GASTeNv2: Generative Adversarial Stress Testing Networks with Gaussian Loss

Cátia Teixeira[1](✉), Inês Gomes[1 [0009−0006−3104−4622]], Luís Cunha[1], Carlos Soares[1,2,3 [0000−0003−4549−8917]], and Jan N. van Rijn[4 [0000−0003−2898−2168]]

[1] Faculdade de Engenharia da Universidade do Porto, Portugal
{up200808037,up201706736}@up.pt, {ines.gomes,csoares}@fe.up.pt
[2] Laboratory for Artificial Intelligence and Computer Science (LIACC), Portugal
[3] Fraunhofer Portugal AICOS, Portugal
[4] Leiden Institute of Advanced Computer Science (LIACS), Netherlands
j.n.van.rijn@liacs.leidenuniv.nl

Abstract. As machine learning technologies are increasingly adopted, the demand for responsible AI practices to ensure transparency and accountability grows. To better understand the decision-making processes of machine learning models, GASTeN was developed to generate realistic yet ambiguous synthetic data near a classifier's decision boundary. However, the results were inconsistent, with few images in the low-confidence region and noise. Therefore, we propose a new GASTeN version with a modified architecture and a novel loss function. This new loss function incorporates a multi-objective measure with a Gaussian loss centered on the classifier probability, targeting the decision boundary. Our study found that while the original GASTeN architecture yields the highest Fréchet Inception Distance (FID) scores, the updated version achieves lower Average Confusion Distance (ACD) values and consistent performance across low-confidence regions. Both architectures produce realistic and ambiguous images, but the updated one is more reliable, with no instances of GAN mode collapse. Additionally, the introduction of the Gaussian loss enhanced this architecture by allowing for adjustable tolerance in image generation around the decision boundary.

Keywords: Synthetic Data Generation · Generative Adversarial Networks · Responsible Artificial Intelligence

1 Introduction

The widespread adoption of Machine Learning (ML) and Artificial Intelligence (AI) technologies has made the development of responsible AI practices crucial to ensure transparency and accountability in their applications [3]. In this context, model cards emerge as a tool to provide standardized documentation to communicate the characteristics and behaviours of deployed AI models [11].

To improve the transparency of ML models, Cunha *et al.* . introduced GASTeN — a deep generative model that generates realistic synthetic data near a

classifier's decision boundary [5]. The authors use the Fréchet Inception Distance (FID) [9] to measure image realism and introduce the Average Confusion Distance (ACD) to assess proximity to the decision boundary. Visual inspection shows that GASTeN can create realistic yet ambiguous images for human observers. However, the results are inconsistent due to a challenging trade-off between ACD and FID, which should be as low as possible. Despite this, GASTeN hand-picked samples can fill model cards, providing insights into the ML model's decision boundary.

To enhance the GASTeN framework, we introduce two modifications in this research: architectural updates and a novel generator loss function approach. On the architectural side, we replaced the generator's initial linear layer and the discriminator's last linear layer with convolutional layers. Then, we updated the generator loss mechanism to include a multi-objective measure that combines the standard generator adversarial loss with a Gaussian loss applied to the classifier's output, targeting a distribution centered at 0.5 with adjustable variance.

Our empirical study shows that GASTeNv2 achieves lower ACD values and consistent FID across low-confidence regions. We conclude that both architectures produce realistic and ambiguous images, being the updated one more reliable, with no instances of GAN mode collapse — problem when the generator outputs non-diverse images despite varying inputs [7].

The contributions of this work are as follows:

1. updated architecture to enhance performance and stability;
2. incorporation of a multi-objective measure with a Gaussian loss to provide more control over the generation process.

These contributions result in more consistent images that remain close to the decision boundary, providing better examples for populating model cards and ultimately enhancing our understanding of the classifiers' decision boundaries.

2 Related Work

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of deep generative models that feature two neural networks: a generator that creates images from noise and a discriminator that assesses their authenticity [8]. GANs have gained significant notoriety in computer vision due to their ability to generate high-quality, realistic images. However, training GANs effectively is challenging. Reaching Nash equilibrium — where simultaneous updates of the two neural networks might not lead to convergence [7] — and mode collapse are prominent issues.

Since the GAN architecture proposal, numerous adaptations have been developed. Research works like WGAN [2], DCGAN [15], and InfoGAN [4] have significantly improved the training processes by addressing issues like training stability and feature representation. These architectures have also enhanced loss functions through various improvements: they integrate adversarial losses [14], application-specific losses [6], or combinations thereof by adopting a multi-objective measure for loss function optimization [1].

Generative Adversarial Stress Test Networks GASTeN [5] uses Deep Convolutional GAN (DCGAN) [15] to synthesize realistic yet challenging data. The method aims to provide insights into deep-learning image classifiers by generating examples near the decision boundary in binary classification. GASTeN introduces a two-step training process: firstly, it pre-trains the original DCGAN to teach the generator; secondly, it integrates classifier predictions, enabling further training of the generator with a new loss function (see Equation 1). The new term of the loss function — confusion distance (cd) — measures the distance from the classifier's prediction to the decision threshold multiplied by α to balance the image quality and ambiguity [5].

$$\mathcal{L}_G^{GASTeN} = \mathcal{L}_G^{GAN} + \alpha \cdot cd(C(G(z))) \tag{1}$$

The authors evaluated their method on binary subsets of MNIST [10] and Fashion MNIST [16] using the FID score as the image quality metric. FID assesses realism by comparing feature distributions with real images. The authors then introduce a new metric, the average confusion distance (ACD), that measures the average cd of all images. By visually inspecting some generated samples, the authors found images visually plausible yet ambiguous, even to human observers. However, The authors found it challenging to identify images that had both low ACD values and low FID scores [5]. Therefore, they concluded that it might not always be feasible to develop a generator that consistently deceives the target classifier while still producing images that are considered realistic.

2.2 Gaussian Negative Log Likelihood Loss

The Gaussian Negative Log Likelihood Loss (GNLL) is used to estimate both the mean (μ) and variance (σ) of a target's probability distribution as a function of the input. It is based on a presumed model of the target's error-distribution [12]. The GNLL formula, as outlined in Equation 2, incorporates these parameters where σ^2 represents the variance, μ indicates the mean of the distribution, and y is the observed actual value. According to this formulation, a lower GNLL value suggests that the actual value is closer to the predicted mean, aligning more closely with the expected target distribution.

$$\mathcal{L}_{\text{GNLL}} = \frac{1}{2}\log(\sigma^2) + \frac{(y-\mu)^2}{2\sigma^2} \tag{2}$$

A study improved deep age estimation using Gaussian loss [13]. The researchers used a Convolutional Neural Network (CNN) with a mean-variance loss to predict age from images. This approach minimized the discrepancy between predicted and actual ages and penalized variance in the age distribution, achieving state-of-the-art results.

3 Gaussian Conditional GASTeN

To develop and deploy responsible models, we are developing methods to investigate the decision boundary of a specified classifier. While exploring state-

of-the-art methods, we identified the GASTeN [5] architecture, which generates realistic images near a classifier's decision boundary. However, GASTeN often struggles to balance FID and ACD scores, meaning achieving realistic images compromises proximity to the decision boundary.

To address this challenge, we suggest two improvements to GASTeN related to the intrinsic architecture and the generator loss function. Cunha *et al.* implemented two modifications to the original DCGAN architecture: they replaced the initial convolutional layer in the generator with a linear layer and, similarly, the last convolutional layer in the discriminator with a linear layer. Our first update was to revert these changes to preserve the architecture of the original DCGAN. Then, we updated the generator loss, introducing a multi-objective measure incorporating the standard generator loss with a Gaussian loss applied to the classifier's output. This Gaussian loss targets a distribution centered at the decision boundary — that we define as 0.5 — with an adjustable variance (σ^2), which acts as a hyperparameter and is weighted by α that balances the data realism and confusion. The improvements to the original GASTeN framework are detailed in Fig. 1, and the new loss is specified in Equation 4.

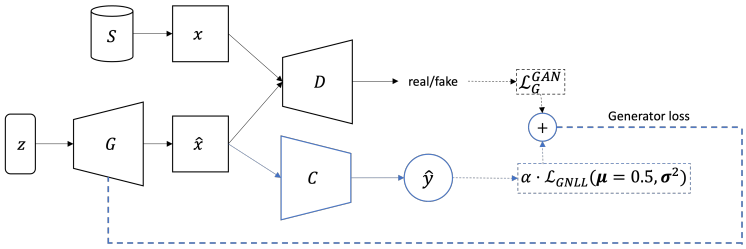


Fig. 1: Schematic overview of GASTeNv2. The generator loss is a multi-objective measure of the generator and Gaussian losses.

$$\mathcal{L}_G^{GASTeN} = \mathcal{L}_G^{GAN} + \alpha \cdot \mathcal{L}_{GNLL} \tag{3}$$

$$= \mathcal{L}_G^{GAN} + \alpha \cdot \frac{1}{2}\log(\sigma^2) + \frac{(C(G(z)) - 0.5)^2}{2\sigma^2} \tag{4}$$

This approach shifts from measuring the average distance of all images from the decision boundary to assessing their deviation from a Gaussian distribution centered at 0.5. GNLL captures the variance around this center, representing the low-confidence region better. Additionally, GNLL allows for fine-tuning the output distribution through the variance parameter, enabling precise adjustments in model predictions and avoiding issues like mode collapse in GANs.

This enhancement maintains GASTeN's principle of generating images based on a differentiable function. Although we set the Gaussian distribution mean at 0.5, other values can be chosen to condition image generation for different

classifier probabilities, addressing various scenarios and objectives, *e.g.*, setting the mean to 1 would generate images with total confidence in the positive class.

4 Experimental Setup

4.1 Dataset

To assess the effectiveness of our approach, we conduct experiments using the MNIST dataset, given that the results can be easily interpreted with general knowledge, as opposed to other datasets that may need expert evaluation. Considerations like dataset size and image characteristics were also taken. As GASTeN exclusively works with binary classification tasks [5], we curate binary subsets of the original dataset by isolating images belonging to two specific classes. In this context, we focus on subsets representing similar concepts: $7vs1$ and $5vs3$.

4.2 Architecture

The base architecture is the same as GASTeN. The used GAN follows the DCGAN [15] architecture, with the non-saturating GAN loss proposed by Goodfellow *et al.* [8]. Optimizer and training hyperparameters are set according to the original DCGAN work [15]. For the binary classification, we use a CNN-based architecture with two convolutional blocks. Each block contains a convolution and a max pooling operation. The first block uses nf number of filters in the convolutions, and the second uses $2 \cdot nf$ blocks. Classifiers with smaller nf values have less capacity. Each classifier was trained for 1 epoch.

4.3 Evaluation Strategy

To measure performance, we use the same metrics as Cunha *et al.* , enabling direct comparison. We use accuracy to assess classification performance. To evaluate GASTeN's performance, we use FID and ACD scores. The FID score measures the realism of generated images, while the ACD score assesses the generator's capacity to produce images with low classification confidence.

Both ACD and FID scores should be minimized. However, since our primary goal with GASTeN is to generate images near the decision boundary, we prioritize the ACD metric, provided the FID scores remain within a reasonable range. To further compare our experiments, we define a low confidence region where $ACD < 0.1$, corresponding to classifier probabilities between 0.4 and 0.6. We then select images generated within this region and measure their FID scores. Finally, we compare these FID scores to the overall FID scores and evaluate how they perform relative to the other experiments.

Additionally, we perform a manual inspection of the generated samples to provide a qualitative assessment. This inspection facilitates the rapid identification of collapsed GANs and allows us to evaluate the ambiguity of the images.

4.4 Experiments

We structured our experimental approach into two distinct phases: initially, we assess whether the updated architecture improves the original; then, we select the best architecture to evaluate the efficacy of the new loss function. To this end, we have designed two specific experiments:

1. **Architecture**: comparing the confusion distance (CD) loss, *i.e.*, the original loss defined by Cunha *et al.* , using the *original* GASTeN architecture against the CD loss using the *updated* GASTeN architecture;
2. **Losses**: comparing the *CD* loss versus the *Gaussian* loss on the winning architecture.

One of GASTeN's objectives is to stress classifiers in a data-driven way by identifying data points that push the model to its limits — in this case, the decision boundary [5]. We continue this goal from the original study by stressing three classifiers of varying capacities. We experimented with CNN using 1, 2, and 4 filters, where fewer filters indicate reduced classifier capacity.

For the original GASTeN configuration, we selected α values of 20, 25, and 30 based on findings from Cunha *et al.* . The authors found that these values are the ones that decrease the ACD score the most. However, it can impact the realness of the images. This strategy is appropriate given our focus on generating images near the decision boundary.

For the Gaussian loss version of GASTeN, the α value also weights the impact of the classifier on the generator loss. Setting α to 0 effectively eliminates the Gaussian loss constraint. After manually assessing that the GNLL and the generator losses are on comparable scales, we opted to test α values of 0.5, 0.8, and 1.0, where 1.0 levels the losses.

Additionally, the GNLL loss requires specifying a variance (σ^2). We experimented with variances of 0.01 and 0.005. A variance of 0.01 corresponds to a standard deviation of 0.1, implying that 68.2% of the values will fall between classifier probabilities of 0.4 and 0.6. A variance of 0.005 leads to a standard deviation of 0.07, meaning that 95.4% will lie between 0.64 and 0.36. We selected these variance values to explore different tolerances on the classifier confidence, where a higher variance allows greater tolerance and a lower variance leads to a closer concentration of images around the decision boundary.

The last hyperparameter we adjusted was the number of GAN pre-training epochs — that we denote by β — for which we explored values of 5 and 10. We chose these values based on observations from the original GASTeN implementation, which the authors found to minimize instances of GAN mode collapse and enhance the FID scores.

All combinations of the mentioned hyperparameters are tested for each classifier, and all GASTeN versions are trained with 40 epochs. Also, all experiments were conducted using a machine equipped with one Tesla T4 GPU, ensuring sufficient computational power for training and inference phases. The source code for the experiments is available on GitHub[1].

[1] https://github.com/crdsteixeira/Thesis-Hubris_Benchmarking_with_AmbiGANS

5 Results

Following the original GASTeN architecture, our first step was to train the classifiers. The accuracies of these classifiers are detailed in Table 1. The selected MNIST subsets exhibit varying levels of complexity, with *7 vs 1* being the easiest and *5 vs 3* the most challenging. Even the least accurate classifier achieves an accuracy of 91.85%, demonstrating that all classifiers can make reliable predictions. The following sections reflect the defined experimental setup.

Table 1: Performance metrics of CNN models on various datasets.

Dataset	Classifier	Accuracy	Loss
7 vs 1	CNN ($nf = 4$)	99.12%	0.036
	CNN ($nf = 2$)	98.15%	0.061
	CNN ($nf = 1$)	97.55%	0.177
5 vs 3	CNN ($nf = 4$)	96.32%	0.116
	CNN ($nf = 2$)	94.37%	0.171
	CNN ($nf = 1$)	91.85%	0.292

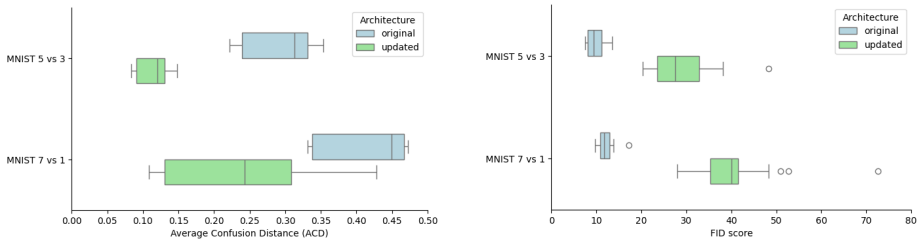
5.1 Architecture

After running all combinations for each architecture, we conducted a total of 36 experiments for each MNIST subset. In the *7 vs 1* MNIST subset, a manual inspection of the generated images revealed two experiments that resulted in GAN mode collapse, both from the original architecture and $\alpha = 30$. In contrast, for the *5 vs 3* MNIST subset, no GANs collapsed.

From the non-collapsed GANs, we compared the experiments' ACD and FID scores. Fig. 2 illustrates the distributions of ACD and FID scores for all experiments, categorized by architecture (*original* versus *updated*) and dataset. We observed that the ACD values are lower for the updated architecture than the original architecture, whereas the FID scores are proportionally higher.

To evaluate the behaviour of GASTeN in regions of low confidence, we generated 10,000 images and computed both their global FID score and the FID score for images with $ACD < 0.1$. This analysis used the selected GASTeN hyperparameters that yielded the lowest ACD scores for each classifier and corresponding architecture. Detailed GASTeN hyperparameters are in Table 2. Table 3 show the average and standard deviation of the results over three runs, along with the percentages of images falling within this ACD range.

We observe that although the original architecture yields the best FID scores overall, there is a significant reduction in the number of images when considering only images close to the decision boundary. In this low-confidence region, the reduction is so drastic that we cannot compute the FID scores for two of the



(a) ACD distribution (b) FID distribution

Fig. 2: Comparison of FID and ACD scores across all hyperparameter configurations, not collapsed, grouped by dataset and architecture.

Table 2: Performance metrics of different classifiers with various architectures and loss functions, for MNIST subset *5 vs 3*.

Classifier	**Architecture**	**Loss**	**Hyperparameters**	**FID**	**ACD**
CNN ($nf = 1$)	original	CD	$\beta = 5$; $\alpha = 20$	7.44	0.221
	updated	CD	$\beta = 10$; $\alpha = 30$	22.37	0.083
	updated	Gaussian	$\beta = 5$; $\alpha = 1$; $\sigma^2 = 0.005$	28.42	0.081
CNN ($nf = 2$)	original	CD	$\beta = 10$; $\alpha = 30$	9.54	0.290
	updated	CD	$\beta = 10$; $\alpha = 25$	23.42	0.116
	updated	Gaussian	$\beta = 10$; $\alpha = 1$; $\sigma^2 = 0.005$	29.95	0.103
CNN ($nf = 4$)	original	CD	$\beta = 10$; $\alpha = 30$	13.46	0.331
	updated	CD	$\beta = 10$; $\alpha = 30$	26.69	0.127
	updated	Gaussian	$\beta = 5$; $\alpha = 1$; $\sigma^2 = 0.005$	34.63	0.110

classifiers, due to a requirement of at least 2048 images. For the CNN with 1 filter on the original architecture, the FID score increases but not as significantly as with the updated version. Contrariwise, for the updated architecture, the FID score remains consistent in the low-confidence region, which is an advantage.

In addition to quantitatively evaluating our approach, we manually inspect the generated images. Fig. 3 resembles the one shown by Cunha *et al.* , where images on the far left are classified with high confidence as negative by the classifier, and those on the far right are classified with high confidence as positive. We conclude that the images are both realistic and ambiguous when the classifier probability is 0.5. Both the original and updated architectures generate images of very similar quality. However, the updated version has the advantage of producing more images close to the decision boundary.

Overall, we conclude that the updated architecture is superior, achieving lower ACD values while maintaining FID scores that result in realistic images.

Architecture	Loss	Classifier	FID (global)	FID ($ACD < 0.1$)	#images ($ACD < 0.1$)
original	CD	CNN (*nf=1*)	5.84 ± 0.11	9.21 ± 0.08	2620.33 ± 29.67
		CNN (*nf=2*)	7.82 ± 0.10	-	2021.00 ± 46.81
		CNN (*nf=4*)	11.82 ± 0.04	-	1578.67 ± 37.00
updated	CD	CNN (*nf=1*)	20.54 ± 0.04	20.10 ± 0.08	6756.00 ± 2.64
		CNN (*nf=2*)	20.68 ± 0.28	20.37 ± 0.61	4777.33 ± 21.12
		CNN (*nf=4*)	24.72 ± 0.11	24.09 ± 0.12	4595.00 ± 24.24

Table 3: Comparison of FID scores for the original and updated architectures with CD loss and different CNN configurations for MNIST subset *5 vs 3*.

5.2 Losses

After selecting the *updated* architecture as the best one, we tested whether the FID and ACD scores improved with the new *Gaussian* loss. By running all combinations of hyperparameters for the new architecture and both losses, we conducted 54 experiments for each MNIST subset. Manual inspection revealed that none of the GANs collapsed.

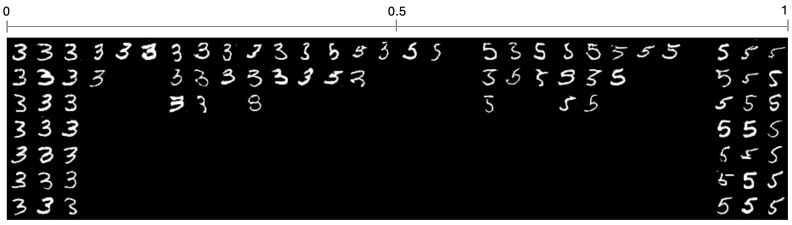
Fig. 4 illustrates the distributions of ACD and FID scores for all experiments, categorized by type of loss (*CD* or *Gaussian*) and dataset. The difference in the distributions is not as apparent as before, with both ACD and FID scores tending to be lower when using the Gaussian loss. When selecting GASTeN hyperparameters to achieve lower ACD or FID values, the version with Gaussian loss consistently attains the minimum values.

When comparing Table 4 with Table 3, we observe that the FID scores are slightly higher than those from GASTeN with CD loss. However, with the Gaussian loss, the FID score remains consistent even when we select only the images close to the decision boundary.

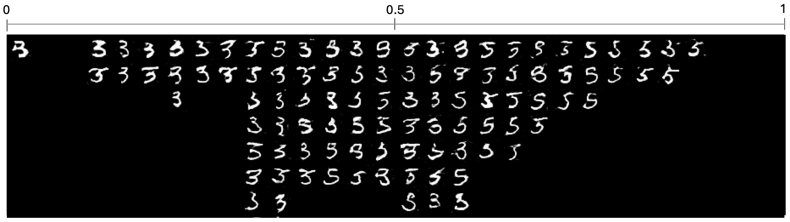
Architecture	Loss	Classifier	FID (global)	FID ($ACD < 0.1$)	#images ($ACD < 0.1$)
updated	Gaussian	CNN (*nf=1*)	26.61 ± 0.09	26.14 ± 0.10	6613.33 ± 61.76
		CNN (*nf=2*)	28.64 ± 0.03	28.56 ± 0.18	5482.33 ± 24.54
		CNN (*nf=4*)	32.77 ± 0.23	32.18 ± 0.36	5073.66 ± 14.22

Table 4: Comparison of FID scores for the updated architecture with Gaussian loss and different CNN configurations for MNIST subset *5 vs 3*.

In addition to quantitatively evaluating our approach, we manually inspect the generated images. Observing Fig. 5, we conclude that the images are both realistic and ambiguous when the classifier probability is 0.5, with quality very similar to the previously shown samples.



(a) Original architecture with CD loss



(b) Updated architecture with CD loss

Fig. 3: Sample of images generated for the best GASTeN ACD score during stress testing of a CNN with 4 filters, trained on the MNIST *5 vs 3* dataset.

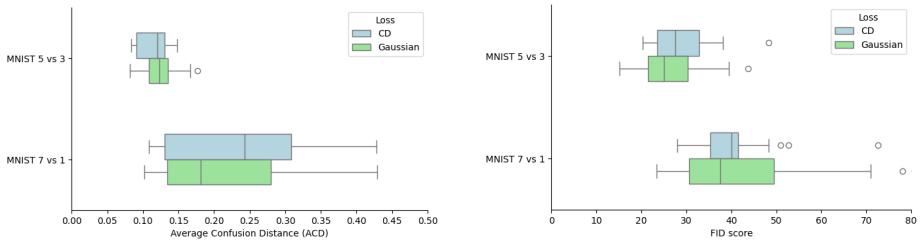
Overall, the new *Gaussian* loss does not immediately outperform the *CD* loss, as the FID and ACD scores are very similar based on our limited experiments. However, the Gaussian loss offers a significant advantage: it allows us to tune the variance hyperparameter, enabling adjustable tolerance for generating images around the mean of 0.5 probability. Therefore, we select the Gaussian loss and the winning loss.

6 Conclusions

This research proposes an improved GASTeN framework with architectural modifications and a new generator loss. This GAN-based framework generates realistic synthetic data in the low-confidence region of a given classifier.

Our study on binary classification tasks using MNIST dataset subsets found that while the original GASTeN architecture yields the highest FID scores, the updated version achieves lower ACD values and consistent performance across low-confidence regions. Both architectures produce realistic and ambiguous images; however, the updated one is more reliable, with no instances of GAN mode collapse. The introduction of Gaussian loss enhances this architecture by allowing for adjustable tolerance in image generation around the decision boundary, thereby improving its robustness and applicability.

Ultimately, the generated images can be integrated into model cards, providing valuable insights into the classifier's decision-making processes and improving transparency in machine learning models. Future work should extend the GASTeN framework to more complex tasks and multi-class classifications.



(a) ACD distribution (b) FID distribution

Fig. 4: Comparison of FID and ACD scores across all hyperparameters, grouped by dataset and optimization measure loss.

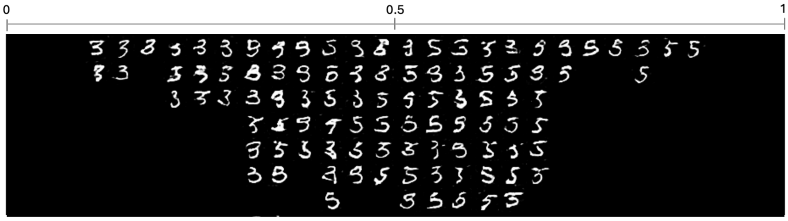


Fig. 5: Sample of images generated by GASTeN (with updated architecture and Gaussian loss) for the hyperparameters that attained the best ACD score during stress testing of a CNN with 4 filters, trained on the MNIST *5 vs 3* dataset.

Acknowledgements. This work was partially funded by projects AISym4Med (101095387) supported by Horizon Europe Cluster 1: Health, ConnectedHealth (n.º 46858), supported by Competitiveness and Internationalisation Operational Programme (POCI) and Lisbon Regional Operational Programme (LISBOA 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF) and NextGenAI - Center for Responsible AI (2022-C05i0102-02), supported by IAPMEI, and also by FCT pluriannual funding for 2020-2023 of LIACC (UIDB/00027/2020_UIDP/00027/2020). The computational resources of Google Cloud Platform were provided by the project CPCA-IAC/AF/594904/2023.

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Albuquerque, I., Monteiro, J., Doan, T., Considine, B., Falk, T.H., Mitliagkas, I.: Multi-objective training of generative adversarial networks with multiple discriminators. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 202–211. PMLR (2019)

2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning. vol. 70, pp. 214–223. PMLR (2017)
3. Arrieta, A.B., Rodríguez, N.D., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)
4. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016. pp. 2172–2180 (2016)
5. Cunha, L., Soares, C., Restivo, A., Teixeira, L.F.: Gasten: Generative adversarial stress test networks. In: Advances in Intelligent Data Analysis XXI - 21st International Symposium on Intelligent Data Analysis. vol. 13876, pp. 91–102. Springer (2023)
6. Dunn, I., Pouget, H., Melham, T.F., Kroening, D.: Adaptive generation of unrestricted adversarial inputs. CoRR (2019)
7. Goodfellow, I.J.: NIPS 2016 tutorial: Generative adversarial networks. CoRR (2017)
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. pp. 2672–2680 (2014)
9. 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: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. pp. 6626–6637 (2017)
10. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
11. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 220–229. ACM (2019)
12. Nix, D., Weigend, A.: Estimating the mean and variance of the target probability distribution. In: Proceedings of 1994 IEEE International Conference on Neural Networks. vol. 1, pp. 55–60 (1994)
13. Pan, H., Han, H., Shan, S., Chen, X.: Mean-variance loss for deep age estimation from a face. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition. pp. 5285–5294. Computer Vision Foundation / IEEE Computer Society (2018)
14. Pan, Z., Yu, W., Wang, B., Xie, H., Sheng, V.S., Lei, J., Kwong, S.: Loss functions of generative adversarial networks (gans): Opportunities and challenges. IEEE Trans. Emerg. Top. Comput. Intell. **4**(4), 500–522 (2020)
15. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations (2016)
16. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. CoRR (2017)