
Invert and Defend: Model-based Approximate Inversion of Generative Adversarial Networks for Secure Inference

Wei-An Lin*
 University of Maryland
 walin@umd.edu

Yogesh Balaji*
 University of Maryland
 yogesh@cs.umd.edu

Pouya Samangouei
 Butterfly Network
 pouya@butterflynetwork.com

Rama Chellappa
 University of Maryland
 rama@umiacs.umd.edu

Abstract

Inferring the latent variable generating a given test sample is a challenging problem in Generative Adversarial Networks (GANs). In this paper, we propose InvGAN - a novel framework for solving the inference problem in GANs, which involves training an encoder network capable of inverting a pre-trained generator network without access to any training data. Under mild assumptions, we theoretically show that using InvGAN, we can approximately invert the generations of any latent code of a trained GAN model. Furthermore, we empirically demonstrate the superiority of our inference scheme by quantitative and qualitative comparisons with other methods that perform a similar task. We also show the effectiveness of our framework in the problem of adversarial defenses where InvGAN can successfully be used as a projection-based defense mechanism. Additionally, we show how InvGAN can be used to implement reparameterization white-box attacks on projection-based defense mechanisms. Experimental validation on several benchmark datasets demonstrate the efficacy of our method in achieving improved performance on several white-box and black-box attacks. Our code is available at <https://github.com/yogeshbalaji/InvGAN>.

1 Introduction

Generative Adversarial Networks (GANs) is one of the most successful frameworks used for generative modeling. Significant research progress in GANs over the last few years has pushed boundaries in generation capabilities and has made possible the synthesis of photo-realistic images of human faces [1, 2] and objects [3]. A fundamental problem involving GANs is the problem of inversion – given a test image, what is the most likely latent code that generates the test sample? The inversion problem is extremely challenging since the generator network in GANs is highly non-linear and non-injective. Inversion has applications in several machine learning problems e.g. domain adaptation [4, 5], compressed sensing [6], adversarial defenses [7], and anomaly detection [8].

In this work, we propose a novel approach for addressing the inversion problem in GANs. Our approach is model-based where the mapping from image space to latent space is represented as a parametric function. We solve for the parameters of this function by sampling the latent codes from the noise distribution of the GAN and making sure that (a) the inversions produced from the generated samples are close to the sampled codes (b) the generated images of the inversions semantically match

*Equal contribution.

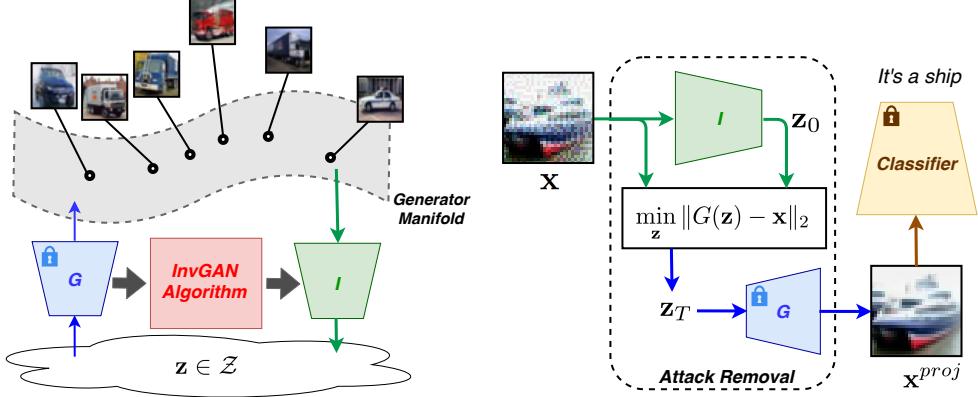


Figure 1: Overview of the proposed InvGAN framework. Left: Given a pre-trained generator G and no data, we solve for I to approximately invert the generator. Right: The application of InvGAN in adversarial defenses, where InvGAN can be used to project an adversarially perturbed sample x onto the generator manifold, and the projected sample x^{proj} can be used to make a robust prediction.

the GAN generations and (c) the distribution of inverted images follows the distribution of the GAN. Our method is a data-free inversion mechanism i.e., given a pre-trained generator network, no access to input dataset is needed. This is particularly important in privacy-preserving learning scenarios in which the data provider does not intend to publicly release the data due to privacy reasons, but instead releases a GAN model trained on this data satisfying several privacy constraints [9]. Our approach can invert such a GAN model.

In addition to comparing the reconstruction performance with previously proposed encoder-GAN models, we demonstrate the effectiveness of our inversion approach for the problem of adversarial defenses. The vulnerability of deep neural networks to small imperceptible perturbations has been demonstrated in several recent papers [10, 11, 12, 13], and this poses a huge threat in security-critical application domains where these networks are used.

One of the successful defense strategies proposed recently is DefenseGAN [7], which uses a GAN as a defense mechanism. In DefenseGAN, the inference step is used to project a test sample onto the GAN’s manifold to remove possible attacks. The inversion is posed as a non-convex optimization problem which is solved using Gradient Descent. However, this needs careful hyper-parameter tuning per dataset and generator, is extremely slow in practice, and does not scale well for deeper generator models. In this work, we propose an inference procedure to speed up DefenseGAN. More specifically, we use our trained encoder network to initialize the optimization problem. Using this initialization, the inference problem can be solved in very few iterations while preserving the quality of reconstructions. This leads to effortless hyperparameter tuning, a dramatic speed up in runtime, and improved reconstruction results.

Finally, we demonstrate how InvGAN can be used to create adversarial attacks for projection-based defense mechanisms that do not allow gradient computation for crafting white-box attacks. The idea, called “reparameterization” [14], is to approximate the projection operation using a differentiable function and derive gradient-based attacks using this approximation. We investigate how InvGAN can be used to implement this attack and analyze its performance in this context.

In summary, the main contributions of this work are as follows:

- 1 A model-based and data-free approach for approximately inverting pre-trained generator networks.
- 2 Significant improvements to Defense-GAN yielding improved defense performance, runtime speed up, and attack detection.
- 3 An implementation of reparameterization white-box attack [14] and analysis of its effectiveness across different datasets and methods.

2 Background

2.1 Inverting generative models

While significant research has focused on improving the quality, stability and diversity of GANs [3, 1, 2], there has been relatively less work on the inversion problem despite its practical significance. The method in [15] poses inversion as a non-convex optimization problem, which is then solved using projected gradient descent with stochastic clipping. A similar optimization with logistic loss has been proposed in [16]. While the above two methods are model-free and work for a pre-trained generator, they are extremely slow and deliver poor reconstructions for harder datasets like CIFAR-10.

Another line of work involves modifying the GAN objective to support generation and inference in a unified framework. They typically involve training an encoder function that maps from the image space to the latent space jointly with the generator function in GANs. Methods presented in [17, 18] trains the encoder network using an adversarial loss on (latent, image) pairs. The method proposed in [19] uses adversarial and reconstruction loss in latent space to train the encoder. While these methods enable fast inference, modifying the GAN objective to support inference affects the quality of generator models. Our approach offers the best of both worlds – fast inference and ability to perform inference on a pre-trained generator, thus preserving the quality of generative models.

2.2 Adversarial attacks and defenses

Adversarial attacks are imperceptible changes crafted to the input samples by an adversary to flip a model’s prediction $\hat{y} = f(\mathbf{x})$. In this work, we focus on classification problems, hence, $f(\mathbf{x}) \in \{1, \dots, c\}$. The most common form of adversarial attacks are additive perturbations where a norm-bounded perturbation δ is added to the input sample $\mathbf{X} \in \mathbb{R}^{w \times h \times c}$ as $\tilde{\mathbf{X}} = \mathbf{X} + \delta$. The strength of the attack correlates inversely with the norm-bound of the perturbation – stronger the attack, lower is the bound [11]. Adversarial attacks can broadly be grouped into two major classes – *white-box* and *black-box* attacks. White-box attacks assume full knowledge of the network structure and parameters of the model to craft an attack. These methods typically find the perturbation by solving an optimization involving the gradients of the loss function with respect to the input $\nabla_{\mathbf{x}} J(\mathbf{x}, y, \theta)$, where J is the classification loss, (\mathbf{x}, y) is a data-point, and θ corresponds to the model parameters [11, 12, 20]. Black-box attacks on the other hand assume no access to the model, and construct attacks just by using the model’s input-output response.

To protect the classifiers from adversarial attacks, one line of research focuses on removing the perturbation from input samples before feeding them to the classifier. MagNet [21] detects and reforms adversarial images using autoencoders. Defense-GAN [7] uses a generative adversarial network to model the image manifold. Adversarial perturbations are removed by projecting the samples onto the learned manifold. A similar idea has been used in PixelDefend [22] where the input distribution is modeled using a PixelCNN, and adversarial perturbations are removed from the input samples using greedy decoding. These projection-based defense methods take advantage of the operations that leads to *obfuscated gradients* which results in the inability to derive gradients to craft white-box attacks. In [14], BPDA and reparameterization attacks are most related to this work. The basic idea of BPDA is to approximate the non-differentiable operations with surrogates (e.g., identity function) during backpropagation. Adversarial attacks can then be crafted on the target classifiers using the gradients derived from the surrogates. Reparameterization uses the change-of-variables trick to circumvent the operations resulting in vanishing gradients.

3 Method

The generator network in a GAN model $G(\mathbf{z}) : \mathbb{R}^d \rightarrow \mathbb{R}^{w \times h \times c}$ maps a latent code $\mathbf{z} \in \mathbb{R}^d$ to an image in $\mathbb{R}^{w \times h \times c}$. The inversion problem is to find the inverse mapping i.e., given a test sample \mathbf{x} , we are interested in finding a latent $\hat{\mathbf{z}}$ such that $G(\hat{\mathbf{z}}) \approx \mathbf{x}$. This problem is extremely hard as most generator networks used in practice are non-convex and non-bijective functions. One common approach to this problem [15, 7] involves solving the following optimization problem:

$$\min_{\mathbf{z}} \|G(\mathbf{z}) - \mathbf{x}\|_2 \quad (1)$$

This optimization is extremely hard due to the non-convexity of the objective function. Solving this requires multiple random initialization of \mathbf{z} , carefully tuned learning rate and number of optimization

steps for each dataset. Also, this optimization is extremely slow, and scales poorly with increasing complexity of the generator network and the input distribution.

To fix these issues, we propose using an inverter network $I_{\theta_I}(\mathbf{x}) : \mathbb{R}^{w \times h \times c} \rightarrow \mathbb{R}^d$ that maps a sample from the image space to the latent space as an initialization for \mathbf{z} in (1). The goal of the inverter network is to approximately invert the generator network. We would like to emphasize that exact inversion is not possible as the generator function is non-bijective.

Theorem 1. Let $\{G(\mathbf{z}^{(i)})\}_{i=1}^N$ represent the generated samples corresponding to a pre-trained generator function G , with the noise vectors $\mathbf{z}^{(i)} \sim \mathcal{N}(0, 1)$. Let $I(\cdot)$ represent an inverter function that is trained to achieve approximate inversion on the training set, i.e.,

$$\|I(G(\mathbf{z}^{(i)})) - \mathbf{z}^{(i)}\| < \epsilon, \quad \forall \mathbf{z}^{(i)}, i \in [n].$$

Let L be the Lipschitz constant corresponding to the composite function $I \circ G(\cdot)$. Then, for $\epsilon' > \epsilon$, with probability $1 - o(1)$,

$$\|I(G(\mathbf{z})) - \mathbf{z}\| < \epsilon', \text{ for } \mathbf{z} \sim \mathcal{N}(0, 1).$$

That is with high probability, the function $I(\cdot)$ approximately inverts the generator $G(\cdot)$.

The above theorem states that under some smoothness conditions on the generator-encoder pair, if the encoder loss is bounded for every training sample, then the encoder approximately inverts the generator with high probability. Please refer to the supplementary material for the proof.

3.1 Encoder training

A natural way to train the encoder is to minimize the following loss function:

$$\min_{\theta_I} \mathbb{E}_{\mathbf{x} \in \mathbf{X}_{train}} \|\mathbf{x} - G(I_{\theta_I}(\mathbf{x}))\|. \quad (2)$$

One issue with this objective arises from the non-surjective nature of the the generator network. There are many samples in the training set that cannot be represented by the generator network as it is not surjective. Enforcing the mean squared error (MSE) loss for such samples per Eq. (2) is not appropriate and leads to blurry reconstructions. Hence, we propose using the following losses for training the encoder.

Approximate semantic consistency: To also make sure that our reconstructions are semantically consistent we add:

$$\mathcal{L}_{semantic} = \mathbb{E}_{\mathbf{z} \sim p_z} \left[\max(\|G(\mathbf{z}) - G(I(G(\mathbf{z})))\|, \eta) \right]. \quad (3)$$

The use of L_2 norm is not a good distance measure between two images, and minimizing the L_2 distance results in blurry reconstructions. Hence, we use hinge loss on L_2 norm in our formulation. The use of hinge loss in combination with adversarial loss (which we define later) searches for sharp reconstructions that are within ηL_2 norm ball of reconstruction error, instead of blurry reconstructions obtained by minimizing just the L_2 reconstruction error.

Latent code recovery: In addition to maintaining semantic consistency between the generated images and the inverted reconstructions, we recover the latent codes by making sure they are close to the sampled \mathbf{z} :

$$\mathcal{L}_{latent} = \|\mathbf{z} - I(G(\mathbf{z}))\|. \quad (4)$$

Inverted image distribution consistency: We want the images that are generated by the inversion $G(I(\mathbf{X}))$ to have the same distribution as the images that are generated from samples of the domain space of the generator $G(\mathbf{z})$. Therefore, we add a discriminator at training time for which the the real samples are the generations $G(\mathbf{z})$ and the fake samples are the inversions $G(I(G(\mathbf{z})))$:

$$\mathcal{L}_{adv}(I, D) = \mathbb{E}_{\mathbf{z} \sim p_z} [\log(D_{\theta_D}(G(\mathbf{z}))) - \log(1 - D_{\theta_D}(G(I(G(\mathbf{z})))))]. \quad (5)$$

This adversarial loss is crucial to improving the quality of the reconstructions.

3.2 Training

The encoder model is trained using a combination of the three loss terms introduced above. The objective function can be written as

$$\min_{\theta_I} \max_{\theta_D} \lambda_1 \mathcal{L}_{adv}(I, D) + \lambda_2 \mathcal{L}_{semantic}(I) + \lambda_3 \mathcal{L}_{latent}(I).$$

We set $\lambda_2 = 100$, $\lambda_1 = \lambda_3 = 1$ so that the semantic consistency gets enforced early on in the training, with the other two losses getting minimized gradually to correct for the distribution mismatch. We train in an iterative adversarial fashion to update the parameters of I and D .

3.3 Adversarial defenses

The objective of adversarial defense mechanisms is to make the classifiers robust to any class of adversarial perturbation. In this work, we consider additive perturbations defined in Section 2 – the most common form of adversarial attacks used till date. However, our framework is general and can be extended to other forms of attacks as well. Given an adversarially perturbed image $\mathbf{x}^{adv} = \mathbf{x} + \boldsymbol{\delta}$, projection-based defense mechanisms project the adversarial sample \mathbf{x}^{adv} to the manifold representing the input dataset. In DefenseGAN [7], the image manifold is first modeled by training a GAN on the input dataset. The perturbed sample \mathbf{x}^{adv} is then projected to the generative manifold by solving the optimization (2).

In this work, we replace the inference step in DefenseGAN using our proposed InvGAN framework. Given a test sample \mathbf{x}^{test} , the approximate latent code is first obtained by passing through the inverter network $I(\cdot)$, which can then be used as an initialization to the optimization (2):

$$\begin{aligned} \mathbf{z}_0 &= I(\mathbf{x}^{adv}), \\ \mathbf{z}_t &= \mathbf{z}_{t-1} - \alpha \nabla_{\mathbf{z}=\mathbf{z}_{t-1}} \|G(\mathbf{z}) - \mathbf{x}^{adv}\|, \quad \forall t \in 1, 2, \dots T, \\ \mathbf{x}^{proj} &= G(\mathbf{z}_T), \end{aligned} \tag{6}$$

where T denotes the number of inference iterations, and α is the learning rate. Let $f(\mathbf{x}) : \mathbb{R}^{w \times h \times c} \rightarrow \mathbb{R}_c^n$ denote a classification network trained to predict a sample \mathbf{x} into one of the n_c classes. The projected sample \mathbf{x}^{proj} computed using Eq. (6) is passed to the trained classification network as $f(\mathbf{x}^{proj})$ to make a prediction for the sample \mathbf{x} .

Attack detection: In addition to robust classification, we provide a framework for detecting adversarially perturbed samples. Let the trained classification network f be decomposed as $f(\mathbf{x}) = C \circ \Phi(\mathbf{x})$ where C denotes the last layer of the network, and Φ denotes all layers except the final layer. $\Phi(\mathbf{x})$ gives a feature representation of the image \mathbf{x} . We define attack detection score $\mathcal{A}(\mathbf{x})$ as

$$\mathcal{A}(\mathbf{x}) = \|\Phi(\mathbf{x}) - \Phi(\mathbf{x}^{proj})\|. \tag{7}$$

By choosing a threshold on the attack detection score, adversarially perturbed samples can be detected. To compute the detection score, we measure the projection distance in feature space and not in image space (i.e., $\|\mathbf{x} - \mathbf{x}^{proj}\|$) as mean squared error is not a good measure of distance between two images. An ablation study showing the advantage of using feature-based distance over image based distance is shown in Section 4.

3.4 Reparameterization white-box attack

The inability to compute gradients for projection-based defense methods such as DefenseGAN makes it hard to craft white-box attacks. To attack these models, [14] constructs a reparameterization attack where the input samples x are reparameterized as $x = h(z)$ for some differentiable function z such that $g(h(z)) = h(z)$, $\forall z$. Then, the gradients of the classifier can be computed using $f(h(z))$ as $h(\cdot)$ is differentiable. Consider $h(\cdot)$ as the generator function, and $g(\cdot)$ as our encoder-generator pair. Through our encoder training, we almost achieve perfect inversion (Section 4) which satisfies $g(h(z)) = h(z)$, $\forall z$. Hence for a given test sample \mathbf{x} , we construct the adversarial attack using the gradient $\nabla_x f(G(I(\mathbf{x})))$.

4 Experiments

Our proposed approach is evaluated on four datasets: MNIST [23], Fashion-MNIST [24], CIFAR-10 [25], and CelebA [26]. We pretrain GANs on all the datasets using the network architectures presented in [27]. We use DCGAN architecture for MNIST and Fasion-MNIST, and GANs with residual blocks for CIFAR-10 and CelebA. The architecture of the inverter I is the mirror image of the generator. The inverter network is trained for 100K iterations using the Adam optimizer with $\beta_1 = 0.5$, and $\beta_2 = 0.999$. Our implementation is based on Tensorflow [28] and Cleverhans [29].²

4.1 Inference

In this experiment, we consider the task of inferring the latent representation \mathbf{z} of an input image and reconstructing the input from the inferred \mathbf{z} . The closeness of the reconstructed image to the input illustrates the strength of the inversion scheme. The following quantitative metrics are considered for evaluation: (1) classification accuracy on a pre-trained classifier to assess whether the semantic information is retained in the reconstructed images, (2) Inception score (IS) [30], (3) Fréchet inception distance (FID) [31] of the reconstructed samples, (4) MSE between the input and reconstruction. The proposed InvGAN is compared with ALI [17] and AGE [19] on the CIFAR-10 test set. We also consider the following baselines:

- Direct optimization: $\mathbf{z}^* = \arg \min_{\mathbf{z}} \|G(\mathbf{z}) - \mathbf{x}\|_2$ is first solved by running gradient descent for 200 iterations. $G(\mathbf{z}^*)$ is then treated as the reconstructed image.
- InvGAN with $T = 0$: The encoder-decoder scheme similar to ALI and AGE.
- InvGAN with $T = 200$: The inference used at defense time.

For fair comparisons, in this experiment, we adopt the simple DCGAN architecture without residual blocks in [27] on CIFAR-10. The results are presented in Table 1 and Figure 3. InvGAN with $T = 0$ achieves the best IS, FID and accuracy than the competing methods, while direct optimization achieves the best MSE. However, lower MSE does not necessarily produce natural looking images (which is reflected in poorer inception and FID scores) since the MSE loss does not take semantic information into account. Also note that InvGAN only suffers slightly from running several steps of MSE updates. However, these optimization updates offer security against common end-to-end attacks as will be discussed in the following sections.

Table 1: Quantitative evaluation of inference on CIFAR-10 test set.

	MSE	IS	FID	Accuracy
ALI	0.32 ± 0.17	6.12 ± 0.15	57.79	0.35
AGE	0.06 ± 0.03	6.43 ± 0.15	39.93	0.43
Direct Optimization	0.03 ± 0.02	6.50 ± 0.20	40.18	0.44
InvGAN ($T = 0$)	0.10 ± 0.06	7.72 ± 0.16	22.35	0.59
InvGAN ($T = 200$)	0.08 ± 0.04	7.36 ± 0.27	23.91	0.59

4.2 Defense against adversarial attacks

In this section, we compare InvGAN ($T = 1000$) with Defense-GAN [7], Cowboy [32], and PixelDefend [22] in defending against white-box attacks. In [7, 22], end-to-end attacks are either not considered or not successful due to the difficulty involved in gradient computation. Following these works, we consider crafting FGSM [11] and CW [12] attacks on the classifier and feeding the adversarial images to our pipeline. In addition, we also experiment on end-to-end attacks using reparameterization and BPDA [14].

Robust classification: We use MNIST, Fashion-MNIST, CIFAR-10, and CelebA for evaluation. For the FGSM attack, we select $\epsilon = 0.3$ on MNIST and Fashion-MNIST, and $\epsilon = 8/255$ on CIFAR-10 and CelebA. For the CW attack, we set the binary search step to 6, learning rate to 0.2, and number

²The source code to reproduce all the experiments will be released after final decisions.

of iterations to 100. Table 2 shows the classification accuracy. We observe that InvGAN achieves improved performance compared to DefenseGAN, and offers defense against the BPDA attack. In Figure 2, we visualize attack removal for DefenseGAN and InvGAN under BPDA attack. For each method, the first row represents reconstruction on clean images, the second row shows the attacked images, and the third row shows the reconstructed attacked images. We observe that DefenseGAN partially reconstructs the attacked images, while InvGAN successfully removes the perturbation produced by BPDA.

Table 2: Classification accuracy under different white-box attacks. We directly report the performance presented by the authors for methods marked in \dagger . BPDA attack introduces intense noise in experiments marked with *. Visualization of the attack is presented in the supplemental material.

	MNIST					Fashion-MNIST				
	Clean	FGSM	CW	RP	BPDA	Clean	FGSM	CW	RP	BPDA
No Defense	0.99	0.18	0.01	0.72	-	0.92	0.06	0.06	0.41	-
Cowboy \dagger	-	0.78	-	-	-	-	0.44	-	-	-
DefenseGAN	0.98	0.83	0.96	0.92	0.79	0.81	0.34	0.80	0.54	0.59
InvGAN	0.99	0.78	0.99	0.92	0.87	0.88	0.28	0.87	0.49	0.71
	CIFAR-10					CelebA				
	Clean	FGSM	CW	RP	BPDA	Clean	FGSM	CW	RP	BPDA
No Defense	0.95	0.16	0.02	0.84	-	0.97	0.05	0.04	0.96	-
Cowboy \dagger	-	0.53	-	-	-	-	-	-	-	-
PixelDefend \dagger	0.85	0.46	-	-	0.09	-	-	-	-	-
DefenseGAN	0.44	0.43	0.44	0.44	n/a*	0.89	0.88	0.88	0.89	n/a*
InvGAN	0.66	0.60	0.58	0.61	n/a*	0.93	0.90	0.92	0.90	n/a*

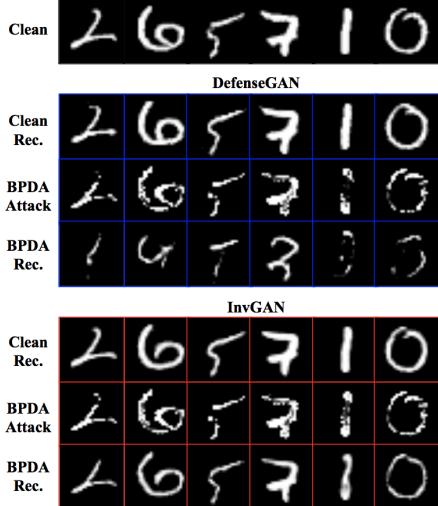


Figure 2: Illustration of adversarial attacks.

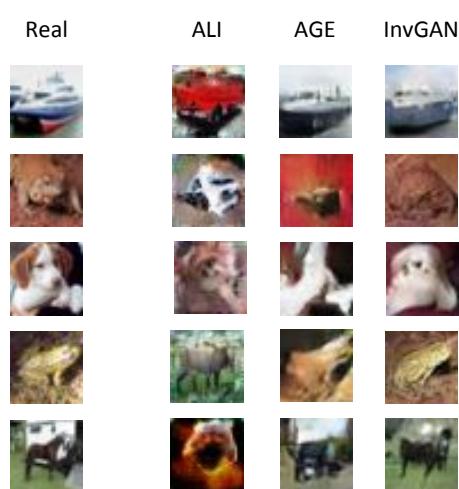


Figure 3: Qualitative inversion results.

Attack detection: In Table 3, we compare the area under ROC curve (AUC) scores for attack detection between DefenseGAN and InvGAN on different adversarial attacks. Comparing Table 2 and Table 3, we observe that for large perturbation settings (e.g. FGSM with $\epsilon = 0.3$) where the classification accuracy of InvGAN is low, attacks can be detected easily. On the other hand, for minute perturbations (e.g. FGSM with $\epsilon = 8/255$ and CW), InvGAN successfully removes perturbation and achieves high accuracy, but the attack becomes more challenging to detect. This suggests that our attack detection and robust classification scheme offers orthogonal benefits – when one fails, the other succeeds.

Table 3: Attack detection performance (AUC) of DefenseGAN and InvGAN.

	MNIST				Fashion-MNIST			
	FGSM	CW	RP	BPDA	FGSM	CW	RP	BPDA
DefenseGAN	0.9878	0.9661	0.9801	0.8706	0.9563	0.7752	0.9364	0.8999
InvGAN	0.9985	0.9880	0.9975	0.9210	0.9932	0.8543	0.9845	0.9334
	CIFAR-10				CelebA			
	FGSM	CW	RP	BPDA	FGSM	CW	RP	BPDA
DefenseGAN	0.6524	0.6823	0.4537	0.5950	0.8753	0.7341	0.4488	0.7491
InvGAN	0.8132	0.8370	0.5469	0.7807	0.9042	0.7764	0.4781	0.7233

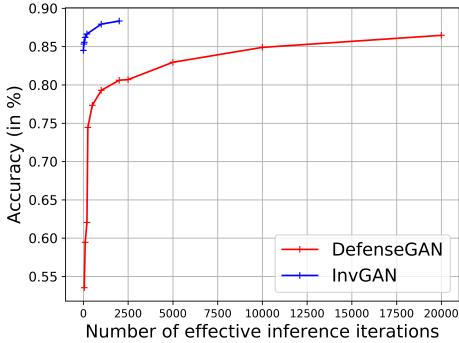


Figure 4: Speed - accuracy trade-off curves.

Table 4: Black-box attack performance comparison on MNIST and Fasion-MNIST.

	MNIST	FMNIST
No Defense	0.6631	0.4988
DefenseGAN	0.9398	0.6320
InvGAN	0.9449	0.6596

Black-box attack: In addition to whitebox attacks, we compare InvGAN with DefenseGAN under black-box attack [33] in Table 4. We use the same procedure described in [33] and [7].

4.3 Run time comparison

Measuring the run time of defense mechanisms is challenging as it depends on the implementation. We propose using the *effective number of inference iterations* as a measure of run time, defined as the product of number of random restarts and the number of iterations per run. For InvGAN, the number of random restarts is always 1 as we initialize it from the encoder. We report the classification accuracy of DefenseGAN and InvGAN on Fashion-MNIST for clean images by varying the effective number of iterations. The result is presented in Figure 4. InvGAN has a significantly shorter run time and offers reconstructions with improved semantic consistency.

4.4 Ablation study

In this section, we study the effect of disabling the adversarial loss. More specifically, we train the encoder model by setting $\eta = 0$, $\lambda_1 = 0$, $\lambda_2 = 1$, and $\lambda_3 = 0$. As can be seen from Table 5, although this network achieves a lower MSE, the images are perceptually less realistic, and is reflected in lower classification accuracy.

Table 5: Analyzing the effect of adversarial loss in InvGAN.

	MSE	IS	FID	Accuracy
w/o \mathcal{L}_{adv}	0.08 ± 0.04	7.54 ± 0.17	28.07	0.603
Ours	0.10 ± 0.06	7.72 ± 0.16	19.85	0.625

5 Conclusion

In this work, we introduce InvGAN – a novel data-free and model-based inversion framework for solving the inference problem in GANs. Our approach involves training an encoder function capable of inverting the generator network back to the latent space. The encoder function is trained using a novel loss function that achieves superior inversion results compared to the contemporary methods performing inference. The usefulness of our inversion scheme is demonstrated in the problem of adversarial defenses, where our inversion scheme has been shown to achieve dramatic improvements in defense performance, run time and attack detection over DefenseGAN – a projection-based defense mechanism using GANs.

References

- [1] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018.
- [2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.
- [3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096, 2018.
- [4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [5] Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.
- [6] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 537–546. JMLR.org, 2017.
- [7] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *International Conference on Learning Representation*, 2018.
- [8] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging - 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, 2017.
- [9] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *CoRR*, abs/1802.06739, 2018.
- [10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.
- [11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
- [12] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.
- [13] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pages 86–94. IEEE Computer Society, 2017.
- [14] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283, 2018.
- [15] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *International Conference on Learning Representation (Workshops)*, 2017.
- [16] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 2018.

- [17] Vincent Dumoulin, Mohamed Ishmael Diwan Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. 2017.
- [18] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.
- [19] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. It takes (only) two: Adversarial generator-encoder networks. In *AAAI*. AAAI Press, 2018.
- [20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.
- [21] Dongyu Meng and Hao Chen. Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS ’17*, pages 135–147. ACM, 2017.
- [22] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.
- [23] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.
- [24] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017.
- [25] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).
- [26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [27] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.
- [28] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [29] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.
- [30] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.
- [31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.
- [32] Gokula Krishnan Santhanam and Paulina Grnarova. Defending against adversarial attacks by leveraging an entire GAN. *CoRR*, abs/1805.10652, 2018.
- [33] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519. ACM, 2017.
- [34] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

A Proof of Theorem 1

Theorem. Let $\{G(\mathbf{z}^{(i)})\}_{i=1}^N$ represent the generated samples corresponding to a pre-trained generator function G , with the noise vectors $\mathbf{z}^{(i)} \sim \mathcal{N}(0, 1)$. Let $I(\cdot)$ represent an inverter function that is trained to achieve approximate inversion on the training set, i.e.,

$$\|I(G(\mathbf{z}^{(i)})) - \mathbf{z}^{(i)}\| < \epsilon, \quad \forall \mathbf{z}^{(i)}, i \in [n].$$

Let L be the Lipschitz constant corresponding to the composite function $I \circ G(\cdot)$. Then, for $\epsilon' > \epsilon$, with probability $1 - o(1)$,

$$\|I(G(\mathbf{z})) - \mathbf{z}\| < \epsilon', \quad \text{for } \mathbf{z} \sim \mathcal{N}(0, 1).$$

That is with high probability, the function $I(\cdot)$ approximately inverts the generator $G(\cdot)$.

Proof. The input samples $\{G(\mathbf{z}^{(i)})\}_{i=1}^N$ correspond to the training data for the inverter network. For any latent $\mathbf{z} \sim \mathcal{N}(0, 1)$,

$$P(\|\mathbf{z} - \mathbf{z}_i\| < \epsilon) \geq 1 - e^{-\frac{d}{18}(\frac{\epsilon^2}{4d}-1)^2} \quad \forall i \in [n].$$

The above inequality follows from the concentration bound for χ^2 distribution [34] since $\frac{1}{4}\|\mathbf{z} - \mathbf{z}_i\|^2$ follows a χ_d^2 distribution with d degrees of freedom, where d is the noise dimension. Then,

$$P(\exists \mathbf{z}^{(i)}, i \in [n] \text{ s.t. } \|\mathbf{z} - \mathbf{z}^{(i)}\| < \epsilon) \geq 1 - e^{-\frac{nd}{18}(\frac{\epsilon^2}{4d}-1)^2}. \quad (8)$$

Eq. (8) says that there exists at least one $\mathbf{z}^{(i)}$ concentrated close to \mathbf{z} . Now, consider $\|I(G(\mathbf{z})) - \mathbf{z}\|$. This can be expanded as

$$\begin{aligned} \|I(G(\mathbf{z})) - \mathbf{z}\| &= \|(I(G(\mathbf{z})) - I(G(\mathbf{z}^{(i)}))) + (I(G(\mathbf{z}^{(i)})) - \mathbf{z}^{(i)}) + (\mathbf{z}^{(i)} - \mathbf{z})\| \\ &\leq \|I(G(\mathbf{z})) - I(G(\mathbf{z}^{(i)}))\| + \|I(G(\mathbf{z}^{(i)})) - \mathbf{z}^{(i)}\| + \|\mathbf{z}^{(i)} - \mathbf{z}\| \\ &\leq (L+1)\|\mathbf{z} - \mathbf{z}^{(i)}\| + \|I(G(\mathbf{z}^{(i)})) - \mathbf{z}^{(i)}\| \\ &\leq (L+1)\|\mathbf{z} - \mathbf{z}^{(i)}\| + \epsilon, \quad \forall i \in [n] \\ &\leq \min_{i \in [n]} (L+1)\|\mathbf{z} - \mathbf{z}^{(i)}\| + \epsilon. \end{aligned} \quad (9)$$

This follows from Triangle inequality and the assumption on training loss. Using (8) and (9) for bounding $\|I(G(\mathbf{z})) - \mathbf{z}\|$, we obtain

$$\begin{aligned} P(\|I(G(\mathbf{z})) - \mathbf{z}\| \leq \epsilon') &\geq P\left(\left[\min_{i \in [n]} (L+1)\|\mathbf{z} - \mathbf{z}^{(i)}\|\right] < \epsilon' - \epsilon\right) \\ &= P(\exists i \|\mathbf{z} - \mathbf{z}^{(i)}\| < \frac{\epsilon' - \epsilon}{L+1}) \\ &\geq 1 - e^{-\frac{nd}{18}(\frac{(\epsilon'-\epsilon)^2}{4d(L+1)^2}-1)^2}. \end{aligned}$$

That is with probability $1 - o(1)$, $\|I(G(\mathbf{z})) - \mathbf{z}\| \leq \epsilon'$. Please note that we assumed that $\epsilon' > \epsilon$. This concludes the proof. \square

Remark: In the above equation, $P(\|I(G(\mathbf{z})) - \mathbf{z}\| \leq \epsilon')$ is taken with respect to $(\{\mathbf{z}^{(i)}\}, \mathbf{z})$. We define the event

$$E_n = \{\omega : \|I_{\mathbf{z}^n(\omega)}(G(\mathbf{z}(\omega))) - \mathbf{z}(\omega)\| > \epsilon'\}. \quad (10)$$

Based on the bound, we have $P(E_n) < \exp(-nC)$, and thus $\sum_{n=1}^{\infty} P(E_n) < \infty$. Applying the Borel-Cantelli lemma, we conclude that with probability 1, the event E_n happens for only finitely many n .

B Additional qualitative results

B.1 Image reconstruction

We visualize the reconstruction results for DefenseGAN and InvGAN on CIFAR-10 and CelebA in Figures 5 and 6. We can observe that with the initialization provided by the inverter, the reconstructed images are semantically more consistent compared to the direct optimization approach adopted by DefenseGAN.



Figure 5: Reconstruction results on CIFAR-10.



Figure 6: Reconstruction results on CelebA.

B.2 Attack removal

We show sample results for InvGAN attack removal in Figure 7. InvGAN effectively removes perturbations added by the attacks.

Remark: We observe that when the attack is mounted solely on the classifier (i.e. FGSM/CW on the classifier), the crafted perturbations are mainly focused on the non-stroke regions. That is, an attacker can easily manipulate the classification results of an unsecured classifier without modifying the semantic nature (i.e. the digit) of an image. In contrast, to mount attack on the end-to-end framework (i.e. BPDA on InvGAN + classifier), the perturbations have to be crafted to ‘erase’ strokes.

B.3 BPDA attack

We visualize the qualitative results of CW attack and BPDA attack on InvGAN model on CIFAR-10 dataset. We observe that while the perturbations produced by the CW attacks are imperceptible, BPDA attacks are very strong. In other words, it’s hard to attack InvGAN with low perturbations. This is the reason why we did not include results on BPDA attack on CIFAR-10 and CelebA in Table 2 of the main paper.

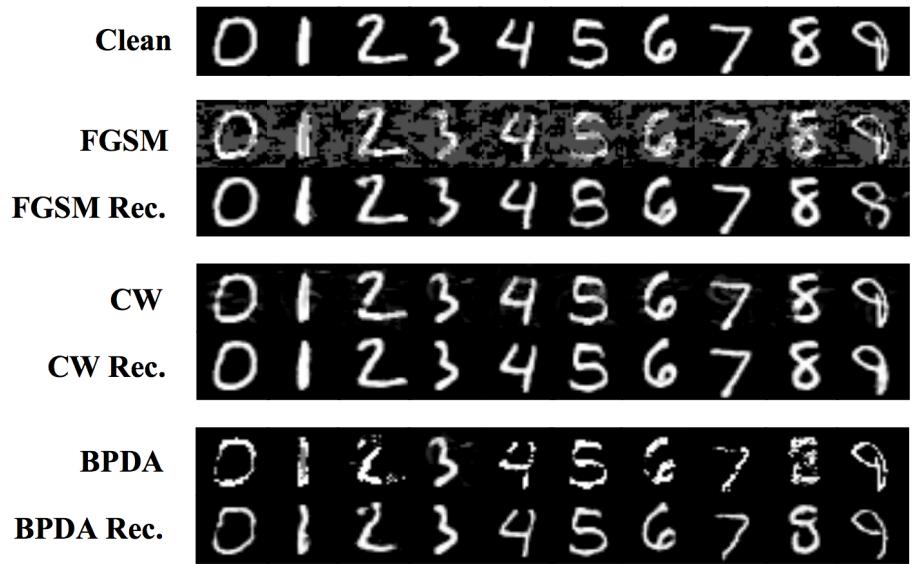


Figure 7: InvGAN attack removal results on MNIST.

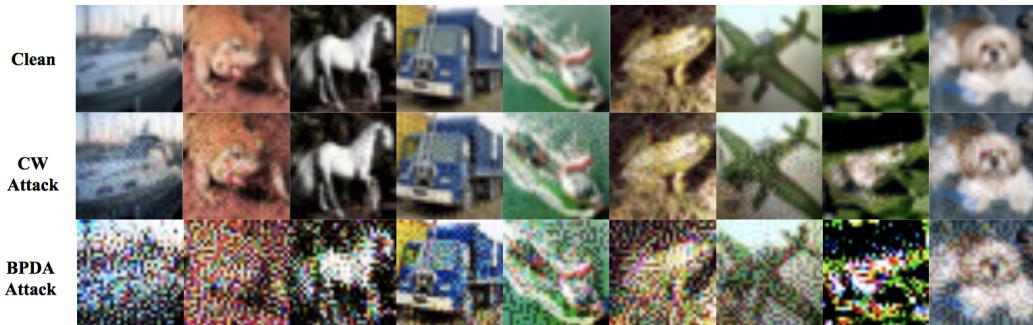


Figure 8: Sample images attacked by CW and BPDA. Clearly, BPDA mounts significant amount of perturbations on clean images.