)	80.9%
MLP (perturbation length = 12)	73.1%
MLP (perturbation length = 16)	66.7%
MLP (perturbation length = 32)	51.1%
CNN (perturbation length = 8)	74.1%
CNN (perturbation length = 12)	68.5%
CNN (perturbation length = 16)	64.4%
CNN (perturbation length = 32)	45.7%

IV. CONCLUSION

Firstly, we verified adversarial examples also exist in the field of HRRP-based RATR. Then we propose RHA, a robust universal adversarial perturbation which could be placed in any position in target region. Our experiments on HRRP recognition show that RHA decrease recognition rate significantly which indicate RHA is an effective adversarial attack method.

REFERENCES

[1] X. Liao, P. Runkle, and L. Carin, “Identification of ground targets from sequential high-range-resolution radar signatures,” *IEEE Transactions on Aerospace and Electronic systems*, vol. 38, no. 4, pp. 1230–1242, 2002.

[2] L. Du, H. Liu, Z. Bao, and M. Xing, “Radar hrrp target recognition based on higher order spectra,” *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2359–2368, 2005.

[3] L. Du, H. Liu, Z. Bao, and J. Zhang, “Radar automatic target recognition using complex high-resolution range profiles,” *IET Radar, Sonar & Navigation*, vol. 1, no. 1, pp. 18–26, 2007.

[4] D. Lan, H. Liu, and B. Zheng, “Radar hrrp statistical recognition: Parametric model and model selection,” *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1931–1944, 2008.

[5] L. Li and Z. Liu, “Noise-robust hrrp target recognition method via sparse-low-rank representation,” *Electronics Letters*, vol. 53, no. 24, pp. 1602–1604, 2017.

[6] X. Zhang, B. Chen, H. Liu, L. Zuo, and B. Feng, “Infinite max-margin factor analysis via data augmentation,” *Pattern Recognition*, vol. 52, pp. 17–32, 2016.

[7] B. Feng, B. Chen, and H. Liu, “Radar hrrp target recognition with deep networks,” *Pattern Recognition*, vol. 61, pp. 379–393, 2017.

[8] K. Liao, J. Si, F. Zhu, and X. He, “Radar hrrp target recognition based on concatenated deep neural networks,” *IEEE Access*, vol. 6, pp. 29211–29218, 2018.

[9] J. Wan, B. Chen, B. Xu, H. Liu, and L. Jin, “Convolutional neural networks for radar hrrp target recognition and rejection,” *EURASIP Journal on Advances in Signal Processing*, vol. 2019, p. 5, Jan 2019.

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.

[11] A. Arnab, O. Miksik, and P. H. Torr, “On the robustness of semantic segmentation models to adversarial attacks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 888–897, 2018.

[12] 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*, pp. 1528–1540, ACM, 2016.

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, 2018.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[15] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, IEEE, 2017.

[16] J. Lu, H. Sibai, E. Fabry, and D. Forsyth, “No need to worry about adversarial examples in object detection in autonomous vehicles,” *arXiv preprint arXiv:1707.03501*, 2017.

[17] R. Nitzberg, “Constant-false-alarm-rate signal processors for several types of interference,” *IEEE Transactions on Aerospace and Electronic Systems*, no. 1, pp. 27–34, 1972.