




Deceptive Diffusion: Generating Synthetic Adversarial Examples

Lucas Beerens^{1,2}, Catherine F. Higham³ , and Desmond J. Higham^{1,2}(✉) 

¹ School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, UK
d.j.higham@ed.ac.uk

² Maxwell Institute, University of Edinburgh, Edinburgh EH8 9BT, UK

³ School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK

Abstract. We introduce the concept of deceptive diffusion—training a generative AI model to produce adversarial images. Whereas a traditional adversarial attack algorithm aims to perturb an existing image to induce a misclassification, the deceptive diffusion model can create an arbitrary number of new, misclassified images that are not directly associated with training or test images. Deceptive diffusion offers the possibility of strengthening defence algorithms by providing adversarial training data at scale, including types of misclassification that are otherwise difficult to find. In our experiments, we also investigate the effect of training on a partially attacked data set. This highlights a new type of vulnerability for generative diffusion models: if an attacker is able to stealthily poison a portion of the training data, then the resulting diffusion model will generate a similar proportion of misleading outputs.

Keywords: Image classification · Generative AI · Stability

1 Motivation

In this work, we combine two types of algorithm that have come to prominence in artificial intelligence (AI): adversarial and generative. Adversarial attack algorithms are designed to reveal vulnerabilities in classification systems; for example by perturbing a chosen image in a way that is imperceptible to the human eye, but causes a change in classification [13, 31]. Generative models are designed to create outputs that are similar to, but not simply copies of, the examples on which they were trained [8, 16]. Here, we show that by training on data that consists of adversarially perturbed images, a generative diffusion model can be made to create fresh examples of adversarial images that do not correspond directly to any underlying real images.

In Sect. 2 we give some background information on the two main ingredients of our work: adversarial attack algorithms and generative diffusion models. Section 3 describes the results of computational experiments where we investigate the idea of training a diffusion model on adversarially-perturbed data. We finish with a brief discussion in Sect. 4.

1.1 Related Work

We refer to [5] for an overview of recent attempts to use generative AI tools to produce adversarial inputs. The AdvDiffuser algorithm of [5] appears to be the first and only approach to generating new, synthesized, examples of adversarial images using a diffusion model. In that work, the authors take an existing, trained diffusion model and adapt the denoising, or backward, process by adding adversarial perturbations at each time step. This change increases computational complexity, since an extra gradient step is required at each time point. Our approach differs by building a new diffusion model, which then generates images with a standard de-noising algorithm. In addition to lowering the computational cost, our *deceptive diffusion* method reveals a new type of security threat that arises when standard generative diffusion models are created on training data that has been attacked. In particular, we find that the drop in classification success is in direct proportion to the fraction of training data that is adversarially perturbed. Hence, if an attacker is able to poison some portion of the training data, the builders of a generative diffusion model may inadvertently create a tool that produces a corresponding proportion of adversarial images.

In the conceptually different, and more traditional, setting of computing an adversarial perturbation to an existing image, we mention that the DiffAttack algorithm [4] also makes use of a diffusion model. We also note that the earlier work [12], which is not concerned with generative AI, showed that adversarial examples can be effective for data poisoning.

2 Background

2.1 Adversarial Attack Algorithms

State of the art image classification tools are known to possess inherent vulnerabilities. In particular, they can be fooled by adversarial attacks, where an existing image undergoes a small perturbation that would not be noticeable to a human, but causes a change in the predicted class. Since this effect was first pointed out, [13, 31], a wide range of attack and defence strategies have been put forward, [1, 2, 24, 25], and bigger picture questions concerning the inevitability of attack success have been investigated, [7, 11, 28, 29, 32, 33]. The susceptibility of AI systems to attack is a serious issue in many application areas and it is pertinent to the recent calls for AI regulation. For example, the amendment of June 2023 [10] to Article 15 - paragraph 4 - subparagraph 1 of the EU AI act [9] requires that: “*High-risk AI systems shall be resilient as regards to attempts by unauthorised third parties to alter their use, behaviour, outputs or performance by exploiting the system vulnerabilities.*”

2.2 Generative Diffusion Models

A generative diffusion model for creating realistic, but synthetic, images can be built by first training a neural network to de-noise a collection of noisy images, and then asking the network to de-noise a new sample of pure noise [3].

In Algorithms 1 and 2 we summarize the basic unconditional diffusion model setting from [16]; see also [15, 23] for detailed explanations of the steps involved. Here, the α_t are parameters taking values between zero and one. They have the form $\alpha_t = 1 - \beta_t$, where the predetermined sequence $\beta_1, \beta_2, \dots, \beta_T$ is known as the *variance schedule*. In [16], linearly increasing values from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ are used. We also let $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, and $\sigma_q^2(t) = (1 - \alpha_t)(1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)$.

In step 5 of Algorithm 1, ϵ_θ denotes the output from a neural network. Given a version of the noisy image, $\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$, corresponding to a time t , the job of the network is to predict the noise ϵ . Here, a simple least-squares loss function is used.

Algorithm 1. Training with the forward process [16]

```

1: repeat
2:    $\mathbf{x}_0 \sim q(\mathbf{x}_0)$  ▷ choose an image from training set
3:    $t \sim \text{Uniform}(\{1, 2, \dots, T\})$ 
4:    $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  ▷ standard Gaussian sample
5:   Take gradient step w.r.t.  $\theta$  on  $\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|_2^2$ 
6: until converged

```

Algorithm 2 from [16] summarizes the sampling process. Here, a set of pure noise pixel values is de-noised from time T to time 0 in order to produce a new synthetic image.

Algorithm 2. Sampling with the backward process [16]

```

1:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  ▷ standard Gaussian sample
2: for  $t = T, T - 1, \dots, 1$  do
3:    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  ▷ standard Gaussian sample
4:    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta \right) + \sigma_q(t) \mathbf{z}$ 
5: end for
6: return  $\mathbf{x}_0$ 

```

3 Experimental Results

We now outline the key components in our computational experiments.

We use the MNIST data set [21], which contains 60,000 training images and 10,000 test images of handwritten digits, with labels indicating the categories: ‘0’, ‘1’, ‘2’, ..., ‘9’.

As a classifier, we use a convolutional neural network (CNN) based on the architecture of LeNet [19, 20]. The exact architecture can be found in our code. After training, this classifier achieves an accuracy of 99.02% on the test images.

For the adversarial attack algorithm we use PGDL2 [17], a PyTorch implementation of the projected gradient method from [24]. This attack algorithm uses a robust optimization approach to seek an optimal perturbation in an ℓ_2 sense, using gradients of the loss function. We use the default setting in PGDL2 where an attack is declared successful if it finds a sufficiently small class-changing perturbation within a specified number of iterations of a first order gradient method. The bound on the ℓ_2 norm of the attack was set to 2 (each of the 784 pixels takes values between 0 and 1). We chose a large bound of 1000 on the number of iterations in order to maximize the size of the attacked image dataset for training the diffusion model. We used PGDL2 in untargeted mode, so that any change of classification is acceptable.

In the diffusion model, we used a neural network with a UNet2DModel architecture from <https://huggingface.co/docs/diffusers/en/api/models/unet2d> which is motivated by the original version in [27].

3.1 Initial Sanity Check

Before moving on to adversarial images, we first report on an initial test which confirms that the diffusion model is capable of producing outputs that are acceptable to the classifier.

In this test, we train the diffusion model using the original MNIST training data. We supply the labels during the training process, so we use a conditional version of Algorithm 1, where in step 5 the network learns to remove noise and produce an image when given both a time t and a label. This is built in to the UNet2DModel. A trainable encoder maps the label into the same space as the timestep. These two quantities are then added and passed to the model in the same way that the time is usually passed [26].

We found that 99.5% of the outputs from the trained model were classified with the intended label.

3.2 Deceptive Diffusion Model

Our aim is now to build a *deceptive diffusion model* that takes a label i and generates a new image that looks like digit i but is misclassified.

Using PGDL2 for untargeted attacks on the 60,000 MNIST training images gave a success rate of 86.5%, thereby producing 51,918 perturbed images that are classified differently to their nearby original images. We trained the diffusion model on these adversarial images, using the original labels. Figure 1 illustrates the process. Here, the image of the three on the left is from the MNIST training set, and the image in the middle arises from a successful attack by PGDL2 (classified as an eight). After training the diffusion model on all 51,918 adversarial images, asking for an output from the ‘3’ category produced the result shown (classified as a five).

After using the trained diffusion model to generate 100 new images from each of the ten categories and passing these through the CNN classifier, we found that

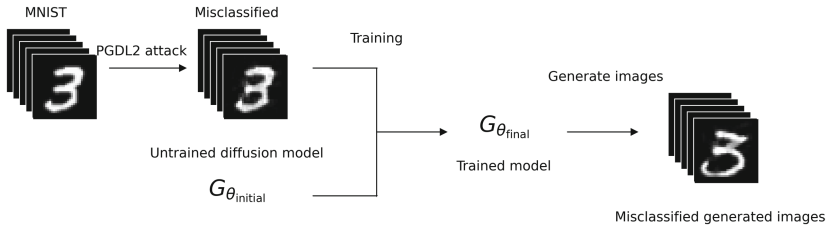


Fig. 1. Building the deceptive diffusion model. Images that were successfully attacked by PGDL2 are used as training data, with the original labels retained. The trained diffusion model, $G_{\theta_{\text{final}}}$, produces adversarial images associated with a given a label. (For the images in this diagram, the image from PGDL2 is classified as an ‘8’ and the image from the deceptive diffusion model, which was supplied with the label ‘3’, is classified as a ‘5’.)

93.6% of the outputs were classified differently to their requested labels. Figure 2 gives a confusion matrix showing the performance by category. For comparison, Fig. 3 shows a confusion matrix for the PGDL2 attacks on the 60,000 training images.

Table 1 shows the correlation between the rows of the confusion matrices in Figs. 2 and 3. The high correlation values indicate that the two confusion matrices are similar. We emphasize that PGDL2 was used in untargeted mode: an image from category i can be perturbed so that the classifier predicts any new category $j \neq i$. From Table 1 we see that although the deceptive diffusion model was not provided with a target class j , it tends to produce new $i \mapsto j$ misclassifications of the same type as PGDL2.

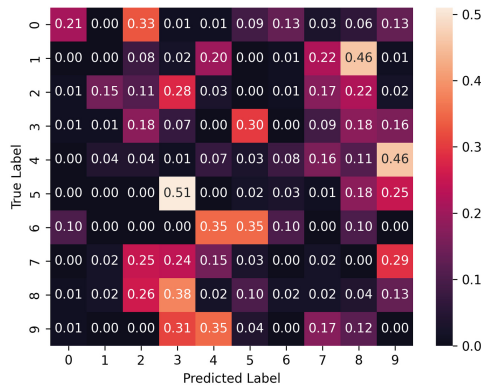


Fig. 2. Confusion matrix for the deceptive diffusion model. For a given label (row) we show the frequency with which the classifier assigned each label (column) to the output. Entries on the diagonal therefore correspond to unsuccessful attempts to create an adversarial image. Overall misclassification rate is 93.6%.

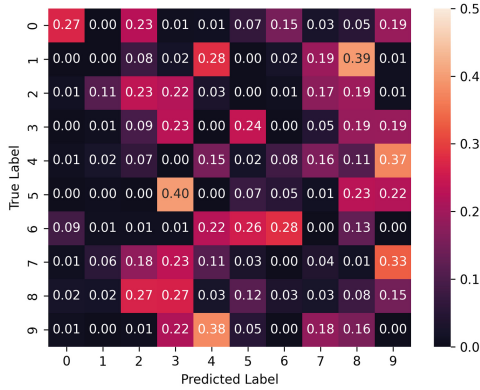


Fig. 3. Confusion matrix for PGDL2 attacks on the 60,000 MNIST training images. With training images corresponding to each label (row) we show the frequency with which the classifier assigned each label (column) after the attack. Entries on the diagonal therefore correspond to unsuccessful attacks. Overall success rate is 86.5%.

Table 1. Correlation of confusion matrix rows for PGDL2 attack and generated data.

Class	0	1	2	3	4	5	6	7	8	9
Correlation	0.90	0.97	0.88	0.79	0.96	0.98	0.82	0.96	0.96	0.96

To give a feel for the outputs from the deceptive diffusion model, Fig. 4 (upper) shows 100 independent outputs corresponding to the label ‘9’. We note from Fig. 2 that 0% of such outputs are classified as nines. Hence, we see that the model is capable of producing convincing adversarial images. For comparison, Fig. 4 (lower) shows the results of PGDL2 on images from the ‘9’ category.

Partial Attacks. So far, we have looked at two options for the training data. Either all training data was attacked, or all training data was clean. Now we look at a third case: partially attacked training data. Again we choose the same MNIST images that were successfully attacked using PGDL2. Consider $p \in \{0, 20, 40, 60, 80, 100\}$. For each class, we replace $p\%$ of the clean images with their successfully attacked counterpart. Now using these six datasets, we train six models.

For each class, 100 images are generated using each of the trained models. In Fig. 5 we show the resulting accuracy of the classifier on these generated images for the models trained on varying levels of poisoned data. We see that the classification accuracy degrades roughly in proportion with the amount of poisoned training data. This result is intuitively reasonable, under the assumption that all training images carry equal weight when the diffusion model is created.

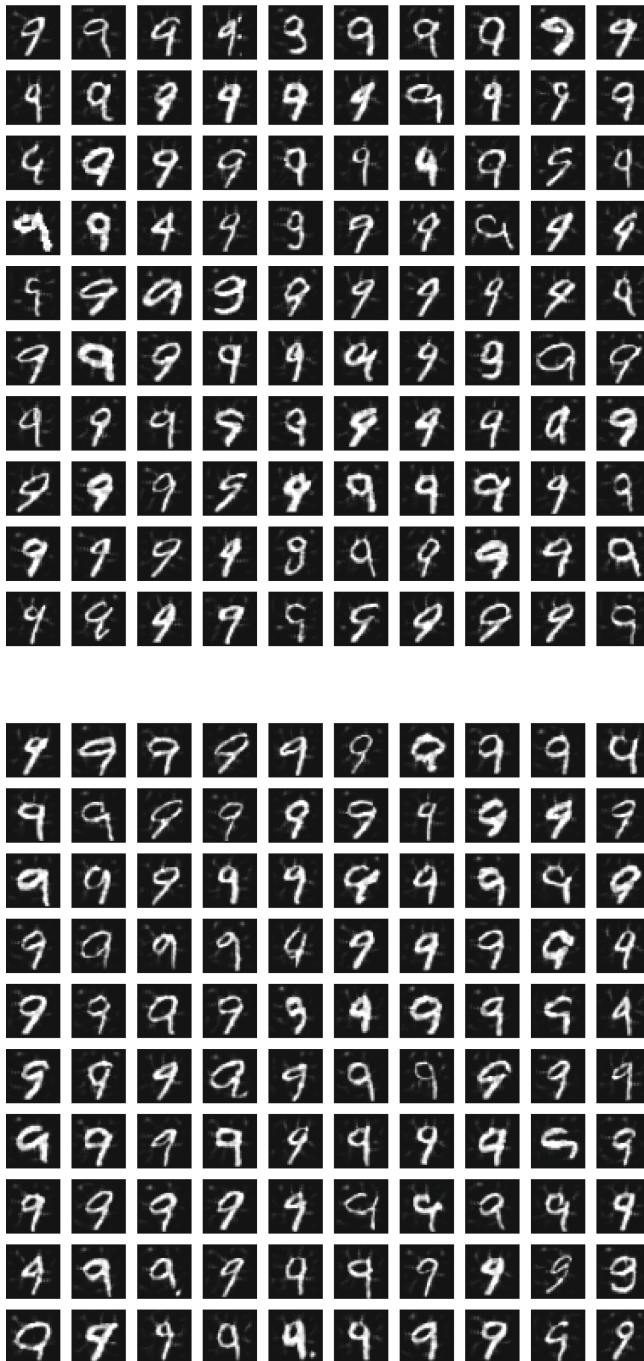


Fig. 4. Upper: example of 100 images arising when the deceptive diffusion model was given the label ‘9’. Lower: example of 100 images arising from successful PGDL2 attacks on images that had label ‘9’.

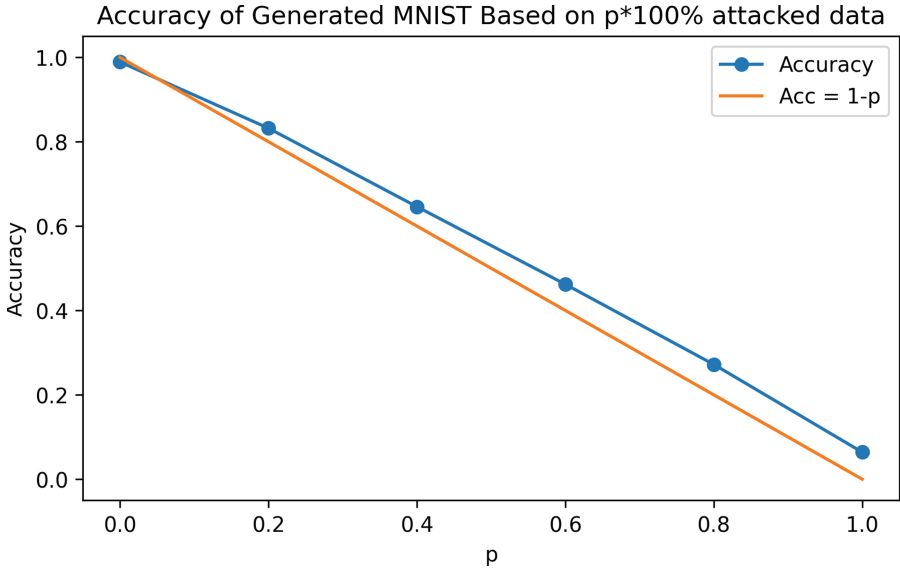


Fig. 5. Classification accuracy (vertical axis) for output from a deceptive diffusion model where a fraction of the training data (horizontal axis) is replaced by its adversarially attacked counterpart. The slope representing linear proportionality is also shown.

Fréchet Inception Distance. A widely used measure for generated image quality is the Fréchet Inception Distance (FID) [14], where lower is better. It compares a generated dataset to a ground truth dataset. First, a classifier is used to extract features. Then the Fréchet distance between these feature sets is computed. Typically the Inception v3 classifier [30] without its last layer is used. To take into account that the generator is conditioned on the class, we use the Class-Aware Fréchet Distance (CAFD), which computes the FID for every class and takes the average [22].

Since our dataset is of low resolution, instead of Inception v3 we use the classifier that we trained earlier, with its last layer removed. This way the output is in \mathbb{R}^{128} .

In Fig. 6, the CAFD is shown for the diffusion models trained with partially poisoned data. These values are compared with the CAFD for the test set and the PGDL2 attacked training set. These are displayed at $p = 0$ and $p = 1$ respectively, because they represent samples from the ground truths for the clean and attacked case respectively. To avoid bias, these two sets are limited to contain the same number of samples as the generated sets, [6].

The results in Fig. 6 show that the CAFD increases monotonically as the level of poisoning increases. This seems reasonable, because, as shown in Fig. 5, higher levels of poisoning lead to higher levels of misclassification. The CAFD relies on the feature extraction of an MNIST classifier. Since the attacks target the classifier, it makes sense that the extracted features are different. The key

observation here is that the fully adversarial model ($p = 1$) corresponds to a CAFD that is similar to that of the PGDL2 attacked data set, indicating that deception diffusion can mimic adversarially attacked data successfully according to this metric.

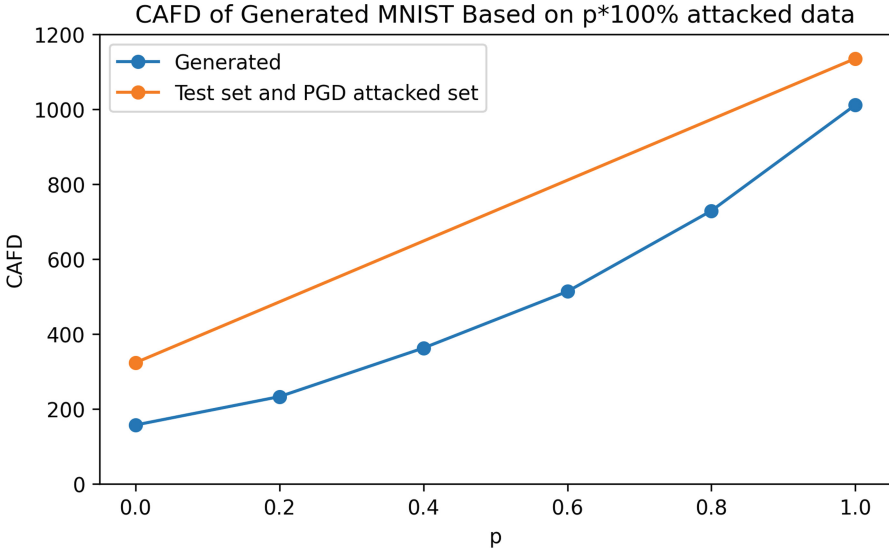


Fig. 6. Class-aware Fréchet Distance for a deceptive diffusion model where a percentage of the training data (horizontal axis) is replaced by its adversarially attacked counterpart. The ground truth dataset is MNIST. The straight line joins the CAFD for the test set at $p = 0$ and the PGDL2 attacked training set at $p = 1$. These two sets contain the same number of samples as the generated sets.

4 Conclusions

A traditional adversarial attack algorithm aims to perturb an existing image across a decision boundary. Instead, by training a generative diffusion model on adversarial data, we are able to create synthetic images that automatically lie on the wrong side of a decision boundary. This observation, which we believe to have been made for the first time in this work, reveals a new type of vulnerability for generative AI: if a diffusion model is inadvertently trained on fully or partially poisoned data then a tool may be produced that generates unlimited amounts of classifier-fooling examples.

In common with the AdvDiffuser algorithm in [5], when deliberately trained on adversarial data, a *deceptive diffusion* model has the potential to

- create effective adversarial images at scale, independently of the amount of training and test data available,

- create examples of misclassification that are difficult to obtain with a traditional adversarial attack; for example, in a healthcare setting when certain classes are underrepresented in the data [18].

This technique has applications for defence as well as attack, since it provides valuable new sources of data for adversarial training algorithms that aim to improve robustness.

There are many directions in which the deceptive diffusion idea could be pursued; notably, testing on other types of labeled image data, generating adversarial images that are successful across a range of independent classifiers, and finding computable signatures with which to identify this new type of threat.

Acknowledgments. LB is supported by the MAC-MIGS Centre for Doctoral Training under Engineering and Physical Sciences Research Council (EPSRC) grant EP/S023291/1. CFH received funding under EPSRC grants EP/T00097X/1, EP/R018634/1, and EP/T021020/1. DJH is supported by a fellowship from the Leverhulme Trust.

Disclosure of Interests. The authors declare no competing interests.

Data Statement. Code for these experiments is available from https://github.com/LucasBeerens/Deceptive_Diffusion.

Licencing Statement. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access* **6**, 14410–14430 (2018)
2. Beerens, L., Higham, D.J.: Adversarial ink: componentwise backward error attacks on deep learning. *IMA J. Appl. Math.* **89**, 175–196 (2024)
3. Cao, H., et al.: A survey on generative diffusion models. *IEEE Trans. Knowl. Data Eng.* **36**, 2814–2830 (2024)
4. Chen, J., Chen, H., Chen, K., Zhang, Y., Zou, Z., Shi, Z.: Diffusion models for imperceptible and transferable adversarial attack. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**(2), 961–977 (2024)
5. Chen, X., Gao, X., Zhao, J., Ye, K., Xu, C.-Z.: AdvDiffuser: natural adversarial example synthesis with diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4562–4572 (2023)
6. Chong, M.J., Forsyth, D.: Effectively unbiased FID and inception score and where to find them. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6070–6079 (2020)
7. Colbrook, M.J., Antun, V., Hansen, A.C.: The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smales 18th problem. *Proceedings of the National Academy of Sciences* (2021)

8. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794. Curran Associates, Inc. (2021)
9. European Comission: Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2021)
10. European Parliament: Amendments adopted by the European parliament on 14 june 2023 on the proposal for a regulation of the european parliament and of the council on laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2023)
11. Fawzi, A., Fawzi, O., Frossard, P.: Analysis of classifiers robustness to adversarial perturbations. *Mach. Learn.* **107**, 481–508 (2018)
12. Fowl, L., Goldblum, M., Chiang, P.-y., Geiping, J., Czaja, W., Goldstein, T.: Adversarial examples make strong poisons. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, pp. 30339–30351. Curran Associates, Inc. (2021)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) *3rd International Conference on Learning Representations*, San Diego, CA (2015)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, pp. 6626–6637 (2017)
15. Higham, C.F., Higham, D.J., Grindrod, P.: Diffusion models for generative artificial intelligence: an introduction for applied mathematicians, *SIAM Review* (to appear)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc. (2020)
17. Kim, H.: Torchattacks: a PyTorch repository for adversarial attacks. *arXiv preprint [arXiv:2010.01950](https://arxiv.org/abs/2010.01950)* (2020)
18. Ktena, I., et al.: Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* **30**, 1166–1173 (2024)
19. LeCun, Y., et al.: Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551 (1989)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998)
21. LeCun, Y., Cortes, C., Burges, C.J.C.: *The MNIST database of handwritten digits* (2010)
22. Liu, S., Wei, Y., Lu, J., Zhou, J.: An improved evaluation framework for generative adversarial networks. *arXiv preprint [arXiv:1803.07474](https://arxiv.org/abs/1803.07474)* (2018)
23. Luo, C.: Understanding diffusion models: a unified perspective. *arXiv:2208.11970* (2022)
24. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: *6th International Conference on Learning Representations*, Vancouver, BC, OpenReview.net (2018)
25. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition*, NV, USA, pp. 2574–2582. IEEE Computer Society (2016)

26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
28. Shafahi, A., Huang, W., Studer, C., Feizi, S., Goldstein, T.: Are adversarial examples inevitable?. In: International Conference on Learning Representations, New Orleans, USA (2019)
29. Sutton, O.J., et al.: Stealth edits for provably fixing or attacking Large Language Models. In: Neural Information Processing Society (NeurIPS), Vancouver, Canada, December 2024
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
31. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint [arXiv:1312.6199](https://arxiv.org/abs/1312.6199) (2013)
32. Tyukin, I.Y., Higham, D.J., Bastounis, A., Woldegeorgis, E., Gorban, A.N.: The feasibility and inevitability of stealth attacks. *IMA J. Appl. Math.* **89**, 44–84 (2024)
33. Tyukin, I.Y., Higham, D.J., Gorban, A.N.: On adversarial examples and stealth attacks in artificial intelligence systems. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2020)