Empirical Investigation of Adversarial Attacks for Semi-Supervised Object Detection

Junhyung Jo*
Graduate School of Artificial Intelligence
Pohang University of Science and Technology (POSTECH)
Pohang, South Korea
jjo22@postech.ac.kr

Joongsu Kim*
Dept. of Computer Science and Engineering
Pohang University of Science and Technology (POSTECH)
Pohang, South Korea
joongsukim@postech.ac.kr

Young-Joo Suh
Graduate School of Artificial Intelligence
Pohang University of Science and Technology (POSTECH)
Pohang, South Korea
yjsuh@postech.ac.kr

* Equal Contribution

***Abstract*—Semi-supervised learning (SSL) techniques have been rapidly developed and adopted in various vision tasks because of their advantage of leveraging unlabeled data. However, existing works have neglected the vulnerability of SSL because most of the existing adversarial attacks are mainly discussed in a supervised learning manner. In this paper, we study the effects of adversarial examples on Semi-Supervised Object Detection (SS-OD) which is the mainstream of SSL techniques. We build our hypothesis that attacks on the supervised learning model are also effective on SSL models. Since the state-of-the-art SS-OD methods borrow the teacher-student network, we prepared two pseudo-label based SS-OD networks to validate our hypothesis. We attempt to attack the inference model with adversarial examples crafted by using pretrained auxiliary model and found that SS-OD networks are more vulnerable to adversarial attacks. In addition, we found that selecting different loss components of SS-OD networks to generate perturbations determines the effect and performance of the attack such as misclassification or mislocalization. Visual examples are provided for a clearer understanding. To the best of our knowledge, this is the first effort to investigate the vulnerability of SS-OD.**

***Index Terms*—Adversarial attack, Semi-supervised object detection**

I. Introduction

Deep Neural Networks (DNNs) have emerged as state-of-the-art solutions in a variety of vision-related tasks such as classification [1]–[4] and object detection [5], [6]. However, one of the major limiting factors in applying these models in practice is the reliance on large labeled data sets that are expensive to collect to train the models. Semi-supervised learning (SSL) techniques have been proposed to handle this issue. SSL model leverages only a small set of labeled data but a large set of unlabeled data to improve performance [7], [8]. Because these techniques can leverage additional unlabeled data, they have improved to the point where they exceed the accuracy of fully supervised learning.

This work was partly supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2022R1A6A1A03052954).

Despite the development of SSL, the vulnerability of this technique has not been studied a lot, unlike many attack methods that have been studied with the development of supervised learning. For the classification task, a *poisoning attack* on unlabeled data has been attempted to show the vulnerability of SSL [9], but the poisoning attack is not realistic enough to be used in practice as it requires directly injecting adversarial data in the training stage of the model.

In this paper, we perform the evaluation in Semi-Supervised Object Detection (SS-OD), which is more difficult than a classification task, to examine the effect of *adversarial examples* on SSL. In particular, we study *evasion attacks* in the *white-box* manner, which are possible attacks in the inference stage of the model. Evasion attacks can affect the inference phase of the model hence they are more realistic than poison attack methods that affect the training phase.

Existing state-of-the-art SS-OD methods [10], [11] apply self-training techniques, specifically, the pseudo-labeling method [12], in which the teacher model generates pseudo-labels and enforces the consistency between unlabeled data with different augmentations. Considering that most of the SS-OD models that record state-of-the-art borrow the common method of pseudo-labeling mentioned above, we first focused on analyzing the structural weaknesses of this model. The structure of the model that borrows the pseudo-labeling method can be divided into a teacher model and a student model. The teacher model trains with a small amount of labeled data set and generates pseudo-labels for a large amount of unlabeled data. The student model is then trained on a large pseudo-labeled data set. In this process, if the teacher model is trained in the wrong direction, it will cause performance degradation in the student model trained with falsely generated pseudo-labeled data [2]. This means that the performance of the teacher model affects the performance of the student model. Based on this evidence, the following hypothesis is formulated.

Verified adversarial attack in the supervised learning model is also working on SS-OD

Therefore, our analysis focuses on verifying the effect of SS-OD on performance by attacking the inference model with the adversarial example generated by the auxiliary model. Here, the auxiliary model refers to a model that generates pseudo-labels, and the inference model refers to a model used in the final inference step. This is to prevent confusion in understanding the explanation because the roles of the teacher model and the student model may change depending on the SS-OD model.

To give a brief overview of the stages that make up the research, in the first step, we begin by implementing an experiment to prove a hypothesis using off-the-shelf attack methods [13]–[15]. We use adversarial examples generated by attacking the pre-trained auxiliary model to attack the inference model. The adversarial examples generated by attacking the pre-trained auxiliary model with different ratios of labeled data sets are used to attack the inference model. In the second experiment, we generate adversarial examples using not only one classification loss, but each loss computed from Faster-RCNN [16] which is the base model of both auxiliary and inference models, and evaluate the impact of each loss and their combination on the attack method. These are evaluated in the same way as in the previous experiment: generating the adversarial example from the auxiliary model and causing the degradation of the performance of the inference model.

We make the following contributions:

- We proved that the attack method verified in the supervised learning model is also valid for SS-OD for the first time. As a result, all SS-OD models [10], [11] used in the experiment showed a performance decrease of at least 80% and a maximum of 95% due to the off-the-shelf attack method [13]–[15].
- We evaluated the influence of adversarial examples generated by each loss of Faster-RCNN [16] and their combination through various attack methods: when total-loss was used, performance improved by at least 2% to 6% compared to other cases.

II. Related Work

A. Adversarial Attacks

Generating adversarial attacks is widely studied in classification. [17] first showed that deep neural network based classifiers are vulnerable to small perturbations to the images and proposed a box-constrained L-BFGS method. [13] proposed fast gradient sign method (FGSM) to generate adversarial examples efficiently by applying one step gradient information. Due to the underfitting issue of FGSM, [18] extended FGSM to an iterative version. [14] improved I-FGSM by starting at a random point. These iterative attacks have a trade-off in that they succeed the attack in high probability in white-box settings, but low transferability leads to poor performance in black-box settings. To resolve this problem, [15] added a momentum term to improve the transferability of adversarial examples and also stabilize the update direction.

Adversarial attacks are also studied in the object detection field. [19] proposed black box attacks by using patches. The attack performs well without the knowledge of the attacked network's architecture, however, the adversarial examples are easily perceptible to humans. [20] proposed DAG that is able to attack both semantic segmentation and object detection, but it requires a large number of iterations which leads to a massive computational overhead.

B. Semi-Supervised Object Detection

Semi-Supervised Object Detection (SS-OD) is proposed to solve the overhead issue of acquiring and labeling data for object detection. [21] applies consistency constraints to improve object classification and localization. In recent works, the pseudo-labeling strategy is widely used in that the model generates pseudo-labels from unlabeled data and includes them in the training data after a confidence check. [10] first used a well-known teacher-student based framework in semi-supervised object detection. Due to its lack of extra training for teacher, STAC has limitations on final detection performance. To deal with this problem, Two networks of Instant-teaching [22] share both parameters. Unbiased teacher [11] uses EMA [23] to train the teacher from the knowledge obtained by the student. Active teacher [24] extends the teacher-student network to an iterative version to maximize the effect of limited label information. In this work, we focus on the pseudo-label based semi-supervised object detection methods of two-stage models.

III. Method

In order to evaluate whether the attack method verified in supervised learning is effective in SS-OD, the attack part of the target model must first be determined. The target models [10], [11] in this paper use the pseudo-labeling method. In pseudo-labeling, if the auxiliary model trains in the wrong direction, an incorrect pseudo-label is generated, which causes the performance of the inference model trained with these pseudo-labeled data to deteriorate. In a similar vein, we generate adversarial examples through off-the-self attacks on auxiliary models and attack inference models with them. For this purpose, it is necessary to set an auxiliary model and an inference model in the target model. In this section, the target model, STAC [10] and Unbiased Teacher's [11] algorithm are analyzed to determine the appropriate attack part. Note that both the teacher model and student model are based on the Faster-RCNN [16].

A. STAC

Teacher model trained on available labeled images. The supervised loss is as follows:

$$\ell_s(x, p^*, t^*) = \sum_i \ell_s(x, p_i^*, t_i^*) \tag{1}$$

$$= \sum_i \left[\frac{1}{N_{cls}} \mathcal{L}_{cls}(p_i, p_i^*) + \frac{\lambda}{N_{neg}} \mathcal{L}_{reg}(p_i, p_i^*) \right] \tag{2}$$

where i is an index of an anchor in a mini-batch. p_i is the predictive probability of an anchor being positive, t_i is the 4-dimensional coordinates of an anchor. p_i^* is the binary label

TABLE I
EXPERIMENT RESULTS ON COCO-STANDARD ON SS-OD METHODS[1].

	COCO-Standard							
	STAC				Unbiased Teacher			
	1%	2%	5%	10%	1%	2%	5%	10%
Benign[2]	11.99	16.56	20.54	24.00	20.16	24.15	27.84	31.39
FGSM	1.59	2.96	2.97	3.55	2.26	2.81	3.09	3.66
PGD	1.56	3.38	2.97	3.41	0.89	0.98	1.07	1.28
MI-FGSM	1.36	3.05	2.73	2.99	0.95	0.96	1.19	1.38

[1] We used the total loss of the auxiliary model to generate adversarial examples.
[2] Benign refers to the performance of the model evaluated with clean images.

of an anchor with respect to ground-truth boxes, t_i^* is the ground-truth box coordinates of the box i for all $p_i^* = 1$. N_{neg} and N_{reg} are regularization factors for classification and regression, respectively. λ is the weight for the regression loss $\mathcal{L}_{reg}$.

In STAC [10], the teacher model is fixed during the entire process of training. Therefore, the teacher model becomes the auxiliary model which is the target of attack. Then, the student model naturally becomes an inference model and is subject to attack by the adversarial example generated by the attacking method.

B. Unbiased Teacher

In this model, *Burn-In* stage is required to make good initialization for both student and teacher models. First, the available supervised data is used to optimize our model θ with the supervised loss $\mathcal{L}_{sup}$. With supervised data $D_s = x_i^s, y_i^{s\,N_s}{}_{i=1}$, the supervised loss consists of four losses: the RPN classification loss $\mathcal{L}_{cls}^{rpn}$, the RPN regression loss $\mathcal{L}_{reg}^{rpn}$, ROI classification loss $\mathcal{L}_{roi}^{cls}$ and ROI regression loss $\mathcal{L}_{reg}^{roi}$.

$$\mathcal{L}_{sup} = \sum_i \mathcal{L}_{cls}^{rpn}(x_i^s, y_i^s) + \mathcal{L}_{reg}^{rpn}(x_i^s, y_i^s) + \mathcal{L}_{cls}^{roi}(x_i^s, y_i^s) + \mathcal{L}_{reg}^{roi}(x_i^s, y_i^s) \quad (3)$$

After Burn-In, we duplicate the trained weights θ for both the teacher and the student models ($\theta_t \leftarrow \theta, \theta_s \leftarrow \theta$). Next, the teacher model trains as the temporal ensemble of the student models in different time steps via EMA [23].

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s \quad (4)$$

This aligns with that the accuracy of the teacher is consistently higher than the student. Hence the teacher model is used to infer the output of the model which means the teacher model is the inference model that is attacked by the adversarial examples generated by the auxiliary model which is the student model in Unbiased Teacher [11].

C. Total-loss Attack

Both STAC [10] and Unbiased Teacher [11] use Faster-RCNN [16] as the base model for the auxiliary models that are targeted for attack. When the auxiliary model is trained on a small labeled data set, the loss is computed as eq.3. In the previous studies [13]–[15], off-the-shelf attack methods update the gradient using only $\mathcal{L}_{cls}^{roi}(x_i^s, y_i^s)$ corresponding to the classification loss in the RoI head, and use this to generate adversarial examples for each attack method. In this paper, an adversarial example is generated using total-loss $\mathcal{L}_{sup}$, which is the sum of all losses, rather than one loss. **One-step gradient-based method**, such as the fast gradient sign method (FGSM) [13], find adversarial examples $\tilde{x}$ by maximizing the loss function $\mathcal{L}_{sup}(\tilde{x}, y)$. where $\mathcal{L}_{sup}$ is the total-loss from auxiliary model mentioned earlier. FGSM generates adversarial examples to satisfy the $\mathcal{L}_\infty$ norm bound $||\tilde{x} - x||_\infty \leq \epsilon$:

$$\tilde{x} = x + \epsilon \cdot sign(\nabla_x \mathcal{L}_{sup}(x, y)) \quad (5)$$

Iterative methods, such as the momentum iterative fast gradient sign method (MI-FGSM) [15] and projected gradient descent (PGD) [14] iteratively apply fast gradient multiple times with a small step size α:

$$\tilde{x}_{t+1} = \tilde{x}_t + \alpha \cdot sign(\nabla_x \mathcal{L}_{sup}(\tilde{x}, y)) \quad (6)$$

To make sure the generated adversarial examples satisfy the bounds of $\mathcal{L}_\infty$, can clip $\tilde{x}_t$ to around ϵ of x, or set $\alpha = \frac{\epsilon}{T}$ and T to the number of iterations. The iterative method was found to be a more robust white box attack than the one-step method, at the cost of less transferability [25], [26].

IV. EXPERIMENT

In this section, we conduct several experiments on the COCO-standard dataset to validate the effectiveness of our proposed method on SS-OD networks. Our experiment settings are specified in Section IV-A. The quantitative results of our experiments are explained in Section IV-B. We showed the difference between student and teacher models in Section IV-C. In addition, an ablation study is provided in Section IV-D.

TABLE II
EXPERIMENT RESULTS FOR THE CROSS-NETWORK ATTACK.

	STAC		Unbiased Teacher	
	Same Model[1]	Cross[2]	Same Model	Cross
FGSM	1.6	3.59	3.14	3.65
PGD	0.24	2.97	0.54	1.37

[1] The adversarial example is crafted by the inference model. It is held by the inference model.
[2] The adversarial example is crafted by the auxiliary model. It is held by the inference model.

(a) Total (b) Loss-cls (c) Loss-reg (d) Loss-rpn

Fig. 1. Illustration of adversarial examples on different loss functions generated by PGD attack. To easily show the impact, we set the number of iterations $N = 40$. The perturbation is still human-imperceptible. Other settings are the same as the previous experiments.

(a) Benign (b) FGSM (c) PGD (d) MI-FGSM

Fig. 2. Illustration of detection in adversarial examples. Benign images are shown in the first column, and three other attack methods are shown in the second to the fourth column. We only visualize the adversarial examples from the Unbiased Teacher since STAC shows a similar pattern.

A. Setup

Networks: We consider two representative SS-OD networks, i.e., STAC [10], and Unbiased Teacher [11]. We adopt the settings of 1%, 2%, 5%, and 10% labeled data of MS-COCO *train2017* for training and *val2017* for testing. The total training iteration steps for each network are 180k. In STAC, we replaced the geometric data augmentation with the data augmentation of the Unbiased Teacher. Other settings are the same as the settings of the original paper.

Implementation details: We prepared three widely used attack methods, i.e., FGSM [13], PGD [14], and MI-FGSM [15]. For a fair comparison, we follow the parameter settings for the attack methods same as the previous studies. We set the maximum perturbation of each pixel to be $\epsilon = 16$. The decay factor is set to be $\mu = 1.0$ for the methods using the momentum term. Maximum iteration for iteration-based methods is set to be $N = 20$. We use mean average precision (mAP) as an evaluation metric, and the performance is cross-evaluated so that the perturbed image is generated by the auxiliary model and the evaluation is held on the inference model. For example, in Unbiased Teacher, we craft adversarial examples only on the student network and test them on the

TABLE III
THE MAP OF USING ADVERSARIAL ATTACKS ON SS-OD METHODS WITH DIFFERENT LOSSES[1].

		STAC				Unbiased Teacher			
		1%	2%	5%	10%	1%	2%	5%	10%
None		11.99	16.56	20.54	24.00	20.16	24.15	27.84	31.39
FGSM	Total-loss	**1.59**	**2.96**	**2.97**	**3.55**	**2.26**	**2.81**	**3.09**	3.66
	Loss-cls	1.65	3.02	3.02	3.63	2.26	2.89	3.28	**3.25**
	Loss-reg	2.32	3.84	3.97	4.61	3.01	3.77	4.44	4.81
	Loss-rpn	3.04	4.45	5.08	6.04	4.43	5.27	6.09	6.72
PGD	Total-loss	**1.56**	**3.38**	**2.97**	**3.41**	**0.89**	**0.98**	**1.07**	**1.28**
	Loss-cls	1.57	3.48	3.01	3.71	1.07	1.08	1.16	1.36
	Loss-reg	2.13	4.22	3.98	4.71	1.68	1.94	2.14	2.70
	Loss-rpn	2.89	5.05	5.09	6.55	3.70	4.87	5.47	6.39

[1] Total loss outperforms the compared single loss function. We observed that classification loss is the most effective loss between its comparisons and even exceeds the performance of total loss in Unbiased Teacher trained with 10% labeled data.

teacher network. Evaluation of STAC works in a reverse manner as the student network is the inference model. This can be considered a black-box attack because student and teacher models have different training data and processes.

B. Quantitative Results

We first evaluate the performance of our attacks on two SS-OD networks that are previously mentioned. Total-loss is adopted to generate perturbation. Results are summarized in Table I. We can observe that three attack methods drop the mAP significantly compared with the mAP evaluated on benign image samples. We also find that the networks trained FGSM show gentle performance on all eight networks as one-step gradient methods are robust on black-box attacks. On the other hand, PGD works well with Unbiased Teacher but underperforms in STAC than FGSM. From the results of PGD attack, we can point out that the iterative attack method is not likely to fit on the black box model which is almost the same as the white box model. Lastly, MI-FGSM is equal to or outperforms the two attack methods as adopting momentum is effective on black-box attacks. In Figure 2, we apply our attacks on four different images to the Unbiased Teacher.

C. Transferability between Teacher and Student Network

As we mentioned earlier, teacher and student networks have the same architecture but the training data and process is different. So applying the perturbations learned from the auxiliary model to the inference network can be considered a black-box attack.

To validate our hypothesis, we generated adversarial examples by using the loss gradient information of the inference model and attacking the same model. This kind of attack is a white-box attack without a doubt. We observed that attacking itself drops mAP more than using perturbed image generated by using the auxiliary model which demonstrates our hypothesis. Additionally, we would like to focus that the performance of PGD drops significantly in cross-attacks compared with FGSM. This is due to the over-fitting of iterative attack methods. Lastly, cross-attacks on Unbiased Teacher are more effective than the attacks in STAC. This could be explained that the two networks in Unbiased Teacher have more similarity than the networks in STAC because the teacher network shares the model weights of the student models by EMA training. We conclude that our attacks can be more effective not only in Unbiased Teacher but also in recent SS-OD methods [11], [24], [27] which applies EMA during the training process [23]. Detailed results are shown in Table II.

D. Ablation Study

We perform a simple ablation study to investigate the impact of adopting total-loss gradient information to generate adversarial examples. The study analyzes the impact of selecting the loss functions on the detector's performance degradation. As shown in Table III, combining all the losses of Faster R-CNN [16] is the most effective choice. One interesting point we observed is that adversarial examples generated by using only RPN loss have weak mAP degradation compared with others. This is because the loss gradient of RPN has small values so more iteration is required. Using the classification loss causes a significant drop however using total loss still performs better.

For clear interpretation, visualization is provided in Figure 1. As can be seen in Figure 1, adversarial samples using total-loss gradient information lead to more misclassification and inaccurate bounding boxes. Furthermore, our observations reveal that selecting various single-loss functions leads to the loss of different capabilities in the detectors. For example, the detectors fail to localize the object when we use box regression loss in the third column. In contrast, adversarial examples using classification loss fool the detector to generate inaccurate bounding boxes and classification. Therefore, we can conclude that choosing different loss functions can attack different abilities of object detectors.

V. CONCLUSION

In this paper, we proved that verified adversarial attack in the supervised learning model is also working on SS-OD, by analyzing the effect of SS-OD on performance by attacking the inference model with the adversarial example generated by

the auxiliary model. A performance of SS-OD decrease of at least 80% and a maximum of 95% by off-the-shelf attacking methods. We found that state-of-the-art SS-OD methods are more vulnerable on adversarial attacks due to their parameter-sharing training strategy. The impact of adversarial examples generated by the type of loss was also analyzed. Experimental results show that adversarial examples generated using total-loss degrade the performance of the target model more than when using single-loss.

REFERENCES

[1] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *International Conference on Machine Learning*, pp. 23965–23998, PMLR, 2022.

[2] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11557–11568, 2021.

[3] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022.

[4] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[5] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3060–3069, 2021.

[6] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022.

[7] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, p. 896, 2013.

[8] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *Advances in neural information processing systems*, vol. 31, 2018.

[9] N. Carlini, "Poisoning the unlabeled dataset of semi-supervised learning," in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1577–1592, 2021.

[10] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020.

[11] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," in *International Conference on Learning Representations*, 2021.

[12] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

[13] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[14] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[15] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," International Conference on Learning Representations, 2014.

[18] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *International Conference on Learning Representations*, 2017.

[19] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen, "Dpatch: An adversarial patch attack on object detectors," arXiv, 2018.

[20] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," arXiv, 2017.

[21] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[22] Q. Zhou, C. Yu, Z. Wang, Q. Qian, and H. Li, "Instant-teaching: An end-to-end semi-supervised object detection framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4081–4090, June 2021.

[23] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," arXiv, 2017.

[24] P. Mi, J. Lin, Y. Zhou, Y. Shen, G. Luo, X. Sun, L. Cao, R. Fu, Q. Xu, and R. Ji, "Active teacher for semi-supervised object detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14462–14471, 2022.

[25] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *International Conference on Learning Representations*, 2017.

[26] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *International Conference on Learning Representations*, 2018.

[27] Y.-C. Liu, C.-Y. Ma, and Z. Kira, "Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.