Computers, Materials & Continua
DOI: 10.32604/cmc.2023.032795
Article

Tech Science Press

Chained Dual-Generative Adversarial Network: A Generalized Defense Against Adversarial Attacks

Amitoj Bir Singh[1], Lalit Kumar Awasthi[1], Urvashi[1], Mohammad Shorfuzzaman[2], Abdulmajeed Alsufyani[2] and Mueen Uddin[3],*

[1]National Institute of Technology, Jalandhar, PB 144001, India
[2]Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif, 21944, Saudi Arabia
[3]School of Digital Science, University Brunei Darussalam, Jln Tungku Link, Gadong, BE1410, Brunei Darussalam
*Corresponding Author: Mueen Uddin. Email: mueenmalik9516@gmail.com

Received: 30 May 2022; Accepted: 05 July 2022

Abstract: Neural networks play a significant role in the field of image classification. When an input image is modified by adversarial attacks, the changes are imperceptible to the human eye, but it still leads to misclassification of the images. Researchers have demonstrated these attacks to make production self-driving cars misclassify Stop Road signs as 45 Miles Per Hour (MPH) road signs and a turtle being misclassified as AK47. Three primary types of defense approaches exist which can safeguard against such attacks i.e., Gradient Masking, Robust Optimization, and Adversarial Example Detection. Very few approaches use Generative Adversarial Networks (GAN) for Defense against Adversarial Attacks. In this paper, we create a new approach to defend against adversarial attacks, dubbed Chained Dual-Generative Adversarial Network (CD-GAN) that tackles the defense against adversarial attacks by minimizing the perturbations of the adversarial image using iterative oversampling and undersampling using GANs. CD-GAN is created using two GANs, i.e., CDGAN's Sub-Resolution GAN and CDGAN's Super-Resolution GAN. The first is CDGAN's Sub-Resolution GAN which takes the original resolution input image and oversamples it to generate a lower resolution neutralized image. The second is CDGAN's Super-Resolution GAN which takes the output of the CDGAN's Sub-Resolution and undersamples, it to generate the higher resolution image which removes any remaining perturbations. Chained Dual GAN is formed by chaining these two GANs together. Both of these GANs are trained independently. CDGAN's Sub-Resolution GAN is trained using higher resolution adversarial images as inputs and lower resolution neutralized images as output image examples. Hence, this GAN downscales the image while removing adversarial attack noise. CDGAN's Super-Resolution GAN is trained using lower resolution adversarial images as inputs and higher resolution neutralized images as output images. Because of this, it acts as an Upscaling GAN while removing the adversarial attak noise. Furthermore, CD-GAN has a modular design such that it can be prefixed to any existing classifier without any retraining or extra effort, and can defend any classifier model against adversarial attack. In this way, it is a Generalized Defense against adversarial attacks, capable of defending any classifier model against any attacks. This enables the user to directly integrate CD-GAN with an existing production deployed classifier smoothly. CD-GAN iteratively removes the adversarial noise using a multi-step approach in a modular approach. It performs comparably to the state of the arts with mean accuracy of 33.67 while using minimal compute resources in training.

Keywords: Adversarial attacks; GAN-based adversarial defense; image classification models; adversarial defense

1 Introduction

Image classification using Neural Networks is an important area of research that now has practical applications spanning multiple industries including Consumer Entertainment in form of filters in apps, the Automobile industry in form of Autonomous Cars, and Industrial Manufacturing with Robotics integrated Object Classification for further processing, etc. This shows how powerful Artificial Intelligence (AI), Machine Learning (ML), and Data Sciences are today, which impact domains like Data Mining [1,2], Big Data [3], Medicine [4], Internet of Things (IOT) [5], etc. Image Classification has benefited from the increased availability of computing resources and newer algorithms, which have enabled us to train massive image classification models capable of handling thousands of classes trained over millions of images. Image classification has many applications across many domains. In the automobile industry, image classification is used in autonomous driving. In gaming, image recognition and classification is used for virtual reality games. In healthcare, it is used for micro surgical procedures and computer guided robotic surgeries. It is also used for diagnosis. Retail industry uses image classification for automation of sales and inventory tracking at cutting edge technologically advanced stores like amazon go. It is being used in cyber security industry for facial recognition and iris recognition based login in windows devices and mobile phones.

However, image recognition models are not perfect. There is a problem that images can sometimes be misclassified. Furthermore, there is a huge vulnerability, which researchers have discovered and can be exploited. If specific noise is added to the images, it leads to misclassification of the images. This noise is insignificant for human vision, and usually ignored by humans as artifact of photography noise, or even remains unnoticed since the quantity of noise is so small/insignificant. Malicious entites can use this vulnerability to interfere with critical systems like self driving cars. These attacks are called adversarial attacks. Adversarial attacks use techniques to add imperceptible perturbations to the input images which results in misclassification of the input image. The changes cannot be noticed by a human eye or are not considered significant as it appears to be random noise, however, the detrimental effect it has on the functioning of Neural Networks is significant [6,7].

Adversarial attacks modify the input image leading to misclassifying the image. The wrong classification can be dangerous in real-world systems that rely on neural networks for object recognition [8]. For example, if a manufacturing robot is not able to detect a human or misclassifies a human as a vacant spot and places a heavy object on it, the risk to human life would exist [9]. Similarly, if an autonomous driving vehicle misclassifies the Stop sign as 45 Miles Per Hour Speed limit sign or red light as a green light sign, the chances that it would cause an accident are high, which again results in a risk to human life [10].

Therefore, without addressing such security risks, a real-world product that uses Neural Networks for image classification would be vulnerable to adversarial attacks, which is unacceptable in critical systems where human life is on the line.

1.1 Motivations

The following are the motivations due to which we focus our research on Defense against Adversarial Attacks. The factors that motivated us to create CD-GAN are as follows:

Image Recognition using Neural Networks is being used in various real-world applications already like Tesla's self-driving cars, Facebook's Auto-Tagging, etc. with more areas and industries adopting Image Recognition every day at a faster rate.

- The vulnerability in Image Recognition deployed in the real world can result in very serious consequences if not safeguarded properly.
- Researchers have already demonstrated the "Stop" road sign being misclassified as a "45 MPH Speed Limit" sign [9] in deployed production cars, and "Turtle" Being misclassified as "AK-47" [11]. Such attacks can have dire consequences in the wrong hands.
- Some specialized defense techniques exist which can defend against some attacks that are known to them, but such defenses are not successful against attacks that are new/unseen to the defense approach.
- A generalized defense technique is needed which can protect against all kinds of attacks.

1.2 Contributions

Our work has the following contributions to the field of Defense against Adversarial attacks on Image Classification:

- Provides an attack-independent and classifier-independent defense approach to defend against adversarial attacks on Image Classification.
- Provides a generalized defense technique that is easy to integrate within existing systems without the extra effort of defense model retraining or classifier retraining.
- Provides a defense technique that performs comparably to the state-of-the-art while being trained with a more lightweight system configuration and lesser computational resources.

Chained-Dual GAN is a completely new defense technique that defends against the adversarial attacks by neutralizing the perturbations of the adversarial image by chaining two independent perturbations minimizing GANs.

2 Existing Adversarial Attacks & Defenses

2.1 Attacks

All adversarial attacks work by inducing minor adversarial noise or Adversarial Perturbations to the input image. The perturbations are not visible to the human eye or might be ignored as noise, but this results in misclassification which is not desirable for the object recognition model. There are primarily three categories of attacks:

- **White-Box Attacks**: These are the attacks where the model information is accessible to the attack algorithm. Such attacks can see the weights, and probabilities of prediction for each attack sample and usually iteratively work on minimizing the prediction probability of the right class or increasing the prediction probability of the wrong class. Some of the white box attacks include the Fast Gradient Sign Method, Basic Iterative Method, Carlini & Wagner Method, etc. [12–15].

- **Black-Box Attacks**: Attacks where the model information, training data, weights, etc. are not accessible to the attack algorithm. Such attacks only have access to the classifier and its predicted output. Some of the attacks are Substitution Model Attack, Zeroth Order Optimization based BlackBox Attack, Query Efficient BlackBox Attacks, etc. [16–19].
- **Gray-Box Attacks**: These techniques have access to the model parameters, data, etc. initially like white-box models, but after a few steps when they have accumulated enough metadata, they no longer need that access and can work like a black-box model from there onwards. Some of the gray-box attacks are Ying & Zhong's Gray-Box Attack, Vivek & Mopuri's Gray-Box Attack, etc. [20].

2.2 Defenses

To safeguard against such attacks, various defense approaches are available. There are three primary types of defenses.

- **Gradient Masking**: This is a technique where the gradient of the model is masked to prevent the attack from understanding the inner working of the model and the weights of the model parameters. It is successful against most white-box attacks but doesn't perform well against black-box attacks. Some of the gradient masking defenses are Defensive Distillation, Shattered Gradient Approach, Randomized Gradients, Exploding and Vanishing Gradients, etc. [21,22].
- **Robust Optimization**: This approach handles the problem of adversarial attacks by re-engineering the classifier model or retraining it with adversarial examples while it is in the training phase. This approach has shown to be successful in many cases against known existing attacks, but it fails when a new attack is used which is previously unseen by the model. Some of the Robust Optimization defenses are Regularization, Adversarial Retraining, etc. [23,24].
- **Adversarial Example Detection:** These methods rely on directly classifying an input as adversarial input and then predicting the actual class. If an example is detected as adversarial, the model will not try to classify it, and instead flag or classify it as Adversarial Class. Some of the Adversarial Example Detection defenses are the Auxiliary model for Adversarial Classifier, Statistical Approaches, Prediction Consistency Check, etc. [25–28].

3 Our Approach: Chained Dual GAN

Our approach provides the defense against adversarial attacks by minimizing the perturbations of the adversarial image using iterative under sampling and oversampling. Using down-scaling chained to up-scaling using a distinct GAN for each of these steps, we can minimize the perturbations of the adversarial image significantly.

The approach relies on the Power of GANs to generate benign samples from adversarial examples. Each of the two GANs is trained with input as Adversarial Image and output as Normal Image. For the Sub-resolution GAN, the input is a 256 × 256 adversarial image, but the output is a 64 × 64 benign image. For the Super-resolution GAN, the input is a 64 × 64 adversarial image, but the output is a 256 × 256 benign image. This makes each GAN independently capable of removing the adversarial perturbations. However, while the GANs are trained independently, for real-world application and use, they are chained together in series before the classifier, such that an image may be passed on to our Chained-Dual GAN. The image is first resized to 256 × 256 and goes to Sub-Resolution GAN, which removes most of the adversarial perturbations and downscales it to 64 × 64. Then, the 64 × 64 output from sub-resolution GAN is passed on to Super-Resolution GAN, which removes the remaining adversarial perturbations and upscales it back to 256 × 256. The Super-Resolution GAN output is then resized as per classifier input parameters and passed on to the classifier. It must be noted that the classifiers are off-the-shelf pre-trained models like ResNet-v2, Inceptionv3, MobileNet-v2, etc., and are not altered in our approach, but chained with our CD-GAN and used as it is with available pre-trained weights available in TensorFlow-hub and torch-hub [29]. CD-GAN can be classified under gradient masking techniques which alters the input image significantly such that the input image is modified with adversarial noise reduction before it is passed to the classifier.

3.1 Datasets Used

3.1.1 ImageNet Large Scale Visual Recognition Challenge-2012

ILSVRC-2012-VAL (ImageNet Large Scale Visual Recognition Challenge-2012-VALIDATION) dataset is a large dataset of images that have been used for image classification [30]. The dataset consists of 50,000 images from 1000 classes. Some of the samples are shown in Fig. 2.

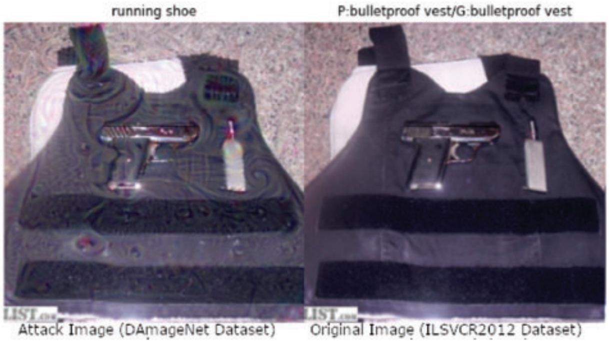
running shoe
P:bulletproof vest/G:bulletproof vest
Attack Image (DAmageNet Dataset)
Original Image (ILSVCR2012 Dataset)

Figure 1: Attack example

Figure 2: ILSVRC-2012-VAL

3.1.2 DAmageNet

DAmageNet dataset contains 50,000 adversarial samples which have proven to provide a generalized zero-query black box attack on multiple models, with post-attack reported error rate between 93% to 100% [31]. Fig. 1 shows an example of a bulletproof vest (From the ILSVRC-2012 Dataset) being misclassified as a running shoe. Some of the adversarial samples are shown in Fig. 3.

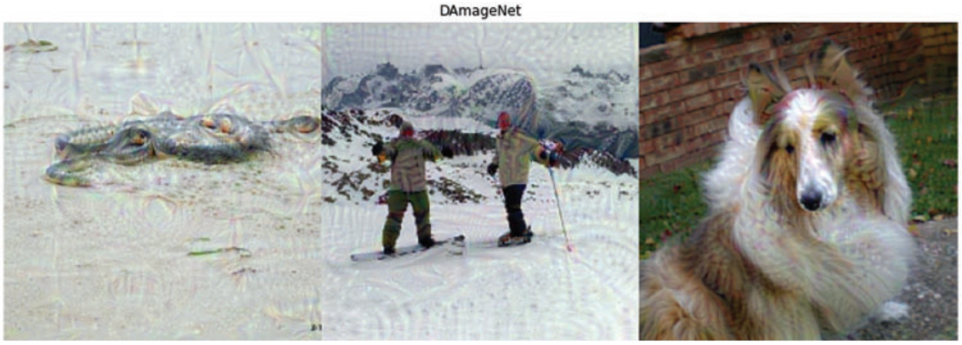

Figure 3: DAmageNet

3.2 Model Architecture

As the name Chained Dual-GAN suggests, our model architecture is based on the following two GANs as shown in Fig. 4, which are chained together in series before the classifier:

- CD-GAN's Sub-Resolution GAN
- CD-GAN's Super-Resolution GAN

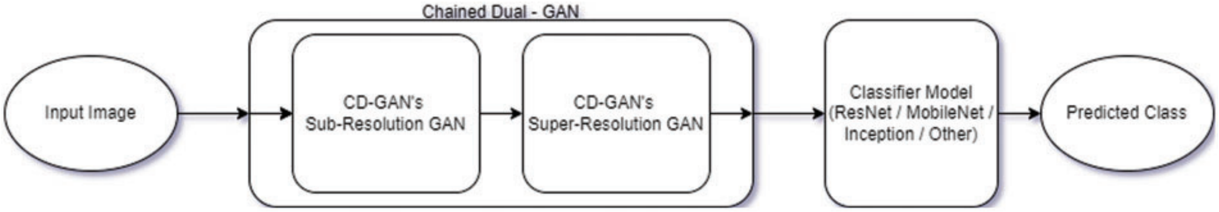
Chained Dual - GAN
Input Image
CD-GAN's Sub-Resolution GAN
CD-GAN's Super-Resolution GAN
Classifier Model (ResNet / MobileNet / Inception / Other)
Predicted Class

Figure 4: CD-GAN's architecture

Each of these takes an adversarial image as an input and generates a benign image. They essentially act as perturbation minimizers. Each GAN is independently trained but is chained together in series for application. A Detailed Breakdown of each of these GAN's architecture is discussed:

3.2.1 CD-GAN's Sub-Resolution GAN

CD-GAN's Sub-Resolution GAN is a perturbation minimizing GAN which takes a high dimension adversarial image and generates a perturbation-free low dimension image. The GAN is trained with input as a 256 × 256 adversarial image, but the output is a 64 × 64 benign image. The GAN's Architecture is shown in Fig. 5 for the Generator.

It uses Parametric Rectified Linear Unit (PReLU), BatchNorm2d, and TanH as activation functions in its various sequential convolution 2d layers.

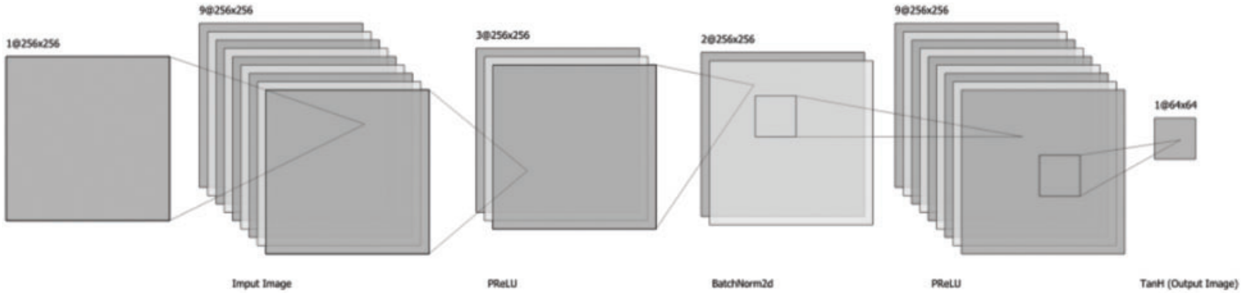

Figure 5: CD-GAN's sub resolution GAN architecture

3.2.2 CD-GAN's Super-Resolution GAN

CD-GAN's Super-Resolution GAN is a perturbation minimizing GAN just like CD-GAN's Sub Resolution GAN, However, it takes a low dimension adversarial image and generates a perturbation-free high dimension image. In this way, it acts as an Upscaling as well as perturbation minimizing GAN. The GAN is trained with input as a 64 × 64 adversarial image, but the output is a 256 × 256 benign upscaled image. The GAN's Architecture is shown in Fig. 6 for the Generator.

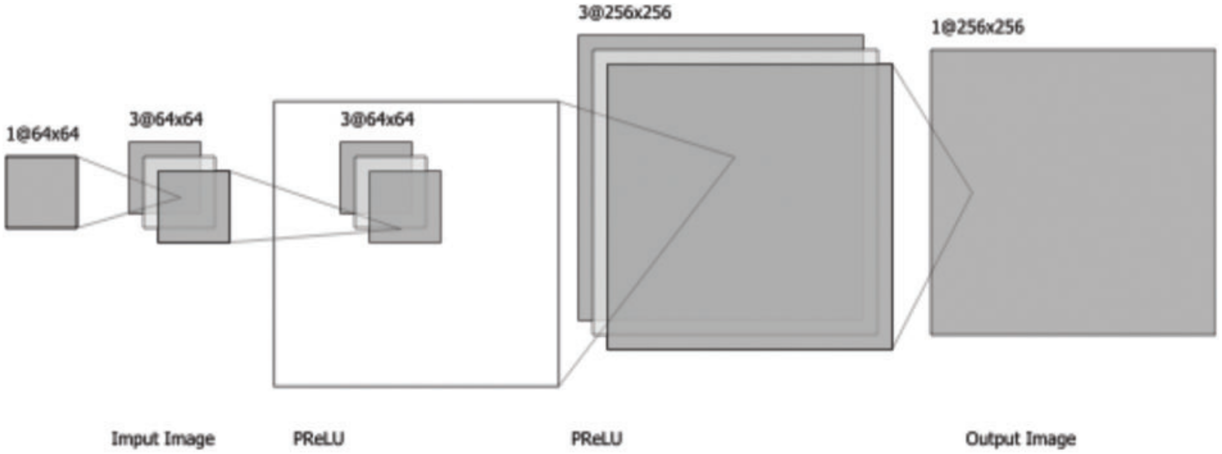

Figure 6: CD-GAN's super resolution GAN architecture

It also uses Parametric Rectified Linear Unit (PReLU), BatchNorm2d, and TanH as activation functions in its various sequential convolution 2d layers.

3.2.3 CD-GAN's Discriminator

For both the CD-GAN's Generators, the Discriminator is a convolutional neural network with a single fully-connected sequential layer. The Discriminator instance is configured with input as 64 × 64 × 3 for CDGAN's Sub-Resolution GAN and another instance with input 256 × 256 × 3 for CD-GAN's Super-Resolution GAN. These Discriminators are used in the training of GANs, but not in the testing/application of our models.

3.3 Experimental Setup

3.3.1 Input

The input of the classifier is a 224 × 224 image. This image can be a normal image or an attack image.

- **Normal Image**: Normal Images have a low probability of being misclassified by the classifier. The normal images are taken from the ILSVRC-2012-VAL dataset as it is without any modification.
- **Attack Images**: Attack Images have a high probability of being misclassified by the classifier. The attack images are generated by adversarial attacks on normal images. For our experimental setup, we use the images from the DAmageNet dataset as the attack input to our models.

3.3.2 Classifier Models

The classifier models are pre-trained models which are trained on ImageNet for image classification. The models are off-the-shelf pre-trained models and are not altered in our approach. These classifier models are used with available pre-trained weights available in TensorFlow-hub and torch-hub. Our approach however does chain these models' input to our CD-GAN output which is then passed on to our classifier. This is done to remove the adversarial perturbations from the input image before it is passed on to the classifier.

- ResNet-v2: ResNet-v2 or Residual Network Version 2 is an extremely successful image deep learning neural network with a great success rate in Image Classification and won the first prize in ILSVRC 2015 Competition. The ResNetV2 model is an improved version of the original ResNet model. We use ResNet152-V2 and ResNet50-V2 from TensorFlow-hub which were trained on ILSVRC Dataset.
- MobileNet-V2: MobileNet-V2 is a lightweight CNN architecture that is designed to be more efficient on mobile devices. It is a general-purpose CNN architecture that has been shown to achieve very good performance on a wide range of tasks, especially image classification. We use MobileNet-V2 from TensorFlow-hub trained on ILSVRC Dataset.
- Inception-V3: Inception-V3 is an extremely successful image deep learning neural network with a great success rate in Image Classification while an emphasis on optimizing the training on GPU, TPU, and distributed computing paradigms. We use Inception-V3 from TensorFlow-hub trained on ILSVRC Dataset.

3.3.3 System Configuration

The System Configuration provided by Google Colab Pro Plus High Ram GPU Instances at the time of training is as follows:

- GPU (Graphics Processing Unit): Tesla P100-PCIE-16GB
- CPU (Central Processing Unit): Intel(R) Xeon(R) CPU @ 2.30 GHz Model 63
- RAM (Random Access Memory): 54.8 GB VRAM (Virtual RAM)
- Local Storage: 170 GB SSD (Solid State Drives)

It took multiple sessions and instances to train and evaluate our work. The data in between the sessions was stored in Google Drive Mounted to the Colab Notebook for persistent storage.

3.3.4 Languages and Frameworks

The implementation of the study was done using Python 3.7 in Jupyter Notebook. Compute resources of Google Colab were used, which uses linux back-end. PyTorch was used for implementing and training the CD-GAN model. The classifiers were used from PyTorch and TensorFlowHub both.

3.4 Training the CD-GAN

For training the CD-GAN, we had to train two independent models, the CDGAN's Sub-Resolution GAN, and the CD-GAN's Super-Resolution GAN. Training for both of these was done on Google Colab Pro Plus GPU Instances with High-RAM and Background Execution options Enabled.

3.4.1 Training the CD-GAN's Sub-Resolution GAN

The CD-GAN's Sub-Resolution GAN is trained with the following steps:

- Train the CD-GAN's Sub-Resolution GAN with the input as a 256 × 256 adversarial image from the DAmageNet dataset.
- Train the CD-GAN's Sub-Resolution GAN with the output as a 64 × 64 benign image corresponding to the DAmageNet attack image from the ILSVRC-2012-VAL dataset.
- The training was monitored visually and manually by us. We continued the training till the output images appeared visually similar to the original image (20 epochs/2.5 h approx.).

The model weights were stored in google drive and mounted to the colab notebook for persistent storage.

3.4.2 Training the CD-GAN's Super-Resolution GAN

The Sub-Resolution GAN is trained with the following steps:

- Train the CD-GAN's Super-Resolution GAN with the input as a 64 × 64 adversarial image from the DAmageNet dataset.
- Train the CD-GAN's Super-Resolution GAN with the output as a 256 × 256 benign image corresponding to the DAmageNet attack image from the ILSVRC-2012-VAL dataset.
- The training was monitored visually and manually by us. We continued the training till the output images appeared visually similar to the original image (250 epochs/30 h approx.).
- The model weights were stored in google drive and mounted to the colab notebook for persistent storage.

It should be noted that while both the models are trained completely independently while using the models in application and evaluation, we chained the CD-GAN's Sub Resolution, CD-GAN's Super-Resolution, and Classifier models in series to get the final output class prediction.

3.5 Performance Evaluation & Results

As described in Experimental Setup (3.3), the performance is evaluated using off-the-shelf classifier models with pre-trained weights. We resized the base input image to 256 × 256 and provided it to CD-GAN, specifically, CD-GAN's SubResolution GAN. We chained the output of CD-GAN's Sub Resolution model to the output of CD-GAN's Super Resolution model. The output of the CD-GAN's Super Resolution model is then resized to Classifier's input size and passed on to the classifier. This is done to remove the adversarial perturbations from the input image before it is passed on to the classifier. As each model is trained to remove perturbations independently, it is theorized that the combined effect of two models within CD-GAN results in a higher quality final output of CDGAN compared to what either of them could provide individually. It is observed that the approach can remove perturbations from the attack image such that it is classified correctly as shown in Fig. 7.

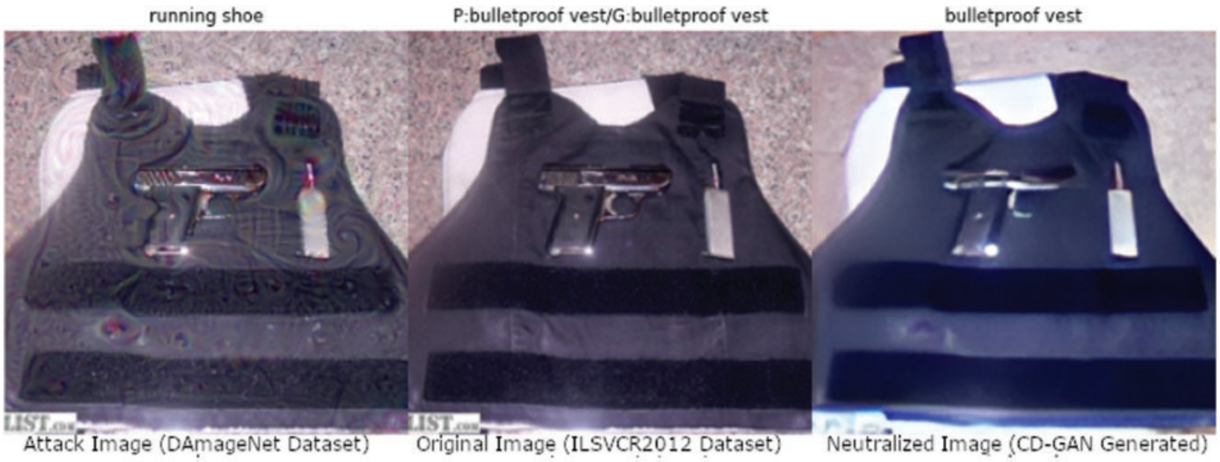
running shoe
P:bulletproof vest/G:bulletproof vest
bulletproof vest
Attack Image (DAmageNet Dataset)
Original Image (ILSVCR2012 Dataset)
Neutralized Image (CD-GAN Generated)

Figure 7: Defense against attack example

It can be seen that the though the original image is a bulletproof vest, and is classified correctly, the attack image is classified as a running shoe, however, the neutralized benign image generated after the attack image passed from CD-GAN is classified correctly as a bulletproof vest. The performance was evaluated on ResNet152-V2, MobileNet-V2, InceptionV3, and ResNet50-V2. For Evaluation, 12,000 random images were chosen from DAmageNet/ILSVRC-2012-VAL Dataset, and the performance was evaluated using the original accuracy of the model for benign images, the accuracy of the DAmageNet Samples, and the accuracy of Neutralized Images Generated by CD-GAN for input DAmageNet. The results are as shown in Tab. 1.

Table 1: Results of CD-GAN

Classifier	Normal accuracy (ILSVRC)	Attack accuracy (DAmageNet)	Defense accuracy (DAmageNet + CD-GAN)
WideResNet50-V2	78.41	8.66	35.44
ResNet152-V2	77.97	8.43	37.77
ResNet50-V2	70.29	9.46	34.71
ResNet18-V2	69.36	2.66	28.58
Inception-V3	70.83	8.34	33.35
MobileNet-V2	70.85	8.36	32.17

Since DAmageNet Samples are using an Untargeted Black-Box Attack which is not even aware of the classifier model and is more close to a real-world situation where the attacker will not have access to the classifier model, the accuracy of our defense CD-GAN against DAmageNet Samples evaluated and compared to the prior state of the art defenses for Untargeted Black-Box Attacks. Tab. 2 shows the performance of states of the arts.

Table 2: Results of prior state of the arts collected from RobustBench [32]

Technique	Classifier	Standard accuracy	Defensive accuracy
Transfer learning based adversarially robust models	WideResNet-50-V2	68.46%	38.14%
	ResNet-50-V2	64.02%	34.96%
	ResNet-18-V2	52.92%	25.32%
Robustness library	ResNet-50-V2	62.56%	29.22%
Fast adversarial training based models	ResNet-50-V2	55.62%	26.24%
Standardly trained model without defense	ResNet-50-V2	76.52%	0.0%

The Data of State-of-the-arts is collected from RobustBench [32] Leaderboard, a competitive repository for robustness evaluation of defense models. The performance of our Defense CD-GAN compared to the state-of-the-art as obtained from RobustBench is compiled in Tab. 3.

Table 3: CD-GAN's performance compared with results of prior works

Technique	Classifier	Standard accuracy	Defensive accuracy
Chained dual-GAN model (Our model)	WideResNet-50-V2	78.41%	35.44%
	ResNet-152-V2	77.97%	37.77%
	ResNet-50-V2	70.29%	34.71%
	ResNet-18-V2	69.36%	28.58%
	Inception-V3	70.83%	33.35%
	MobileNet-V2	70.85%	32.17%
Transfer learning based adversarially robust models	WideResNet-50-V2	68.46%	38.14%
	ResNet-50-V2	64.02%	34.96%
	ResNet-18-V2	52.92%	25.32%
Robustness library	ResNet-50-V2	62.56%	29.22%
Fast adversarial training based models	ResNet-50-V2	55.62%	26.24%
Standardly trained model without defense	ResNet-50-V2	76.52%	0.0%

It can be observed that the performance of our approach in terms of accuracy is comparable to other states of the arts for black-box untargeted attacks. The results are analysis of our approach on multiple classifiers with mean accuracy of 33.67% on the analysed classifiers. Our approach performs slightly less than other state of the arts for WideResNet-50-V2 with a delta of approx. 3%. It performs comparably to state of the art for ResNet-50-V2 with delta of approx. 0.2%. It performs slightly better than state of the arts for ResNet-18-V2 with delta of approx. 3%. We have also analysed our approach for MobileNet-V2 and Inception-V3, for which, defense results were not available for comparison at the time of the study.

3.5.1 Advantages of CD-GAN Over Prior Approaches

Our proposed approach has a few advantages compared to the old models. The advantages are as follows:

- Our model is a modular approach which can be retrofitted with any existing classifier directly.
- Accuracy of our model is in the same range as existing state of the art, however the compute resources used to accomplish the same are significantly less. The other authors have used 4 × NVidia 2080 Ti Graphic cards in SLI, while our approach accomplished similar results using only 1 × Nvidia Tesla P100.
- Our model is a generalized defense, i.e., it protects the models against all the attacks, The other works only defend against a specific attack, and fail against new attacks. Our model can defend against known as well as previously unknown/unseen attacks.
- Our approach is a gradient masking technique which performs well against black-box attacks. Existing gradient masking techniques do not perform well against black-box attacks.

These are the few benefits of our approach compared to existing state of the arts.

3.5.2 Limitations of CD-GAN and our Study

Our work has significant contribution to the field, but it is not without limitations. The following are limitations of our work:

- We have only analyzed performance of our approach against black-box attacks. Performance of the approach against white-box attacks is undetermined.
- Performance of approach is comparable to state of the arts with mean accuracy of 33.67%, but it is still significantly less from standard accuracy mean 72.95% without any attacks. Ideal defense would be to have defense accuracy comparable to standard accuracy.

These are the biggest limitations of our work, and we will improve upon them in our future works.

3.5.3 Impact of CD-GAN

CD-GAN brings a new approach which performs admirably by combining the powers of GAN with a multi-layered and redundancy enforced safety architecture. It intervenes before the extraction phase [33,34] of the image recognition. This approach works by removing attack noise from all images instead of detecting an attack [35,36] like in some other approaches. It therefore provides security to the solution against adversarial cyber-attacks [37,38]. It is an AI based technique which protects the classifier from attacks using GANs [39,40] by using iterative adversarial attack noise elimination. It can have significant impact in domains of image classification, image processing and image transmission over networks [41,42]. It will also have cyber-security impact in the domains of health care [43], cloud computing [44,45], image transmission using optical fibre [46–48] and deep learning [49].

4 Future Scope & Conclusion

4.1 Future Scope

The Approach is performing well in terms of accuracy and robustness compared to the state of the arts for black-box untargeted defense. However, there are still many things that can be done to approach better and its analysis more comprehensive:

- The defense accuracy of 35%–38% compared to the Natural Accuracy of 70%–77% still leaves us with a lot of room for improvement.

- The paper primarily focuses on Black-Box Attacks and the defensive results. It does not take into account the performance of white-box attacks.
- A bigger training dataset can be utilized to see if the performance improves.
- The CD-GAN's Input/Output Dimensions can be generalized and taken dynamically instead of using the fixed input/output sizes of 256 × 256 and 64 × 64 in various intermediate stages. Other Applications of CD-GAN in different domains can be explored using Transfer Learning to use CD-GAN for some unique applications in Image Processing.

4.2 Conclusion

In this paper, we have implemented the approach CD-GAN (Chained Dual-GAN) using two distinct GANs (CD-GAN's Sub-Resolution GAN & CD-GAN's Super-Resolution GAN) attached in series. This CD-GAN in itself is prepended to any standard classifier model and is used to defend the model against adversarial attacks. The CD-GAN is a defense that performs comparably to state-of-the-art defense against Black-Box adversarial attacks. Moreover, it does not depend on the attack type or the classifier model for its functioning. Rather, the modular design of CD-GAN increases its applicability very strongly across multiple domains. The core work of this paper demonstrates how a defense can be created which is classifier-blind and attack-blind. The CD-GAN is a defense that can be used to defend any classifier model effectively as shown by experimental results on the benchmark datasets of DAmageNet and ILSVRC. There is still a lot to be done in the domain of adversarial defense, as it is a relatively new research area, but our approach is a promising one.

Acknowledgement: We deeply acknowledge Taif University for Supporting this study through Taif University Researchers Supporting Project Number (TURSP-2020/115), Taif University, Taif, Saudi Arabia.

Funding Statement: This study was funded by Taif University, Taif, Saudi Arabia through Taif University Researchers Supporting Project Number (TURSP-2020/115). Grant was received by A. B. Singh. Link: https://www.tu.edu.sa/En/Deanship-of-Scientific-Research/83/Pages/22310/Researchers-Supporting-Project-(TURSP).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] K. Lakshmanna and N. Khare, "Constraint-based measures for DNA sequence mining using group search optimization algorithm," *International Journal of Intelligent Engineering & Systems*, vol. 9, pp. 91–100, 2016.

[2] K. Lakshmanna and N. Khare, "Mining DNA sequence patterns with constraints using hybridization of firefly and group search optimization," *Journal of Intelligent Systems*, vol. 27, no. 3, pp. 349–362, 2018.

[3] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput *et al.,* "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[4] A. Priyanka, M. Parimala, K. Sudheer, R. Kaluri, K. Lakshmanna *et al.,* "BIG data based on healthcare analysis using IOT devices," *IOP Conference Series: Materials Science and Engineering*, vol. 263, no. 4, pp. 042059, 2017.

[5] R. Kaluri, D. S. Rajput, Q. Xin, K. Lakshmanna, S. Bhattacharya *et al.,* "Roughsets-based approach for predicting battery life in IoT," in arXiv preprint arXiv:2102.06026, 2021.

[6] H. Sun, T. Zhu, Z. Zhang, D. J. Xiong and W. Zhou, "Adversarial attacks against deep generative models on data: A survey," in arXiv preprint arXiv:2112.00247, 2021.

[7] L. Ye, "Thundernna: A white box adversarial attack," in arXiv preprint arXiv:2111.12305, 2021.

[8] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," in ICLR'15. arXiv preprint arXiv:1412.6572, 2015.

[9] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati *et al.,* "Robust physical-world attacks on deep learning visual classification," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, pp. 1625–1634, 2018.

[10] N. Morgulis, A. Kreines, S. Mendelowitz and Y. Weisglass, "Fooling a real car with adversarial traffic signs," arXiv preprint arXiv:1907.00374, 2019.

[11] A. Athalye, L. Engstrom, A. Ilyas and K. Kwok, "Synthesizing robust adversarial examples," in *Int. Conf. on Machine Learning PMLR*, Stockholm Sweden, pp. 284–293, 2018.

[12] T. Muncsan and A. Kiss, "Transferability of fast gradient sign method," in *Proc. of SAI Intelligent Systems Conf.*, Virtual Event, Springer, Cham, pp. 23–34, 2020.

[13] J. S. Pang, "On the convergence of a basic iterative method for the implicit complementarity problem," *Journal of Optimization Theory and Applications*, vol. 37, no. 2, pp. 149–162, 1982.

[14] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proc. of the 10th ACM Workshop on Artificial Intelligence and Security*, Dallas Texas, USA, pp. 3–14, 2017.

[15] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symp. on Security and Privacy (SP)*, San Jose, California, USA, pp. 39–57, 2017.

[16] C. Guo, J. Gardner, Y. You, A. G. Wilson and K. Weinberger, "Simple black-box adversarial attacks," in *Int. Conf. on Machine Learning PMLR*, Long Beach, California, USA, pp. 2484–2493, 2019.

[17] N. Papernot, P. McDaniel and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," in arXiv preprint arXiv:1605.07277, 2016.

[18] P. Y. Chen, H. Zhang, Y. Sharma, J. Yi and C. J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. of the 10th ACM Workshop on Artificial Intelligence and Security*, Dallas, USA, pp. 15–26, 2017.

[19] A. Ilyas, L. Engstrom, A. Athalye and J. Lin, "Query-efficient black-box adversarial examples (superceded), " in arXiv preprint arXiv:1712.07113, 2017.

[20] B. S. Vivek, K. R. Mopuri and R. V. Babu, "Gray-box adversarial training," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 203–218, 2018.

[21] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symp. on Security and Privacy (SP)*, San Jose, California, pp. 582–597, 2016.

[22] H. Qiu, Y. Zeng, Q. Zheng, T. Zhang, M. Qiu *et al.,* "Mitigating advanced adversarial attacks with more advanced gradient obfuscation techniques," in arXiv preprint arXiv:2005.13712, 2020.

[23] F. Tram`er, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh *et al.*, "Ensemble adversarial training: Attacks and defenses," in arXiv preprint arXiv:1705.07204, 2017.

[24] Y. Song, T. Kim, S. Nowozin, S. Ermon and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," in arXiv preprint arXiv:1710.10766, 2017.

[25] B. Li, Y. Vorobeychik and X. Chen, "A general retraining framework for scalable adversarial classification," in arXiv preprint arXiv:1604.02606, 2016.

[26] J. H. Metzen, T. Genewein, V. Fischer and B. Bischoff, "On detecting adversarial perturbations," in arXiv preprint arXiv:1702.04267, 2017.

[27] K. Grosse, P. Manoharan, N. Papernot, M. Backes and P. McDaniel, "On the (statistical) detection of adversarial examples," in arXiv preprint arXiv:1702.06280, 2017.

[28] W. Xu, D. Evans and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in arXiv preprint arXiv:1704.01155, 2017.

[29] A. B. Singh, L. K. Awasthi and U. Bansal, “Defense against adversarial attacks using chained dual-GAN approach,” in *Proc. of ICSMDI*, Trichy, India, 2022.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh *et al.,* “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[31] S. Chen, Z. He, C. Sun, J. Yang and X. Huang, “Universal adversarial attack on attention and the resulting dataset damagenet,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2188–2197, 2020.

[32] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion *et al.,* “Robustbench: A standardized adversarial robustness benchmark,” in arXiv preprint arXiv:2010.09670, 2020.

[33] H. Sun and R. Grishman, “Lexicalized dependency paths based supervised learning for relation extraction,” *Computer Systems Science and Engineering*, vol. 43, no. 3, pp. 861–870, 2022.

[34] H. Sun and R. Grishman, “Employing lexicalized dependency paths for active learning of relation extraction,” *Intelligent Automation & Soft Computing*, vol. 34, no. 3, pp. 1415–1423, 2022.

[35] V. Priya, I. Sumaiya Thaseen, T. Reddy Gadekallu, M. K. Aboudaif and E. Abouel Nasr, “Robust attack detection approach for iiot using ensemble classifier,” *Computers, Materials & Continua*, vol. 66, no. 3, pp. 2457–2470, 2021.

[36] A. R. Javed, S. U. Rehman, M. U. Khan, M. Alazab and T. R. G, “CANintelliIDS: Detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU,” *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 1456–1466, 2021.

[37] C. Iwendi, Z. Jalil, A. R. Javed, T. G. Reddy, R. Kaluri *et al.,* “Keysplitwatermark: Zero watermarking algorithm for software protection against cyber-attacks,” *IEEE Access*, vol. 8, pp. 72650–72660, 2020.

[38] M. Hedabou and A. Y. Simpa, “Efficient and secure implementation of BLS multisignature scheme on TPM,” in *18th Annual IEEE Int. Conf. on Intelligence and Security Informatics*, Virginia, SA, 2020.

[39] E. Tcydenova, T. W. Kim, C. Lee and J. H. Park, “Detection of adversarial attacks in AI-based intrusion detection systems using explainable AI,” *Human-Centric Computing and Information Sciences*, vol. 11, no. 35, pp. 1–13, 2021.

[40] L. Zhao, Y. Zhang and Y. Cui, “A multi-scale u-shaped attention network-based GAN method for single image dehazing,” *Human-Centric Computing and Information Sciences*, vol. 11, no. 38, pp. 1–18, 2021.

[41] B. Xiong, K. Yang, J. Zhao and K. Li, “Robust dynamic network traffic partitioning against malicious attacks,” *Journal of Network and Computer Applications*, vol. 87, pp. 20–31, 2017.

[42] W. Wang, Y. Li, T. Zou, X. Wang, J. You *et al.,* “A novel image classification approach via dense-MobileNet models,” *Mobile Information Systems*, vol. 2020, no. 7602384, pp. 1–8, 2020.

[43] G. S. Gaba, M. Hedabou, P. Kumar, A. Braeken, M. Liyanage *et al.,* “Zero knowledge proofs based authenticated key agreement protocol for sustainable healthcare,” *Sustainable Cities and Society*, vol. 80, pp. 103776, 2022.

[44] M. Hedabou “Cloud Key management based on verifable secret sharing,” in *15th Int. Conf. on Network and System Security*, Tianjin, China, Lecture Notes in Computer Science, Cham, Springer, pp. 289–303, 2021.

[45] Y. S. Abdulsalam and M. Hedabou, “Security and privacy in cloud computing: Technical review,” *MDPI Future Internet*, vol. 14, no. 1, pp. 11, 2021.

[46] E. M. Amhoud, M. Chafii, A. Nimr and G. Fettweis, “OFDM with index modulation in orbital angular momentum multiplexed free space optical links,” in *IEEE 93rd Vehicular Technology Conf.*, Helsinki, Finland, pp. 1–5, 2021.

[47] E. M. Amhoud, G. R. B. Othman, L. Bigot, M. Song, E. R. Andresen *et al.,* “Experimental demonstration of space-time coding for MDL mitigation in few-mode fiber transmission systems,” in *2017 European Conf. on Optical Communication*, Gothenburg, Sweden, pp. 1–3, 2017.

[48] E. M. Amhoud, G. R. Othman and Y. Jaouën, “Capacity enhancement of few-mode fiber transmission systems impaired by mode-dependent loss,” *Applied Sciences*, vol. 8, no. 3, pp. 326, 2018.

[49] K. Zerhouni, E. M. Amhoud and M. Chafii, “Filtered multicarrier waveforms classification: A deep learning-based approach,” *IEEE Access*, vol. 9, pp. 69426–69438, 2021.