Transferability Analysis of Adversarial Examples in CNN-based SAR Image Classification

Minjae Kim, Haksu Han, Gyeongsup Lim, Yehyeong Lee, Sanghun Sim, Junbeom Hur
Department of Computer Science and Engineering,
Korea University,
Seoul, South Korea
{mjkim, hshan, gslim, yhlee, shsim, jbhur}@isslab.korea.ac.kr

***Abstract*—Recently, CNN-based Synthetic Aperture Radar (SAR) Automatic Target Recognition (ATR) systems have received increasing attention for adversarial examples as a cybersecurity threat. SAR images consist of target, shadow, and speckle regions that are different from optical images due to their unique imaging mechanism. Recent studies on adversarial examples for SAR-ATR have developed black-box attacks suitable for real-world environment by focusing perturbation on target-region using surrogate model. However, if an attack is focused only on the target area, the attack may be overfitted to surrogate model, and may become unsuitable for the target model. In this paper, we derived three research questions. RQ1: "Does an attack focused only on a target show superior performance in transferability?", RQ2: "Is the transferability also affected by the other areas in SAR images?", and RQ3: "If physical attacks are feasible not only on the target but also on the shadows, how does transferability change?" Seeking the answers to these research questions, we conducted comparative experiments of attacks focused on both each individual region and their combinatorial regions diversely. For the analysis, we used 8 models (including 1 surrogate model and 7 target models) trained under the MSTAR dataset. In addition, we used four algorithms (FGSM, CW, DeepFool, and PGD) to create adversarial example and SAR-Bake dataset to divide SAR regions. Specifically, we measured the transfer attack success rate when perturbations were applied to each specific region (all pixels, target, shadow, speckle and target+shadow). Furthermore, we utilized Grad-CAM to visualize the impact of these specific regions on model outcomes to interpret the experimental result intuitively. Our findings highlight that targeting only the target area is insufficient and extending perturbations to shadow regions effectively enhances transferability.**

***Index Terms*—Adversarial example, Synthetic Aperture Radar, Convolutional Neural Networks, Transferability.**

I. Introduction

Synthetic Aperture Radar (SAR) image is a high-resolution radar image created by emitting microwaves into the ground or surface of a target, and obtaining the reflected signal [1]. The SAR Automatic Target Recognition (ATR) system is designed to automatically identify and classify targets from SAR images, and actively used in military, geography, and civilian applications [2]. The SAR-ATR mainly have relied on features extracted from experts' knowledge [3], which thus restricted its efficiency and applicability in practice [4]. Due to the recent improvement of feature extraction ability using Convolutional Neural Networks (CNNs), the performance of target recognition has significantly enhanced [5]. However, CNN was found to be vulnerable to the attack with adversarial examples [6], which injects small and imperceptible perturbations into the input image to mislead a well-trained model, making CNN-based SAR-ATR also vulnerable to the attack [7].

Unlike optical images, SAR images can be divided into three areas: *target*, *shadow* and *speckle* [8]. Target is the object in the input image; shadow is the black area caused by the physical shadow of the target; and speckle is the portion of the background. Each region affects the performance of CNN-based SAR-ATR differently. Especially, the target area affects its performance most significantly, because the amount of information in the target is more highly related to the object compared to the other regions [9].

Recent studies have focused on two main directions to develop attacks suitable for real-world environments. One is the attacks focusing on the target region, since it is the only part in which the attacker can physically create the transformation required for the adversarial example [10]–[12]. In the real world, however, it is a very challenging task because the attacker must have the domain knowledge of how the SAR images are acquired in order to actually implement the desired adversarial example. Since the target region is less affected by the physical environments than the other regions, recent attacks have focused on the target region.

The other direction is the black-box attacks. Because gaining information (such as model structure or image preprocessing process) about the CNN model of a SAR-ATR can hardly be achieved by attackers in the real world [7], [13], [14], black-box attacks are considered more practical senarios. In the black-box attack, specifically, the attacker trains another model called a surrogate model using the inputs and outputs of the target model [15], and create an adversarial example of the surrogate model to compromise the target model. In this setting, *transferability*, the ability that adversarial examples generated from the surrogate model are effectively applied to the target model, is the vital factor in a black box attack [16]-[19].

Concerning the fact that recent researches on adversarial examples in the SAR-ATR domain focus on target region and

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No.2022-0-00411, IITP-2024-2021-0-01810), and also supported by Basic Science Research Program through the National Research Foundation (NRF) funded by the Korea government (NRF-2021R1A6A1A13044830).

transferability, we pose three new research questions (RQs).

1) RQ1: "Does an attack focused only on a target show superior performance in transferability?"
2) RQ2: "Is the transferability also affected by the other areas in SAR images?"
3) RQ3: "If physical attacks are feasible not only on the target but also on the shadows, how does transferability change?"

These research questions stem from the environment when utilizing a surrogate model in the black box setting. If an attack is focused only on the target area, the attack may be overfitted to the surrogate model, and may become unsuitable for the target model. For RQ1, we compare attacks on the entire SAR image versus target-focused attacks under black-box conditions to see how transferability is changed. For RQ2, we also examine transferability for attacks on the other regions to analyze how each region affects transferability. For RQ3, we analyzed the transferability of the attacks combining target and shadow areas. For the analysis, we trained eight different models (including 1 surrogate model and 7 target models) on the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset [20], and divided the SAR image into target, shadow, and speckle regions using SAR-Bake dataset [8]. To create adversarial example, we utilize the four most common attacks in the AI vision domain, i.e. Fast Gradient Sign Method (FGSM) [6], DeepFool [21], Projected Gradient Descen (PGD) [22], and Carlini&Wagner (CW) [23].

Grad-CAM effectively visualizes how different regions of an image contribute to the classification process, thereby enhancing the interpretability of models. By applying adversarial perturbations to each distinct region (i.e., target, shadow, and the entire image), we examined how these contributions shifted to better understand the results of our experiments by observing Grad-CAM. Our analysis reveals that targeting perturbations solely at the target area is inadequate for improving transferability. Instead, we discovered that including shadow areas in the perturbations significantly boosts transferability. This observation highlights the necessity of applying broader spatial perturbations to facilitate more effective adversarial attacks.

II. Related Work

In this section, we introduce the recent work and background of adversarial examples in SAR-ATR.

A. SAR-ATR

SAR-ATR is designed based on feature-based methods [24] or model-based methods [25]. The former requires expert knowledge to extract features for target recognition, while the latter identifies the target by simulating the electromagnetic scattering process. Upon the development of CNN, SAR-ATR has greatly been improved by employing a robust feature extraction capability [26]. Recently, the performance of SAR-ATR has been further improved using various deep learning techniques such as unsupervised methods [27] and transfer learning [28].

B. Adversarial example

An adversarial example aims to cause a model to misclassify samples by inserting a perturbation that is too small to be perceived by the human eye into the input image [29]. An adversarial attack is a method to generate adversarial examples. Although numerous adversarial attack methods have been proposed so far, we select the four most commonly used attack schemes in AI vision, i.e., Fast Gradient Sign Method (FGSM) [6], DeepFool [21], Projected Gradient Descent (PGD) [22], and Carlini&Wagner (CW) [23] in this study.

The above attack methods considers white-box settings, where the attackers are allowed to access information about the structure, input, output, and weights of the target model. However, in a black-box attack scenario, the attacker can only access the input and output of the target model [15]. Thus, instead of attacking the target model directly, the attacker first trains a surrogate model separately. A white-box attack is then carried out on the surrogate model, and the target model is compromised using the generated adversarial example.

C. Adversarial examples in SAR images

Recently, several studies have been conducted on adversarial example in the SAR-ATR system in two main directions to develop attacks suitable for real-world environments [7], [10]-[14].

1) Focusing on target region: Several studies [10]–[12] proposed attack methods with a focus on physically vulnerable target areas. These methods involve the generation of adversarial perturbations in the form of electromagnetic wave scattering responses, taking into account the characteristics of the actual attack environment.

For example, a Scattering Model Guided Adversarial (SMGA) attack generates adversarial examples in the form of electromagnetic scattering responses [10]. An Attributed Scattering Center Spatial Transformation Attack (ASC-STA) method utilizes the SAR image characterization function of ASC model [11]. Also, an attack that generates perturbations near the target using a parametric model of the camouflage structure was developed [12].

2) Transferability in SAR Image: Adversarial examples generated for SAR-ATR should have high transferability across the CNN models, since there is no chance to modify perturbations after they are captured on the device [7]. Additionally, transferability is desperately required in SAR-ATR as it is difficult for attackers to directly access CNN-based SAR-ATR in the real world. Therefore, in recent efforts, attack methods have been developed to improve transferability. For instance, the speckle-variant-attack improved transferability by reconstructing the speckle-noise pattern [13], and the Scattering Center Model Attack (SCMA) enhanced transferability by designing it based on the target scattering distribution [14].

III. Empirical evaluation

In this section, we conduct an experiment to evaluate transferability of each attack based on different SAR regions, and discuss their implications.

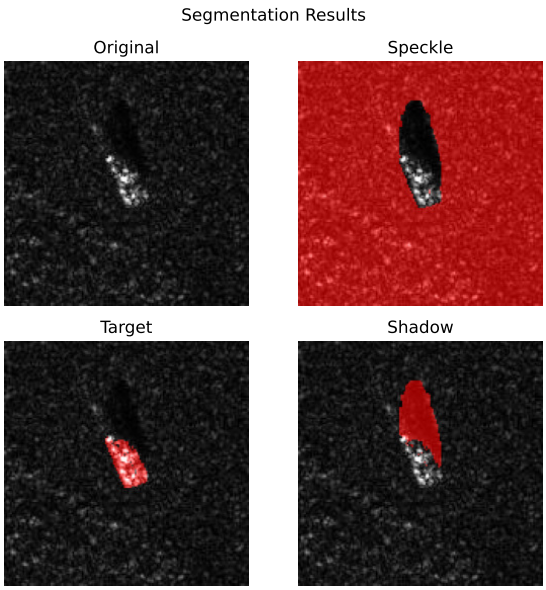

Fig. 1. Divided SAR image regions using SARBake

A. Datasets and Experimental Setup

1) Dataset and Models: We carried out our experiments on the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset released by the Defense Advanced Research Projects Agency (DARPA) [20]. It contains SAR images for 10 different types of armored vehicles and tanks.

We trained eight CNN models using MSTAR dataset. For the eight models, specifically, we used VGG16 [30] as a surrogate model; and Alexnet [31], Densenet [32], Inception [33], Mobile [34], Resnet50 [35], Resnext [36], and Shuffle [37] as target models.

SARBake [8], a public MSTAR segmentation annotation, was used to divide the SAR image into target, shadow, and speckle as Fig.1.

2) Metric and Details: Considering the black-box attack, we created adversarial example for surrogate model using four most commonly used attacks, FGSM, CW, DeepFool, and PGD. We used default value for attack parameter, which determines the strength of the attack. (ϵ = 8/255, iteration = 1 in FGSM; iteration = 50 in CW; ϵ = 8/255, iteration = 10 in PGD; and iteration = 50 in DeepFool.)

We selected 100 adversarial examples per class which are misclassified in the surrogate model, and calculated attack success rate for seven target models as transfer attack success rate to assess transferability.

B. Transferability Comparison between Target and All Pixels

Responding to RQ1, we compare the average transfer attack success rate between the attack focused on target and the attack focused on all pixels as shown in Fig. 2. Each attack algorithm shows different transfer attack success rates. In FGSM and PGD attacks, the average transfer attack success rate for all pixels was 61.05% higher than the one focused on only the target. This result shows that when the attack focuses on the target region, the adversarial example is overfitted to the surrogate model, thereby the transfer attack success rate falls. In CW and DeepFool attacks, the average transfer attack success rate for all pixels was 2.23% higher than the one focused on only the target. It implies that these attacks were overfitted to the surrogate model while going around more iteration than the previous attacks. As a result, even in the case of all pixels, the transfer attack success rate was low, and even lower in the target-focused case. A detailed analysis of this experiment result will be given in Section III-E1.

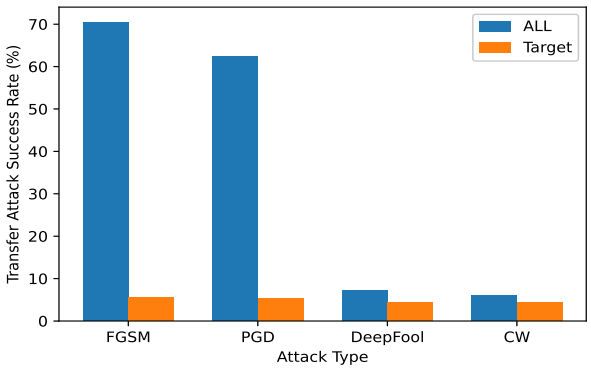

Fig. 2. Transfer attack success rate (all pixels and target) (%)

C. Impact of Target, Shadow, Speckle on Transfer Rate

Responding to RQ2, we compare the impact of each region on transferability individually. Fig. 3 shows the averaged transfer attack success rates of the four attacks in each of the 7 target models ('Alexnet', 'Desenet', 'Inception', 'Mobile', 'Resnet50', 'Resnext', and 'Shuffle') as well as the averaged success rates of all models ('Average'). In terms of the transfer attack success rate on average, Speckle shows the highest rate of 33.32%, followed by Shadow (8.64%) and Target (4.88%). It implies that more information about objects are included in the order of target, shadow, and speckle; and actually affects model performance in the same order in SAR images [9]. Since the adversarial example also generates perturbation according to the model performance, the more information the region has, the more likely it is to be overfitted. Through this experiment, we confirmed that the attack focused on the target area with more information was actually overfitted, resulting in the lowest transfer attack success rate.

D. Transferability of Target and Shadow

Responding to RQ3, we conducted an experiment on the transferability of a new attack that creates perturbations focused on the target and shadow regions simultaneously. Fig. 4 shows the result of the experiment, which was measured in the same manner as Fig. 3.

As the figure shows, performing adversarial attacks against both target and shadow increases transferability. It implies that

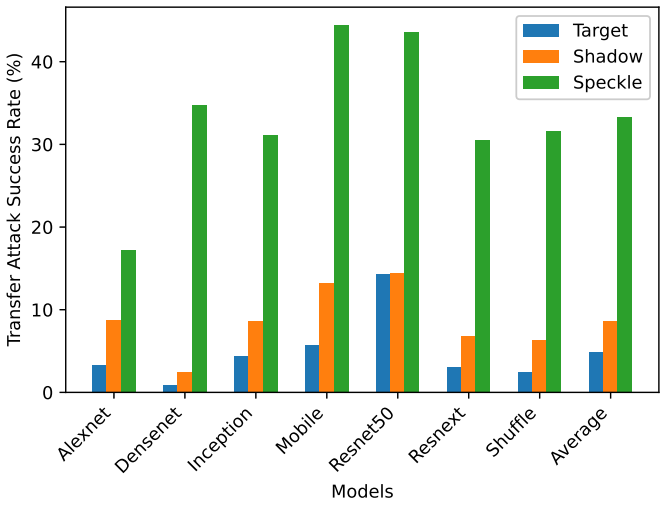

Fig. 3. Transfer attack success rate (target, shadow, and speckle) (%)

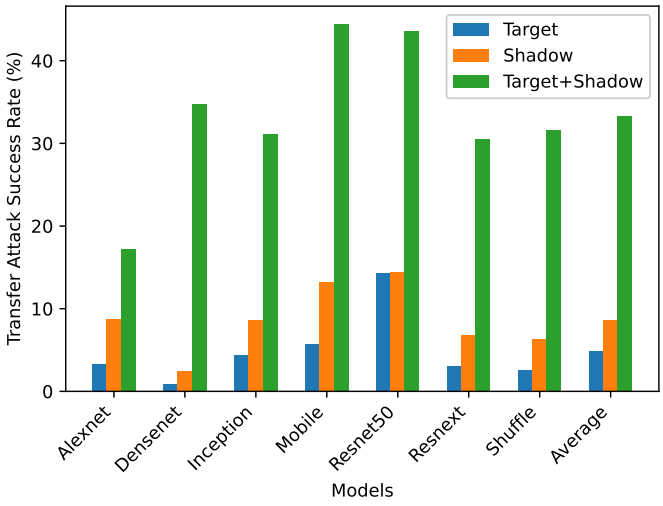

Fig. 4. Transfer attack success rate (target + shadow) (%)

the transferability of adversarial attacks can be enhanced by focusing on target and shadow at the same time. A detailed analysis of this result will be given in Section III-E2.

E. Analysis of Transferability by Grad-CAM

Grad-CAM [38] is a technique utilized to visually represent the regions in an image that significantly influence the prediction of a specific class within a CNN. Its algorithm is designed to generate an activation map, commonly referred to as a heatmap, utilizing gradient information. This technique facilitates the interpretation of the decision-making process of the CNN model, providing insights into which image regions significantly influenced the class prediction. The diverse colors in the heatmap serve as indicators of the influence of a specific region on the decision outcome, where red color signifies a significant contribution from that region to the decision, while blue color denotes a contrasting effect.

Using Grad-CAM, we conducted an in-depth analysis in order to understand the experiment results in Section III-B and III-D. Specifically, we compared the changes of model decision with different attacks in order to check how region-specific perturbations of each attack affect the model decisions differently. In addition, we compared model decision changes of surrogate model (vgg) and target model (Alex) to check how the same perturbation affects the model decisions differently. Fig. 5 and Fig. 6 contain the original image and the attacked images by region in different rows; and the model's decisions for each image and the corresponding Grad-CAM.

1) All-pixel and Target: As shown in Fig. 5, both the surrogate model (vgg) and the target model (Alex) classified the original image with a label of 2S1 correctly. In the attack that applied perturbation on all pixels ('adv-all' in the figure), it can be seen that both models' decisions were changed to T62. When comparing Grad-CAM in the original label (2S1) and the misclassified label (T62) with the original image's Grad-CAM (easily, comparing the first and second rows), we could observe the color-changes in overall pixels. It demonstrates the perturbation applied to all of the pixels changed both models' results by changing the contributions of many pixels.

On the other hand, as for the attack that applied perturbation only on target region ('adv-target' in the figure), we can see that the results of the surrogate model (vgg) and the target model (Alex) are conspicuously different, as well as the Grad-CAM (first and third rows). In the case of the surrogate model, changes in the target area are noticeable; while in the case of the target model, there is little change. This means that the perturbation applied to the target region affected only the surrogate model, not the target model. In other words, the adversarial example ('adv-target' in the figure) is overfitted to the surrogate model, resulting in poor transferability.

2) Target and Combining Target and Shadow: When comparing the original image with the adversarial example which focused on target in Fig. 6 (first and second rows), as done in Fig. 5, we can see that the adversarial example is overfitted to the surrogate model. To understand the experiment result in Section III-D, it is necessary to analyze Grad-CAM of target model (Alex) in the target+shadow case (third row). Compared to the original image (first row), there was a big change in both the original label (2S1) and the wrong label (BRDM_2), unlike the previous case (first and second rows). It implies that when perturbation was applied only to the target (second row), it was overfitted such that it affected only the surrogate model. However, by applying perturbation to the target and shadow simultaneously (third row), it was not overfitted such that it could affect the target model (Alex), increasing the transferability.

IV. Conclusion

In this paper, we investigated how the transferability of adversarial examples can be affected by each different region in SAR images. As a result of our experiments, we found (1) the attack focused on only the target region is overfitted to

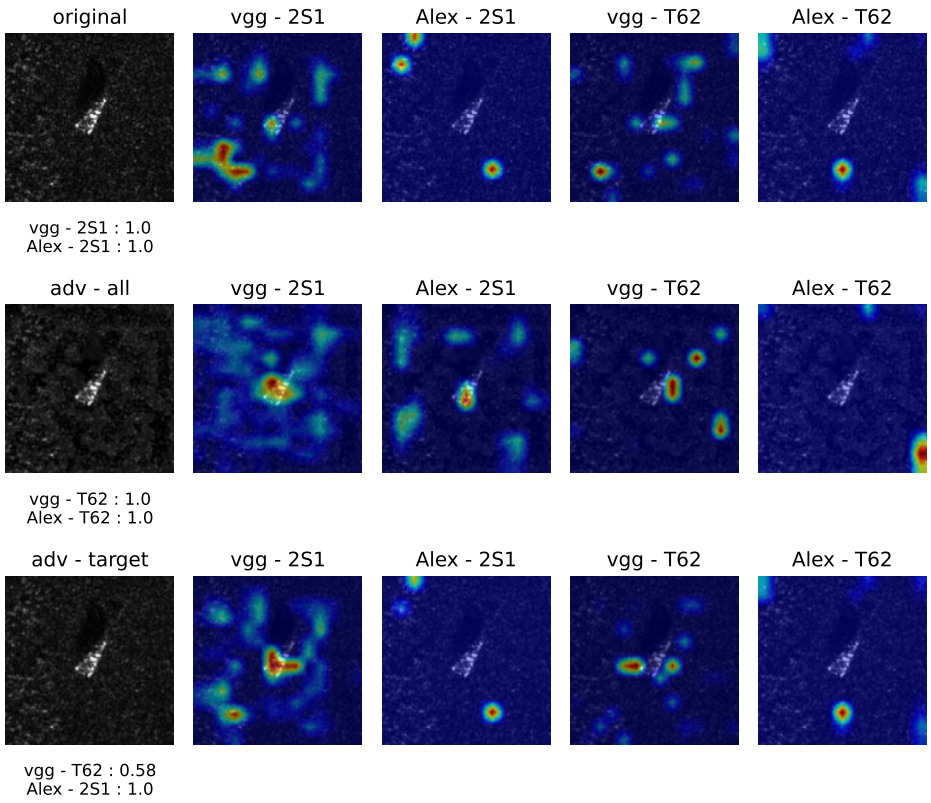

Fig. 5. Grad-CAM images with individually different perturbed regions in surrogate model and target model

the surrogate model, resulting low transfer attack success rate, (2) transferability varies depending on the regions including information about object, (3) the transferability increases by attacking the target and shadow areas simultaneously compared to attacking only the target area. To the best of our knowledge, this is the first work that investigated the effect of each different region of SAR image on the transferability.

REFERENCES

[1] G. C. Anagnostopoulos, "Svm-based target recognition from synthetic aperture radar images using target region outline descriptors," Nonlinear Analysis: Theory, Methods & Applications, vol. 71, no. 12, pp. e2934–e2939, 2009.

[2] O. Kechagias-Stamatis and N. Aouf, "Automatic target recognition on synthetic aperture radar imagery: A survey," IEEE Aerospace and Electronic Systems Magazine, vol. 36, no. 3, pp. 56–81, 2021.

[3] Y. Li, L. Du, and D. Wei, "Multiscale cnn based on component analysis for sar atr," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–12, 2021.

[4] Z. Feng, M. Zhu, L. Stankovic, and H. Ji, "Self-matching cam: A novel accurate visual explanation of cnns for sar image interpretation," Remote Sensing, vol. 13, no. 9, p. 1772, 2021.

[5] X. X. Zhu, S. Montazeri, M. Ali, Y. Hua, Y. Wang, L. Mou, Y. Shi, F. Xu, and R. Bamler, "Deep learning meets sar: Concepts, models, pitfalls, and perspectives," IEEE Geoscience and Remote Sensing Magazine, vol. 9, no. 4, pp. 143–172, 2021.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[7] H. Li, H. Huang, L. Chen, J. Peng, H. Huang, Z. Cui, X. Mei, and G. Wu, "Adversarial examples for cnn-based sar image classification: An experience study," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 1333–1347, 2020.

[8] D. Malmgren-Hansen, M. Nobel-J et al., "Convolutional neural networks for sar image segmentation," in 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, 2015, pp. 231–236.

[9] P. Li, X. Hu, C. Feng, X. Shi, Y. Guo, and W. Feng, "Sar-adbagnet: An interpretable model for sar image recognition based on adversarialdefense," IEEE Geoscience and Remote Sensing Letters, vol. 20, pp.1–5, 2022.

[10] B. Peng, B. Peng, J. Zhou, J. Xie, and L. Liu, "Scattering model guided adversarial examples for sar target recognition: Attack and defense,"IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–17, 2022.

[11] J.Zhou, S. Feng, H. Sun, L. Zhang, and G. Kuang, "Attributed scattering center guided adversarial attack for dcnn sar target recognition," IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1–5, 2023

[12] X. Dang, H. Yan, L. Hu, X. Feng, C. Huo, and H. Yin, "Sar image adversarial samples generation based on parametric model," in 2021 International Conference on Microwave and Millimeter Wave Technology (ICMMT). IEEE, 2021, pp. 1–3.

[13] B. Peng, B. Peng, J. Zhou, J. Xia, and L. Liu, "Speckle-variant attack: Toward transferable adversarial attack to sar target recognition," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2022.

[14] W. Qin, B. Long, and F. Wang, "Scma: A scattering center model attack on cnn-sar target recognition," IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1–5, 2023.

[15] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," arXiv preprint arXiv:1810.00069, 2018.

[16] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation, vol. 23, no. 5, pp. 828–841, 2019.

[17] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," arXiv preprint arXiv:1801.02610, 2018.

[18] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille,

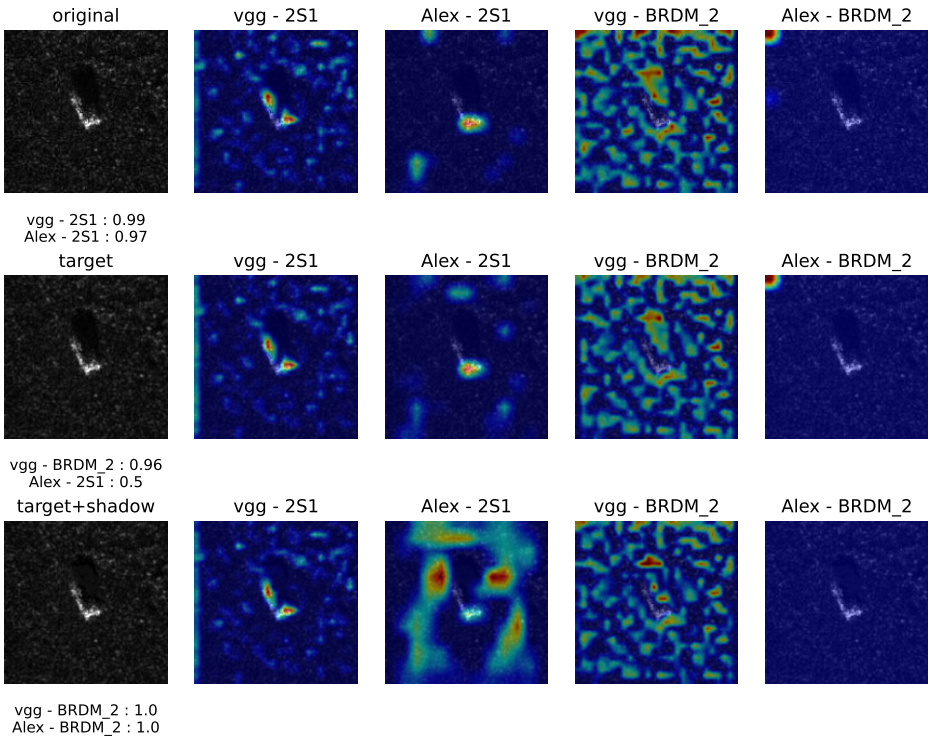

Fig. 6. Grad-CAM images with combinatorially different perturbed regions in surrogate model and target model

"Improving transferability of adversarial examples with input diversity," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2730–2739.

[19] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1924–1933.

[20] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for sar images," IEEE transactions on geoscience and remote sensing, vol. 54, no. 8, pp. 4806–4817, 2016.

[21] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2574–2582.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in ICLR, 2018.

[23] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 ieee symposium on security and privacy (sp). Ieee, 2017, pp. 39–57.

[24] L. M. Novak, G. J. Owirka, W. S. Brower, and A. L. Weaver, "The automatic target-recognition system in saip," Lincoln Laboratory Journal, vol. 10, no. 2, 1997.

[25] Z. Jianxiong, S. Zhiguang, C. Xiao, and F. Qiang, "Automatic target recognition of sar images based on global scattering center model," IEEE transactions on Geoscience and remote sensing, vol. 49, no. 10, pp. 3713–3729, 2011.

[26] J. Li, Z. Yu, L. Yu, P. Cheng, J. Chen, and C. Chi, "A comprehensive survey on sar atr in deep-learning era," Remote Sensing, vol. 15, no. 5, p. 1454, 2023.

[27] Y. Tian, J. Sun, P. Qi, G. Yin, and L. Zhang, "Multi-block mixed sample semi-supervised learning for sar target recognition," Remote Sensing, vol. 13, no. 3, p. 361, 2021.

[28] Li, Sen, et al. "Automatic target recognition of SAR images based on Transformer." 2021 CIE International Conference on Radar (Radar). IEEE, 2021.

[29] S. Bubeck, Y. T. Lee, E. Price, and I. Razenshteyn, "Adversarial examples from computational constraints," in International Conference on Machine Learning. PMLR, 2019, pp. 831–840

[30] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[31] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

[32] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[33] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[34] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[35] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[36] Xie, Saining, et al. "Aggregated residual transformations for deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[37] Ma, Ningning, et al. "Shufflenet v2: Practical guidelines for efficient cnn architecture design." Proceedings of the European conference on computer vision (ECCV). 2018.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.