2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428437

BIAS FIELD POSES A THREAT TO DNN-BASED X-RAY RECOGNITION

Binyu Tian[1] *Qing Guo*[2*] *Felix Juefei-Xu*[3] *Wen Le Chan*[2] *Yupeng Cheng*[2]
Xiaohong Li[1*] *Xiaofei Xie*[2] *Shengchao Qin*[4]

[1]College of Intelligence and Computing, Tianjin University, China
[2]Nanyang Technological University, Singapore
[3]Alibaba Group, USA [4]Teesside University, UK

ABSTRACT

Chest X-ray plays a key role in screening and diagnosis of many lung diseases including the COVID-19. Many works construct deep neural networks (DNNs) for chest X-ray images to realize automated and efficient diagnosis of lung diseases. However, *bias field* caused by the improper medical image acquisition process widely exists in the chest X-ray images while the robustness of DNNs to the bias field is rarely explored, posing a threat to the X-ray-based automated diagnosis system. In this paper, we study this problem based on the adversarial attack and propose a brand new attack, *i.e.*, *adversarial bias field attack* where the bias field instead of the additive noise works as the adversarial perturbations for fooling DNNs. This novel attack poses a key problem: how to locally tune the *bias field* to realize high attack success rate while maintaining its spatial smoothness to guarantee high realisticity. These two goals contradict each other and thus has made the attack significantly challenging. To overcome this challenge, we propose the *adversarial-smooth bias field attack* that can locally tune the bias field with joint smooth & adversarial constraints. As a result, the adversarial X-ray images can not only fool the DNNs effectively but also retain very high level of realisticity. We validate our method on real chest X-ray datasets with powerful DNNs, *e.g.*, ResNet50, DenseNet121, and MobileNet, and show different properties to the state-of-the-art attacks in both image realisticity and attack transferability. Our method reveals the potential threat to the DNN-based X-ray automated diagnosis and can definitely benefit the development of bias-field-robust automated diagnosis system.

Index Terms— Medical image analysis, bias field, X-ray recognition, adversarial attack

1. INTRODUCTION

Medical image diagnosis and recognition is starting to be automated by DNNs with a clear advantage of being very efficient in diagnosing the disease outcomes. However, unlike human experts, such automated methods based on DNNs still have some caveats. For example, with the presence of image-level degradations during the image acquisition process, the recognition accuracy can be dramatically suppressed. Sometimes, such DNN-based medical image recognition system can even become entirely vulnerable when maliciously attacked by an adversary or an abuser that is financially incentivized.

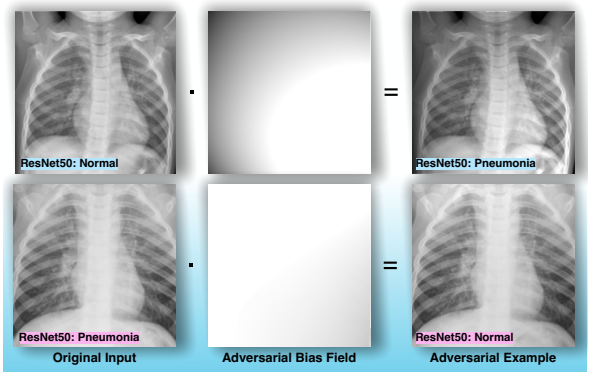

Fig. 1: Two cases of our adversarial bias field examples. Our proposed adversarial-smooth bias field attack can adversarially but imperceptibly altered the bias field, misleading the advanced DNN models, *e.g.*, ResNet50, to diagnose the normal X-ray image as the pneumonia one. More troubling, the DNN could be fooled to diagnose the pneumonia X-ray image as the normal one, having higher risk of delaying patients' treatment.

There are mainly two types of image perturbations or degradations in medical imagery: (1) image noise, and (2) image bias field. The image noise is primarily caused by the image sensor noise and the image bias field is caused by the spatial variations of radiation [1], which is common among medical imaging, ranging from magnetic resonance imaging (MRI) [2], computed tomography (CT) [3], to X-ray imaging, *etc*. The bias field appears as the intensity inhomogeneity in the MRI, CT, or X-ray images. For consumer digital imaging, the bias field shows up as the illumination changes or vignetting effect.

In this work, we want to reveal this vulnerability caused by image bias field. To the best of our knowledge, this is the very first attempt to adversarially perturb the bias field, in order to attack DNN-based X-ray recognition. Contrary to the additive noise-perturbation attack on DNN-based recognition systems, the attack on the bias field is multiplicative in nature [4], which is fundamentally different from the noise attack. What is more

*Qing Guo and Xiaohong Li are the corresponding authors (tsingqguo@ieee.org and xiaohongli@tju.edu.cn).

important is that in order to make the bias field attack realistic and imperceptible, the successful attacks need to maintain the smoothness property of the bias field, which is genuinely more challenging because local smoothness usually contradicts with high attack success rates.

To overcome this challenge, we capitalize on this proprietary degradation surrounding X-ray imagery and initiate adversarial attacks based on imperceptible modification on the bias field itself. Specifically, we have proposed the adversarial-smooth bias field generator that can locally tune the bias field with joint smooth and adversarial constraints by tapping into the bias field generation process based on a multivariate polynomial model. As a result, the adversarially perturbed bias field applied to the X-ray image can not only fool the DNN-based recognition system effectively, but also retain high level of realisticity. We have validated our proposed method on several chest X-ray classification datasets with the state-of-the-art DNNs such as ResNet, DenseNet, and MobileNet, by showing superior performance in terms of both image realisticity and high attack success rates. A careful investigation into which bias field region contributes more significantly to the adversarial nature of the attack can lead to a better interpretation and understanding of the DNN-based recognition system and its vulnerability, which, we believe, is of utmost importance. The ultimate goal of this work is to reveal that the bias field does pose a potential threat to the DNN-based automated recognition system, and can definitely benefit the development of bias-field-robust automated diagnosis system in the future.

2. RELATED WORK

X-Ray imagery recognition. X-ray radiography is widely used in the medical field for diagnosis or treatment of diseases. [5] releases the ChestX-ray14 dataset and evaluates the performance of 4 classic convolutional neural network (CNN) on the multi-label image classification of diseases. [6] proposes the use of a CNN backbone with a variant of DenseNet model. [7] presents a three-branch attention guided CNN that combines local cues and global features. CNNs are also explored for COVID-19 detection [8, 9], motivated by the need of quick and convenient screening, since abnormalities can be found in some patients' chest X-Ray images.

Despite considerations made to address data irregularities like class imbalance in dataset, the effect of medical image degradation is rarely addressed. For example, bias field could adversely affect quantitative image analysis [10]. Though many inhomogeneity correction strategy are proposed [11, 12], the possible detrimental effect on disease identification, location or segmentation by bias field is rarely explored, possibly reducing DNN's robustness. To the best of authors' knowledge, this paper is very first work that looks at the effect of bias field from the view of adversarial attack.

General adversarial attack. DNNs in image, speech or natural language processing application are susceptible to adversarial attacks [13, 14, 15, 16, 17, 18, 19]. Specifically, fast gradient sign method (FGSM) proposed by [13] involves only one back propagation step when calculating the cost function's gradient, allowing fast adversarial example generation. [20] proposes basic iteration method (BIM), an iterative version of FGSM. [16] proposes to use margin loss instead of entropy loss during attacks. [15] proposes the exploitation of transferability of adversarial examples.

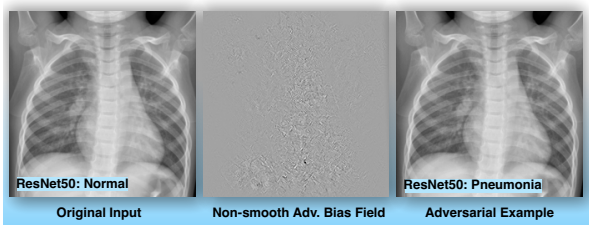

Fig. 2: An example of using Eq. (3) to generate the non-smooth adversarial bias field.

Adversarial attack on medical imagery. There are existing literature that look into adversarial attack against deep learning system for medical imagery. [21] shows that both black box and white box PGD attack and adversarial patch attack can affect the classifiers' performance on fundoscopy, chest X-ray and dermoscopy, respectively. By producing crafted mask, an adaptive segmentation mask attack (ASMA) is proposed to fool DNN model [22].

However, very few literature has leveraged on and conducted adversarial attack based on the inherent characteristics of the targeted medical imagery. For example, common noise degradation used for general adversarial attacks are rarely found in X-ray imagery. Hence in this work, we capitalize on the proprietary degradation surrounding X-ray imagery, bias field, and initiate adversarial attacks based on imperceptible modification on the bias field itself.

3. METHODOLOGY

3.1. Adversarial Bias Field Attack and Challenges

Given a X-ray image, *e.g.*, $\mathbf{X}^{\mathrm{a}}$, we can assume it is generated by adding a bias field $\mathbf{B}$ to a clean version, *i.e.*, $\mathbf{X}$, with the widely used imaging model

$$\mathbf{X}^{\mathrm{a}} = \mathbf{X}\mathbf{B}. \tag{1}$$

Under the automate diagnosis task where a DNN is used to recognize the category (*i.e.*, normal or abnormal) of $\mathbf{X}^{\mathrm{a}}$, it is necessary to explore a totally new task, *i.e.*, *adversarial bias field attack* aiming to fool the DNN by calculating an adversarial bias field $\mathbf{B}$, with which we can study the influence of the bias field as well as the potential threat of utilizing it to fool the automate diagnosis.

A simple way is to take logarithm on Eq. (1) and transform the multiplication to additive operation

$$\hat{\mathbf{X}}^{\mathrm{a}} = \hat{\mathbf{X}} + \hat{\mathbf{B}}, \tag{2}$$

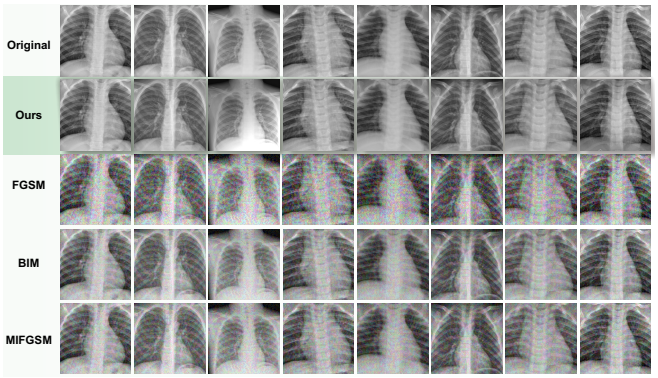

Fig. 3: Examples of adversarial examples generated with different techniques.

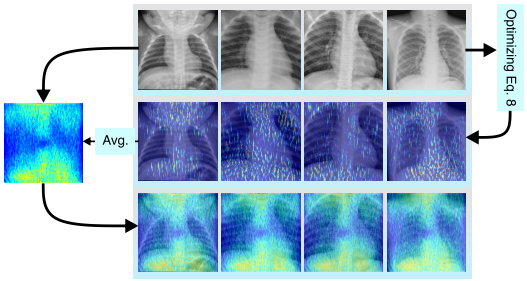

Fig. 4: Pipeline and examples of exploring bias-field-sensitive regions. A subject model, *i.e.*, ResNet50, is employed to generate adversarial bias field examples for 240 X-ray images and we then use Eq. (8) to produce the interpretable map $\mathbf{M}$ for each image (*i.e.*, the images at the second row where the maps are blended with the raw X-ray images for better understanding.). Finally, we can calculate an averaging map covering all interpretable maps and blend it with raw images (*i.e.*, the images at the third row.)

where we use the '$\hat{\cdot}$' to represent the logarithm of a variable. With Eq. (2), it seems that all existing additive-based adversarial attacks, *i.e.*, FGSM, BIM, MIFGSM, DIM, and TIM-IFGSM, could be used for the new attack. For example, we can calculate $\hat{\mathbf{B}}$ to realize attack by solving

$$\arg\max_{\hat{\mathbf{B}}} J(\hat{\mathbf{X}} + \hat{\mathbf{B}}, y), \quad \text{subject to} \quad \|\hat{\mathbf{B}}\|_{\mathrm{p}} \leq \epsilon, \tag{3}$$

where $J(\cdot)$ is the loss function for classification, *e.g.*, the cross-entropy loss, and y denotes the ground truth label of $\mathbf{X}$. Nevertheless, we argue that such solution cannot generate the real 'bias field' since the optimized $\hat{\mathbf{B}}$ violated the basic property of bias field, *i.e.*, *spatially smooth changes* resulting in intensity inhomogeneity. For example, as shown in Fig. 2, when we optimize Eq. (3) to produce a bias field, we can attack the ResNet50 successfully while the bias field is noise-like and far from the appearance in the real world.

As a result, due to requirement of spatial smoothness of bias field, the *adversarial bias field attack* posts a totally *new challenge* to the field of adversarial attack: how to generate the adversarial perturbation that can not only achieve high attack success rate but maintain its spatial smoothness for the realisticity of bias field. Actually, since the high attack success rate relies on the pixel-wise tunable perturbation and violates the smoothness requirement of bias field, the two constraints contradicts each other and make the *adversarial bias field attack* significantly challenging.

3.2. Adversarial-Smooth Bias Field Attack

In this section, we propose the *distortion-aware multivariate polynomial model* to represent the bias field whose inherit property guarantees the spatial smoothness of the bias field while the distortion helps achieve effective attack. Then, we define a new objective function for effective attack by combining the constraints of spatially smooth bias field, sparsity of the original image with the adversarial loss. Finally, we introduce the optimization method and attack algorithm.

Distortion-aware multivariate polynomial model. We model the bias filed $\hat{\mathbf{B}}$ as

$$\hat{\mathbf{B}}_i = \sum_{t=D_0}^{D} \sum_{l=D_0}^{D-t} a_{t,l} \mathrm{T}_\theta(x_i)^t \mathrm{T}_\theta(y_i)^l \tag{4}$$

where T_θ represents the distortion transformation and we use the thin plate spline (TPS) transformation with θ being the control points. We denote i as the i-th pixel with its coordinates (x_i, y_i) while $(\mathrm{T}_\theta(x_i), \mathrm{T}_\theta(y_i))$ means the pixel has been distorted by a TPS. In addition, $\{a_{t,l}\}$ and D are the parameters and degree of the multivariate polynomial model, respectively, and the number of parameters are $|\{a_{t,l}\}| = \frac{(D-D_0+1)(D-D_0+2)}{2}$. For convenient representations, we concatenate $\{a_{t,l}\}$ and obtain a vector $\mathbf{a}$.

Adversarial-smooth objective function. With Eq. (4), we can tune $\mathbf{a}$ and θ for adversarial attack and the multivariate polynomial model can help preserve the smoothness of bias field. Intuitively, on the one hand, the lower degree D leads to less model parameters $|\{a_{t,l}\}|$ and smoother bias field. On the other hand, the distortion $(\mathrm{T}_\theta(x_i), \mathrm{T}_\theta(y_i))$ can be locally tuned with different θ and can help achieve effective attack. The key problem is how to calculate $\{a_{t,l}\}$ and θ to balance the spatial smoothness and adversarial attack. To this end, we define a new objective function to realize the attack.

$$\arg\max_{\mathbf{a},\theta} J(\hat{\mathbf{X}} + \hat{\mathbf{B}}(\mathbf{a}, \theta), y) - \lambda_a \|\mathbf{a}\|_1 - \lambda_\theta \|\theta - \theta_0\|_1, \tag{5}$$

where θ_0 denotes parameters of the identify TPS transformation, *i.e.*, $x_i = \mathrm{T}_{\theta_0}(x_i)$. The first term is to tune the $\mathbf{a}$ and θ to fool a DNN. The second term encourages the sparse of $\{a_{t,l}\}$ and would let the bias field smooth. The final term is to let the TPS transformation not go far away from the identity version. Two hyper-parameters, *i.e.*, λ_a and λ_θ control the balance between the smoothness and adversarial attack.

Optimization Like the optimization methods used in general adversarial noise attack, we solve Eq. (3) and (5) via sign gradient descent where $\mathbf{a}$ and θ are updated via fixed rate

$$\mathbf{a}_t = \mathbf{a}_{t-1} + \epsilon_a \mathrm{sign}(\nabla \mathbf{a}_{t-1}), \tag{6}$$

$$\theta_t = \theta_{t-1} + \epsilon_\theta \mathrm{sign}(\nabla \theta_{t-1}), \tag{7}$$

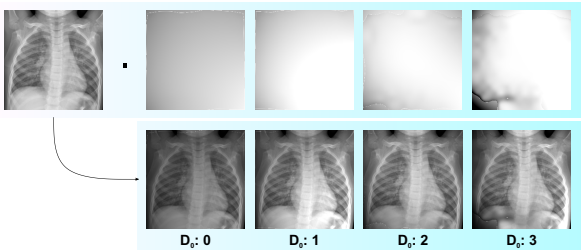

Fig. 5: Effects of the multivariate polynomial model with different number of degrees, *i.e.*, D_0 and D in Eq. (4).

where $\nabla \mathbf{a}_{t-1}$ and $\nabla \theta_{t-1}$ denote the gradient of $\mathbf{a}_{t-1}$ and θ_{t-1} with respect to the objective function in Eq. (5), respectively. For Eq. (3), we use the same to update $\hat{\mathbf{B}}$ directly. We fix $\epsilon_a = \epsilon_\theta = 0.06$ with the iteration number being 10.

4. EXPERIMENTS

4.1. Setup and Dataset

Dataset. We carry out our experiments on a chest-xray dataset about pneumonia, which contains 5863 X-ray images[1]. These images were selected from retrospective cohorts. The dataset is divided into two categories, *i.e.*, pneumonia and normal.

Models. In order to show the effect of the attack on different models, we finetune three pre-trained models on the chest-xray dataset. The three models are ResNet50, MobileNet and Densenet121 (Dense121).The accuracy of ResNet50, MobileNet and Densenet121 is 88.62%, 88.94% and 87.82%.

Metrics. We choose the attack success rate and image quality to evaluate the effectiveness of the bias field attack. The image quality measurement metric is BRISQUE [23]. BRISQUE is an unsupervised image quality assessment method. A high score for BRISQUE indicates poor image quality.

Baselines. We select five adversarial attack methods as our baselines, including basic iterative method (BIM) [20], Carlini & Wagner L2 method (C&W$_{L2}$) [16], saliency map method (SaliencyMap) [24], fast gradient sign method (FGSM) [13] and momentum iterative fast gradient sign method (MIFGSM) [25]. For the setup of hyperparameters of these baselines, we use the default setup of foolbox [26]. We set max perturbation to be $\epsilon = 0.1$ relative to [0,1] range in basic experiments. Besides, we set iterations as 10 for MIFGSM and BIM.

4.2. Comparison with Baseline Methods

For our method, we set the size of the control points, D and D_0 as (16*16), 10, and 1, respectively. Table 1 shows the quantitative results with our method and the baseline methods, which are conducted with different settings. Specifically, we conduct two different attacks, i.e., the white-box attack and the transfer attack. The white-box attack aims to attack the target DNN directly while the transfer attack attacks the target DNN with the adversarial examples generated from other models. For example, for the transfer attack in Table 1, the attack is performed on DNNs in the first row, and the generated adversarial examples are used to attack DNNs in the first two columns of the second row.

As we can see, for the white-box attack (i.e., the third column for each model), we could find that the success rate of our method is lower than the existing baselines. For example, on ResNet50, our method achieves 38.69% success rate while most of the baselines achieves 100% success rate. The main reason is that the existing attacking techniques could add arbitrary noises on the image, which is not realistic. However, our method has a strict smooth limitation such that the generated adversarial examples look more realistic. As shown in Fig. 3, we show some examples generated by different attacks. The first row shows the original images while the following rows list the corresponding adversarial examples. It is clear that our method could generate high-quality adversarial examples that are smooth and realistic. In most cases, the change between original image and the generated image is imperceptible. However, we could find obvious noises in the adversarial examples generated by the baseline methods. Such noises are difficult to appear in X-rays in the real world.

For the transfer attack (*i.e.*, the first two columns), we found that our method achieves much higher success rate than others. For example, the attack on ResNet50 achieves 7.57% and 14.05% transfer success rate on MobileNet and DenseNet121, respectively. However, the the best results of the baseline are only 1.08% and 0.18%. It is because that existing techniques calculate the ad-hoc noise, which may be only effective on the target DNN but not on other models. However, our attack considers the smoothness such that the generated adversarial examples are more realistic. Such adversarial examples are more robust and could reveal the common weakness of different DNNs (*i.e.*, higher success rate of the transfer attack). The results indicate that our method could generate high-quality adversarial examples. We also compare the image quality with the BRISQUE score (*i.e.*, the forth column). The results show that our method could achieve competitive results with the-state-of-the-arts.

In summary, our method aims to generate high-quality and realistic adversarial examples. To generate such adversarial examples, the attack success rate is naturally lower than the noise-based adversarial attack techniques.

4.3. Understanding Effects of Bias Field

In this subsection, we aim to explore how the bias field affect the DNN-based X-ray recognition. [27] proposes a method for understanding DNNs with the adversarial noise attack and generates an interpretable map indicating the classification-sensitive regions of a DNN. Inspired this idea, we can study which regions in the chest X-ray images are sensitive to the bias filed and affect the X-ray recognition. Specifically, given

[1]Please find more details about the dataset in https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

Crafted from	ResNet50				Dense121				MobileNet			
Attacked model&BRISQUE	MobileNet	Dense121	ResNet50	BRISQUE	ResNet50	MobileNet	Dense121	BRISQUE	ResNet50	Dense121	MobileNet	BRISQUE
BIM	0.36	0	100	30.0249	0.54	0.36	100	29.6599	0	0	100	29.9947
C&W$_{L2}$	0.36	0	100	30.1128	1.08	0.72	100	29.6455	0	0	100	30.051
SaliencyMap	1.08	0.18	100	28.7108	2.53	1.26	100	28.4046	0.72	0.18	100	30.8351
FGSM	0	0.18	67.8	67.0028	0.72	0.72	29.38	28.5753	0	0	30.09	28.5404
MIFGSM	0.36	0	100	30.0578	0.54	0.36	94.34	29.6094	0	0	100	30.0134
AdvSBF (Ours)	7.57	14.05	38.69	28.5703	7.78	5.95	34.49	28.9535	20.07	18.98	33.51	29.5475

Table 1: Adversarial comparison results on chest-Xray dataset with five attack baselines and our method. It contains the success rates (%) of transfer & whitebox adversarial attack on three normally trained models: ResNet50, Dense121, and MobileNet. For each four columns, whitebox attack results are shown in the third one. The first two columns display the transfer attack results. And the last column shows the BRISQUE score.

$(gridsize, gridsize)$, D_0	ResNet50				Dense121				MobileNet			
	MobileNet	Dense121	ResNet50	BRISQUE	ResNet50	MobileNet	Dense121	BRISQUE	ResNet50	Dense121	MobileNet	BRISQUE
(4,4), 0	10.84	15.33	37.97	32.4873	14.65	8.29	31.39	31.331	21.52	20.44	35.68	34.9368
(8,8), 0	9.91	14.05	37.79	32.5778	13.2	6.49	31.57	31.3609	21.7	20.26	35.68	34.0957
(12,12), 0	9.73	14.23	37.61	32.097	12.84	6.49	31.2	31.9176	21.7	20.44	35.86	34.3194
(16,16), 0	10.81	14.42	38.34	32.3661	13.56	6.85	31.02	31.4455	21.34	20.26	36.04	34.0944
(16,16), 1	11.35	13.5	36.89	31.3312	14.65	9.37	32.12	30.6853	17	19.34	32.79	31.7842
(16,16), 2	8.11	7.85	29.48	29.0977	12.84	8.83	26.09	30.0223	12.12	11.86	26.85	29.5885
(16,16), 3	4.15	2.19	18.81	28.606	4.7	4.68	16.24	29.0152	4.7	3.47	15.32	29.2909

Table 2: Adversarial comparison results on chest-Xray dataset with different setup of hyper-parameters in our method. It contains the success rates (%) of transfer& whitebox adversarial attacks. For each model, the first two columns display the blackbox attack results, the third one shows the attack results and the last column shows the BRISQUE score.

an adversarial bias field example $\mathbf{X}^{\mathrm{a}}$ generated by our method and the original image $\mathbf{X}$, we can calculate an interpretable map $\mathbf{M}$ for a DNN $\mathrm{DNN}(\cdot)$ by optimizing

$$\arg\min_{\mathbf{M}} \ \mathrm{DNN}_y(\mathbf{M}\odot\mathbf{X}^{\mathrm{a}}+(1-\mathbf{M})\odot\mathbf{X}) + \lambda_1\|\mathbf{M}\|_1+\lambda_2\mathrm{TV}(\mathbf{M}) \tag{8}$$

where $\mathrm{DNN}_y(\cdot)$ denotes the score at label y that is the ground truth label of $\mathbf{X}$ and $\mathrm{TV}(\cdot)$ is the total-variation norm. Intuitively, optimizing Eq. (8) is to find the region that causes misclassification. We optimize Eq. (8) via gradient decent in 150 iterations and fix $\lambda_1 = 0.05$ and $\lambda_2 = 0.2$.

With Eq. (8), given a pre-trained model, *i.e.*, $\mathrm{DNN}(\cdot)$, and a dataset $\mathcal{X}$ containing the successfully attacked X-ray images, we calculate a $\mathbf{M}$ for each X-ray image and then average all interpretable maps to show the statistical regions that are sensitive to the bias field. For example, we adopt ResNet50 as the subject model and construct $\mathcal{X}$ with 240 attacked X-ray images that can fool ResNet50 successfully. Then, we calculate the interpretable maps for all images in $\mathcal{X}$ (*e.g.*, the second row in Fig. 4) and average them, achieving a statistical mean map (*e.g.*, the left image shown in Fig. 4). With the visualization results, we observe that: ❶ Our method helps identify the bias-field-sensitive regions in each attacked image and we observe that these regions are related to the organ positions. *This demonstrates that the effects of the bias field to the DNN stems from intensity variation around organs.* ❷ According to the statistical mean map, we see that *the bias-field sensitive regions mainly locate at the top and bottom positions across all attacked images*, suggesting that future designed DNN should consider the spatial variations within in X-ray images. We observe similar results on other DNNs (Please find more results in the supplementary material), hinting that these are common phenomenons in the DNN-based X-ray recognition and demonstrating the potential applications of this work.

4.4. Effects of Hyper-parameters

We also evaluate the effects of hyper-parameters in our attack, *i.e.*, θ and D in Equation 4. Specifically, we change θ for TPS transformation by changing the number of control points. $(gridsize \times gridsize)$ is denoted to represent the control points in the TPS transformation. Then we select different $gridsize$ to conduct the attack. For the parameter D, we set the fixed D as 10 and change the value of D_0, *i.e.*, observe part of the sample display of the bias field by ignoring the lowest D_0 degree in the multivariate polynomial model.

Table 2 shows the results with different configurations. In the second row, we fix the D_0 as 0 and change value of $gridsize$ as 4, 8, 12 and 16, respectively. As we can see, there seems to be no clear difference in the attack success rate when the parameter $gridsize$ varies. We conjecture that the attack could easily reach the upper bound in terms of the success rate with different $gridsize$.

Then we fix the $gridsize$ as 16 and change the parameter D_0 as 0, 1, 2 and 3 (in the third row). As we can see, as D_0 increases (*i.e.*, more lower degree are ignored), the success rate of our method decreases and the BRISQUE score decreases. It is reasonable as ignoring more low degree in Equation 4 may reduce the space of the manipulation, resulting in higher image quality and lower attack success rate. The visualization results are shown in Fig. 5. When more lower degree is ignored (*i.e.*, larger D_0), the bias field samples tend to be less smooth.

5. CONCLUSION

Deep learning has been used in chest X-ray image recognition for the diagnosis of lung diseases (*e.g.*, COVID-19). It is especially important to ensure the robustness of the DNN in this scenario. To tackle this problem, this paper proposed a new adversarial bias field attack, which aims to generate more realistic adversarial examples by adding more smooth pertur-

bations instead of noises. We demonstrated the effectiveness of our attack on the widely used DNNs. The results show that our method can generate high quality adversarial examples, which achieve high success rate of the transfer attack. The generated realistic images can reveal issues of the DNN, which calls for the attention of robustness enhancement of the deep learning-based healthcare system.

In the future, we will extend the proposed attack against other tasks, *e.g.*, visual object tracking [28, 29, 30] and DeepFake evasion [31, 32], and also in tandem with other natural modalities such as [33, 34, 35]. In addition, we can regard the adversarial bias field as a new kind of mutation for DNN testing [36, 37, 38, 39].

Acknowledgement. This work has partially been sponsored by the National Science Foundation of China (No. 61872262) and the Natural Science Foundation of Tianjin (No. KJZ40420200017).

6. REFERENCES

[1] Uro Vovk, Franjo Pernus, and Botjan Likar, "A review of methods for correction of intensity inhomogeneity in mri," *IEEE transactions on medical imaging*, vol. 26, no. 3, pp. 405–421, 2007.

[2] Mohamed N Ahmed, Sameh M Yamany, Nevin Mohamed, Aly A Farag, and Thomas Moriarty, "A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data," *IEEE transactions on medical imaging*, vol. 21, no. 3, pp. 193–199, 2002.

[3] Qing Guo, Shuifa Sun, Fangmin Dong, Wei Feng, Bruce Zhi Gao, and Siyu Ma, "Frequency-tuned acm for biomedical image segmentation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 821–825.

[4] Yuanjie Zheng and James C Gee, "Estimation of image bias field with sparsity constraints," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 255–262.

[5] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *CoRR*, vol. abs/1705.02315, 2017.

[6] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *CoRR*, vol. abs/1710.10501, 2017.

[7] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *CoRR*, vol. abs/1801.09927, 2018.

[8] Linda Wang and Alexander Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *arXiv preprint arXiv:2003.09871*, 2020.

[9] Parnian Afshar, Shahin Heidarian, Farnoosh Naderkhani, Anastasia Oikonomou, Konstantinos N Plataniotis, and Arash Mohammadi, "Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images," *arXiv preprint arXiv:2004.02696*, 2020.

[10] Jaber Juntu, Jan Sijbers, Dirk Van Dyck, and Jan Gielen, "Bias field correction for mri images," in *Computer Recognition Systems*, pp. 543–551. Springer, 2005.

[11] Ayres Fan, William M Wells, John W Fisher, Müjdat Cetin, Steven Haker, Robert Mulkern, Clare Tempany, and Alan S Willsky, "A unified variational approach to denoising and bias correction in mr," in *Biennial international conference on information processing in medical imaging*. Springer, 2003, pp. 148–159.

[12] David L Thomas, Enrico De Vita, Ralf Deichmann, Robert Turner, and Roger J Ordidge, "3d mdeft imaging of the human brain at 4.7 t with reduced sensitivity to radiofrequency inhomogeneity," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 53, no. 6, pp. 1452–1458, 2005.

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[14] Matt Fredrikson, Somesh Jha, and Thomas 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.

[15] 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, pp. 506–519.

[16] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[17] Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Jian Wang, Wei Feng, and Yang Liu, "Abba: Saliency-regularized motion-based adversarial blur attack," *arXiv preprint arXiv:2002.03500*, 2020.

[18] Qing Guo, Xiaofei Xie, Felix Juefei-Xu, Lei Ma, Zhongguo Li, Wanli Xue, Wei Feng, and Yang Liu, "Spark: Spatial-aware online incremental attack against visual tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[19] Run Wang, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Yihao Huang, and Yang Liu, "Amora: Black-box adversarial morphing attack," in *ACM Multimedia Conference (ACMMM)*, 2020.

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[21] Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam, "Adversarial attacks against medical deep learning systems," *arXiv preprint arXiv:1804.05296*, 2018.

[22] Utku Ozbulak, Arnout Van Messem, and Wesley De Neve, "Impact of adversarial examples on deep learning models for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 300–308.

[23] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[24] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[25] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[26] Jonas Rauber, Wieland Brendel, and Matthias Bethge, "Foolbox: A python toolbox to benchmark the robustness of machine learning models," *arXiv preprint arXiv:1707.04131*, 2017.

[27] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *ICCV*, 2017, pp. 3449–3457.

[28] Qing Guo, Ruize Han, Wei Feng, Zhihao Chen, and Liang Wan, "Selective spatial regularization by reinforcement learned decision making for object tracking," *IEEE Transactions on Image Processing*, vol. 29, pp. 2999–3013, 2020.

[29] Qing Guo, Wei Feng, Ce Zhou, Chi-Man Pun, and Bin Wu, "Structure-regularized compressive tracking with online data-driven sampling," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5692–5705, 2017.

[30] Ce Zhou, Qing Guo, Liang Wan, and Wei Feng, "Selective object and context tracking," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 1947–1951.

[31] Yihao Huang, Felix Juefei-Xu, Run Wang, Qing Guo, Lei Ma, Xiaofei Xie, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu, "Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1217–1226.

[32] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *arXiv:2103.00218*, 2021.

[33] Ruijun Gao, , Qing Guo, Felix Juefei-Xu, Hongkai Yu, Xuhong Ren, Wei Feng, and Song Wang, "Making Images Undiscoverable from Co-Saliency Detection," *arXiv preprint arXiv:2009.09258*, 2020.

[34] Yupeng Cheng, Felix Juefei-Xu, Qing Guo, Huazhu Fu, Xiaofei Xie, Shang-Wei Lin, Weisi Lin, and Yang Liu, "Adversarial Exposure Attack on Diabetic Retinopathy Imagery," *arXiv preprint arXiv:2009.09231*, 2020.

[35] Liming Zhai, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Lei Ma, Wei Feng, Shengchao Qin, and Yang Liu, "It's Raining Cats or Dogs? Adversarial Rain Attack on DNN Perception," *arXiv preprint arXiv:2009.09205*, 2020.

[36] Xiaofei Xie, Lei Ma, Felix Juefei-Xu, Minhui Xue, Hongxu Chen, Yang Liu, Jianjun Zhao, Bo Li, Jianxiong Yin, and Simon See, "DeepHunter: A Coverage-Guided Fuzz Testing Framework for Deep Neural Networks," in *ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*, 2019.

[37] Lei Ma, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Felix Juefei-Xu, Chao Xie, Li Li, Yang Liu, Jianjun Zhao, and Yadong Wang, "DeepMutation: Mutation Testing of Deep Learning Systems," in *The 29th IEEE International Symposium on Software Reliability Engineering (ISSRE)*, 2018.

[38] Xiaoning Du, Xiaofei Xie, Yi Li, Lei Ma, Yang Liu, and Jianjun Zhao, "Deepstellar: Model-based quantitative analysis of stateful deep learning systems," in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2019, pp. 477–487.

[39] Lei Ma, Felix Juefei-Xu, Jiyuan Sun, Chunyang Chen, Ting Su, Fuyuan Zhang, Minhui Xue, Bo Li, Li Li, Yang Liu, Jianjun Zhao, and Yadong Wang, "DeepGauge: Multi-Granularity Testing Criteria for Deep Learning Systems," in *The 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2018.