Combo-Breaker: Combining GAN algorthims to improve adversarial attack resistance for Intrusion Detection Systems

James Watson
Faculty of Science
Queensland University of Technology
2, George Street, Brisbane, 4001, QLD, Australia

***Abstract* - Machine learning (ML) based Intrusion detection systems (IDS) are commonly used within network security to protect against increasing threats to connected devices in the modern world. Whilst extremely successful in achieving this, over reliance of machine learning methods have come at the immense cost of a lack of overall robustness. This has led to a rise in adversarial attacks of these ML frameworks. Whilst these traditional convolutional perturbations have been thwarted in various ways, most unfortunately share vast similarities, leading to increasing susceptibility towards emerging attacks. To counter this, investigations were undertaken to see if combing features found in existing generative models would strengthen adversarial attack resistance for use in intrusion detection systems. In this paper it is proposed a novel generative adversarial network (GAN), called Combo-breaker GAN (CB-GAN). Combining class leading features found within state-of-the-art GAN algorithms. Through a combination of traditional convolutional and emerging transformer network features. This brand-new approach uses improvements within latent space through RBF networks, along with multiple instances of traditional GAN features, new additions of dual encoders and classifiers to ensure result rigidity. Coupled with convolutional vision transformers (CvT) and convolutional based subsampling for both power and reduced computational overhead. Theoretical results on CIFAR10 and MNIST datasets show significantly improved resistance of 25% against traditional convolutional neural network (CNN) based adversarial attacks, Fast Gradient Sign Method (FGSM) and Integrity attacks. Finally, the use of self-supervised learning, greatly reduces traditional requirements of the need for high training data, greatly improving on existing machine learning methods.**

Keywords— Generative Adversarial Networks, Adversarial Attacks, Intrusion Detection Systems, Combined Networks, Convolutional Vision Transformers

I. Introduction

Cyber-attacks are an increasing threat to connected systems today (Piplai et al., 2020),from IoT (internet of things) to super computers, these are all possible targets.

From DOS Attacks on a game server (Ruppert, 2018) to advanced ransomware attacks on commercial systems (Barrett, 2021.; Glenny, 2021), cyber-attacks are an increasing and evolving threat. To counter this, different strategies are proposed to stop unwanted network intrusion (Biswas, 2018.). The most popular of these systems is known as Intrusion Detection Systems (IDS) (Anderson,1980.). IDS's are a combination of various techniques and technologies to provide a wall of defense against internal and external attacks (Biswas, 2018.).

Traditional IDS are based on different approaches. Anomaly and misuse detection (Denning, 1987; Helman et al., n.d.; W. Lee & Stolfo, 2000) (Figure 1) both offer different advantages and weaknesses with hybridization proving effective for IDS applications (Bangui et al., 2021).

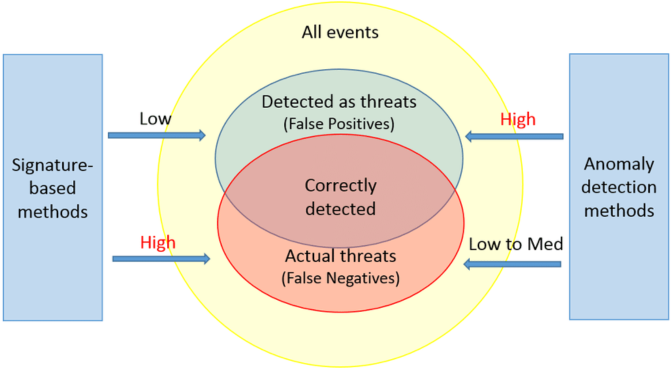


Figure 1: Anomaly vs misuse detection (Degeler et al., 2016)

Technological developments have given rise to machine learning (Samuel,1967;1959) assisted IDS's due to the increasingly complex and fast-moving nature of modern systems. Generative Neural Networks (GAN) (I. J. Goodfellow et al., 2014), changed the way in which data generation was possible (Aggarwal et al., 2021).

With these deep neural network (DNN) frameworks (Pouyanfar et al., 2019), GANs can learn from an input of training data, resulting in a new output with properties of the input data in which the two parts compete to reach Nash equilibrium (I. J. Goodfellow et al., 2014; Nash, 1950). This form of mimicry is a vital tool in the success of GAN algorithms.

This generative approach led to wide use within multiple fields (Aggarwal et al., 2021),with success leading to GANs increasing appliance within the information security sector

(Yinka-Banjo & Ugot, 2020), via machine learning based IDSs (ML-IDS). With these systems used for various tasks including detecting malicious network traffic for instance viruses and malware (Debar et al., 1999).

Despite this success, these networks are vulnerable to adversarial examples, perturbations of the input data (Szegedy et al., 2014).

Since then, the number of adversarial attacks have increased significantly (Ren et al., 2020).Multiple solutions for defence against adversarial attacks have emerged with a range of techniques developed using GAN based systems (Deldjoo et al., 2021; Piplai et al., 2020; Sabuhi et al., 2021; Yinka-Banjo & Ugot, 2020). (Figure 2).

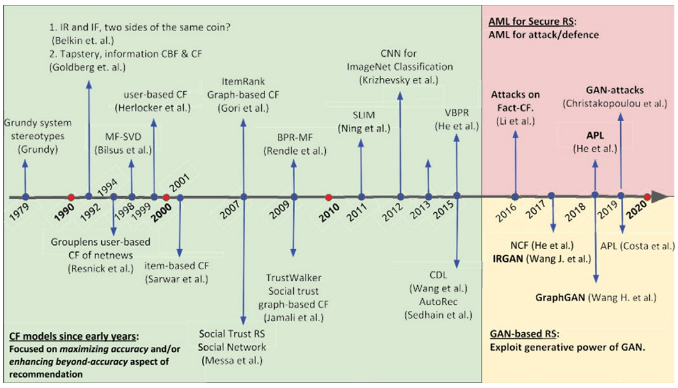


Figure 2: CF recommender models Milestones (Deldjoo et al., 2021)

The use of techniques including adversarial training (I. J. Goodfellow et al., 2015; Madry et al., 2019), show while effective, these systems are still vulnerable due to the linear nature of these networks (Araujo et al., 2020; Kotsiantis et al., 2006).

IDS based systems are typically attacked by either white (attacker has knowledge of network) or black box attack (attacker does not have knowledge of network) methods (Y. Zhang et al., 2020).

Attacks including gradient based attacks e.g. FGSM (Fast Gradient Signal Method) (I. J. Goodfellow et al., 2015) and data integrity attacks (poising attacks) (Jagielski et al., 2018) are common Whitebox attacks also effective within Blackbox situations (Liu et al., 2017)

FGSM Attacks are successful with limited datasets when tested against traditional GAN models (Sabuhi et al., 2021) due to higher success with reduced datasets making perturbation more effective (Madry et al., 2019).

This is thanks to overfitting (Rice et al., 2020), with counter training been only effective against similar attacks (Madry et al., 2019), so transferability of different attacks makes traditional training methods ineffective (Liu et al., 2017). Alternative networks such as Radial Basis function Neural networks (RBFNN) have found to be highly effective against traditional adversarial attacks(I. J. Goodfellow et al., 2015) but have limited use within most GAN models due to higher training cost (Zadeh et al., 2018)

Current GAN models lack strength in two major areas. Robustness and Transferability resilience. With various models using multiple discriminators (D) (Shieh et al., 2021) or generators (Fang et al., 2020) to fix these issues, existing research into the use of hybrid classification models (Bangui et al., 2021) is ongoing. With these limitations of GAN models (Aggarwal et al., 2021), it is proposed that a combination GAN algorithm to both strengthen and harden this new model to provide maximum flexibility increasing effectiveness against the changing nature of IDS attacks or implementation into non cyber security systems.

This paper proposes a new solution. When combining current algorithms, multiple weaknesses are found within present models (Chakraborty et al., 2018) are corrected or improved.

With the combination of leading algorithm features including both transformers (Vaswani et al., 2017) and RBFNN networks (Broomhead,1988.; Lecun, 1998; Zadeh et al., 2018), this provides a flexible robust framework model for application within a multitude of advanced fields. This novel approach is called combo breaker GAN (*CB-GAN*).

This paper follows a simple structure, beginning with (2) an overview of related works in both GAN and IDS applications related to adversarial attacks. This is followed by (3). The construction of CB-GAN. This is concluded (4) with theoretical results of the initial testing of this algorithm on MNIST (LeCun et al., 1989) and CIFAR10 datasets (Krizhevsky et al., 2010.) concluding (5) with future applications of the work.

II. Related works

Generative adversarial Networks (GAN) (I. J. Goodfellow et al., 2014) (Figure 3) are a form of Deep Generative modelling (DGM) (Rawat et al., 2021) (Figure 4), like variational Auto Encoders (VAE) (Kingma & Welling, 2014), are based on probabilistic modelling (Saxena & Cao, 2021), GAN is a minimax game using a Generator (G) and Discriminator (D) with a goal of achieving Nash equilibrium (I. J. Goodfellow et al., 2014).

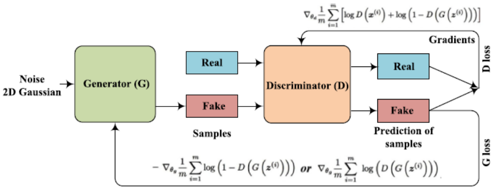

$\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}\left[\log D\left(x^{(i)}\right)+\log\left(1-D\left(G\left(z^{(i)}\right)\right)\right)\right]$
Gradients
D loss
Noise
2D Gaussian
Generator (G)
Real
Fake
Samples
Discriminator (D)
Real
Fake
Prediction of samples
G loss
$-\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log\left(1-D\left(G\left(z^{(i)}\right)\right)\right)$ or $\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log\left(D\left(G\left(z^{(i)}\right)\right)\right)$

Figure 3: Architecture of GAN (Saxena & Cao, 2021)

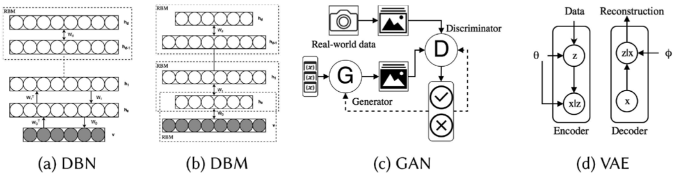


Figure 4: Forms of generative networks (Pouyanfar et al., 2019)

Showing success in various tasks including multimedia and healthcare (Engel et al., 2019; Karras et al.,2021.; Kumar & Tsvetkov, n.d.; Sharma et al., 2021; Vondrick et al., 2016).

Despite success, difficulties remain (Ferdowsi & Saad, 2021; Saxena & Cao, 2021). With improvement (Figure 5) (I. Goodfellow et al., 2020; Karras et al., 2018; Radford et al., 2016; Salimans et al., 2016; Zhu et al., 2020) greatly improving the original model.

Figure 5: Evolution of GAN 2014-2018 (Saxena & Cao, 2021)

Wasserstein metric gradient loss functionality (Arjovsky et al., 2017) greatly improved loss distributions compared to previous work of (I. J. Goodfellow et al., 2014) and increased resistance to mode collapse issues. While ffurther improvements from (Gulrajani et al., 2017; J. Wu et al., 2018) showed success (Hsu et al., 2021).

Though vulnerabilities within this new method are apparent, (J. Li, Cao, Zhang, Xu, et al., 2021), so further research is required. The use of Wasserstein distance is a major step forward compared to traditional methods (Lloyd & Weedbrook, 2018), becoming the standard in modern GANs.

To overcome dueling natures of GAN, various methods have been proposed. Combination methods from (Bitaab & Hashemi, 2017) proved successful in reducing classification problems within IDS.Despite these setbacks, this showed success in *SGAN* (Chavdarova & Fleuret, 2017). With alternative methods showing resistance against adversarial attacks (Arora et al., 2017; Bangui et al., 2021; Ghosh et al., 2016; Shu et al., 2020).

Cross-integration of GAN with other generative frameworks has been proposed, improving GANs generative abilities (Saxena & Cao, 2021). Having been ssuccessfully applied in the works of (Larsen et al., 2016) and (Y. Wang et al., 2020). The use of both VAE and GAN provides strengths and weakness of each generative network, whilst data clustering using Gaussian mixture models proved fruitful (Pandeva & Schubert, 2019.

The use of bidirectional mapping to achieve mapping between data and latent space (Donahue et al., 2017) proved to be unsuitable for certain applications requiring multiple data sources.. Further resistance with Transformer networks (Vaswani et al., 2017) have proven increasingly robust against traditional adversarial attacks (Benz et al., n.d.; Naseer et al., 2021), but training difficulties remain (Naseer et al., 2021).

Latent spaces between D+G are critical in the success or failure of adversarial attacks (de Alfaro, 2018), with RBF based networks (Broomhead, 1988.; Powell, 1977) showing success in hardening against such attacks (L. Hu et al., 2020; Zadeh et al., 2018).

As shown (Marrs & Webb, 1998) latent spaces are useful in discriminator processes and have been successfully applied to GAN (L. Hu et al., 2020).

Directional editing of the latent space provides interpretable paths with disentangled changes in the image space, with increased accuracy and nonlinear results (Tzelepis et al., 2021). This tightening of latent space hardens against adversarial attacks, shorting the distance between the two points (Gan et al., 2012).

Providing a rejection class (Figure 6) RBFs are faced with implementation difficulties (Crecchi et al., 2020),but are successful when applied towards anomaly detection showing increased resistance to poisoning attacks (M. Burruss et al., 2021; M. P. Burruss, 2020.).

Gradient optimization within latent spaces are the underpinning of artificial intelligence but also the overall goal for model success (I. Goodfellow et al., 2020).

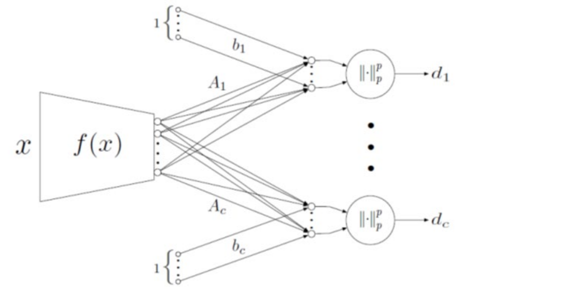


Figure 1: The diagram of a deep-RBF network with c classes. As the diagram shows, there is an RBF unit for every class on the transformation of the input, i.e., $f(x)$.

Figure 6: Class rejection in Deep RBF networks (Zadeh et al., 2018)

Model based poising attacks (Mei & Zhu,2014.) are used extensibility in IDS systems (Creswell et al., 2017; Koh et al., 2018; Suya et al., 2021) with limited use in GAN based systems (Suya et al., 2021).

Gradient based poisoning attacks are popular (Biggio et al.,2012.) whlist sanitation (Cretu et al., 2008) is ineffective in cleaning data of these attacks (Koh et al., 2018), latent space poisoning is shown to be effective (Creswell et al., 2017).

Alternative methods, including current Activation (AC) clustering (Chen et al., 2018) are unsuitable for sparsely poisoned datasets, (M. P. Burruss,2020.) with deep RBF networks showing increased resistance due to strict data representation (M. Burruss et al., 2021).

Integration of transformers with traditional GANs (Dai et al., 2021; K. Lee et al., 2021; H. Wu et al., 2021) show benefits of transformer designs (Child et al., 2019) with advantages of traditional GANs deliver optimum results, but with residue connections been prone to classifier degradation (Xu et al., 2021).

Robustness is a major cause of training issues, though GANs possesses augmentation strengths, weakness persists (Carlini & Wagner, 2017; Kurach et al., 2019; H. Wang et al., 2021) with the recent methods of (H. Wang et al., 2021) showing that both diversity and hardness create an optimal result.

Large datasets are traditionally required for optimum results (Brock et al., 2019; Karras et al.,2021.), with benefits in both training stability and data diversity (Roth et al., 2020).

Whilst limiting via core-set selection methods (Sinha et al.,2020.) are initially effective, this lacks helpful data characteristics (Roth et al., 2020).Improving via ranked based methods for data compression accuracy (Roth et al., 2020).

Cyber threats are not a novel concept, with IDS been just one of many (G. Kumar et al., 2010) with ML frameworks extremely effective (Chakraborty et al., 2018; W. Lee & Stolfo, 2000) compared to previous methods including access control and firewalls (Hamid et al., 2016), with hybrid classification techniques (Bangui et al., 2021; Haq et al., 2015; Kim et al., 2014) including multiple discriminators (Shieh et al., 2021) also showing success.

Common threats for IDS systems include malware (G. Kumar et al., 2010). Detection algorithms like MALGAN (W. Hu & Tan, 2017) (Figure 7) use feature quantity through malware

APIs by incorporating results into a substitute detector and MALGAN itself for creating non executable examples. Although updated methods (Kawai et al., 2019; Zhao et al., 2021) (figure 8) remove some of these initial requirements, further research is required

.

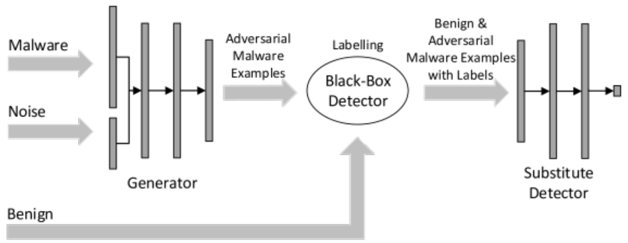

Malware
Noise
Benign
Generator
Adversarial Malware Examples
Labelling
Black-Box Detector
Benign & Adversarial Malware Examples with Labels
Substitute Detector

Figure 7: Architecture of MALGAN (W. Hu & Tan, 2017)

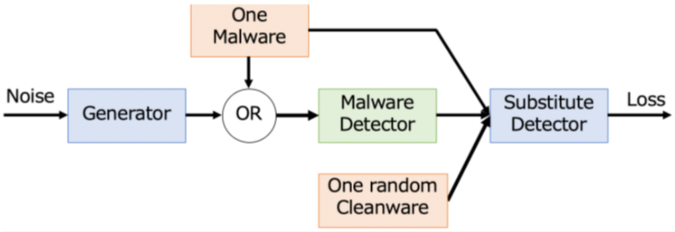
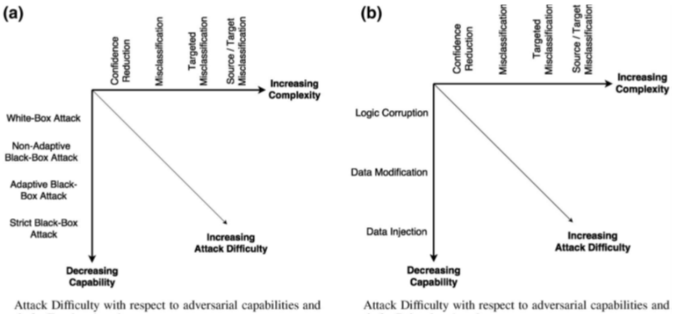


Figure 8: Improved MALGAN (Kawai et al., 2019)

Adversarial trained IDS are vulnerable to label leaking problems when trained via FGSM (Kurakin et al., 2017), as training on these methods greatly diminishes successful outcomes (Deldjoo et al., 2021) (Figure 9) With the nature of training examples and optimizations, difficulty levels increase, regulating attack success regardless of model used (Ahmad et al., 2021), showing further optimizations are required (Bai et al., 2021).



Attack Difficulty with respect to adversarial capabilities and goals for Evasion Attacks

Attack Difficulty with respect to adversarial capabilities and goals for Poisoning Attacks

Figure 9: Taxonomy of Adversarial Model for a) Evasion Attacks and b) Poisoning Attacks with respect to adversarial capabilities and goals(Chakraborty et al., 2018)

Recent studies show limited applications due to testing on retired datasets (Ahmad et al., 2021) and use generative samples quantitatively (Sabuhi et al., 2021), with over 50% using image-based datasets for testing against adversarial attacks (Sabuhi et al., 2021). This greatly diminishes real-world success. This restricted testing means real world applications are unknown (Ahmad et al., 2021).

III. Methods

This section describes the artefact-based approach for constructing CB-GAN. Current research aims to adapt the field of combination Machine learning algorithms into being further applicable to IDS implementations(Ahmad et al., 2021), with current work focusing on classifiers (Bangui et al., 2021; Kim et al., 2014)Whilst some algorithms exist (D. Li et al., 2019), issues experienced are present within standard designs. . Traditional GANs require large datasets for successful training (Saito et al., 2020) so reduction methods are developed as found in (Roth et al., 2020).

This paper will use a curated combination of the GAN models in (Figure 10). All chosen parts of this model are agnostic and as such not platform dependent so are easily added separately to existing frameworks. The use of multi part GANs have been successful (Baioletti et al., 2020; Saxena & Cao, 2021; Shieh et al., 2021) due to scalability of adding extra classifiers to the model showing higher success rates (Alqahtani et al., 2021) compared to singular models (G. Kumar et al., 2010).

Paper	Key Strength	Key Weakness	Chosen Part
Wasserstein Divergence for GAN[6]	Loss of 1-Lipshchitz Constraint	No two-time update scale	Loss function
CvT: Introducing Convolutions to vision transformers[3]	Transformer based; convolution inputs	High training requirements	Primary Generator and Discriminator
Adaptable GAN Encoders for Image Reconstruction via Multi-type Latent Vectors with Two-scale Attentions[4]	StyleGAN2 level Encoder	Fine tuning required	Encoder (E)
MMGAN: Generative Adversarial Networks for Multi-Modal Distributions	Multi-modal Gaussian clusters	Requires retraining	Multi-modal clusters for inputs (G, D, C)
Train Sparsely, Generate Densely: Memory-efficient Unsupervised Training of High-resolution Temporal GAN (T-GAN2)[7]	Subsampling layers	Large training requirements	sub-generator and discriminator setup
Detection of Adversarial DDoS Attacks Using Generative Adversarial Networks with Dual Discriminators	Dual Discriminator	Large training requirements	Primary Discriminator setup framework
Flow Field Reconstructions with GANs based on Radial Basis Functions[7]	Radial Basis Function discriminator	Training difficulty	Basis for CNN Discriminators
WarpedGANSpace Finding non-linear RBF paths in GAN latent space[5]	Latent space optimization – creates non-linear paths (learned RBF)	Untested on non-facial based datasets	Latent space (Z)
TAC-GAN[8]	Sub - classifier	Tied to generator	Classifier (C2)
Defense-GAN[11]	Classifier protector	Requires trained and tuned GAN	Classifier protector (CP)

Figure 10: Chosen GAN Models for inclusion into model, unnumbered means code is currently unavailable..

Small datasets e.g.: MNIST (LeCun et al., 1989) are typically chosen for wide use and ease of trainability (Saxena & Cao, 2021). As seen in (I. J. Goodfellow et al., 2015) the linear nature and ease of training increase adversarial attack success. Methods proposed of combining networks found within (figure 10), that when tested against FGSM attacks will show increased resistance or immunity.

FGSM attacks (I. J. Goodfellow et al., 2015) are increasingly successful against limited sized datasets e.g.: MNIST with larger datasets proving more robust (Carlini & Wagner, 2017). With traditional adversarial training methods proving ineffective (de Alfaro, 2018), alternative frameworks such as RBF and transformers, show increased resistance to these attacks (Benz et al.,2021.; Zadeh et al., 2018).

Data integrity attacks (poisoning attacks) (Nelson et al., 2021.) are famously used against emails to bypass spam filters (Suya et al., 2021). GAN based Poisoning attacks like KKT (Koh et al., 2018) show limited effectiveness (Suya et al., 2021) in both scaling and indiscriminate attack settings with RBFNN having shown immunity towards typical adversarial attacks (I. J. Goodfellow et al., 2015) with strengths in non-stationary data.

While the use of deep RBF networks remains unsolved against GAN based integrity attacks, they show great reduction in comparison to traditional defense methods in Deep Neural Networks.(M. Burruss et al., 2021). This is due to the gradient rejection imposed by these networks (Figure 6).

Pytorch version 1.9 was used the basis for construction of the model, with each selected part downloaded from the official GitHub repository listed in each paper[3-8]. Chosen datasets of MNIST[1] (LeCun et al., 1989) and CIFAR10[2] (Krizhevsky et al., n.d.) were downloaded from their official websites.

Theoretical implementation was done via deconstructing each part of the model, selecting the chosen part (e.g.: generator or discriminator). This model was constructed with the primary goal of robustness against two traditional CNN based adversarial attacks: FGSM and integrity attacks. Due to the agnostic nature of the selected models, no prior models (e.g.: *styleGAN* (Karras, Laine, & Aila., 2019) where required. *CB-GAN* was designed with the assumptions that combining multiple GAN parts would be successful in attack resistance.

CB-GAN was designed by deconstructing parts of CvT (H. Wu et al., 2021) (Figure 11).

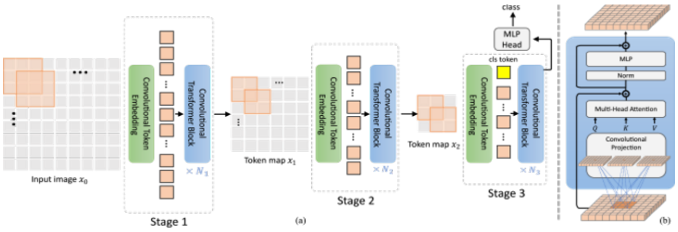


Figure 11: CvT model Diagram: (H. Wu et al., 2021)

Integration of chosen parts is based on (Feng et al., 2021) (Figure 12), showing it is possible to have convolutional encoders/decoders with both multi-modal attention layers and dual discriminators. The following parts (figure 10) are chosen for integration.

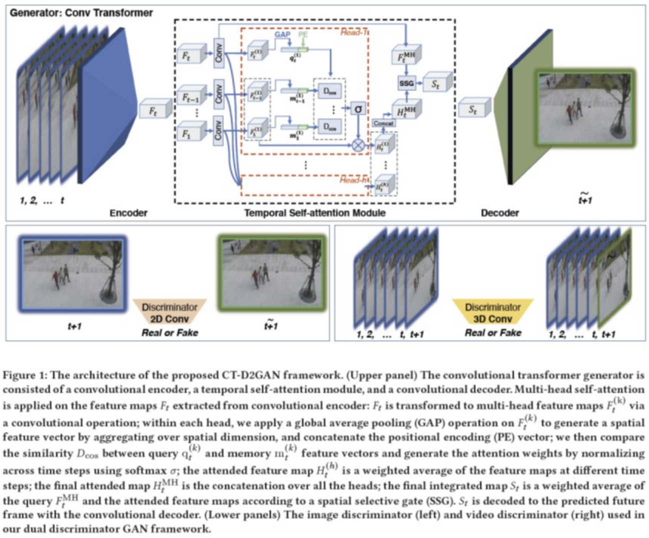


Figure 1: The architecture of the proposed CT-D2GAN framework. (Upper panel) The convolutional transformer generator is consisted of a convolutional encoder, a temporal self-attention module, and a convolutional decoder. Multi-head self-attention is applied on the feature maps F_t extracted from convolutional encoder: F_t is transformed to multi-head feature maps $F_t^{(k)}$ via a convolutional operation; within each head, we apply a global average pooling (GAP) operation on $F_t^{(k)}$ to generate a spatial feature vector by aggregating over spatial dimension, and concatenate the positional encoding (PE) vector; we then compare the similarity $D_{\cos}$ between query $\mathrm{q}_t^{(k)}$ and memory $\mathrm{m}_t^{(k)}$ feature vectors and generate the attention weights by normalizing across time steps using softmax σ; the attended feature map $H_t^{(h)}$ is a weighted average of the feature maps at different time steps; the final attended map H_t^{MH} is the concatenation over all the heads; the final integrated map S_t is a weighted average of the query F_t^{MH} and the attended feature maps according to a spatial selective gate (SSG). S_t is decoded to the predicted future frame with the convolutional decoder. (Lower panels) The image discriminator (left) and video discriminator (right) used in our dual discriminator GAN framework.

Figure 12: Architecture of CT-D2GAN (Feng et al., 2021)

This was done via the structure of (d'Ascoli et al., 2021) (figure 13) so input data chooses the optimum input based on the data source.

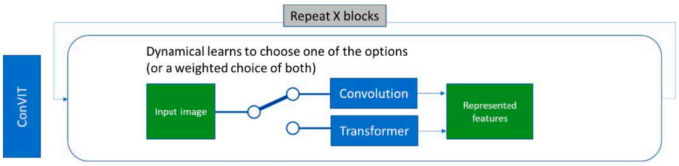

Repeat X blocks
ConVIT
Dynamical learns to choose one of the options
(or a weighted choice of both)
Input image
Convolution
Transformer
Represented features

Figure 13:ConVit Architecture (d'Ascoli et al., 2021)

This model was designed with the assumption each chosen part is theoretically possible for connection. Further research is required for accurate confirmation.

This model is based on 3 main parts. A transformer-based generator and discriminator based on (H. Wu et al., 2021) (Figure 11) with a LSTM based cconvolutional temporal GAN based attention module (*TGAN2*) (Saito et al., 2020) (figure 14).

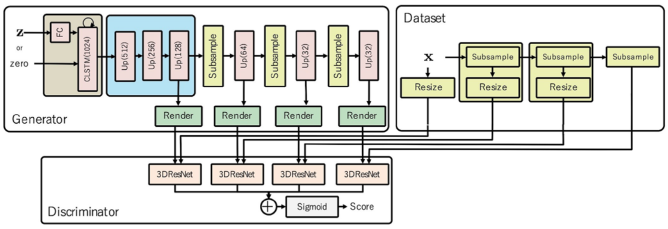


Figure 14: T-GAN2 Network configuration. CLSTM(C)" represents the convolutional LSTM with C channels and 3 × 3 kernel. "Up(C)" means the upsampling block that returns a feature map with C channels and twice the resolution of the input (Saito et al., 2020)

Finally, a CNN based dual discriminator was used for local consistency of inputs. To preserve memory efficiency, Tokenization was done via methods of (H. Wu et al., 2021) as previous work of *T2T* (token 2 token) (Yuan et al., 2021) that adopts a deep narrow structure to increase layer depth proved computationally intensive.

T-GAN2 (Saito et al., 2020) positional LSTM are added to the Temporal self-attention module in (Feng et al., 2021) as being a hyperparameter-less method means reduced batch size when scaling for output data. This sub-generator input is split L sub-generators (eq1).

For increased $\mathbf{x} = \_\, g_{RL} \circ g_{AL} \circ g_{AL-1} \circ \cdot \cdot \cdot \circ g_{A1}$ $\_\,(\mathbf{z}).$

Eq 1: L sub generator and rendering with singular G (Saito et al., 2020)

For hardening, RBF based discriminators are used on the CNN based classifiers to ensure accurate results. As shown in (H. Zhang et al., 2019) replacement of convolutions with attention-based methods gave more accurate results. Adapting the method of (Feng et al., 2021) shows this is indeed possible, by replacing the convolutional encoder and decoder with that of (Yu & Wang, 2021) (Figure 15).

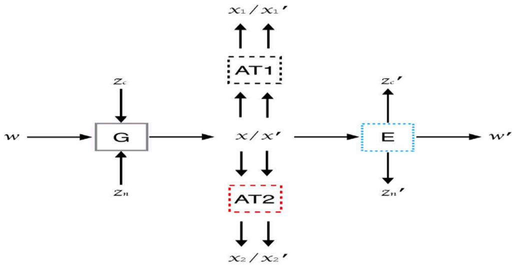
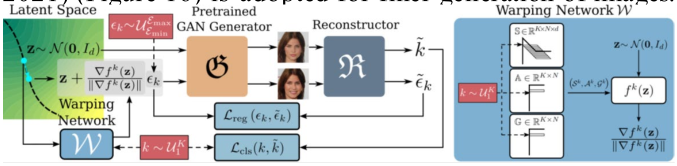


Figure 15: Architecture of adaptable GAN encoders (MTV-TSA) (Yu & Wang, 2021)

To further improve the latent paths, the method (Tzelepis et al., 2021) (Figure 16) is adopted for finer generation of images.



Latent Space
$\epsilon_k \sim \mathcal{U}_{\mathcal{E}_{\min}}^{\mathcal{E}_{\max}}$
Pretrained GAN Generator
Reconstructor
Warping Network $\mathcal{W}$
$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_d)$
$\mathbf{z} + \frac{\nabla f^k(\mathbf{z})}{\|\nabla f^k(\mathbf{z})\|} \epsilon_k$
$\mathfrak{G}$
$\mathfrak{R}$
$\tilde{k}$
$\tilde{\epsilon}_k$
Warping Network
$\mathcal{W}$
$\mathcal{L}_{\text{reg}}(\epsilon_k, \tilde{\epsilon}_k)$
$k \sim \mathcal{U}_1^K$
$\mathcal{L}_{\text{cls}}(k, \tilde{k})$
$S \in \mathbb{R}^{K \times N \times d}$
$A \in \mathbb{R}^{K \times N}$
$G \in \mathbb{R}^{K \times N}$
$(\mathcal{S}^k, \mathcal{A}^k, \mathcal{G}^k)$
$f^k(\mathbf{z})$
$\frac{\nabla f^k(\mathbf{z})}{\|\nabla f^k(\mathbf{z})\|}$

Figure 16: Architecture of WarpedGANspace (Tzelepis et al., 2021)

As the discriminator of (Feng et al., 2021) is based on CNN network of (Isola et al., 2017), discriminator replacement of *AG-IDS* (Shieh et al., 2021) (Figure 17) is done to ensure output accuracy and resistance.

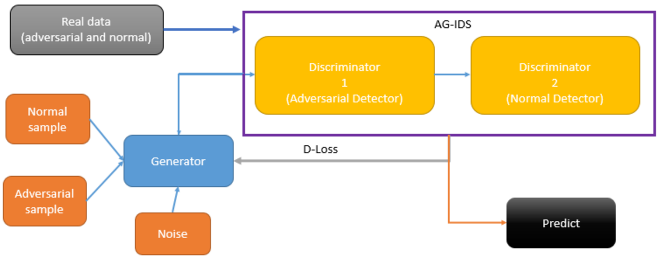


Figure 17: Architecture of *AG-IDS* (Shieh et al., 2021)

This is matched with 2D dual stage classifiers of *TAC-GAN* (Gong et al., 2019) (Figure 18) to ensure accuracy and attached to each model layer.

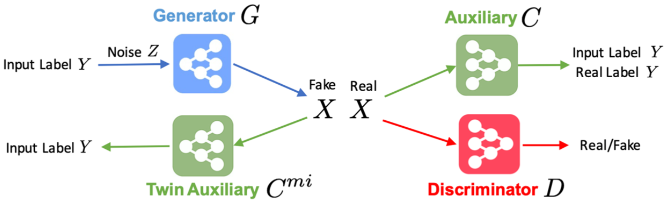


Figure 18: The architecture of TAC-GAN (Gong et al., 2019)

Loss function is based on W-GAN (Arjovsky et al., 2017) variant W-GAN-adv (J. Wu et al., 2018) (eq 2) chosen because of scalability thanks to the removal of the 1-Lipschitz constraint. This is combined with *Adam* (Kingma & Ba, 2017) for optimization.

$$LDIV = E_{x \sim P_r}[f(x)] - E_{\tilde{x} \sim P_g}[f(\tilde{x})] + k\, E_{\hat{x} \sim P_u}[\|\nabla f(\hat{x})\|^p],$$

Eq2: Wasserstein divergence GAN (J. Wu et al., 2018)

For comparison to other networks numerical results were collected from chosen literatures (figures 19 and 20) official results on both MNIST and CIFAR10 datasets. These were chosen on both FGSM and poisoning attacks.

Results were calculated by comparing the given result in original literature out of 100 to obtain percentages with deviances (e.g.: success is reduced if poisoning obtains over 10%) added to calculations to leverage averaging to successfully calculate theoretical results.

Results were extrapolated via successful calculations of this data. Indepth analysis including loss calculations and joint distribution matching were not included in result calculations and is something for future work.

Literature Name	**Data set**							
	MNIST	CIFAR10	FGSM	**Deep RBF**	CNN GAN		T-GAN	
(Zadeh et al., 2018)	🟩	🟩	🟦	76%	2.22			
(Aldahdooh et al., 2021)	🟩	🟩	🟦	77%				
(de Alfaro, 2018)	🟩	🟩	🟦	94.90				
(H. Wu et al., 2021)	🟩	🟩	🟦	N/A			99.39	
(Carlini & Wagner, 2017)	🟩	🟩	🟦	N/A	42	50		
(J. Li, Cao, Zhang, Chen, et al., 2021)	🟩	🟩	🟦	N/A				92.34
(Madry et al., 2019)	🟩	🟩	🟦	N/A	95.6	56.1		

Figure 19: Reference literature for FGSM attacks

Data from the selected original literature was analyzed based on if literature referenced one or both attacks against chosen datasets. Success of additions such as RBF networks was added towards result calculation based on results against chosen datasets.

Literature	Success rate
(Carnerero-Cano et al., n.d.)	24
(Suya et al., 2021)	96.9
(Burruss, 2021.)	0.25

Figure 20: Reference Literature for Poisoning attacks

Compared to previous literature, calculations are based purely on a numerical result comparison instead of in-depth outcomes as seen in (Benz et al., 2021.) as further analysis is required.

IV. Results and discussion

Results show that the combination as per section 10 are successful with high resistance against adversarial attacks. Confirming results of (Aldahdooh et al., 2021) that increasing attention blocks reduces attack transferability. The addition of deep RBF networks within the model significantly improves resistance to variants of FGSM attacks such as I-FGSM (Kurakin et al., 2017).

Showing fantastic results of 98% successful defense against using FGSM on the MNSIT dataset. The CIFAR10 dataset also showed well against the transformer-based networking showing with the input of the transfer CB-GAN scored a fantastic 99.6 FGSM attack resistance compared to (Wu et al., 2021) at only 99.3%. This shows the increased robustness of this new model outperforms previous works significantly. Small datasets such as MNIST show increased resistance even after initial targeting with more runs showing improved results.

These results confirm adding RBF networks are initially more susceptible to adversarial attacks of FGSM but greatly improved after significant distortion has been achieved by rejecting the adversarial inputs. Attack transferability is greatly reduced via the use of transformers mixed with convolutional networks of this model as this model shows a 18% success rate of FGSM attack transfer compared to previous works of only 20%.

When compared to previous CNN based architectures that suffer from shift invariance (Wiles et al., 2021) making them more susceptible towards adversarial attacks. The inclusion of latent space upgrades thanks to *WarpedGANspace* (Tzelepis et al., 2021) greatly improves resistance to FGSM but not with initial training as due to the linear nature of FGSM this makes attacks more successful.

Autoencoders of standard transformer models diminish under normal FGSM attacks showing only 63% success on clean datasets (Li et al., 2021) but due to the dual encoder/decoder setup of *MTV-TSA* (Yu & Wang, 2021) and double classifiers of *TAC-GAN* (Gong et al., 2019) this is greatly improved by 5% with a score of 98% on the *CIFAR10* dataset.

Feature representation via the use of *TGAN2* (Saito et al., 2020) improves robustness by decreasing the requirements of learning via the original feature space is tightened with *WarpedGANspace* (Tzelepis et al., 2021).

As found in (Li et al., 2021) the combination of Auto-encoders (Kramer, 1991) and U-net (Ronneberger et al., 2015) is shown to greatly improve defense transformation learning. This combination of features greatly improves resistance to FGSM attacks (Goodfellow et al., 2015) with the reduction of transfer learning-based attacks common against FGSM attacks and variants.

Defense against poisoning attacks is improved greatly via the use of deep RBF frameworks within the model.
Compared to previous methods such as AC that requires large amounts of data to be poisoned. The inclusion of RBF network included on the discriminators greatly improves poisoning resistance. This is further strengthened via the dual stage classifiers and improved latent paths of CB-GAN. The RBF outlier detection method (Burruss,2021.) used within CB-GAN cleaned the data when coupled with dual stage classifiers of (Hu et al., 2019) which are used as identifiers for poisoned data.

Traditional CNN are interconnected fully and thus more susceptible towards poisoning attacks with over a 95% success rate (Burruss, 2021.). The implementation into *CBGAN* shows a 98% reduction rate when poisoned with 20% of input data. Significant increase of latent path smoothing via (Tzelepis et al., 2021) reduces susceptibility of poisoning attack on both the CIFAR10 and MNIST datasets.

The poisons points within models are also the key points within the latent vector space and under normal network conditions would be susceptible to attacks. This is shown an improvement in comparison to both (Koh et al., 2018) and (Suya et al., 2021) results of only 97.5 and 96.9 respectively.

Because data is collected via theoretical results it is unknown if increased target model infection rates or increased linear paths used as the nonlinear paths used via the RBF networks within this model would show increased error rates given larger data poison samples of input. The RBF outlier detection method used within this model traditionally shows great performance on small, poisoned data samples of less than 10% but it is unknown the success rate if poisoned with large amounts of data such as 50% or more.

This is significantly countered due to higher false positive success rates of RBF network embedded with the discriminators of *CB-GAN.*

Due to the large computational requirements, full testing was not undertaken. Preliminary setup testing on a basic GAN network against the MNIST dataset was undertaken on an Intel core i7 3770 (4 core/8 thread 3.5ghz) with 32GB of memory and a Nvidia GTX Titan X 12GB on ubuntu variant popOS! Version 20.04, this proved to be unsuccessful.

Implementation of this model was extremely difficult with comparison results and datasets of selected parts been tested against 4-8 Nvidia GPUs including high-capacity models, Telsa V100 and Geforce GTX Titan Xp. This proved to be unfeasible to replicate in any modest capacity. Theoretical initial testing has shown great success on MSNT and CIFAR10 datasets.

Showing high resistance to both FGSM and integrity poisoning attacks with encouraging results, with a 25% increase in defense compared to both Li et al. and Carnerero-Cano et al in both FGSM and poisoning attacks.

The integration of multimedia focused generative networks into the model was chosen due to the underlying basis of input data been images or sometimes video with images been in 50% of all popular datasets (Alqahtani et al., 2021). This decision of using multimedia datasets provides a basis for the addition of IDS focused inclusions such as improved forms of MALGAN12

Flexibility of the model is significantly improved with integration of *TGAN2* compression.

This came at a significant increase to training time by up to guesstimated 55% in comparison to standard training time using MNST dataset estimated after extensive literature review. As larger datasets are usually required, the success on small datasets of MNIST proved to be a great success especially against fast gradient signal method attacks.

Showing that combing features with selected strengths proved high resistance towards FGSM based attacks is possible as the weakness of traditional generative adversarial networks are overcome.

The integration of RBF networks and latent space improvement immensely strengthened poisoning attack resistance. As shown in comparison results of figure 20, improvements are consistent with others results but require more accurate testing and alternative datasets for accurate conclusions.
Based on the above results (Figure 21), *CB-GAN* shows immense potential for implantation into cyber security systems as a defense against these common attacks.

The use of Deep RBF within the models shows great potential for further success when properly tested removing the requirements of clean datasets, meaning datasets can be used without issue within many different environments. Showing fast gradient signal method and poisoning integrity attacks that whilst common, are defeatable adversarial attacks.

With the success of CB-GAN, this has shown hybrid approaches are suitable for the new world of ever evolving cyber threats or training measures for new attack variants.

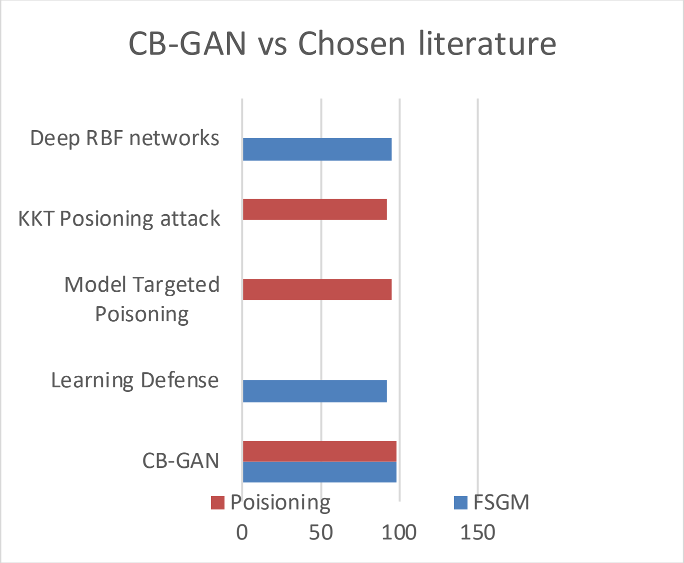


Figure 21: CB-GAN vs chosen literature against FGSM and poisoning attacks

Several flaws are present within this new model. The use of repeated datasets of MNIST and CIFAR10 which while useful are not scalable and so it is unknown what the true performance of the model is, as transformer based generative adversarial networks traditionally prefer large datasets including CIFAR100 or others used in systems such as other convolutional transfer-based models such as *TransGAN* (Jiang et al., 2020) or resistance to transformer adversarial attacks such as token attacks (Joshi et al., 2021).

V. Conclusion

Successful theoretical implementation of this new state of the art GAN model showed that a wide range of applications for a range of uses are possible.

With the evolving threat of malware and ransomware attacks (Shu et al., 2020) use of this model as integration into a first line of defense in government or corporate applications could be applied against ML based IDS attacks. The chosen combinations within this algorithm proved the initial goal of strengths beating out weaknesses was successfully done.

Other areas such as healthcare, with applications in cancer diagnoses and research or medication resistance (Jeong et al., 2018) and reactions along with resistance to untested attacks. Both adversarial attacks are significantly reduced, while transformer based adversarial attacks (Benz et al., 2021.) remain untested.

In further versions of this model, transformers could be added to more parts of the model (S. Li, Chen, He, & Hsieh, 2021) replacing existing CNN based networks for further resistant to new attacks or applications.

Other methods based on the use of unrolled GAN (Metz et al., 2017) could be used to reduce the chances mode collapse. Improved MALGAN(Kawai et al., 2019) could be implemented for malware and virus detection but further research is required for successful implementation, Image generation for detection testing could be further improved with the use of GIRAFFE (Niemeyer & Geiger,2021.), PaWs (Assran et al., 2021) and DINO (Caron et al., 2021) for improved self-learning capabilities.

Further research is required for successful practical implementation of the combined GAN algorithm as theoretical results look promising.

High training cost contributed significantly for implementation failure. For confirmed testing within IDS environments, this new algorithm of *CB-GAN* could be tested on IDS datasets mentioned in (Ahmad et al., 2021) including *NSL-KDD* (Revathi & Malathi, 2013) and *UNSW-NB15* (Moustafa, n.d.) or more current datasets including CSE-CIC-IDS2018 and *CIC-Bell-DNS-EXF-2021* (Ghurab et al., 2021; Sharafaldin et al., 2018)

Acknowledgment

The Author would like to thank the following people for assistance with this paper. Shantanu pal and Zahra jadidi for excellent discussion points, contributing for idea in the paper and Levy Hu, PhD for assistance in understanding RBF network implementation in the radial basis flow paper.

References

[1] Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights, 1(1), 100004. https://doi.org/10.1016/j.jjimei.2020.100004

[2] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 32(1). https://doi.org/10.1002/ett.4150

[3] Aldahdooh, A., Hamidouche, W., & Deforges, O. (2021). Reveal of Vision Transformers Robustness against Adversarial Attacks. ArXiv:2106.03734 [Cs]. http://arxiv.org/abs/2106.03734

[4] Alqahtani, H., Kavakli-Thorne, M., & Kumar, G. (2021). Applications of Generative Adversarial Networks (GANs): An Updated Review. Archives of Computational Methods in Engineering, 28(2), 525–552. https://doi.org/10.1007/s11831-019-09388-y

[5] Anderson, J. P. (n.d.). Computer Security Threat Monitoring and Surveillance. 56.

[6] Araujo, A., Meunier, L., Pinot, R., & Negrevergne, B. (2020). Advocating for Multiple Defense Strategies against Adversarial Examples. ArXiv:2012.02632 [Cs]. http://arxiv.org/abs/2012.02632

[7] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. ArXiv:1701.07875 [Cs, Stat]. http://arxiv.org/abs/1701.07875

[8] Arora, S., Ge, R., Liang, Y., Ma, T., & Zhang, Y. (2017). Generalization and Equilibrium in Generative Adversarial Nets (GANs). ArXiv:1703.00573 [Cs, Stat]. http://arxiv.org/abs/1703.00573

[9] Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Ballas, N., & Rabbat, M. (2021). Semi-Supervised Learning of Visual Features by Non-Parametrically Predicting View Assignments with Support Samples. ArXiv:2104.13963 [Cs, Eess]. http://arxiv.org/abs/2104.13963

[10] Bai, T., Zhao, J., Zhu, J., Han, S., Chen, J., Li, B., & Kot, A. (2021). AI-GAN: Attack-Inspired Generation of Adversarial Examples. 2021 IEEE International Conference on Image Processing (ICIP), 2543–2547. https://doi.org/10.1109/ICIP42928.2021.9506278

[11] Baioletti, M., Coello, C. A. C., Di Bari, G., & Poggioni, V. (2020). Multi-objective evolutionary GAN. Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion, 1824–1831. https://doi.org/10.1145/3377929.3398138

[12] Bangui, H., Ge, M., & Buhnova, B. (2021). A hybrid machine learning model for intrusion detection in VANET. Computing. https://doi.org/10.1007/s00607-021-01001-0

[13] Barrett, B. (n.d.). A New Kind of Ransomware Tsunami Hits Hundreds of Companies. Wired. Retrieved November 10, 2021, from https://www.wired.com/story/kaseya-supply-chain-ransomware-attack-msps/

[14] Benz, P., Zhang, C., Ham, S., & Karjauv, A. (n.d.). Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs. 6.

[15] Biggio, B., Nelson, B., & Laskov, P. (n.d.). Poisoning Attacks against Support Vector Machines. 8.

[16] Biswas, S. K. (n.d.). Intrusion Detection Using Machine Learning: A Comparison Study. 15.

[17] Bitaab, M., & Hashemi, S. (2017). Hybrid Intrusion Detection: Combining Decision Tree and Gaussian Mixture Model. 2017 14th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC), 8–12. https://doi.org/10.1109/ISCISC.2017.8488375

[18] Brock, A., Donahue, J., & Simonyan, K. (2019). Large Scale GAN Training for High Fidelity Natural Image Synthesis. ArXiv:1809.11096 [Cs, Stat]. http://arxiv.org/abs/1809.11096

[19] Broomhead, D. S. (n.d.). Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Nctworks. 39.

[20] Burruss, M. P. (n.d.). Enhancing the Robustness of Deep Neural Networks Against Security Threats Using Radial Basis Functions. 49.

[21] Burruss, M., Ramakrishna, S., & Dubey, A. (2021). Deep-RBF Networks for Anomaly Detection in Automotive Cyber-Physical Systems. ArXiv:2103.14172 [Cs]. http://arxiv.org/abs/2103.14172

[22] Carlini, N., & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. ArXiv:1608.04644 [Cs]. http://arxiv.org/abs/1608.04644

[23] Carnerero-Cano, J., Pfitzner, B., & Lupu, E. C. (n.d.). Poisoning Attacks with Generative Adversarial Nets. 22.

[24] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. ArXiv:2104.14294 [Cs]. http://arxiv.org/abs/2104.14294

[25] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial Attacks and Defences: A Survey. ArXiv:1810.00069 [Cs, Stat]. http://arxiv.org/abs/1810.00069

[26] Chavdarova, T., & Fleuret, F. (2017). SGAN: An Alternative Training of Generative Adversarial Networks. ArXiv:1712.02330 [Cs, Stat]. http://arxiv.org/abs/1712.02330

[27] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating Long Sequences with Sparse Transformers. ArXiv:1904.10509 [Cs, Stat]. http://arxiv.org/abs/1904.10509

[28] Crecchi, F., Melis, M., Sotgiu, A., Bacciu, D., & Biggio, B. (2020). FADER: Fast Adversarial Example Rejection. ArXiv:2010.09119 [Cs]. http://arxiv.org/abs/2010.09119

[29] Creswell, A., Bharath, A. A., & Sengupta, B. (2017). LatentPoison—Adversarial Attacks On The Latent Space. ArXiv:1711.02879 [Cs]. http://arxiv.org/abs/1711.02879

[30] Cretu, G. F., Stavrou, A., Locasto, M. E., Stolfo, S. J., & Keromytis, A. D. (2008). Casting out Demons: Sanitizing Training Data for Anomaly Sensors. 2008 IEEE Symposium on Security and Privacy (Sp 2008), 81–95. https://doi.org/10.1109/SP.2008.11

[31] d'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G., & Sagun, L. (2021). ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. ArXiv:2103.10697 [Cs, Stat]. http://arxiv.org/abs/2103.10697

[32] Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). CoAtNet: Marrying Convolution and Attention for All Data Sizes. ArXiv:2106.04803 [Cs]. http://arxiv.org/abs/2106.04803

[33] de Alfaro, L. (2018). Neural Networks with Structural Resistance to Adversarial Attacks. ArXiv:1809.09262 [Cs, Stat]. http://arxiv.org/abs/1809.09262

[34] Debar, H., Dacier, M., & Wespi, A. (1999). Towards a taxonomy of intrusion-detection systems. Computer Networks, 31(8), 805–822. https://doi.org/10.1016/S1389-1286(98)00017-6

[35] Degeler, V., French, R., & Jones, K. (2016). Self-Healing Intrusion Detection System Concept. 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 351–356. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.27

[36] Deldjoo, Y., Noia, T. D., & Merra, F. A. (2021). A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial Networks. ACM Computing Surveys, 54(2), 1–38. https://doi.org/10.1145/3439729

[37] Denning, D. E. (1987). An Intrusion-Detection Model. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, SE-13, 17.

[38] Donahue, J., Krähenbühl, P., & Darrell, T. (2017). Adversarial Feature Learning. ArXiv:1605.09782 [Cs, Stat]. http://arxiv.org/abs/1605.09782

[39] Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). GANSYNTH: ADVERSARIAL NEURAL AUDIO SYNTHESIS. 17.

[40] Fang, H., Deng, W., Zhong, Y., & Hu, J. (2020). Triple-GAN: Progressive Face Aging with Triple Translation Loss. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 3500–3509. https://doi.org/10.1109/CVPRW50498.2020.00410

[41] Feng, X., Song, D., Chen, Y., Chen, Z., Ni, J., & Chen, H. (2021). Convolutional Transformer based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection. Proceedings of the 29th ACM International Conference on Multimedia, 5546–5554. https://doi.org/10.1145/3474085.3475693

[42] Ferdowsi, A., & Saad, W. (2021). Brainstorming Generative Adversarial Networks (BGANs): Towards Multi-Agent Generative Models with Distributed Private Datasets. ArXiv:2002.00306 [Cs, Stat]. http://arxiv.org/abs/2002.00306

[43] Gan, M., Peng, H., & Dong, X. (2012). A hybrid algorithm to optimize RBF network architecture and parameters for nonlinear time series prediction. Applied Mathematical Modelling, 36(7), 2911–2919. https://doi.org/10.1016/j.apm.2011.09.066

[44] Ghosh, A., Kulharia, V., & Namboodiri, V. (2016). Message Passing Multi-Agent GANs. ArXiv:1612.01294 [Cs]. http://arxiv.org/abs/1612.01294

[45] Ghurab, M., Gaphari, G., Alshami, F., Alshamy, R., & Othman, S. (2021). A Detailed Analysis of Benchmark Datasets for Network Intrusion

Detection System. Asian Journal of Research in Computer Science, 14–33. https://doi.org/10.9734/ajrcos/2021/v7i430185

[46] Glenny, M. (2021, May 17). Colonial Pipeline cyber attack a warning of worse to come. Australian Financial Review. https://www.afr.com/companies/energy/the-colonial-pipeline-cyber-attack-is-a-warning-of-worse-to-come-20210517-p57skj

[47] Gong, M., Xu, Y., Li, C., Zhang, K., & Batmanghelich, K. (2019). Twin Auxiliary Classifiers GAN. ArXiv:1907.02690 [Cs, Stat]. http://arxiv.org/abs/1907.02690

[48] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. ArXiv:1406.2661 [Cs, Stat]. http://arxiv.org/abs/1406.2661

[49] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. ArXiv:1412.6572 [Cs, Stat]. http://arxiv.org/abs/1412.6572

[50] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139–144. https://doi.org/10.1145/3422622

[51] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved Training of Wasserstein GANs. ArXiv:1704.00028 [Cs, Stat]. http://arxiv.org/abs/1704.00028

[52] Hamid, Y., Sugumaran, M., & Balasaraswathi, V. (2016). IDS Using Machine Learning—Current State of Art and Future Directions. British Journal of Applied Science & Technology, 15(3), 1–22. https://doi.org/10.9734/BJAST/2016/23668

[53] Haq, N. F., Onik, A. R., & Shah, F. M. (2015). An ensemble framework of anomaly detection using hybridized feature selection approach (HFSA). 2015 SAI Intelligent Systems Conference (IntelliSys), 989–995. https://doi.org/10.1109/IntelliSys.2015.7361264

[54] Helman, P., Liepins, G., & Richards, W. (n.d.). Foundations of Intrusion Detection. 7.

[55] Hsu, G.-S., Xie, R.-C., & Chen, Z.-T. (2021). Wasserstein Divergence GAN With Cross-Age Identity Expert and Attribute Retainer for Facial Age Transformation. IEEE Access, 9, 39695–39706. https://doi.org/10.1109/ACCESS.2021.3062499

[56] Hu, L., Wang, W., Xiang, Y., & Zhang, J. (2020). Flow Field Reconstructions with GANs based on Radial Basis Functions. ArXiv:2009.02285 [Physics]. http://arxiv.org/abs/2009.02285

[57] Hu, W., & Tan, Y. (2017). Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN. ArXiv:1702.05983 [Cs]. http://arxiv.org/abs/1702.05983

[58] Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5967–5976. https://doi.org/10.1109/CVPR.2017.632

[59] Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. 2018 IEEE Symposium on Security and Privacy (SP), 19–35. https://doi.org/10.1109/SP.2018.00057

[60] Jiang, L., Qiao, K., Qin, R., Wang, L., Yu, W., Chen, J., Bu, H., & Yan, B. (2020). Cycle-Consistent Adversarial GAN: The Integration of Adversarial Attack and Defense. Security and Communication Networks, 2020, 1–9. https://doi.org/10.1155/2020/3608173

[61] Joshi, A., Jagatap, G., & Hegde, C. (2021). Adversarial Token Attacks on Vision Transformers. ArXiv:2110.04337 [Cs]. http://arxiv.org/abs/2110.04337

[62] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION. 26.

[63] Karras, T., Laine, S., & Aila, T. (n.d.). A Style-Based Generator Architecture for Generative Adversarial Networks. 10.

[64] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (n.d.). Training Generative Adversarial Networks with Limited Data. 11.

[65] Kawai, M., Ota, K., & Dong, M. (2019). Improved MalGAN: Avoiding Malware Detector by Leaning Cleanware Features. 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 040–045. https://doi.org/10.1109/ICAIIC.2019.8669079

[66] Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 41(4), 1690–1700. https://doi.org/10.1016/j.eswa.2013.08.066

[67] Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [Cs]. http://arxiv.org/abs/1412.6980

[68] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. ArXiv:1312.6114 [Cs, Stat]. http://arxiv.org/abs/1312.6114

[69] Koh, P. W., Steinhardt, J., & Liang, P. (2018). Stronger Data Poisoning Attacks Break Data Sanitization Defenses. ArXiv:1811.00741 [Cs, Stat]. http://arxiv.org/abs/1811.00741

[70] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

[71] Kramer, M. A. (1991). Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. AIChE Journal, 37(2), 11.

[72] Krizhevsky et al., A. (n.d.). CIFAR-10 and CIFAR-100 datasets. Retrieved November 13, 2021, from https://www.cs.toronto.edu/~kriz/cifar.html

[73] Kumar, G., Kumar, K., & Sachdeva, M. (2010). The use of artificial intelligence based techniques for intrusion detection: A review. Artificial Intelligence Review, 34(4), 369–387. https://doi.org/10.1007/s10462-010-9179-5

[74] Kumar, S., & Tsvetkov, Y. (n.d.). End-to-End Differentiable GANs for Text Generation. 11.

[75] Kurach, K., Lucic, M., Zhai, X., Michalski, M., & Gelly, S. (2019). A Large-Scale Study on Regularization and Normalization in GANs. 10.

[76] Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial Machine Learning at Scale. ArXiv:1611.01236 [Cs, Stat]. http://arxiv.org/abs/1611.01236

[77] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. ArXiv:1512.09300 [Cs, Stat]. http://arxiv.org/abs/1512.09300

[78] Lecun, Y. (1998). Gradient-Based Learning Applied to Document Recognition. PROCEEDINGS OF THE IEEE, 86(11), 47.

[79] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation, 1(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

[80] Lee, K., Chang, H., Jiang, L., Zhang, H., Tu, Z., & Liu, C. (2021). ViTGAN: Training GANs with Vision Transformers. ArXiv:2107.04589 [Cs, Eess]. http://arxiv.org/abs/2107.04589

[81] Lee, W., & Stolfo, S. J. (2000). A framework for constructing features and models for intrusion detection systems. ACM Transactions on Information and System Security, 3(4), 227–261. https://doi.org/10.1145/382912.382914

[82] Li, D., Chen, D., Shi, L., Jin, B., Goh, J., & Ng, S.-K. (2019). MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. ArXiv:1901.04997 [Cs, Stat]. http://arxiv.org/abs/1901.04997

[83] Li, J., Cao, J., Zhang, S., Xu, Y., Chen, J., & Tan, M. (2021). Internal Wasserstein Distance for Adversarial Attack and Defense. ArXiv:2103.07598 [Cs]. http://arxiv.org/abs/2103.07598

[84] Li, J., Cao, J., Zhang, Y., Chen, J., & Tan, M. (2021). Learning Defense Transformers for Counterattacking Adversarial Examples. ArXiv:2103.07595 [Cs]. http://arxiv.org/abs/2103.07595

[85] Li, S., Chen, X., He, D., & Hsieh, C.-J. (2021). Can Vision Transformers Perform Convolution? ArXiv:2111.01353 [Cs]. http://arxiv.org/abs/2111.01353

[86] Liu, Y., Chen, X., Liu, C., & Song, D. (2017). Delving into Transferable Adversarial Examples and Black-box Attacks. ArXiv:1611.02770 [Cs]. http://arxiv.org/abs/1611.02770

[87] Lloyd, S., & Weedbrook, C. (2018). Quantum generative adversarial learning. Physical Review Letters, 121(4), 040502. https://doi.org/10.1103/PhysRevLett.121.040502

[88] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019). Towards Deep Learning Models Resistant to Adversarial Attacks. ArXiv:1706.06083 [Cs, Stat]. http://arxiv.org/abs/1706.06083

[89] Marrs, A. D., & Webb, A. R. (1998). Exploratory Data Analysis Using Radial Basis Function Latent Variable Models. 7.

[90] Mei, S., & Zhu, X. (n.d.). Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners. 17.

[91] Metz, L., Poole, B., Pfau, D., & Sohl-Dickstein, J. (2017). Unrolled Generative Adversarial Networks. ArXiv:1611.02163 [Cs, Stat]. http://arxiv.org/abs/1611.02163

[92] Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2021). Intriguing Properties of Vision Transformers. ArXiv:2105.10497 [Cs]. http://arxiv.org/abs/2105.10497

[93] Nash, J. F. (1950). Equilibrium Points in n-Person Games. Proceedings of the National Academy of Sciences of the United States of America, 36(1), 48–49.

[94] Nelson, B., Barreno, M., Chi, F. J., Rubinstein, B. I. P., Saini, U., Sutton, C., Joseph, A. D., Tygar, J. D., & Xia, K. (n.d.). Exploiting Machine Learning to Subvert Your Spam Filter. 10.

[95] Niemeyer, M., & Geiger, A. (n.d.). GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. 12.

[96] Piplai, A., Chukkapalli, S. S. L., & Joshi, A. (2020). NAttack! Adversarial Attacks to bypass a GAN based classifier trained to detect Network intrusion. 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), 49–54. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS49724.2020.00020

[97] Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2019). A Survey on Deep Learning: Algorithms, Techniques, and Applications. ACM Computing Surveys, 51(5), 1–36. https://doi.org/10.1145/3234150

[98] Powell, M. J. D. (1977). Restart procedures for the conjugate gradient method. Mathematical Programming, 12(1), 241–254. https://doi.org/10.1007/BF01593790

[99] Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. ArXiv:1511.06434 [Cs]. http://arxiv.org/abs/1511.06434

[100]Rawat, A., Levacher, K., & Sinn, M. (2021). The Devil is in the GAN: Defending Deep Generative Models Against Backdoor Attacks. ArXiv:2108.01644 [Cs]. http://arxiv.org/abs/2108.01644

[101]Revathi, S., & Malathi, D. A. (2013). A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection. International Journal of Engineering Research, 2(12), 6.

[102]Rice, L., Wong, E., & Kolter, J. Z. (2020). Overfitting in adversarially robust deep learning. ArXiv:2002.11569 [Cs, Stat]. http://arxiv.org/abs/2002.11569

[103]Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv:1505.04597 [Cs]. http://arxiv.org/abs/1505.04597

[104]Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., & Cohen, J. P. (2020). Revisiting Training Strategies and Generalization Performance in Deep Metric Learning. ArXiv:2002.08473 [Cs]. http://arxiv.org/abs/2002.08473

[105]Ruppert, L. "LiLi." (2018, October 24). Ubisoft Servers Hit With DDoS Attack, Multiplayer Games Having Issues [Blog Post]. GAMING. https://comicbook.com/gaming/news/ubisoft-servers-down-ddos-attack/

[106]Sabuhi, M., Zhou, M., Bezemer, C.-P., & Musilek, P. (2021). Applications of Generative Adversarial Networks in Anomaly Detection: A Systematic Literature Review. ArXiv:2110.12076 [Cs]. http://arxiv.org/abs/2110.12076

[107]Saito, M., Saito, S., Koyama, M., & Kobayashi, S. (2020). Train Sparsely, Generate Densely: Memory-Efficient Unsupervised Training of High-Resolution Temporal GAN. International Journal of Computer Vision, 128(10–11), 2586–2606. https://doi.org/10.1007/s11263-020-01333-y

[108]Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved Techniques for Training GANs. ArXiv:1606.03498 [Cs]. http://arxiv.org/abs/1606.03498

[109]Samuel, A. L. (n.d.-a). Some Studies in Machine Learning Using the Game of Checkers. 20.

[110]Samuel, A. L. (n.d.-b). Some Studies in Machine Learning Using the Game of Checkers. II-Recent Progress. MACHINE LEARNING, 17.

[111]Saxena, D., & Cao, J. (2021). Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. ACM Computing Surveys, 54(3), 1–42. https://doi.org/10.1145/3446374

[112]Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization: Proceedings of the 4th International Conference on Information Systems Security and Privacy, 108–116. https://doi.org/10.5220/0006639801080116

[113]Sharma, U. C., Zhao, K., Mentkowski, K., Sonkawade, S. D., Karthikeyan, B., Lang, J. K., & Ying, L. (2021). Modified GAN Augmentation Algorithms for the MRI-Classification of Myocardial Scar Tissue in Ischemic Cardiomyopathy. Frontiers in Cardiovascular Medicine, 8, 726943. https://doi.org/10.3389/fcvm.2021.726943

[114]Shieh, C.-S., Lin, W.-W., Nguyen, T.-T., Huang, Y.-L., Horng, M.-F., Lo, C.-C., & Tu, K.-M. (2021, August). Detection of Adversarial DDoS Attacks Using Generative Adversarial Networks with Dual Discriminators. The 7th World Congress on Electrical Engineering and Computer Systems and Science. https://doi.org/10.11159/cist21.121

[115]Shu, D., Leslie, N. O., Kamhoua, C. A., & Tucker, C. S. (2020). Generative adversarial attacks against intrusion detection systems using active learning. Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning, 1–6. https://doi.org/10.1145/3395352.3402618

[116]Sinha, S., Zhang, H., Goyal, A., Bengio, Y., Larochelle, H., & Odena, A. (n.d.). Small-GAN: Speeding up GAN Training using Core-Sets. 11.

[117]Suya, F., Mahloujifar, S., Suri, A., Evans, D., & Tian, Y. (2021). Model-Targeted Poisoning Attacks with Provable Convergence. ArXiv:2006.16469 [Cs, Stat]. http://arxiv.org/abs/2006.16469

[118]Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions. ArXiv:1409.4842 [Cs]. http://arxiv.org/abs/1409.4842

[119]Tzelepis, C., Tzimiropoulos, G., & Patras, I. (2021). WarpedGANSpace: Finding non-linear RBF paths in GAN latent space. ArXiv:2109.13357 [Cs]. http://arxiv.org/abs/2109.13357

[120]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. ArXiv:1706.03762 [Cs]. http://arxiv.org/abs/1706.03762

[121]Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating Videos with Scene Dynamics. ArXiv:1609.02612 [Cs]. http://arxiv.org/abs/1609.02612

[122]Wang, H., Xiao, C., Kossaifi, J., Yu, Z., Anandkumar, A., & Wang, Z. (2021). AugMax: Adversarial Composition of Random Augmentations for Robust Training. ArXiv:2110.13771 [Cs]. http://arxiv.org/abs/2110.13771

[123]Wang, Y., Zhou, L., Wang, M., Shao, C., Shi, L., Yang, S., Zhang, Z., Feng, M., Shan, F., & Liu, L. (2020). Combination of generative adversarial network and convolutional neural network for automatic subcentimeter pulmonary adenocarcinoma classification. Quantitative Imaging in Medicine and Surgery, 10(6), 1249–1264. https://doi.org/10.21037/qims-19-982

[124]Wiles, O., Gowal, S., Stimberg, F., Alvise-Rebuffi, S., Ktena, I., Krishnamurthy, Dvijotham, & Cemgil, T. (2021). A Fine-Grained Analysis on Distribution Shift. ArXiv:2110.11328 [Cs]. http://arxiv.org/abs/2110.11328

[125]Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). CvT: Introducing Convolutions to Vision Transformers. ArXiv:2103.15808 [Cs]. http://arxiv.org/abs/2103.15808

[126]Wu, J., Huang, Z., Thoma, J., Acharya, D., & Van Gool, L. (2018). Wasserstein Divergence for GANs. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), Computer Vision – ECCV 2018 (Vol. 11209, pp. 673–688). Springer International Publishing. https://doi.org/10.1007/978-3-030-01228-1_40

[127]Xu, R., Xu, X., Chen, K., Zhou, B., & Loy, C. C. (2021). STransGAN: An Empirical Study on Transformer in GANs. ArXiv:2110.13107 [Cs]. http://arxiv.org/abs/2110.13107

[128]Yinka-Banjo, C., & Ugot, O.-A. (2020). A review of generative adversarial networks and its application in cybersecurity. Artificial Intelligence Review, 53(3), 1721–1736. https://doi.org/10.1007/s10462-019-09717-4

[129]Yu, C., & Wang, W. (2021). Adaptable GAN Encoders for Image Reconstruction via Multi-type Latent Vectors with Two-scale Attentions. ArXiv:2108.10201 [Cs, Eess]. http://arxiv.org/abs/2108.10201

[130]Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E., Feng, J., & Yan, S. (2021). Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. ArXiv:2101.11986 [Cs]. http://arxiv.org/abs/2101.11986

[131]Zadeh, P. H., Hosseini, R., & Sra, S. (2018). Deep-RBF Networks Revisited: Robust Classification with Rejection. ArXiv:1812.03190 [Cs, Stat]. http://arxiv.org/abs/1812.03190

[132]Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-Attention Generative Adversarial Networks. ArXiv:1805.08318 [Cs, Stat]. http://arxiv.org/abs/1805.08318

[133]Zhang, Y., Song, Y., Liang, J., Bai, K., & Yang, Q. (2020). Two Sides of the Same Coin: White-box and Black-box Attacks for Transfer Learning. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2989–2997. https://doi.org/10.1145/3394486.3403349

[134]Zhao, S., Li, J., Wang, J., Zhang, Z., Zhu, L., & Zhang, Y. (2021). attackGAN: Adversarial Attack against Black-box IDS using Generative

Adversarial Networks. Procedia Computer Science, 187, 128–133. https://doi.org/10.1016/j.procs.2021.04.118

[135]Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. ArXiv:1703.10593 [Cs]. http://arxiv.org/abs/1703.10593