The Visual Computer (2022) 38:1929–1944
https://doi.org/10.1007/s00371-021-02256-6

ORIGINAL ARTICLE

Adversarial defenses for object detectors based on Gabor convolutional layers

Abdollah Amirkhani[1] · Mohammad Parsa Karimi[1]

Accepted: 6 July 2021 / Published online: 24 July 2021

Abstract

Despite their many advantages and positive features, the deep neural networks are extremely vulnerable against adversarial attacks. This drawback has substantially reduced the adversarial accuracy of the visual object detectors. To make these object detectors robust to adversarial attacks, a new Gabor filter-based method has been proposed in this paper. This method has then been applied on the YOLOv3 with different backbones, the SSD with different input sizes and on the FRCNN; and thus, six robust object detector models have been presented. In order to evaluate the efficacy of the models, they have been subjected to adversarial training via three types of targeted attacks (TOG-fabrication, TOG-vanishing, and TOG-mislabeling) and three types of untargeted random attacks (DAG, RAP, and UEA). The best average accuracy (49.6%) was achieved by the YOLOv3-d model, and for the PASCAL VOC dataset. This is far superior to the best performance and accuracy and obtained in previous works (25.4%). Empirical results show that, while the presented approach improves the adversarial accuracy of the object detector models, it does not affect the performance of these models on clean data.

Keywords Machine vision · Object detection · Adversarial attack · Robust detector

1 Introduction

The object detection technique, whose aim is to detect specific objects and their positions in an image, constitutes one of the most applicable fields in machine vision [1]. In recent years, very good object detection models, whose backbones are the deep neural networks (DNNs), have been introduced. Some of the more important of these models include the different versions of the YOLO [2] and SSD [3]. These models are called the single-shot object detection models, because they detect the target objects and determine their positions in a single step. Of course, for detecting the salient objects, these models need to be robustified. A sample activity in this field is the work of Liu et al. [4], who have presented a robust technique for the detection of salient objects.

Despite all the known advantages of the DNNs, it was discovered in 2013, that these networks are vulnerable to adversarial attacks, and that the images corrupted by such attacks can fool and mislead the DNNs [5]. Since then, the robustification of the deep neural networks has become one of the most important concerns of the researchers in this field [6]. The adversarial attacks can be combined with images in the form of small perturbations; and although these perturbations and contaminations cannot be detected by human eye [7], they can mislead the DNNs and severely reduce the accuracy of the models that are based on deep learning. Unfortunately, the DNNs report the images they have misrecognized as being highly reliable and least erroneous [8].

Since the introduction of this drawback, many efforts have been made to improve the adversarial accuracy of models and to enhance their robustness against adversarial attacks in machine vision tasks. In spite of all these efforts, still the adversarial accuracy of the models, especially those in the field of object detection, has not reached an acceptable level [9]. Considering these facts, the application of the deep learning knowledge in various fields, especially in the real-world applications (e.g., the autonomous vehicles) would be a challenging task [10]. The adversarial accuracy is a parameter that shows the accuracy of a model when it is tested on the images perturbed by adversarial attacks [11]. Another problem reported by the research works conducted on this subject is the reduction of the

✉ Abdollah Amirkhani
amirkhani@iust.ac.ir

[1] School of Automotive Engineering, Iran University of Science and Technology, 16846-13114 Tehran, Iran

models' clean accuracy. The defenses presented in the literature against adversarial attacks have led to a significant decline in the clean accuracy of the object detection models, which means a reduction in the accuracy of object detectors when they deal with the input images that are not perturbed by adversarial attacks [12].

The adversarial attacks can be divided into targeted and untargeted classes. The targeted adversarial attacks are those that perturb the input images in such a way that for a defined class of objects, a specific label is presented as the output. Of course, in the targeted attacks, sometimes the images are perturbed in such a way that no labels are designated for them by the object detectors. The actions of the targeted adversarial attacks are controlled by the attack designer. Conversely, the untargeted adversarial attacks simply combine with the input images and they have no particular control over the outcome of the object detectors as they examine such contaminated images [13].

The adversarial attacks of different forms can be devised by different techniques and algorithms. However, to be able to compare the results of various relevant research works, a standardized benchmark of targeted and untargeted attacks was presented in [14] for the first time; and in this benchmark, standardized attacks were used for the object detection models. Although, there have been very few attacks and defenses in the field of object detection so far, it is necessary to have this benchmark, if we want to better compare the results of various research works.

The Gabor filters are the most frequently used filters in the conventional machine vision tasks. These filters are based on a sinusoidal plane wave of a specific frequency and direction; which enables them to extract the spatial structures from images [15]. Combining these filters with the DNNs for different purposes has become a research interest. In [16], the Gabor filters have been combined with the deep learning models in classification tasks and have improved the robustness of these networks against adversarial attacks. Also, the Gabor filters have been combined with the DNNs in [17] in order to reduce the complexity and increase the learning speed of these networks.

In this paper, we introduce five object detectors that are robust to adversarial attacks. These robust models have been obtained by combining the Gabor filters with the backbones of the different versions of famous models (YOLO.v3, SSD, and Faster R-CNN) [18]. Then, each of these robust object detectors has received adversarial training by means of the perturbed images from the MSCOCO (2017 version) and the PASCAL VOC (2012 version) datasets. Adversarial training means training a network with the images that have been perturbed by adversarial attacks [19].

In recent years, newer adversarial attacks based on more novel techniques have also been introduced, such as the Evaporate attacks [20]. This type of attacks, which are categorized as the black-box attacks, can successfully mislead the detection models without having to know the architecture of a destination system. For example, Wang et al. [20] introduced an effective type of attack that could successfully mislead the object detection models such as the YOLO and the FRCNN. They effectively combined the Evaporate, Boundary and the Gaussian Noise Attacks and formed a black-box type of attack. In our paper, in addition to the 6 adversarial attacks considered, the presented method has been evaluated again on all the object detection models by applying this combined attack. In our paper, we have abbreviated this attack as EBG and we have reexamined the models for the case in which the images are disturbed by the mentioned type of attack. Also, Lee and Kolter [21] have presented an adversarial patch for deceiving the object detectors. These authors claim that this type of attack is quite effective on the object detection models and totally disrupts their object detection ability. Some attacks are developed exclusively for a specific image [22], and some attacks are designed to mislead a system on a particular class of images. These attacks have recently attracted the attention of the attack developers. Wang et al. [23] have presented a patch which is aimed at deceiving the object detection systems on specific classes of images. In its maximum state of performance, this patch has been able to reduce the accuracy of the detection systems by 81%.

For perturbing the images in this paper, we have used 6 of the more famous adversarial attacks in the field of object detection (TOG-vanishing, TOG-fabrication, TOG-mislabeling [14], DAG [24], RAP [25] and UEA [26]). Finally, the results of implementing the considered models on different datasets and adversarial attacks have been obtained and the performances of these models have been compared with each other. Some of the attacks used in this paper are new, and there are no reports in the literature about the performances of other defensive techniques against these attacks. Therefore, for comparing the models presented in this paper with those in other papers, we have also evaluated and compared the performances of other defensive techniques against the attacks used in this paper. The results of these comparisons have been presented at the end of this manuscript.

The next sections of this paper contain the following: Sect. 2 briefly introduces some of the works carried out on the robustification of DNNs against adversarial attacks. Section 3 describes the proposed technique and explains its application on some well-known object detectors. The results of our model are given in Sect. 4 and compared with those of the other models. And finally, the Conclusion and the Discussion are presented in Sect. 5. The main contributions of our paper are as follows:

- In this paper, a novel method based on the Gabor filters has been presented for robustifying the object detectors against adversarial attacks. This approach improves the adversarial accuracy of the VOD models much more than the former methods considered.
- The proposed method has been implemented on the most famous object detection models (YOLOv3-m, YOLOv3-d, SSD300, SSD512 and FRCNN) and the results have been evaluated and compared extensively on different models.
- To verify the proper performance of the defense technique presented in this paper, the newest and the most common adversarial attacks (both targeted and untargeted) have been used in this work, and the proposed model has been evaluated by considering 7 different types of attacks.
- Finally, the proposed method has been compared with the most recent techniques in this field, and it has been successfully demonstrated that the performance of this approach is better than that of the other state-of-the-art methods introduced in the literature.

2 Related works

Numerous research works have been conducted on the robustification of DNNs against adversarial attacks in different tasks, with most of them related to the classification field. To make the DNNs robust in the classification tasks, Gabor filters have been combined with several well-known architectures including the ALEX NET and the VGG16. The adversarial training method has been employed in [26–30] to make the DNNs robust in the classification tasks. This technique is not sufficiently effective against strong attacks [9]. The denoising auto encoders have been used in [29] to deal with the adversarial attacks. The authors of that paper claim that using a denoising auto encoder can boost the network robustness against such attacks. Nevertheless, combining a denoising auto encoder with a main network can create more problems for that network [9]. A combination of gradient regularization and DNNs has been used in [30] to robustify the deep learning models against adversarial attacks. Although the adversarial accuracy increases in this approach, the clean accuracy of the model diminishes, which is not desirable.

Some researchers have tried to devise new attacks in other tasks as well and to robustify the DNNs against them. For example, the “spare aware online incremental attack” technique has been employed in [31] to create online attacks; which can pose a serious challenge to object tracking efforts. Various defense strategies in the field of semantic segmentation have been explored in [32]. This paper has revealed that the methods used in the classification tasks cannot be applied effectively to network robustification in the semantic segmentation tasks.

Most of the efforts undertaken to robustify the object detection models are based on adversarial training; and less attention has been paid to making changes to these models and their backbones [34]. Considering the similarity in the backbones of object detection models and famous classification architectures, it seems that by relying on the research efforts related to classification and by making changes to network architectures and improving their robustness, some techniques for robustifying the object detectors could be devised.

A multitask method for model training as well as various techniques for the adversarial training of models have been used in [33]. The model achieved in this work has been tested on the images perturbed by the DAG and RAP attacks.

3 The proposed method

The method proposed in this paper exploits the Gabor filter banks in the first layer of famous object detectors. As we know, the backbones of the well-known detectors (e.g., the YOLO and the SSD) are based on famous architectures; and we can generate the convolutional Gabor layers by combining the starting filters of such detectors with the Gabor filter banks [16]. The Gabor filters used to be very common in traditional machine vision applications, and they were placed at the start of machine vision systems in order to detect the edges and curves [35]. These filters were used for the first time in classification tasks in [16] and yielded very promising results. In our proposed approach, we attempt to match the Gabor filter banks with the backbone structure of the object detector models and then to replace the ordinary convolutional layers in these backbones with the convolutional Gabor layers.

It is a known fact that the Gabor filters are able to extract the spatial features of images quite successfully [17]. Hence, it is assumed that the extraction of these spatial features could make the object detection systems more robust. We will prove this hypothesis in the next section by means of several experimental results. Subsequently, we will explore the Gabor filter equations and the matching of these filters with the backbone structure of object detectors. Our method of generating the filters and using them in the first layer of the DNNs has been illustrated graphically in Fig. 1.

$$G_{\theta}(x', y'; \alpha, \beta, \delta, \eta) = e^{-\alpha^2(x'^2+\beta y'^2)} \cos(\delta x' + \eta) \tag{1}$$

The Gabor filter is a complex sinusoidal function in the form of Eq. 1.

In the above formula, x' and y' are defined as

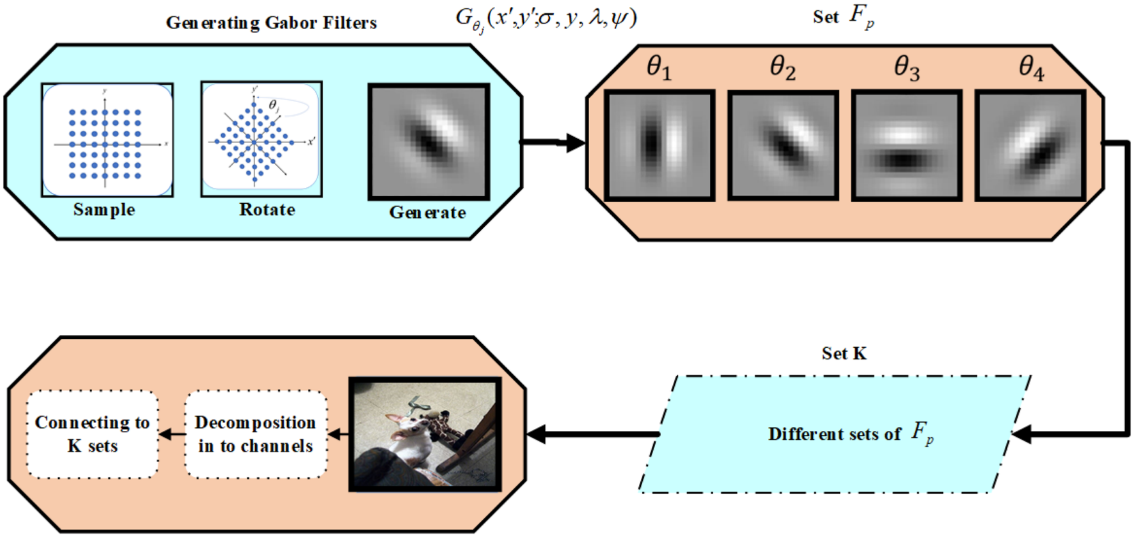


Fig. 1 The algorithm proposed in this paper

$$x' = x\cos\theta - y\sin\theta \tag{2}$$

$$y' = x\sin\theta + y\cos\theta \tag{3}$$

In order to make a discrete Gabor filter (Fig. 1), we uniformly resolve the x and y parameters in our DNN model. The filter size is determined according to the number of network samples. To construct the filter in the $\{(x_i,y_i)\}_{i=1}^{k^2}$ network with dimensions $k \times k$, these parameters are inserted into the network as a set of trainable parameters and are trained via a conventional learning method. Here, the $\{\alpha, \beta, \delta, \eta\}$ is considered as the set of trainable parameters. According to [17], the learnable parameters of the Gabor layer are trained exactly like the vector coefficients and weights of a network. Like any other learnable parameter, these parameters are trained, within a specific range, during the learning process. The effect of these parameters on the final accuracy is exactly like the influence of network weights; i.e., these parameter values are updated in each succeeding epoch so as to raise the final adversarial accuracy. These parameters are trained at different rotation angles (θ), and according to Fig. 1, they form a set of Gabor filters which are eventually applied to the input images. Equation 4 shows the procedure for constructing the F_p. The value of θ indicates the filter rotation angle. And since a Gabor filter detects features such as image edges in the direction of its theta angle, a Gabor filter bank must include different values of theta (from 0 to 2π) in order to cover various rotation angles. Therefore, the filter bank used in our paper contains a large number of Gabor filters with different θ values so that the edges and the low-level features of images can be detected at different rotation angles.

$$F_p = \{G_{\theta_1}, G_{\theta_2}, G_{\theta 3}, \ldots.., G_{\theta_n}\} \tag{4}$$

Using this equation, different filters can be made with various θ angles in the range of $[0, 2\pi]$. The K set is eventually constructed by producing several F sets for different p values.

In the classical machine vision, the frequency and the rotation angle of the Gabor filters are, respectively, obtained from Eqs. 5 and 6.

$$\omega_n = \frac{\pi}{2}\sqrt{2}^{-(n-1)} \tag{5}$$

$$\theta_m = \frac{\pi}{8}(m-1) \tag{6}$$

These equations can also be used here to obtain the F set.

After completing the steps shown in Fig. 1 and producing the set of filters, the convolutional Gabor layer is finally obtained. Now, we can add this layer to the first layer of the backbone of object detectors. In this method, the activation function of ReLU has been used in the convolutional Gabor layer so that the output of this layer can be connected to the next layer.

As is shown in Fig. 2, in this approach, an image is first divided into its constituent RGB channels. These channels are then fed to a Gabor filter bank as a tensor. As the input layer of the detection system, the Gabor filter bank extracts the image's low-level features. Based on the explained technique and according to Fig. 1, each filter in the filter bank is constructed with a specific theta angle ($0 \le \theta \le 2\pi$) and it can extract the edges and the other low-level features of images corresponding to this theta angle.

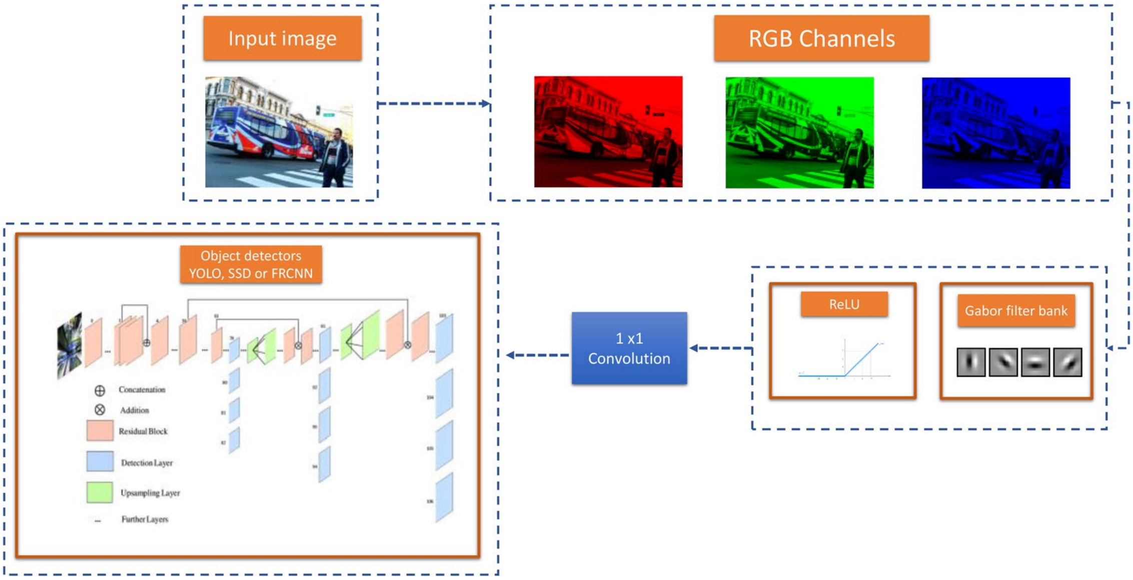


Fig. 2 The block diagram of the proposed method

In order to cover different theta angles, our filter bank includes various filters with different rotation angles. Every channel of an input image is convoluted with every existing filter in the filter bank. After applying this filter bank, it would be necessary to prepare the output tensor to be fed into the main section of the detector. The object detector will be selected after matching the input channels with the output of the Gabor layer. In this work we have used the YOLOv3-m object detector with the MobileNet backbone, the YOLOv3-d with the Darknet backbone, and the FRCNN. The reason for choosing these models is to evaluate the presented method on the models with different architectures and backbones. After exiting from the convolutional Gabor filters, the tensor is fed to an activation function in order to prepare the output tensor for input into the 1×1 convolution block. After applying the convolution procedure, the tensor obtains the number channels that have to be fed to the main part of the detector model.

4 Experimental results

In this research, all the networks have been trained, under similar conditions, in 70 epochs. After evaluating the number of epochs in the training process, we found out that the adversarial accuracy does not increase significantly and has little fluctuation after epoch 70. So it was decided that in this work and for the model considered, it is sufficient to use 70 epochs for training purposes. As an example, the average adversarial accuracy of the YOLOv3-d model for the TOG-mislabeling type of attack and on the PASCAL VOC dataset has been reported in Fig. 3. By examining this figure, we can see that no significant improvement has occurred in the accuracies following epoch 70; therefore, the accuracy obtained at this epoch number can be trusted.

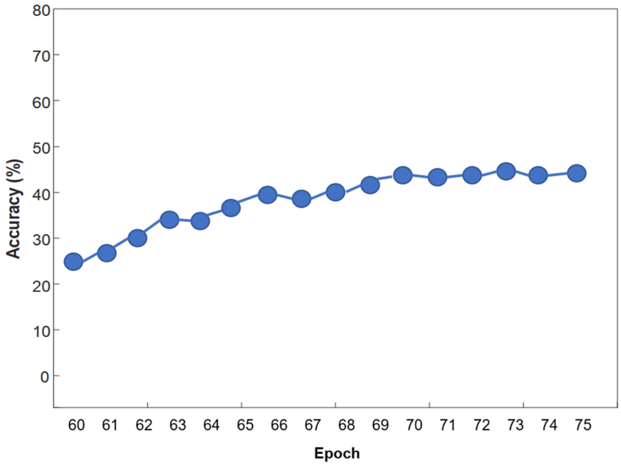


Fig. 3 The average adversarial accuracy of the YOLOv3-d model for the TOG-mislabeling type of attack and for the PASCAL VOC dataset

Also, the batch size has been considered as 64. Different batch sizes were also evaluated in this paper and based on the algorithm used and the existing hardware, the best accuracy and the most suitable computation speed were achieved at the batch size of 64. For example, using a batch size of 128, the average adversarial accuracies are reduced by about 2% at the same number of epochs.

The networks and the learnable parameters of the Gabor layer have been trained by the stochastic gradient descent approach. To evaluate the performance of the method

Table 1 The models used in this paper and their specifications

Object detector	Backbone	Input image size
YOLOv3-m	MobileNetV1	448×448
YOLOv3-d	Darknet53	448×448
SSD300	VGG16	300×300
SSD512	VGG16	512×512
FRCNN	VGG16	1000×600

presented in the preceding section, it has been applied on 5 well-known object detectors: YOLOv3-m, YOLOv3-d, SSD300, SSD512, and the Faster R-CNN (henceforth called FRCNN in this paper). The specifications of these detectors have been listed in Table 1.

The reason for choosing these object detector models is to evaluate the efficacy of the presented method in various models with different inputs and backbones. The datasets of MSCOCO (v. 2017) and PASCAL VOC (v. 2012) have been used to test the robust object detectors obtained. The MSCOCO dataset is one of the most famous datasets in the field of object detection. This dataset includes 4000 images for model training, 5000 images for validation, and 5000 images for testing. All the images of this dataset have been used in this paper. The images in this dataset cover 80 classes of objects.

The PASCAL VOC dataset includes 20 object classes. This dataset consists of 1464 images for training as well as 1464 images for validation and testing. In evaluating our proposed method, we have used all the images of this dataset.

For perturbing the database image, different techniques have been proposed in recent years. In this research, we have employed six of the most famous attacks that exist in the field of object detection. The adversarial attacks must perturb the images in such a way that these perturbations are not recognizable by human eye. The targeted attacks used in this paper comprise the TOG-fabrication, TOG-vanishing, and the TOG-mislabeling attacks and the untargeted attacks are the DAG, RAP, and UEA. A sample image perturbed by the targeted TOG-mislabeling attack has been illustrated in Fig. 4.

As is observed in Fig. 4, the perturbed image is not recognizable by human eye, but it has been able to mislead the object detector and cause it to miss the considered object in the output image. In Fig. 5, the same image has been perturbed via the untargeted random UEA attack in magnified form so that it is visible to human eye. This elevated degree of image perturbation is so that it can be recognized by human eye; otherwise, a much lower perturbation level can completely fool the detection system. Each of the targeted attacks has been designed to mislead the detection network in a particular way. Figures 6 and 7 respectively show the performances of these targeted and untargeted attacks and their effects on the recognition ability of object detectors.

The efficacy of various attacks can also be evaluated by means of two parameters: the false negative increase (FNI) and the mean square error (MSE). The FNI parameter indicates the ratio of the false negative detections of objects (ΔN) by the system to the total number of positive detections (N). This parameter is defined as follows [21]:

$$FNI = \frac{\Delta N}{N + 1} \tag{7}$$

These parameters are not usually reported by the attack developers for their devised attacks. However, to shed more light on the performance and effectiveness of these attacks, the FNI and MSE parameters have been calculated for the attacks analyzed in this research and the results have been tabulated in Table 2.

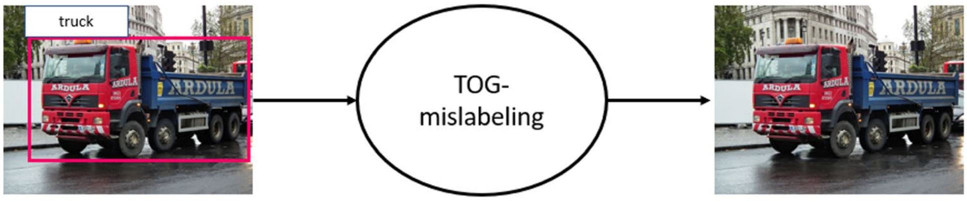


Fig. 4 A sample image perturbed by the TOG-mislabeling attack

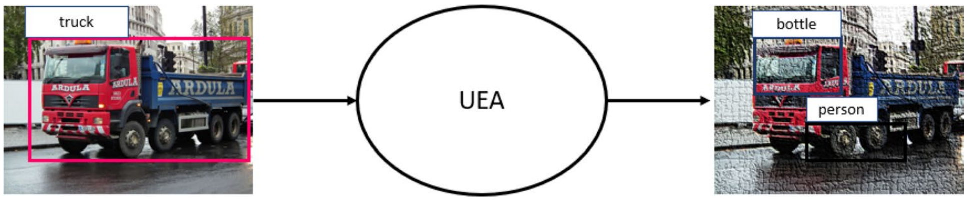


Fig. 5 A sample image perturbed by the UEA attack in magnified form

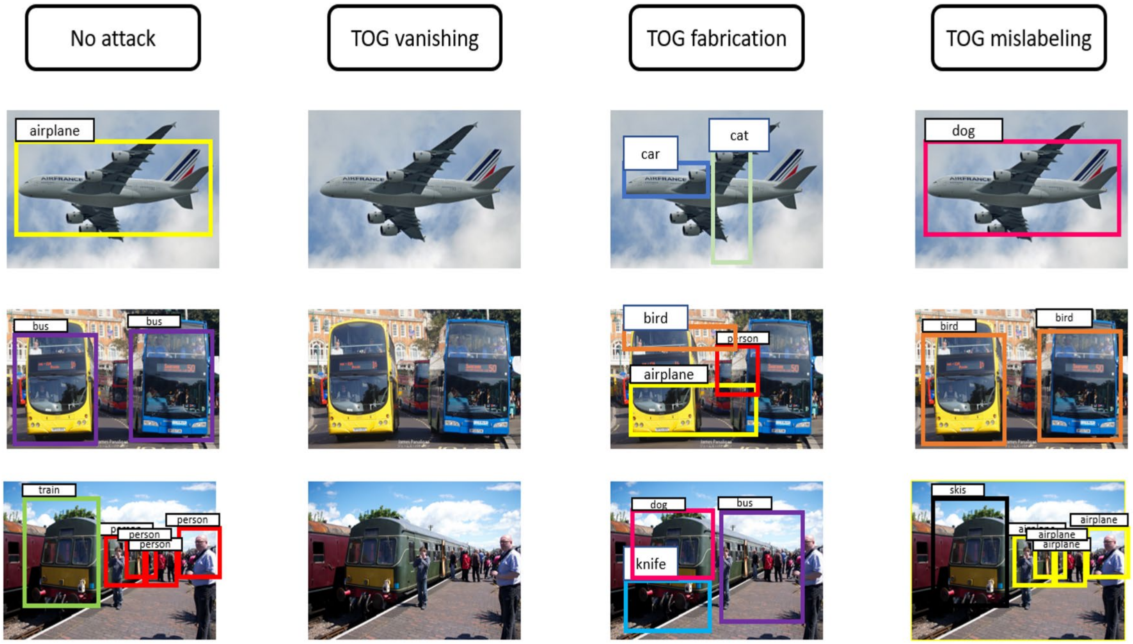


Fig. 6 The effects of targeted attacks on the recognition performance of the object detector

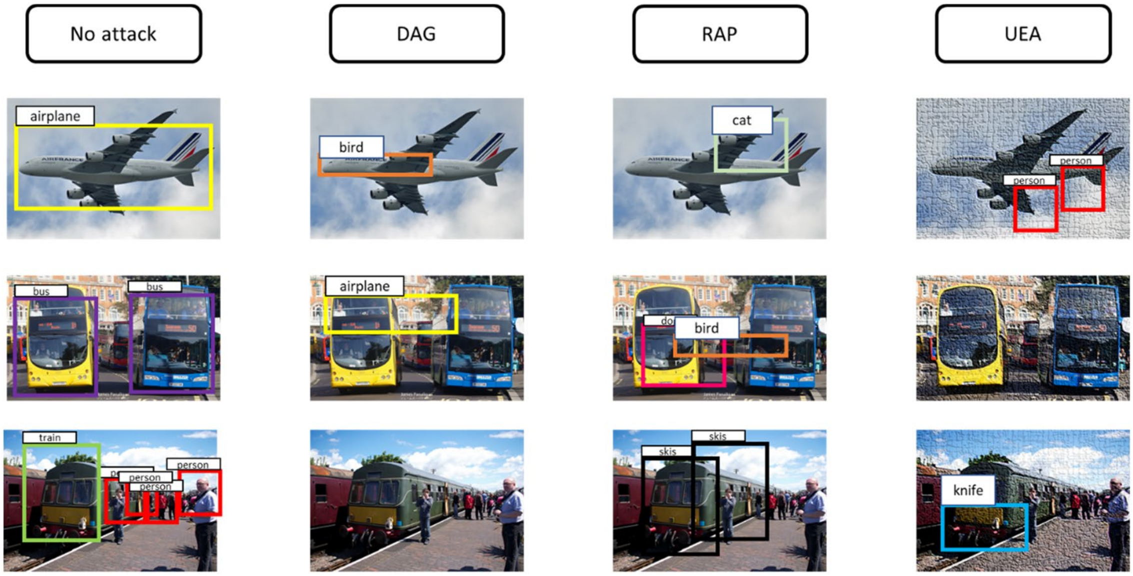


Fig. 7 The effects of untargeted attacks on the recognition performance of the object detector (The UEA attack has been magnified to make it visible.)

As is observed, the TOG series of attacks mislead the object detectors more effectively, and it is harder to formulate a defense strategy against them. To actually test the presented algorithm, the existing models are robustified first, according to the procedures given in Sect. 2. Next, by implementing the introduced attacks, the input images are perturbed independently. Then, by employing two GPUs with specifications (NVIDIA GEFORCE 1080 TI and NIVIDIA GEFORCE 2060 SUPER) and using the perturbed images obtained, each of the networks is subjected to adversarial training. Adversarial training means training a model with perturbed images. It should be pointed out that every network in this paper has been trained and tested by each of the attacks considered. In this research, the training data of each dataset have been used to train the networks, the validation data have been used to evaluate the networks during

Table 2 The FNI and MSE values for the attacks used in this paper

Attack		YOLOv3-m	YOLOv3-d	SSD300	SSD512	FRCNN
MSE	TOG-V	5.21E-03	3.25E-03	3.16E-03	3.74–03	8.31E-03
FNI		94.12%	94.17%	97.12%	95.09%	92.19%
MSE	TOG-F	3.17E-03	3.14E-03	3.52E03	6.14E-03	5.12E-03
FNI		97.69%	94.19%	98.17%	95.52%	98.82%
MSE	TOG-M	3.32E-03	3.97E-03	4.76E-03	4.12E-03	6.19E-03
FNI		98.61%	99.88%	97.34%	97.94%	96.18%
MSE	DAG	5.92E-03	5.88E-03	6.84E-03	7.97E-03	8.20E-03
FNI		84.12%	87.90%	86.13%	85.12%	83.18%
MSE	RAP	6.99E-03	7.25E-03	9.19E-03	7.29E-03	9.95E-03
FNI		91.56%	92.19%	89.19%	90.32%	87.62%
MSE	UEA	5.92E-03	6.17E-03	8.12E-03	8.82E-03	5.96E-03
FNI		81.26%	84.61%	83.25%	85.52%	79.63%
MSE	EBG [21]	3.61E-03	4.12E-03	3.96E-03	3.78E-03	8.65E-03
FNI		99.15%	93.78%	97.36%	97.89%	97.83%

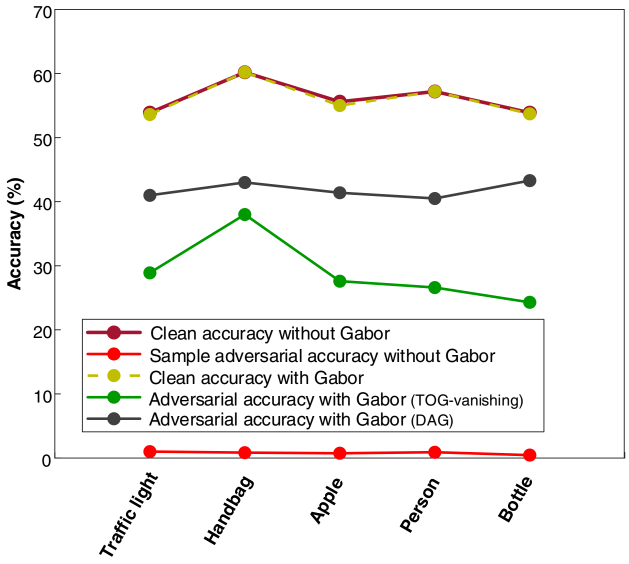


Fig. 8 Comparison of clean and adversarial accuracy for 5 classes of MSCOCO datasets in the presence of a targeted attack and an untargeted attack for the YOLOv3-m model

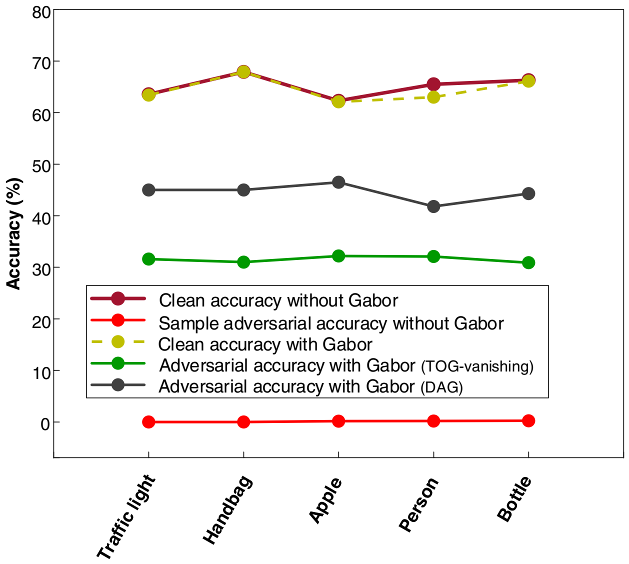


Fig. 9 Comparison of clean and adversarial accuracy for 5 classes of MSCOCO datasets in the presence of a targeted attack and an untargeted attack for the YOLOv3-d model

the training process, and the test data have been used for the final evaluation of the detection models.

Now, to examine the performance of the introduced method more closely, the accuracy obtained for each class has been computed. The graphs in Figs. 8, 9, 10, 11, and 12 illustrate some examples of this evaluation. These diagrams show the clean accuracy of the models in the absence of any defense, the adversarial accuracy of the models subjected to an arbitrary random attack in the absence of a defense, and the accuracies of the models following their robustification via clean data and the data perturbed by adversarial attacks. The results of various classes are illustrated in these graphs for the TOG-vanishing and the DAG attacks, as examples of targeted and untargeted attacks, respectively. The accuracies obtained for each class can provide valuable information, which can be used to evaluate the effectiveness of the presented method for each class of the dataset. Figures 8, 9, 10, 11, and 12show the results of this analysis on the MSCOCO dataset. Due to the large number of classes in this dataset, sample results have been presented in this paper for just 5 of these classes.

The results obtained by applying the algorithms in this paper on the datasets of PASCAL VOC and MSCOCO have been analyzed and compared with the results of other works in Tables 3 and 4, respectively. An important point to consider when trying to robustify the DNNs against.

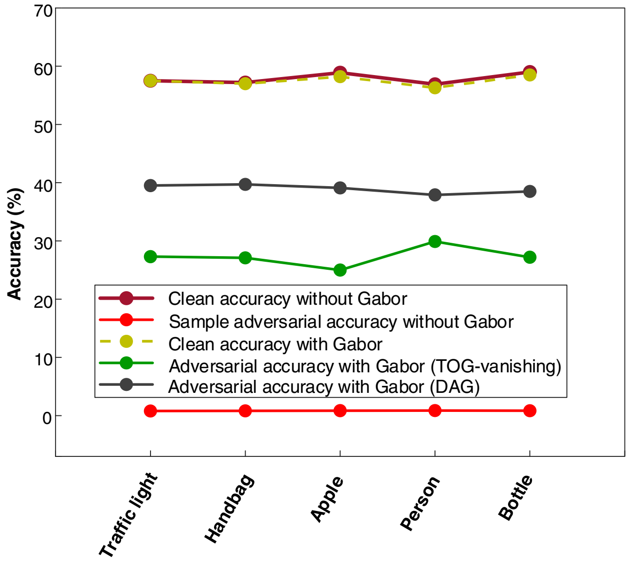


Fig. 10 Comparison of clean and adversarial accuracy for 5 classes of MSCOCO datasets in the presence of a targeted attack and an untargeted attack for the SSD300 model

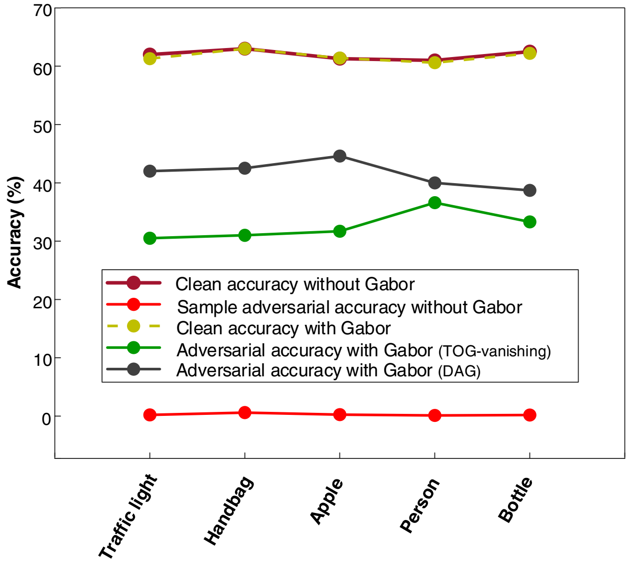


Fig. 11 Comparison of clean and adversarial accuracy for 5 classes of MSCOCO datasets in the presence of a targeted attack and an untargeted attack for the SSD512 model

adversarial attacks is to make sure that the clean accuracy of these networks does not drop significantly relative to the accuracy of their undefended state. By examining Tables 3 and 4, it is observed that in all the states and models, the clean accuracy drop, in our method, relative to the undefended state is negligible and much lower than that in former works. The accuracy results for the defended states of networks and for different attacks on each of the mentioned datasets have been plotted in Figs. 13, 14, and 15. By inspecting these figures, we can see that the adversarial accuracies of the models presented in this paper are better than those of the previous works and substantially improved against all the considered attacks.

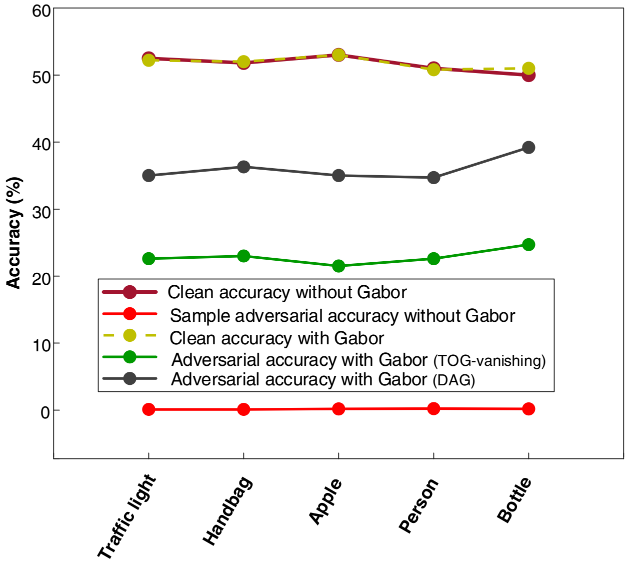


Fig. 12 Comparison of clean and adversarial accuracy for 5 classes of MSCOCO datasets in the presence of a targeted attack and an untargeted attack for the FRCNN model

Also, by comparing the graphs in Fig. 15, it is confirmed that the combined model of “YOLOv3-d + Gabor” has the best performance in both the PASCAL VOC and MSCOCO datasets. A closer examination of Tables 3 and 4 shows that the considered models perform much better in the PASCAL VOC dataset than in the MSCOCO. This superiority is due to the higher clean accuracy achieved by the models of this paper in their undefended state in the PASCAL VOC dataset than in the MSCOCO dataset, which can be attributed to the simpler data contained in the PASCAL VOC dataset and the smaller number of classes that exist in this dataset. Also, the accuracies of all the models in their undefended state is very low and less than 2% on the average, which shows the serious vulnerability of all the existing models against adversarial attacks. Of course, by analyzing the results, one can see that the model accuracies.

are reduced much more by the newer targeted attacks. This is due to the more complex and precise design of these attacks compared to the older untargeted attacks.

Moreover, the information obtained from Tables 3 and 4 clearly shows that the former defense strategies perform much better against the untargeted attacks than the targeted ones.

Table 3 The clean and the adversarial accuracies obtained by different models (with/without a defense) by considering all the attacks in the PASCAL VOC dataset

State	Attack	YOLOv3-m (%)	YOLOv3-d (%)	SSD300 (%)	SSD512 (%)	FRCNN (%)	SSD + VGG16 [36] (%)	SSD + RESNET [36] (%)
Clean accuracy without defense	-	71.84	83.43	76.11	79.83	67.37	71.4	69.5
Clean accuracy with defense	-	71.47	83.39	76	79.23	67.2	64.5	61.8
Adversarial accuracy without defense	TOG-V	0.85	0.56	0.43	0.39	0.21	0.11	0.19
	TOG-F	5.52	2.12	2.18	2	1.76	1.1	1.23
	TOG-M	1.5	1.9	1.4	1.2	1.9	0.9	0.85
	DAG	2.2	1.8	2	2.3	2	0.3	0.4
	RAP	2.9	1.9	1.5	1.3	1.1	5.4	5.5
	UEA	1.58	1	0.56	1.2	1.73	2.3	3
	Average	2.45	1.54	1.34	1.39	1.45	1.68	1.69
Adversarial accuracy with defense	TOG-V	43.2	45.9	38.7	41.1	33.2	17.3	21
	TOG-F	33.5	47.1	39.9	42.4	35.1	16.5	20.5
	TOG-M	21	41.5	45	47.7	30.2	15.5	18.3
	DAG	51.1	57	48.8	49.9	38.9	28.5	22.9
	RAP	48.8	54.3	43.3	52	41.1	44.9	39.1
	UEA	52.9	52.2	47.9	51.6	39.7	29.8	27.6
	Average	41.75	49.6	43.9	47.45	36.36	25.4	24.9

Table 4 The clean and the adversarial accuracies obtained by different models (with/without a defense) by considering all the attacks in the MSCOCO dataset

State	Attack	YOLOv3-m (%)	YOLOv3-d (%)	SSD300 (%)	SSD512 (%)	FRCNN (%)	SSD + VGG16 [36] (%)	SSD + RESNET [36] (%)
Clean accuracy without defense	-	58.2	65.5	53.2	59.1	49	55.6	59.8
Clean accuracy with defense	-	57.9	65	53.06	58.5	48.8	50.1	49.9
Adversarial accuracy without defense	TOG-V	0.33	1.2	0.77	1.5	0.11	0.14	0.19
	TOG-F	1.8	1.3	5.2	0.43	1.1	0.21	1.1
	TOG-M	2.2	2.8	0.55	0.21	1.1	0.25	0.15
	DAG	1.3	3	0.83	0.19	0.58	1	1.8
	RAP	1.1	3.26	1.9	0.3	1	1.2	1
	UEA	0.59	1.95	0.95	0.11	0.11	1.5	0.52
	Average	1.22	2.25	1.7	0.45	0.66	0.71	0.79
Adversarial accuracy with defense	TOG-V	30	34.4	27	31	23	11.6	8.5
	TOG-F	31.2	36.1	28.1	34.2	21.5	8.95	7.45
	TOG-M	25	30.5	23.3	39.5	26.6	10.2	10.66
	DAG	40.5	45.2	37	37.8	34.6	18	16.6
	RAP	48.8	51.2	33.1	34.6	32.2	20.5	18.6
	UEA	43.3	47.5	34.2	44.1	30	29.8	27.6
	Average	36.46	40.81	30.45	36.86	27.98	16.5	14.9

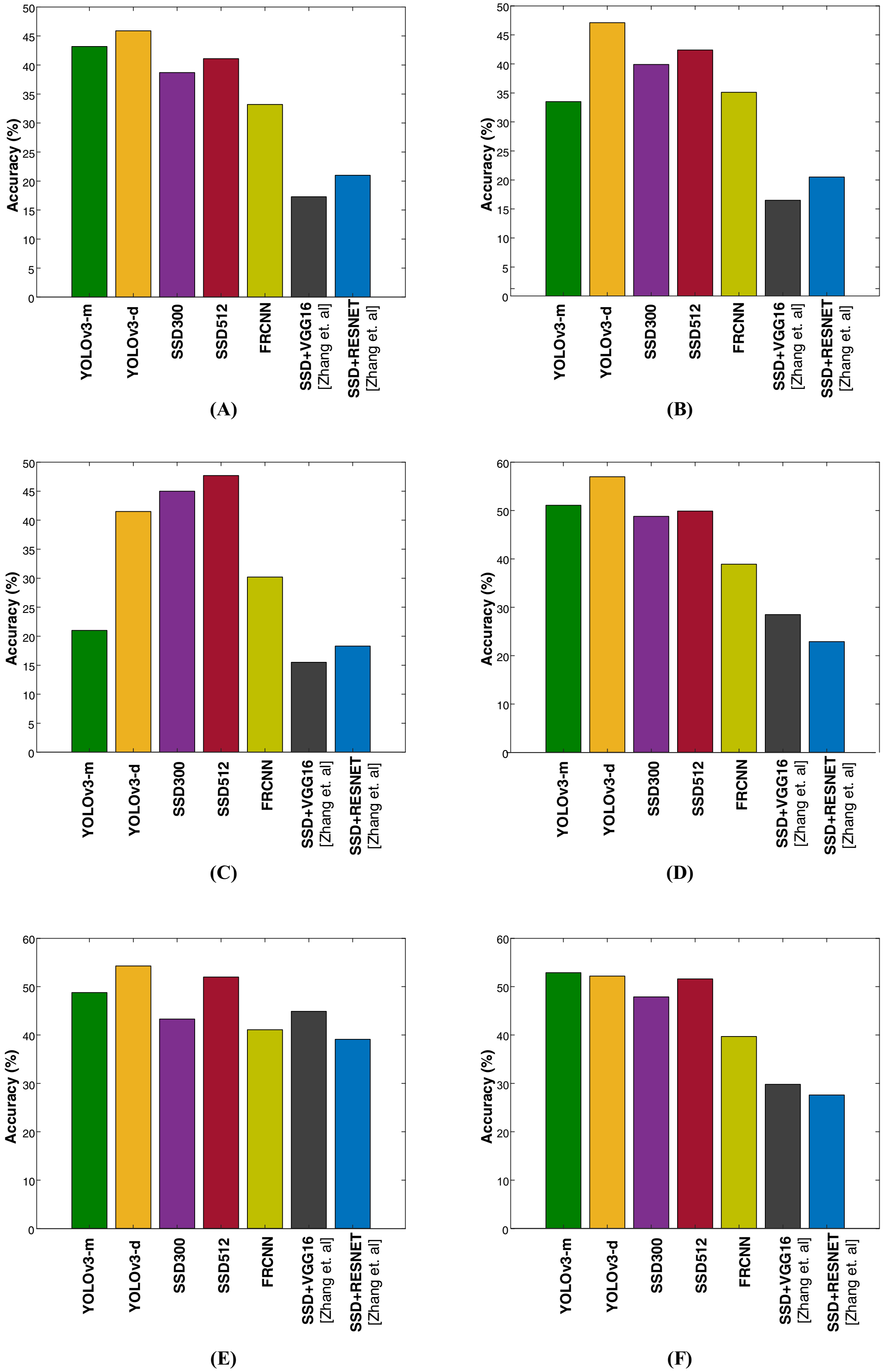
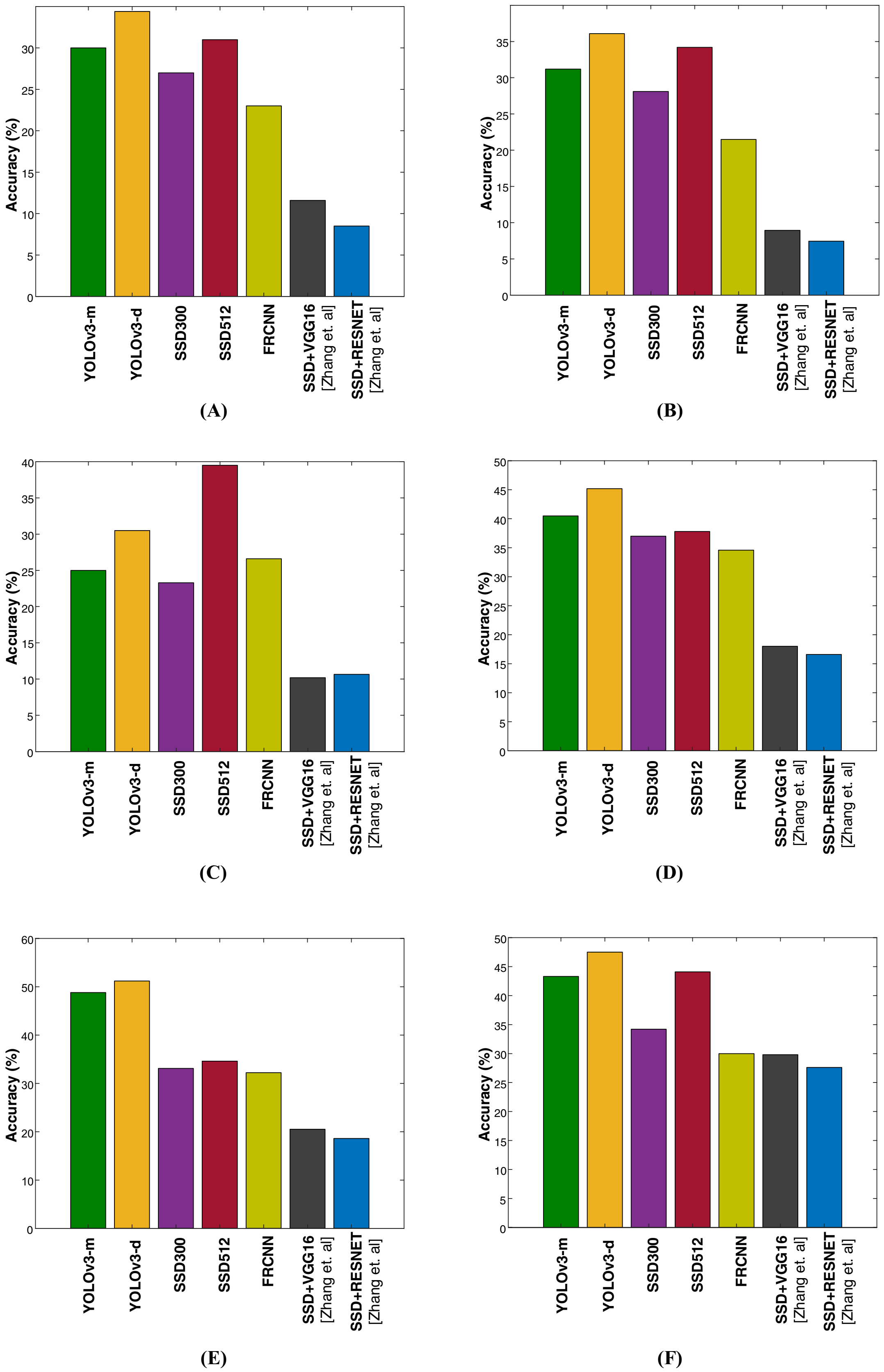


Fig. 13 Comparing the performances of different models against the A) TOG-vanishing, B) TOG-fabrication, C) TOG-mislabeling, D) DAG, E) RAP and F) UEA in the PASCAL VOC dataset



(A)

(B)

(C)

(D)

(E)

(F)

Fig. 14 Comparing the performances of different models against the A) TOG-vanishing, B) TOG-fabrication, C) TOG-mislabeling, D) DAG, E) RAP and F) UEA in the MSCOCO dataset

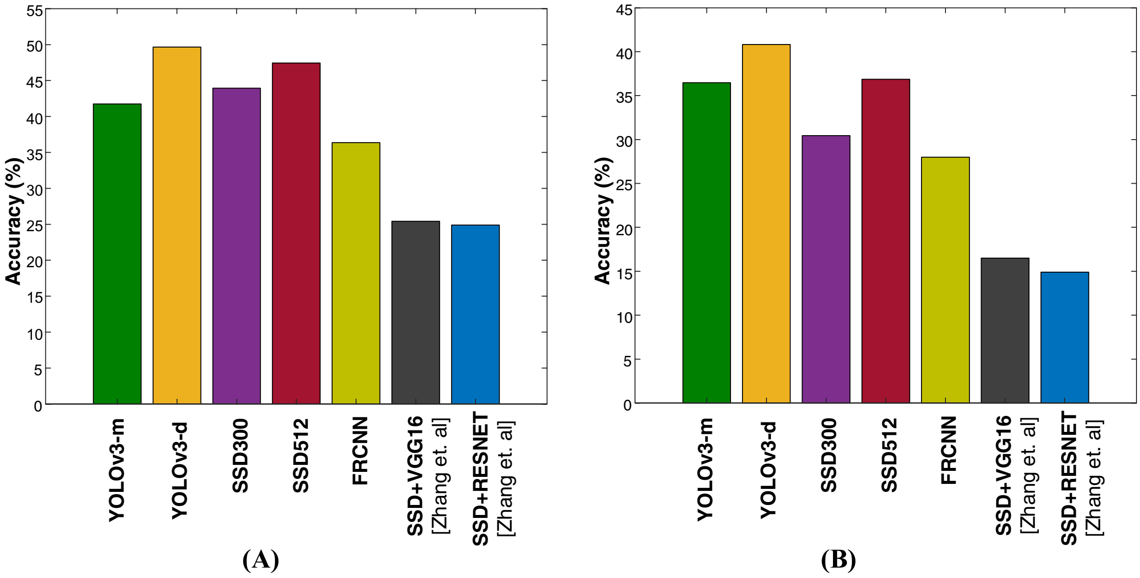


Fig. 15 Comparing the average performances of different models against adversarial attacks in the A) PASCAL VOC dataset and B) MSCOCO dataset

The closest work to our research in terms of the applied conditions and the datasets used is the work of Zheng et al. [36]. We had already compared these two works in the context of graphs and tables. However, for comparing the proposed algorithm with similar works, we tried to simulate, once again, some of the existing algorithms for the conditions close to the test conditions of that paper. A method based on pre-training has been introduced in [37] for face recognition applications. In this approach, the image features are extracted first and the dimensions are reduced. We applied the method presented in [37] to the models used in this paper, called the technique "SDF" for short, and compared it with the results of our work in Table 5. Another robustification technique based on the detection and mitigation of adversarial attacks has been proposed by Goswami et al. [38] for boosting the system robustness in face recognition tasks. We also evaluated this method by using our datasets and tabulated the results.

of these comparisons in Table 5. This table clearly shows that our proposed method has been able to improve the adversarial accuracy better than the other approaches.

For a better assessment of the method proposed, the introduced models are evaluated once again by using a new attack strategy that combines the Evaporate, Boundary, and the Gaussian Noise Attacks (the new hybrid attack was abbreviated as the EBG attack and was fully described in the

Table 5 Comparing the presented methods with the similar adapted approaches

Defense		Best adversarial accuracy with TOG-v	Best adversarial accuracy with TOG-f	Best adversarial accuracy with TOG-m
YOLO v3-m + Gabor (ours)	PASCAL	43.2%	33.5%	21%
	COCO	30%	31.2%	25%
YOLO v3-d + Gabor (ours)	PASCAL	45.9%	47.1%	41.5%
	COCO	34.4%	36.1%	30.5%
FRCNN + Gabor (ours)	PASCAL	33.2%	35.1%	30.2%
	COCO	23%	21.5%	26.6%
SDF [37]	PASCAL	11.2%	12.8%	8.8%
	COCO	8.2%	7.14%	6.2%
Goswami et al. [38]	PASCAL	14.3%	16.7%	11.3%
	COCO	11.8%	5.5%	5.9%
SSD + RESNET + defense [36]	PASCAL	21%	20.5%	18.3%
	COCO	8.5%	7.45%	10.6%

Table 6 The performances of the presented models against the EBG attack on the PASCAL VOC dataset

FRCNN (%)	SSD512 (%)	SSD300 (%)	YOLOv3-d (%)	YOLOv3-m (%)	State
67.37	79.83	76.11	83.43	71.48	Clean accuracy without defense
67.20	79.23	76	83.39	71.47	Clean accuracy with defense
7.11	9.85	9.35	10.25	8.61	Adversarial accuracy without EBG defense
31.33	39.87	35.81	46.17	44.21	Adversarial accuracy with EBG defense

Table 7 The performances of the presented models against the EBG attack on the MSCOCO dataset

State	YOLOv3-m (%)	YOLOv3-d (%)	SSD300 (%)	SSD512 (%)	FRCNN (%)
Clean accuracy without defense	71.48	83.43	76.11	79.83	67.37
Clean accuracy with defense	71.47	83.39	76	79.23	67.20
Adversarial accuracy without EBG defense	8.61	10.25	9.35	9.85	7.11
Adversarial accuracy with EBG defense	44.21	46.17	35.81	39.87	31.33

Introduction). The exact results of this evaluation for the datasets of PASCAL VOC and MSCOCO have been listed in Tables 6 and 7, respectively. This attack has a high FNI value and it can also efficiently measure our defense performance against the hybrid black-box types of attack. By examining Tables 6 and 7, it is realized that not only the presented defense strategy can adequately deal with older attacks as well as the targeted TOG attacks, but it also can perform effectively against the new types of hybrid black-box attacks. This shows the credibility of the mentioned defense strategy in dealing with various types of adversarial attacks under different conditions.

5 Conclusion

Using the Gabor filter banks in different model backbones, a new method was introduced in this paper for robustifying the visual object detectors. This approach was applied on five models and the obtained results were tested by means of the PASCAL VOC and the MSCOCO datasets. In this study, the input images were perturbed by three types of targeted and three types of untargeted attacks and the results were reported for all the considered states. Here, six models that are robust to adversarial attacks and suitable for object detection applications have been proposed and their results for different states have been compared. Based on the findings of this research, the introduced models perform well against the adversarial attacks, and the best performance among these models belongs to the robust YOLOv3 model with the DARKNET backbone and convolutional layers.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

1. Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., Shi, J.: FoveaBox: Beyound anchor-based object detection. IEEE Trans. Image Process. **29**, 7389–7398 (2020)
2. Wu, F., Jin, G., Gao, M., He, Z. and Yang, Y.: "Helmet detection based on improved YOLO V3 deep Model," *IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, Canada, pp. 363–368, 2019.
3. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C.: "Ssd: Single shot multibox detector, " *European Conference on Computer Vision (ECCV)*, 2016.
4. Liu, Z., Xiang, Q., Tang, J., Wang, Y., Zhao, P.: Robust salient object detection for RGB images. Vis. Comput. **36**, 1823–1835 (2020)
5. Naseer, M., Khan, S. and Porikli, F.: "Local gradients smoothing: Defense against localized adversarial attacks," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, USA, pp. 1300–1307, 2019.
6. Ramanathan, A., Pullum, L., Husein, Z., Raj, S., Torosdagli, N., Pattanaik, S. and Jha, S. K.: "Adversarial attacks on computer vision algorithms using natural perturbations," *2017 Tenth International Conference on Contemporary Computing (IC3),* Noida, 2017, pp. 1–6.
7. Chow, K.-H., Liu, L., Gursoy, M. E., Truex, S., Wei, W., and Wu, Y.: "Understanding object detection through an adversarial lens," *Computer Security–ESORICS 2020 Lecture Notes in Computer Science*, pp. 460–481, 2020.
8. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access **6**, 14410–14430 (2018)
9. Li, H., Li, G., Yu, Y.: ROSA: Robust salient object detection against adversarial attacks. IEEE Trans. Cybern. **50**(11), 4835–4847 (2020)
10. Kamboj, A., Rani, R., and Nigam, A.: "A comprehensive survey and deep learning-based approach for human recognition using ear biometric," *The Visual Computer*, 2021, https://doi.org/10.1007/s00371-021-02119-0.
11. Yadav, K. and Singh, A.: "Comparative analysis of visual recognition capabilities of CNN architecture enhanced with Gabor filter," *International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2020, pp. 45–50,.
12. Cho, S., Jun, T. J., Oh, B. and Kim, D.: "DAPAS : Denoising autoencoder to prevent adversarial attack in Semantic Segmentation," *International Joint Conference on Neural Networks (IJCNN)*, Glasgow, United Kingdom, 2020, pp. 1–8.
13. Yahya, Z., Hassan, M., Younis, S., Shafique, M.: Probabilistic analysis of targeted attacks using transform-domain adversarial examples. IEEE Access **8**, 33855–33869 (2020)
14. Chow, K.H., Liu, L., Loper, M., Bae, J., Gursoy, M.E., Truex, S., Wei, W. and Wu, Y: Adversarial objectness gradient attacks in real-time object detection systems. 2020 [Online]. Available: https://khchow.com/media/TPS20_TOG.pdf
15. Naghdy, G., Ros, M., Todd, C. and Norahmawati, E.: "Cervical cancer classification using Gabor filters," *IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*, San Jose, CA, 2011, pp. 48–52.
16. Pérez, J. C., Alfarra, M., Jeanneret, G., Bibi, A., Thabet, A., Ghanem, B. and Arbeláez, P.:"Gabor layers enhance network robustness," *Computer Vision – ECCV 2020 Lecture Notes in Computer Science*, pp. 450–466, 2020.
17. Alekseev, A. and Bobe, A.: "GaborNet: Gabor filters with learnable parameters in deep convolutional neural network," *International Conference on Engineering and Telecommunication (EnT)*, Dolgoprudny, Russia, 2019, pp. 1–4.
18. Bansal, A., Ranjan, R., Castillo, C. D. and Chellappa, R.: "Deep features for recognizing disguised faces in the wild", *Computer Vision and Pattern Recognition Workshops (CVPRW) IEEE/CVF Conference on*, pp. 10–106, 2018.
19. Miyato, T., Maeda, S., Koyama, M. and Ishii, S.: "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2019.
20. Wang, Y., Tan, Y., Zhang, W., Zhao, Y. and Kuang, X.: "An adversarial attack on DNN-based black-box object detectors," *Journal of Network and Computer Applications*, vol. 161, 2020.
21. Lee, M. and Kolter, Z.: "On physical adversarial patches for object detection", 2019, [online] Available: https://arxiv.org/abs/1906.11897.
22. Li, D., Zhang, J. and Huang, K.: "Universal adversarial perturbations against object detection", *Pattern Recognition*, vol. 110, 2021.
23. Wang, Y., Lv, H., Kuang, X., Zhao, G., Tan, Y., Zhang, Q., Hu, J.: Towards a physical-world adversarial patch for blinding object detection models. Inf. Sci. **556**, 459–471 (2021)
24. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L. and Yuille, A.: “Adversarial examples for semantic segmentation and object detection,” *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
25. Li, Y., Tian, D., Bian, X., Lyu, S.: "Robust adversarial perturbation on deep proposal-based models", *British Machine Vision Conference (BMVC)*, 2018.
26. Wei, X., Liang, S., Chen, N. and Cao, X.: "Transferable adversarial attacks for image and video object detection", *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 954–960, 2019.
27. Aprilpyone, M., Kinoshita, Y., Kiya, H.: Adversarial robustness by one Bit double quantization for visual classification. IEEE Access **7**, 177932–177943 (2019)
28. Carlini, N. and Wagner, D.: "Towards evaluating the robustness of neural networks", *Proc. IEEE Symp. Secur. Privacy (SP),* pp. 39–57, May 2017.
29. Moosavi-Dezfooli, S., Fawzi, A. and Frossard, P.: "DeepFool: A simple and accurate method to fool deep neural networks", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2574–2582, Jun. 2016.
30. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B. and Swami, A.: "The limitations of deep learning in sdversarial settings," *IEEE European Symposium on Security and Privacy (EuroS&P)*, Germany, 2016, pp. 372–387.
31. Ross, A. and Doshi-Velez, F.: “Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients”, *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
32. Guo, Q., Xie, X., Ma, L., Li, Z., Xue, W., Feng, W. and Liu, Y.: "SPARK: Spatial-aware online incremental attack against visual tracking," *Proc. of the European Conference on Computer Vision (ECCV)*, 2019.
33. Arnab, A., Miksik, O. and Torr, P. H. S.:"On the robustness of semantic segmentation models to adversarial attacks", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 888–897, 2018.
34. Sarwar, S. S., Panda, P. and Roy, K.: "Gabor filter assisted energy efficient fast learning convolutional neural networks," *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Taipei, 2017, pp. 1–6..
35. Song, D., Eykholt, K., Evtimov, I. and Fernandes, E.: "Physical adversarial examples for object detectors," *12th Workshop on Offensive Technologies (WOOT),* 2018.
36. Zhang, H. and Wang, J.: "Towards adversarially robust object detection", *Proc. IEEE Int. Conf. Computer Vision*, pp. 421–430, 2019.
37. Arora, S., Bhatia, M. P. S. and Mittal, V.: "A robust framework for spoofing detection in faces using deep learning, " *The Visual Computer*, 2021, https://doi.org/10.1007/s00371-021-02123-4.

38. Goswami, G., Agarwal, A., Ratha, N., Singh, R., Vatsa, M.: Detecting and mitigating adversarial perturbations for robust face recognition. Int. J. Comput. Vision **127**(6), 719–742 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abdollah Amirkhani received the MSc and PhD degrees (with honors) in electrical engineering from Iran University of Science and Technology (IUST), Tehran, in 2012 and 2017, respectively. He earned the Outstanding Student Award (2015) form the First Vice President of Iran. In 2016, he was conferred award by the Ministry of Science, Research and Technology. He is an Assistant Professor in the school of automotive engineering at IUST. He is the Associate Editor of the "Engineering Science and Technology, an International Journal". He has been actively involved in several National R&D projects, related to the development of new methodologies and learning algorithms based on AI techniques. His research interests are in machine vision, fuzzy cognitive maps, data mining and machine learning.

Mohammad Parsa Karimi received the BSc degree in Electrical and Electronic engineering from Shahid Beheshti University and he is currently a master's degree student in digital electronics at Iran University of Science and Technology. Parsa's research interest is in the field of computer vision and robustifying deep neural networks against adversarial attacks. He is currently researching in the field of robust object detectors, especially in self-driving cars application.