ADVERSARIAL DEFENSE VIA PERTURBATION-DISENTANGLEMENT IN HYPERSPECTRAL IMAGE CLASSIFICATION

*Cheng Shi, Ying Liu, Minghua Zhao *, Zhenzhen You*

School of Computer Science and Engineering
Xi'an University of Technology
Xi'an, Shannxi, China

Ziyuan Zhao

Institute of Infocomm Research (I2R)
A*STAR
Singapore

ABSTRACT

In recent years, deep neural networks (DNNs) have been widely used in hyperspectral image (HSI) classification. However, it has a strong vulnerability to crafted adversarial examples. Therefore, defense against adversarial examples is an urgent problem to be solved. To date, most defense methods are difficult to defend against unknown attacks. In this paper, we propose a perturbation-disentanglement-based adversarial defense method (PD-Defense) to protect HSI classification networks from unknown attacks. In the proposed method, the adversarial examples are decoupled into attack-invariant features and perturbation features, and the defense is conducted on the attack-invariant feature to defend against unknown attacks. Extensive experiments are performed on two benchmark HSI datasets, including PaviaU and HoustonU 2018. The results indicate that the proposed PD-Defense method achieves an excellent defense performance compared to four state-of-the-art defense methods.

***Index Terms*—** adversarial attack, adversarial defense, hyperspectral image.

1. INTRODUCTION

Deep neural networks (DNNs) have achieved breakthroughs on hyperspectral image (HSI) classification tasks [1–3]. However, recent studies have found that DNNs are sensitive to perturbations, and imperceptible perturbations added to the input can misclassify DNNs [4–6]. Vulnerability has also been reported by researchers in HSI classification networks [7, 8]. An HSI has higher spectral dimensions and more complex scene information than natural images; therefore, the defense of HSI classification networks is more challenging.

To defend against adversarial attacks, researchers have tried to improve the robustness of DNNs by utilizing adversarial examples [9–11]. Goodfellow et al. [12] proposed an adversarial training (AT) strategy, and the adversarial examples crafted by the fast gradient sign method (FGSM) attack method are combined with the clean examples to retrain the target model to improve its robustness. Jin et al. [13] proposed an adversarial perturbation erasing defense model (APE-GAN), which generates clean examples based on generative adversarial networks. However, these methods are highly dependent on the given types of adversarial examples and lack generalization for unknown attacks. Other researchers have tried to reduce the dependence of the defense model on adversarial examples. Cheng et al. [14] tried to generate adaptive adversarial perturbations and defended against unknown attacks by an AT strategy. Chen et al. [15] proposed a salient feature extractor (SFE) and performed adversarial defense on the extracted salient features. From a different perspective, we tried to decouple the adversarial examples into attack-invariant features and perturbation features and use the attack-invariant features to defend against unknown features.

To ensure the invisibility of the perturbations, there are still many features that are not attacked in adversarial examples, which are named attack-invariant features and can be used for defense. In this paper, we propose a perturbation-disentanglement-based defense model (PD-Defense) to decouple the attack-invariant features from the adversarial examples. A similar work was reported by Zhou et al. [16], who proposed an adversarial noise removing network (ARN) to defend against unknown attacks. The ARN extracts the attack-invariant features via an attack discriminator. In contrast to this work, we propose a perturbation-disentanglement learning mechanism for extracting the attack-invariant features. The reason is that the training of the attack discriminator is unstable with the limited training examples in HSI; therefore, it is difficult to effectively decouple the attack-invariant features and perturbations. The proposed PD-Defense model can obtain strong generalization ability to defend against unknown attacks with limited labelled examples. We implement tested experiments on two benchmark HSI datasets, and the results show that the proposed PD-Defense has obvious advantages in defending against un-

*Corresponding author (E-mail: mh_zhao@126.com)

This work was supported by the National Natural Science Foundation of China (61902313, 62002272) and Young Talent Fund of Association for Science and Technology in Shaanxi, China.

known attacks compared with other state-of-the-art defense methods.

The rest of the paper is organized as follows: Section 2 describes the details of the proposed PD-Defense model. Section 3 introduces the experimental results and analysis. Section 4 summarizes the work of this paper.

2. METHOD

The framework of the proposed PD-Defense model is shown in Fig. 1. The PD-Defense model has an encoder-decoder structure. The encoder aims to decouple the attack-invariant features and perturbation features, and the decoder reconstructs the extracted attack-invariant features into clean examples. In detail, the encoder consists of three subnetworks: one is an attack-invariant feature extraction (AIF) subnetwork, and two are specific-perturbation (SP) subnetworks. The SP subnetwork learns the attack-invariant features from the adversarial input space to achieve perturbation decoupling, and the AIF subnetwork has a parameter-sharing structure, aiming at extracting more generalized attack-invariant features from different types of adversarial examples. These three branches have the same CNN structure. The decoder has a deconvolution structure and is combined with a pixel discriminator to achieve effective reconstruction.

Assume x is the clean HSI example, and the corresponding adversarial examples are crafted by two types of adversarial attack methods, which are denoted as $\tilde{x}_1$ and $\tilde{x}_2$. The AIF subnetworks take the adversarial examples $\tilde{x}_1$ and $\tilde{x}_2$ as inputs and predict the attack-invariant feature f_c, i.e.,$f_c = (f_{c1} + f_{c2})/2 = (N_c(\tilde{x}_1) + N_c(\tilde{x}_2))/2$, where $N_c(\cdot)$ denotes the AIF subnetworks. Inputs of the SP subnetworks and the AIF subnetworks are the same, and an independence criterion is used to enable two specific branches to learn different feature informations. The extracted attack-invariant feature f_c is sent to a reconstruction subnetwork and a discriminator for example reconstruction, and $\tilde{x}$ is the reconstructed examples.

2.1. Model training

We jointly train the AIF subnetwork N_{AIF}, SP subnetwork N_{PS1} and N_{PS2}, reconstruction subnetwork N_C, and discriminator N_D by optimizing the following objective function.

$$\underset{N_{AIF},N_{PS1},N_{PS2},N_C,N_D}{\arg\min} \{L_{MSE} + L_D + \theta L_c + \beta L_d\}, \quad (1)$$

where L_{MSE} and L_D are pixel-level reconstruction loss and discriminator loss, respectively; L_c and L_d are used to constrain the learning of attack-invariant features and specific perturbation features, respectively; and θ and β are hyperparameters, which are discussed in section 3. In the following, we will describe each term in detail.

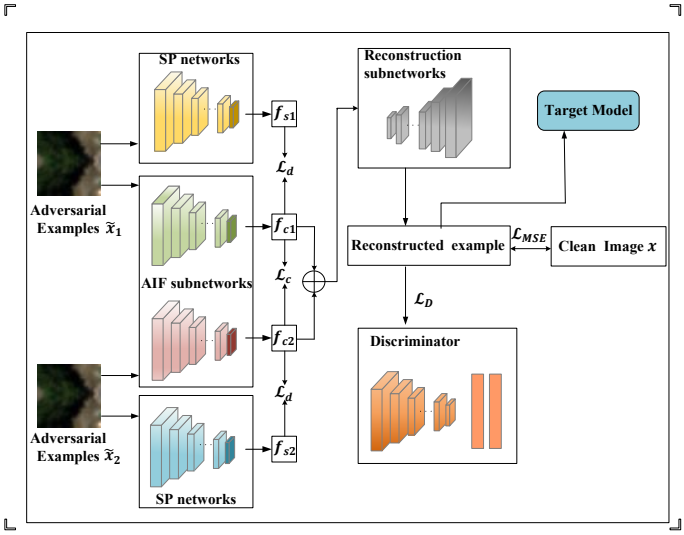

Fig. 1. The overall architecture of our proposed PD-Defense model.

2.2. Model structure

2.2.1. *Optimization of SP subnetwork*

To remove the perturbations information in the input adversarial examples, two SP subnetworks are applied for perturbation disentangling. The specific perturbation features should be independent of the attack-invariant features; therefore, a HilbertSchmidt independence criterion (HSIC) [17] is applied to constrain the learning of SP subnetworks, as shown Eq. (2).

$$L_d = (HSIC(f_{c1}, f_{s1}) + HSIC(f_{c2}, f_{s2}))/2, \quad (2)$$

where

$$HSIC(f_{c1}, f_{s1}) = (n-1)^{-2} tr(RK_{c1}, RK_{s1}), \quad (3)$$

and K_{c1} and K_{s1} are Gram matrix, $R = I - (1/n)ee^T$, I represents an identity matrix, and e is an all-one column matrix. Similarly, another HSIC constraint of f_{c2} and f_{s2} is defined in Eq. (4).

$$HSIC(f_{c2}, f_{s2}) = (n-1)^{-2} tr(RK_{c2}, RK_{s2}). \quad (4)$$

2.2.2. *Optimization of AIF subnetworks*

The perturbation distribution in different types of adversarial examples is different. To extract more generalized attack-invariant feature, the attack-invariant features are extracted by capturing the common features of different types of adversarial examples. Based on this motivation, a consistency constraint is used to capture the common features of adversarial examples $\tilde{x}_1$ and $\tilde{x}_2$, as shown in Eq. (5).

$$L_c = \|f_{c1} \cdot f_{c1}^T - f_{c2} \cdot f_{c2}^T\|_2^2, \quad (5)$$

where $f_{c1} \cdot f_{c1}^T$ is to normalize the feature f_{c1}.

Table 1. Target model in the PD-Defense method

Perturbation-disentanglement network (encoder)	(128,3,1,1) *	(256,3,1,1)	(512,3,1,1)	(1024,3,1,1)	(2048,3,1,1)
Reconstruction network (decoder)	(1024,3,1,0)	(512,3,1,1)	(256,4,2,1)	(128,3,2,0)	(3,4,2,1)
Discriminator	(128,4,2,2)	(256,2,2,1)	(512,4,2,1)	(1024,4,2,1)	FC(1024, 1)
Target model	(32,3,1,1)	(64,3,1,1)	(128,3,1,1)	FC(128,class)	

*(a,b,c,d) = (output feature channel, filter kernel size, stride, padding)

2.2.3. Optimization of reconstruction subnetworks and discriminator

The reconstruction object consists of two parts: pixel-level reconstruction and adversarial reconstruction. The mean square error (MSE) metric is used for the pixel-level reconstruction loss, as shown in Eq. (6):

$$L_{MSE} = \|\tilde{x} - x\|_2^2. \tag{6}$$

Adversarial reconstruction is implemented by an adversarial learning mechanism, and the adversarial reconstruction constraint is defined in Eq. (7).

$$L_D = E_{\hat{x}} \log D(\tilde{x}). \tag{7}$$

For training the discriminator, the adversarial loss is defined in Eq. (8).

$$L_{adv} = E_x \log D(x) + E_{\hat{x}} \log(1 - D(\tilde{x})). \tag{8}$$

3. EXPERIMENTS

3.1. Experimental setting

Datasets and target model: We evaluate the proposed defense model on two HSI datasets, PaviaU and HoustonU 2018, which are shown in Fig.2. For each dataset, a principal component analysis (PCA) transform is applied to the original HSI, and the first three principal components (PCs) are used for experiments. The target model used in the experiments is a basic CNN model, which is shown in Table 1.

Training example: For the PaviaU dataset, we randomly selected 900 pixels per class as training pixels, and the rest of the pixels are used for testing. For the HoustonU 2018 dataset, 400 pixels per class are selected as training pixels, except for class 7 and class 14. Due to the limited labelled examples for class 7 and class 14, 200 pixels and 100 pixels are randomly selected from class 7 and class 14 as training pixels, respectively. For each pixel, a neighborhood window with size 26×26 is applied to construct the HSI examples.

Hyperparameter: For both datasets, the learning rate is 0.01, the iterative number is 300 epochs, and the dropout learning is 0.5. The consistency constraint weight θ of the PaviaU dataset is 0, and the disentanglement constraint weight β is 1e-4; the consistency constraint θ of the HoustonU 2018 dataset is 1e-4, and the disentanglement constraint β is 1e-6.

Attack methods: Three well-known attack methods are used in the experiments, including FGSM (Fast Gradient Sign Method) [12], PGD (Project Gradient Descent) [18] and C&W (Carlini and Wagner) [19, 20] attacks. The default perturbation budget is set as 0.3 for all the attack methods.

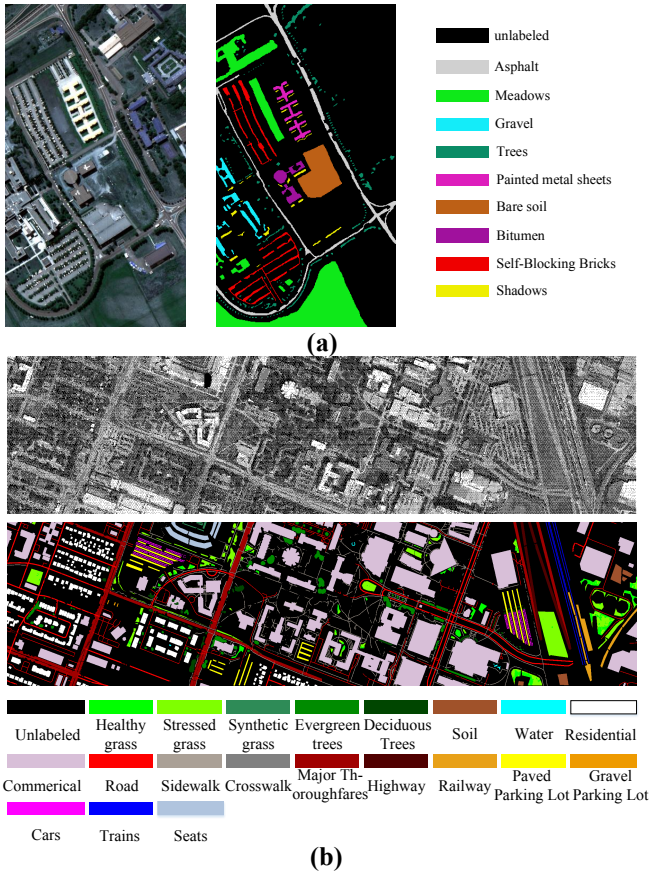

Fig. 2. HSI datasets. (a) PaviaU dataset and ground-truth map. (b) HoustonU 2018 dataset and ground-truth map.

3.2. Comparison results

In this section, four state-of-the-art defense methods are compared with the proposed DP-Defense method, including AT [12], APE-GAN [13], SFE [15] and ARN [16]. Among all the compared methods, the adversarial examples crafted by FGSM and PGD are used to train the defense model; there-

Table 2. Defense accuracy on the PaviaU dataset and HoustonU 2018 dataset.

Dataset	Defense	Attack Methods			
		None	FGSM	PGD	C&W
PaviaU	None	0.9996	0.0691	0.0000	0.0000
	AT	0.8722	0.8490	0.8346	0.6849
	APE-GAN	0.3915	0.9135	0.9444	0.3958
	SFE	0.9262	0.6420	0.4126	0.1461
	ARN	0.9606	0.8217	0.8687	0.5169
	Ours	0.9623	0.9368	0.9018	0.6664
HoustonU 2018	None	0.9365	0.0221	0.0000	0.0000
	AT	0.8479	0.6560	0.7469	0.5247
	APE-GAN	0.6327	0.5772	0.7884	0.2575
	SFE	0.8675	0.1993	0.1735	0.1262
	ARN	0.7176	0.4864	0.6243	0.2201
	Ours	0.8290	0.7157	0.7376	0.5221

Table 3. Defense accuracy with a combination of different types of adversarial examples.

Dataset	Inputs	Attack Methods			
		None	FGSM	PGD	C&W
PaviaU	None	0.9996	0.0961	0.0000	0.0000
	FGSM+PGD	0.9623	0.9368	0.9018	0.6664
	PGD+C&W	0.9667	0.5595	0.9042	0.9047
	FGSM+C&W	0.9588	0.8328	0.8363	0.8956
HoustonU 2018	None	0.9365	0.0221	0.0000	0.0000
	FGSM+PGD	0.8290	0.7157	0.7376	0.5221
	PGD+C&W	0.8589	0.4327	0.6823	0.6092
	FGSM+C&W	0.8060	0.7263	0.6573	0.6218

fore, the FGSM and PGD are considered known attacks, and C&W is the unknown attack. For a fair comparison, the target model has the same CNN structure for all the compared methods. Table 2 shows the defense accuracy on the PaviaU dataset and HoustonU 2018 dataset. It can be observed that the classification accuracies are sharply reduced in the face of adversarial attacks. Most defense methods can achieve better performance in defending against known attacks but fail to defend against unknown attacks, especially on the HoustonU 2018 dataset. However, the proposed DP-Defense method still achieves a better defense effect on defending against unknown attacks.

3.3. Generalization experiment

We train the defense model by using different types of adversarial examples to verify the generalization ability of the proposed PD-Defense method. The results are shown in Table 3. In Table 3, we construct three different combinations of adversarial examples, including a combination of FGSM and PGD, a combination of PGD and C&W, and a combination of FGSM and C&W. For instance, in the combination of PGD and C&W, the adversarial examples crafted by PGD and C&W are the known attacks, and FGSM is the unknown attacks. We can observe that the proposed PD-Defense method can effectively defend against adversarial examples in both known and unknown attacks.

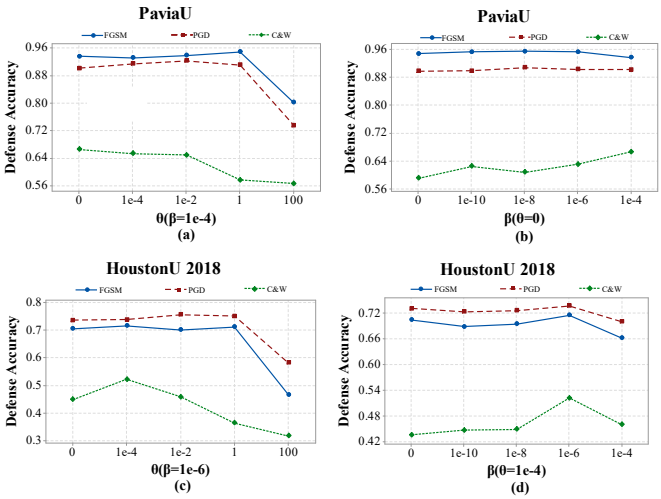

Fig. 3. Hyperparameter study. (a) θ (PaviaU). (b) β (PaviaU). (c) θ (HoustonU 2018). (d) β (HoustonU 2018).

3.4. Ablation study

The two hyperparameters of consistency constraint weight θ and disentanglement constraint weight β are set to balance the extraction of attack-invariant feature and perturbation features, which are analyzed in this section. The hyperparameters θ and β range from [0,100] and [0,1e-4]. Fig. 3 shows the results of the PaviaU dataset and HoustonU 2018 dataset. In Figs. 3(a) and (c), better defense performance is obtained when θ is less than 1e-2, and a larger θ will sharply reduce the defense performance. In contrast, a larger β can effectively improve the defense ability of the proposed model.

4. CONCLUSION

In this paper, we propose a per-processing adversarial defense model to enhance the robustness of the classification model. By decoupling the adversarial examples into attack-invariant features and perturbation features, the proposed defense model can effectively defend against known and unknown attacks. Experimental results show that the proposed model achieves superior performance compared to several state-of-the-art defense methods.

5. REFERENCES

[1] J. Feng, X. Wu, R. Shang, and et al., "Attention multi-branch convolutional neural network for hyperspectral image classification based on adaptive region search," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, pp. 5054–5070, 2020.

[2] X. Zhang, S. Chen, P. Zhu, X. Tang, J. Feng, and L. Jiao, "Spatial pooling graph convolutional network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[3] C. Xing, Y. Cong, C. Duan, Z. Wang, and M. Wang, "Deep network with irregular convolutional kernels and self-expressive property for classification of hyperspectral images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 1, pp. 1–15, 2022.

[4] W.-B. Qin and F. Wang, "A universal adversarial attack on cnn-sar image classification by feature dictionary modeling," *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1027–1030, 2022.

[5] G. R. Machado, E. Silva, and R. R. Goldschmidt, "Adversarial machine learning in image classification: A survey towards the defender's perspective," *ACM Computing Surveys*, vol. 55, pp. 1–38, 2021.

[6] Y. Duan, X. Zhou, J. Zou, J. Qiu, J. Zhang, and Z. Pan, "Mask-guided noise restriction adversarial attacks for image classification," *Computers and Security*, vol. 100, 2021.

[7] Y. Xu, B. Du, and L. Zhang, "Self-attention context network: addressing the threat of adversarial attacks for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 8671–8685, 2021.

[8] C. Shi, Y. Dang, L. Fang, Z. Lv, and M. Zhao, "Hyperspectral image classification with adversarial attack," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[9] J. Wang, C. Wang, Q. Lin, C. Luo, C. Wu, and J. Li, "Adversarial attacks and defenses in deep learning for image recognition: A survey," *Neurocomputing*, vol. 514, pp. 162–181, 2022.

[10] Q. Liang, Q. Li, and W. Nie, "Ld-gan: Learning perturbations for adversarial defense based on gan structure," *Signal Processing: Image Communication*, vol. 103, p. 116659, 2022.

[11] Y. Xiao and C.-M. Pun, "Improving adversarial attacks on deep neural networks via constricted gradient-based perturbations," *Information Sciences*, vol. 571, pp. 104–132, 2021.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Proceedings - 3th International Conference on Learning Representations*, pp. 1–11, 2015.

[13] G. Jin, S. Shen, D. Zhang, F. Dai, and Y. Zhang, "Ape-gan: adversarial perturbation elimination with gan," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2019-May, pp. 3842–3846, 2019.

[14] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, "Perturbation-seeking generative adversarial networks: a defense framework forrremote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[15] R. Chen, J. Chen, H. Zheng, Q. Xuan, Z. Ming, W. Jiang, and C. Cui, "Salient feature extractor for adversarial defense on deep neural networks," *Information Sciences*, vol. 600, pp. 118 – 143, 2022.

[16] D. Zhou, T. Liu, B. Han, N. Wang, C. Peng, and X. Gao, "Towards defending against adversarial examples via attack-invariant features," *ArXiv*, vol. abs/2106.05036, 2021.

[17] X. Wang, M. Zhu, D. Bo, P. Cui, C. Shi, and J. Pei, "Am-gcn: adaptive multi-channel graph convolutional networks," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1243–1253, 2020.

[18] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, vol. abs/1607.02533, 2017.

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

[20] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, pp. 2805–2824, 2019.