Conventional Augmentation is More Effective than ImageGPT and GANs? A Comparison of Synthetic Data Evaluation Methods

Andrew Kennedy
Defence Data Research Centre
University of Exeter
Exeter, United Kingdom
a.kennedy3@exeter.ac.uk

Richard Everson
Defence Data Research Centre
University of Exeter
Exeter, United Kingdom
r.m.everson@exeter.ac.uk

***Abstract*—Generative synthetic data models have proven effective in tasks like image synthesis and data augmentation. This study aims to better understand the effectiveness and performance of generative synthetic data generators (SDGs), specifically ImageGPT and GANs, against conventional methods of data augmentation for downstream supervised learning. We evaluate downstream image classifier accuracy and Fréchet Inception Distance (FID) as measures of task efficacy and data fidelity. While FID was useful, it was not directly predictive of downstream classifier performance, with its relationship to classification outcomes varying across different SDGs, since each generator introduced different types of dissimilarity. We found that conventional, geometric and colour-space transformations to generate synthetic data achieved greater classification accuracy than the ImageGPT model and only the best GAN model could match this accuracy with moderate amounts of synthetic data. While GANs were shown to be a useful method, and are applicable to a wide variety of data types, image augmentation with conventional transformations resulted in comparable or higher downstream classifier accuracy. In addition, these transformations also provide greater control over the extent of augmentation and can produce a larger amount of useful data for the downstream task.**

***Index Terms*—Synthetic Data, validation methods, classification, FID, transformations, GAN, ImageGPT.**

I. Introduction

Synthetic data is gaining increasing interest and development, and is being proposed as a solution to issues like limited data, privacy concerns, and bias in datasets [1]–[3]. While some of these challenges can be addressed during model training, the ability to provide a synthetic dataset overcoming them enables more open sharing of data, models and results.

The objective of this study is to improve image classifier performance in the situation of limited data by augmenting an original dataset with synthetic data. We explore whether synthetic data can enhance the performance and generalisability of image classifiers, especially when training data is scarce, imbalanced, or lacks diversity [4]. While traditional image augmentation methods, such as affine transformations of training images have been effective in computer vision tasks, they maybe be limited in capturing deeper variations or complex structural relationships within data. Emerging generative approaches, such as Generative Adversarial Networks (GANs) [5] and ImageGPT [6], offer promising alternatives by learning intricate patterns directly from data, enabling the generation of realistic samples. Here we investigate the effectiveness of these newer synthetic data generators (SDGs) compared to conventional augmentation techniques for improving image classification performance under limited data conditions [7].

To assess the quality and utility of synthetic data, we compare two evaluation methods: classifier accuracy, which measures task-specific performance, and Fréchet Inception Distance (FID), a fidelity metric to quantify the similarity between synthetic and real datasets. While classifier accuracy provides a direct measure of utility, it requires labelled data for training. In contrast, FID offers an approach to evaluating synthetic data without relying on labels. This comparison raises an important question: How reliably can fidelity metrics like FID predict performance on downstream tasks?

This paper makes the following key contributions:

- A comparison of generative synthetic data methods (GANs and ImageGPT) with conventional augmentation techniques for image classification under limited data conditions, demonstrating that conventional augmentations are more flexible and often yield better performance.
- An evaluation of the relationship between classification accuracy and FID as synthetic data assessment measures, finding that, while there is a link between the two, FID is not a strong predictor of classifier performance.

The remainder of this paper is organised as follows. Section II introduces the assessment and generation methods. Section III compares the assessment methods, and Section IV the SDGs. In Section V, we discuss these findings and conclude the study.[1]

This work was supported by the Defence Science and Technology Laboratory.

[1]Supplementary material is available on our github repository: `https://github.com/ar929/syn-data-comp`.

II. Techniques

A. How to Assess Synthetic Image Data?

It can be difficult to quantitatively assess generated synthetic image quality or its likeness or similarity to the original images, since a simple geometric distance between pixel values fails to capture higher level feature similarities. Even if very high perceived image quality is achieved, the synthetic images may not necessarily be desirable or suitable for downstream tasks such as classification or object detection [8].

To address this, Jordon *et al.* [9] discuss three key attributes of synthetic data that can be considered for their quality: *utility*, *fidelity* and *privacy*, which are often balanced against each other. We do not consider privacy in this study, concentrating on the capability of synthetic data to alleviate data scarcity. Utility refers to the efficacy of synthetic data for a given task, whereas fidelity refers more directly to the degree of (statistical) similarity of synthetic data to the data from which it was generated.

Fidelity and utility are linked and often considered together – an SDG with high fidelity will typically have high utility. However, the converse is not necessarily true – high utility does not always imply high fidelity. Good synthetic data generators aim to sample from the same parent distribution as the original data, ideally capturing aspects of this distribution not fully seen in the original dataset. A high-utility synthetic dataset may sacrifice fidelity by re-weighting the parent distribution, perhaps by prioritising areas of the data near a decision boundary, or even including data outside the parent distribution. For training a classifier, additional dissimilarity in the synthetic data may improve its generalisability thus improving the utility while decreasing fidelity. In this study, we consider two methods of evaluating synthetic data: the accuracy of a classifier trained on the synthetic-augmented dataset, and the Fréchet Inception Distance (FID) [10] between the original and synthetic datasets.

1) Utility – Classification Accuracy: For our use-case and our utility measure, we consider image classification on the well-known CIFAR-10 dataset Krizhevsky *et al.* [11]. The CIFAR-10 dataset consists of $60,000$ 32×32 colour images in 10 classes, with $6,000$ images per class, split into $50,000$ training images and $10,000$ test images.

To establish a baseline, we trained a simple CNN model, based on [12], using the full dataset. The goal was to create a model for a limited data regime, so that it could benefit from additional synthetic training images. This model, consisting of five convolutional layers, four fully connected layers, and a softmax classifier, achieved a top-1 classification accuracy of over 0.8 after parameter optimisation when trained on the full training set (see Fig. 1 for the corresponding learning curve).

For a limited data regime, we selected a smaller training subset of $n_{orig} = 1024$, chosen from the left side of the learning curve to clearly show the improvements made by synthetic data in scenarios where data is scarce.

As can be seen from Fig. 1, with this reduced amount of data and a standard number of training epochs of 100, the test's top-1 classification accuracy was 0.422.

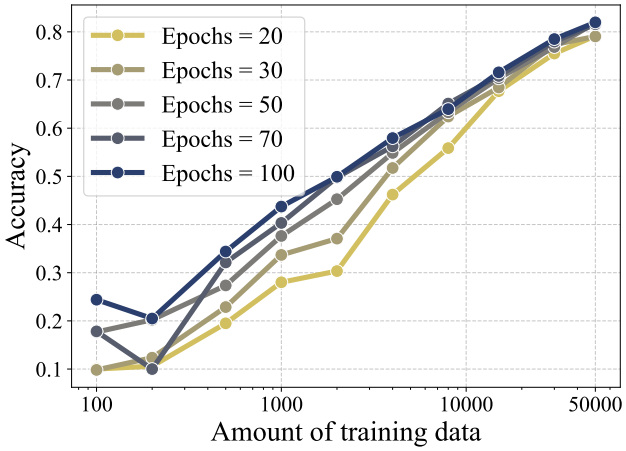


Fig. 1: Learning curve showing top-1 test accuracy versus amount of training data for our CNN model on CIFAR-10 dataset, varying number of training epochs.

To better compare synthetic-augmented datasets with original-only data, we compute the *relative accuracy* by dividing the classifier accuracy by the baseline of 0.422. A relative accuracy above 1 indicates that improvement has been made by augmenting with a given synthetic dataset.

2) Fidelity – Fréchet Inception Distance: The Fréchet Inception Distance (FID) was introduced by Heusel *et al.* [10] as a metric for evaluating the quality of generated images, primarily in the context of training GANs. FID aims to quantify both the realism and diversity of generated images by employing a pre-trained Inception-v3 image classification model [10], [13] to obtain a lower-dimensional feature space tailored for computer vision. In this process, feature representations are extracted from both the original and synthetic images. The mean and covariance statistics capture the distributions of these features, and their similarity is assessed using the Fréchet distance – a measure of similarity between two curves [14], [15]. A lower FID value indicates a higher similarity in distribution, suggesting that the synthetic images exhibit similar quality and diversity to the original images.

FID is a commonly used metric, particularly for GAN evaluation [16], and has been shown to correspond well in many cases to human judgement of image quality [10]. Given its widespread application in assessing synthetic image generation, FID has been suggested as a potential indicator of classifier performance because use of a smaller, tailored feature space could allow better representation of the underlying image characteristics and semantics. Note that FID does not use image labels and can be fast, since it does not require training of a downstream classifier. We therefore focus on the capability of FID to predict test classification accuracy.

There are some technical limitations to the use of FID, as noted by Borji [16]. We have mitigated these in our study by carrying out all experiments with the same PyTorch

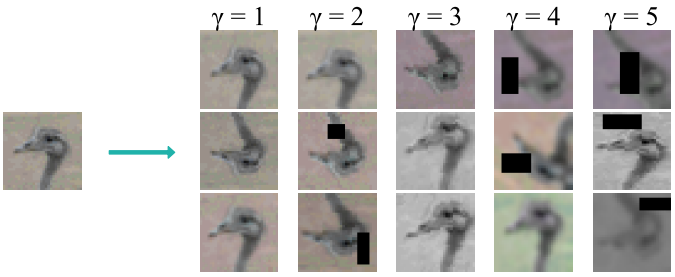


Fig. 2: Samples of conventional image transformations for different intensity parameter values, γ, using combinations of random horizontal flip, resizing, rotation and shearing, rectangle erasure, blurring, colour-jittering & grey-scaling.

implementation and processing the images in the same way to ensure the scores are comparable [17], and by implementing the extrapolation approach of Chong *et al.* [18] to obtain an unbiased estimate of FID with respect to the sample size of synthetic images, FID_∞.[2]. Other limitations are mentioned in Section V-A

B. Conventional Transformation Augmentation

Image augmentation by transformation of images is a fundamental technique in computer vision tasks [19], [20], improving model robustness and generalisability. Augmentation of training data is commonly used as part of the pipeline of training a neural network or other ML model on image data. Transformations can include compositions of affine transforms, such as translation, rotation, scaling, shearing and horizontal reflection, and other transforms such as blurring, random erasures, colour-jittering or random grey-scaling.

In this study, we applied standard conventional image augmentation techniques using PyTorch's `torchvision` transforms library [21]. We combined these transforms in series, using a single heuristically-defined intensity parameter γ to control the extent or probability of each transform. γ is defined such that mild transformations are produced for $\gamma = 1$, extreme transformations are produced for $\gamma = 5$, and $\gamma = 0$ is the identity transformation.[3] An illustrative example of the effect of γ is shown in Fig. 2. To optimise use of these transformations, we varied two key parameters: transformation intensity γ and the quantity of transformed data used to augment the original dataset, and this is shown in Fig. 3.

Fixing $n_{orig} = 1024$, we calculated the accuracy of our CNN classifier, trained on different amounts and intensities of augmented data together with the original data. We define α as the ratio of additional conventional synthetic data with which the original data is augmented:

$$n_{syn} = n_{orig} \times (\alpha - 1)$$

[2]We reproduced the results of [18] and verified that this method did achieve a less biased estimate.

[3]Details of our transformation function are available at: `https://github.com/ar929/syn-data-comp`.

where we subtract 1 since the original data is retained in our augmentation regime. Thus in these experiments, the classifier was trained on a total of

$$n_{orig} + n_{syn} = \alpha \times n_{orig}$$

examples.

For several fixed values of α, we compared accuracy against transformation intensity γ. As can be seen in Fig. 3a, when a sufficient amount $\alpha \gtrsim 4$ of augmented data was used, the accuracy lies on a roughly n-shaped curve – very small γ yields little improvement over using only original data; moderate γ yields the best results; and as γ increases further, the accuracy begins to decrease again. We know that for small γ the synthetic data is very similar to the original data, and so provides little additional information or benefit. On the other hand, for larger γ the augmenting images are too noisy and dissimilar to the training data and thus they are less representative of the original data distribution. Consequently, a moderate $\gamma = 3$ was chosen for the remainder of our experiments.

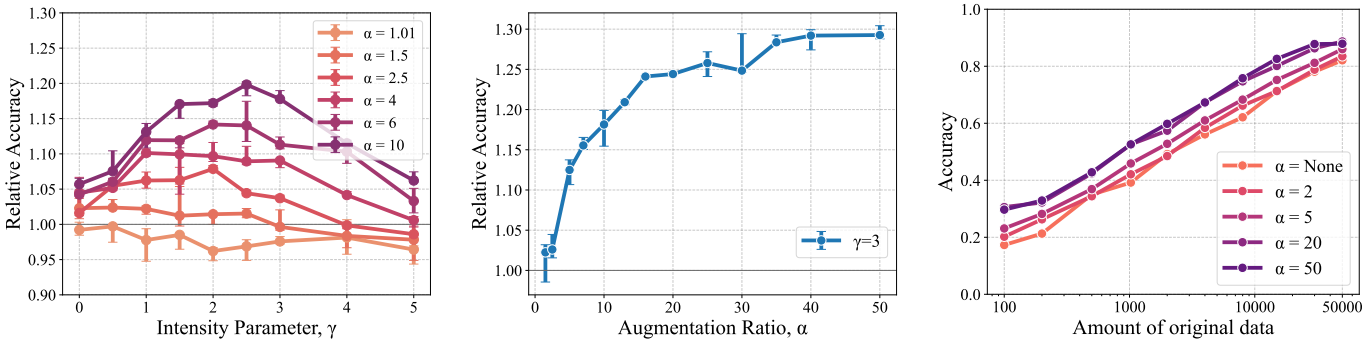
Fig. 3b shows the variation of the classifier accuracy with the quantity of data augmentation, with $\gamma = 3$ fixed. Continual improvement was seen to a surprisingly large α, with some accuracy improvement from up to 20-40 augmented images per original image. Fig. 3c shows learning curves on the full CIFAR-10 dataset with differing amounts of conventional augmentation. The classifier accuracy continues to improve as more original data is added, even when using significant additional synthetic augmentation (when α is large). This verifies that this conventional augmentation is of significant value for both large and small amounts of original data.

C. ImageGPT as a Synthetic Data Generator

This experiment investigates ImageGPT [6], a novel method of synthetic data generation. Developed by OpenAI, ImageGPT employs a large transformer framework similar to that of their ChatGPT LLM, but trained on pixel sequences rather than text.

We compared conventional synthetic augmentation to newer generative methods. Many generative methods such as GANs (discussed in Section II-D) require expensive training on the available data; whereas ImageGPT is provided as a pre-trained model, trained on the commonly used ImageNet dataset [6], which can be run directly to produce images from given input image data. Other popular and often openly accessible image generators such as DALL-E [22] have been highly successful, but these take textual prompts to describe the desired image; whereas for our synthetic image generation, the objective is to generate new images from existing images.

ImageGPT generates images by taking a certain proportion of a given original image, and extrapolating to complete the image based on observed patterns from the ImageNet dataset [23]. In our study, we used prompts containing the top $\beta = \frac{1}{4}/\frac{1}{2}/\frac{3}{4}$ of the original CIFAR-10 images' pixels. We also refer to the $\frac{1}{4}/\frac{1}{2}/\frac{3}{4}$ prompts as small, medium and large, respectively. These prompt sizes β control the degree of



(a) Accuracy against intensity of conventional augmentation γ for $n_{orig} = 1024$. $\gamma = 0$ is the identity transformation. Error bars show median and IQR across 5 repeats.

(b) Accuracy against augmentation ratio α with $\gamma = 3$ and $n_{orig} = 1024$. Error bars show median and IQR across 5 repeats.

(c) Learning curve on full CIFAR-10 data, varying α with $\gamma = 3$ fixed.

Fig. 3: Optimisation of classification accuracy from augmentation with conventional image transformations.

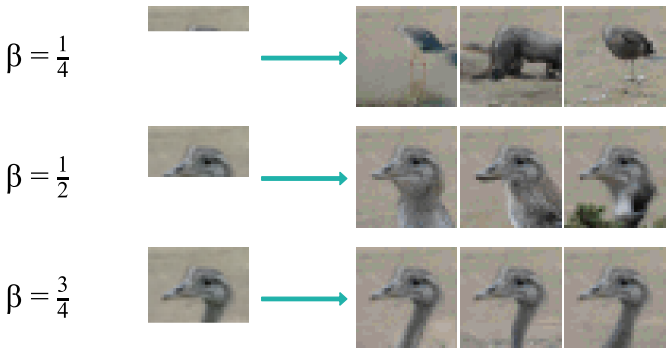


Fig. 4: Sample of different images from 3 invocations of ImageGPT generation for each prompt size β.

similarity to the original data, and so are somewhat analogous to the γ parameter of our conventional synthetic data. This process is stochastic, since ImageGPT does not extrapolate the images consistently between repetitions, thus allowing different synthetic images to be generated from each prompting original image. An illustrative example of this process is shown in Fig. 4. The left side of the figure shows the proportion of the original image used as a prompt, and the right-hand panels show the extrapolation ImageGPT produces. This procedure was repeated for $n_{orig} = 1024$ images sampled from CIFAR-10. Various synthetic datasets were created by repeating this process a different number of times per original image (for different α), and for different prompt sizes β.

D. Generative Adversarial Networks

Generative Adversarial Networks (GANs) were proposed by Goodfellow *et al.* [5] to create realistic and high-resolution synthetic data [24] (initially of images, though this has been generalised to other domains [16]). GANs operate by training two neural networks, an image generator and a discriminator, in parallel as a min-max game. The generator maps random noise to images, while the discriminator classifies images as real or synthetic. As both networks are trained in parallel, the generator learns to create increasingly realistic images, and the discriminator learns to better classify between real and synthetic images.

The discriminator classifies based on the probability each image is real or synthetic, with its loss calculated from prediction accuracy. The discriminator aims to minimise this loss, while the generator seeks a loss of 0.5, meaning the discriminator cannot tell between real and synthetic images. If one network converges too quickly, it may hinder the other: If the discriminator learns too fast, it will too easily distinguish any generated images, so will not provide a useful gradient for the generator (vanishing gradient [25]); conversely, if the generator converges too quickly, it risks mapping different inputs to the same output and so not represent the full diversity of the original data (mode collapse [26]).

GANs are inherently an unsupervised method, simply generating images from a given dataset. However, extensions such as Conditional GANs (CGANs) [27] condition outputs on additional data, such as class labels. This enables generating images of specific classes, which is useful for synthetic data in classification tasks.

GANs are notoriously difficult to train [28], due to these complexities, the large number of hyperparameters and particularly the need to balance the learning rates for the generator and discriminator. Data augmentation, often using transformations like those in Section II-B, is also common [29].

A key difference of using models such as GANs instead of a pre-trained model like ImageGPT is that they can be trained specifically for a given dataset. While this means that training the model can be an intensive task, it means that the model will be tuned to the specific problem. Furthermore, while training the GAN model itself is a slow process, it is relatively very quick to sample from its output once a trained model is produced.

We selected GANs as an example of a self-trained generative model to compare the performance of a deep model trained on our original dataset against conventional synthetic data and

Parameter	Model 1	Model 2	Model 3	Model 4
Conventional aug. α	15.6	18	10	10
Conventional aug. γ	0	1.8	1.0	0
Epochs	791	500	400	400
Dropout (Class.)	0.407	0.68	0.25	0.25
Dropout (Disc.)	0.453	0.40	0.25	0.25
$LR_{gen.}$	2.6e-5	3e-5	5e-5	5e-5
$LR_{disc.}$	2.4e-5	3e-4	5e-5	5e-5
Scores				
Relative Accuracy	1.20	1.13	1.19	1.21
FID	128	115	124	168

TABLE I: Parameters and scores for several GAN models. Model 1 was directly selected from the Bayesian optimisation output, Models 2, 3 & 4 from heuristic experimentation of partial distributions from the optimisation.

ImageGPT. We used a Deep Convolutional GAN (DCGAN), a common extension of the GAN framework [30], incorporating convolutional layers into the structure of the neural network to better capture the features of the images, based on [31]. We trained one sub-model per CIFAR-10 class, allowing our generator to produce labelled images. This might alternatively be achieved using a CGAN.

Hyperparameters were initially varied individually to identify the most sensitive ones, then optimised using Bayesian optimisation [32], focusing on classification accuracy,[4] though also considering FID, loss functions and visual inspection. While results were good with the full CIFAR-10 dataset, training with a smaller subset ($n_{orig} = 1024$) proved more challenging.

Augmenting with conventional synthetic data made GAN training much easier, but could complicate comparisons to other SDGs. To address this, in our Bayesian optimisation experiments, we sought a region of the hyperparameter space with augmentation intensity $\gamma = 0$ – the identity transformation.[5] This adjustment, equivalent to training for more epochs with a slower learning rate, ensured fair comparisons between GAN-generated and conventionally generated images, and yielded results comparable to other augmentation regimes.

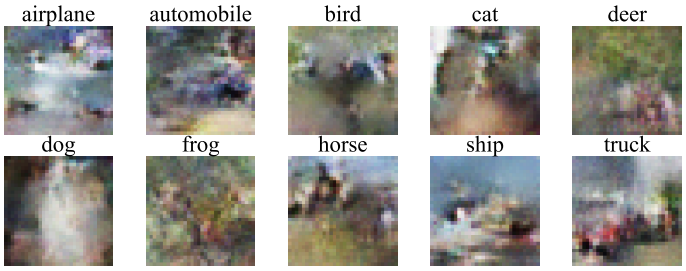
After optimisation, we further assessed models using FID and visual inspection to compare different solutions. Table I summarises key hyperparameters alongside relative accuracy and FID scores for selected models. Sample outputs from two of those models, those that achieved lowest FID and highest relative accuracy respectively, are shown in Fig. 5.

Note that, for synthetic data from our GAN models, relative accuracy and FID were not strongly correlated; higher relative accuracy did not always coincide with lower FID. As shown in Fig. 5, models with lower FID produced more vivid variation of colour contrast and shape, whereas the models with higher relative accuracy had a less varied colour palette, though sometimes clearer objects.

[4]Augmenting our $n_{orig} = 1024$ original data with data generated from the GAN, using augmentation ratio $\alpha = 4$.

[5]Optimisation often finds multiple "good" solutions in different regions of the parameter space [32]. In our case, intrinsic uncertainty in GAN training and classifier testing often exceeded the variance between these solutions.

(a) GAN Model 2



(b) GAN Model 4

Fig. 5: Some examples of images sampled from the output of select GAN models from Table I. GAN Model 2 had the lowest FID score, whereas GAN Model 4 had the highest relative accuracy score.

We choose GAN Models 2 and 4 for further investigation, since these were the two models that optimised FID and relative accuracy scores, respectively. Note that there is not an easy way to control the dissimilarity of GAN outputs, like the γ and β parameters for conventional transformations and ImageGPT. Therefore, the only way that we can adjust the dissimilarity is by using a different model, though this is not as robust as for the other methods.

III. Experiment 1 - Classification Accuracy vs FID

A. Experiment Details

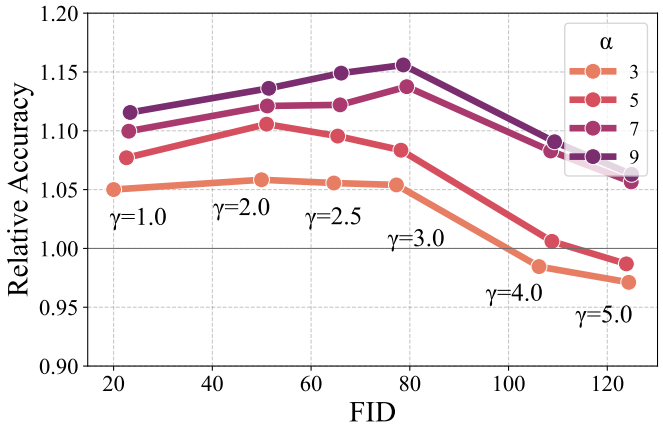
We first examine the relationship between classification accuracy (a utility measure) and FID (a fidelity measure). FID has been proposed as a potential predictor of classification accuracy [33], [34] and, as discussed in Section II-A2, it would be beneficial if it were a strong predictor, since evaluation of FID does not require labelled data.

We compare classification accuracy and FID over a range of synthetic data generated for different diversity parameters β (ImageGPT) and γ (conventional augmentation) as well as different ratios of synthetic-to-real data (α).

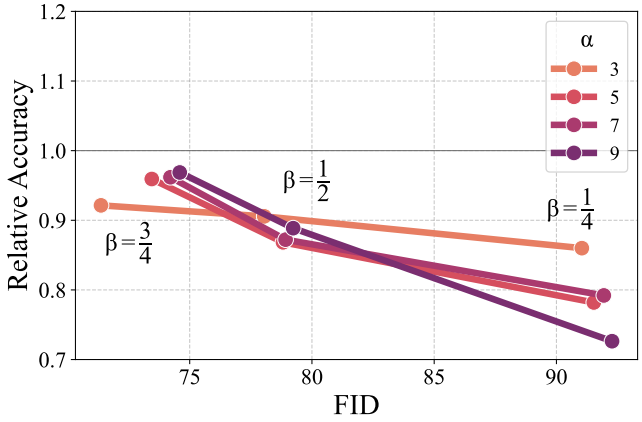
B. Results

Plots of relative accuracy against FID for conventional synthetic data and ImageGPT with varying augmentation parameters held constant are shown in Fig. 6.

1) Conventional Synthetic Data: For conventional synthetic data, Fig. 6a shows relative accuracy plotted against FID for varying γ, with the amount of synthetic data (α) held constant. As γ increases, FID increases, reflecting greater diversity



(a) Relative Accuracy against FID for conventional synthetic data. Lines show constant α for varying γ.



(b) Relative Accuracy against FID for ImageGPT. Lines show constant α for small ($\beta = \frac{1}{4}$), medium ($\beta = \frac{1}{2}$) and large ($\beta = \frac{3}{4}$) prompts.

Fig. 6: Comparison of Relative Accuracy against FID for different synthetic datasets.

in the generated data. This agrees with intuition that FID is a fidelity measure which captures dissimilarity from our synthetic augmentation.

The relationship between relative accuracy and FID forms an n-shaped curve, consistent with earlier observations in Fig. 3a. At moderate FID values, relative accuracy is optimal, suggesting unsurprisingly that a balance between augmentation intensity and fidelity results in the best utility.

FID is almost independent of α here, increasing only slightly with increasing α. This supports the hypothesis that, given enough data to extract meaningful features, FID should be constant between original and synthetic datasets. This effect may be partially due to our use of the FID_∞ estimator [18] to extrapolate FID to very large data sets.

2) ImageGPT Synthetic Data: For ImageGPT-generated data (as shown in Fig. 6b), the relationship between relative accuracy and FID differs significantly from the conventional case. As for conventional synthetic data, FID increases as β (the proportion of original data used in the prompt) decreases, which again reflects greater dissimilarity. However, the relationship between relative accuracy and FID is monotonic-negative for each fixed ratio of original to synthetic data, α; the smallest FID at largest prompt size β, yields the highest relative accuracy.

The impact of α (amount of synthetic data) on FID is more significant for ImageGPT. Particularly for $\beta = \frac{3}{4}$, increasing α significantly increases FID. This trend likely occurs because for larger prompts, each synthetic image contributes less new information, allowing relatively more information to be added from the Inception model as more synthetic data is used.

Furthermore, the relationship between relative accuracy and FID varies with prompt size, β. For $\beta = \frac{1}{4}$, relative accuracy decreases as both FID and α increase. In contrast, for $\beta = \frac{3}{4}$, relative accuracy increases with FID and α. These contrasting trends demonstrate that even for the same SDG, the relationship between utility and fidelity may vary significantly depending on the individual parameters.

Although further exploration with a wider range of β values would be valuable, the computational costs of generating large numbers of images with ImageGPT limited this study's scope.

3) GAN Models: We also compared classification accuracy and FID for synthetic data generated by our GAN models. As noted above, there is there is no single dissimilarity parameter, like γ or β, to control similarity of GAN outputs and the output from the trained GAN depends solely on the random noise vector that is input. When comparing between different model versions in our hyperparameter optimisation, we found no statistically significant correlation between classification accuracy and FID for the GAN models.

C. Summary

In conclusion, while FID serves as a reliable measure of dissimilarity in synthetic data, it is not a consistent predictor of classification accuracy. While more synthetic data generally improves accuracy (except for small-prompt ImageGPT), this is not clearly reflected in the FID score. For conventional synthetic data, a moderate FID optimises accuracy, whereas for ImageGPT, the lowest FID (achieved with the largest prompt) yields the best performance. In contrast, GAN models exhibit no clear relationship between FID and accuracy. These findings suggest that the utility of synthetic data cannot be inferred directly or predicted from FID alone. Therefore, for our comparison between different SDGs, we focus on classification accuracy.

IV. Experiment 2 - Conventional vs Generative Synthetic Data

Selecting (relative) classification accuracy as the primary measure for assessing synthetic data, we move to comparisons between our synthetic sets. Fig. 7 shows the relative classification accuracy achieved by each SDG as the ratio of synthetic to original data is increased: $1 \leq \alpha \leq 20$. Note that for $\alpha = 1$, only the real data is used, so by definition, the relative accuracy is 1. For comparison, we also include scores from a model trained on additional real data sampled

from the rest of the 50,000 CIFAR-10 images, representing the theoretical optimum. Error bars indicate the inter-quartile range in relative accuracy over 5 data generation and training of the downstream classifier runs.

A. Conventional vs ImageGPT synthetic data

First, we compare ImageGPT (yellow/orange/red dotted lines in Fig. 7) to conventional augmentation (purple solid line) with $\gamma = 3$. As shown earlier in Fig. 3, conventional synthetic data improves the classifier over the baseline of real-only data. However, this is not true for ImageGPT data. The classifier performed worse with every ImageGPT-generated datasets, especially for smaller prompts ($\beta = \frac{1}{4}$). Larger synthetic datasets generated from smaller prompts tend to further dilute the original images with low-quality images that are dissimilar to the original images. This dilution causes a further drop in classification accuracy (cf., examples of this class of image in Fig. 4)

The large ImageGPT prompt sets ($\beta = \frac{3}{4}$) did perform better than the other ImageGPT sets, achieving relative accuracy only a little lower than 1. However, this means that this synthetic data was no better than using no synthetic data, so it was not useful for classification.

In this experiment, the ImageGPT synthetic data proved less effective for augmenting our classifier, even compared to no augmentation at all, let alone compared to the conventional transformation techniques. While careful tuning might yield marginal improvements, generating images with ImageGPT is computationally expensive and our results indicate that this would be far less useful than conventional methods.

B. Conventional vs GAN synthetic data

Next, we compare the top two GAN models identified in Section II-D (the green and blue dashed lines in Fig. 7) to the conventional synthetic data. Recall that GAN Model 2 achieved optimal FID, and GAN Model 4 optimised classification accuracy.

For both Model 2 and Model 4, the relative accuracy is greater than 1 for all $\alpha > 1$, which means the synthetic data makes a significant improvement for augmenting the classifier. Model 4 was able to achieve greater accuracy than conventional augmentation on experiments with low and moderate α. This is particularly impressive, since Model 4 was trained using $\gamma = 0$ for the conventional augmentations of its training data (recall this is the identity transformation), so it did not rely on additional information gained from the conventional augmentations for its training. However, for larger $\alpha \gtrsim 12$, the conventional synthetic data outperformed the GAN generative data, since the performance of the GAN did not significantly improve after $\alpha \approx 10$, whereas continuing improvement was observed for much higher α for the conventional augmentations (cf. Fig. 3b). This could indicate that, while some additional semantic information can be provided by the GAN, there is a limit to how much new information it can generate, particularly in comparison to the conventional synthetic data.

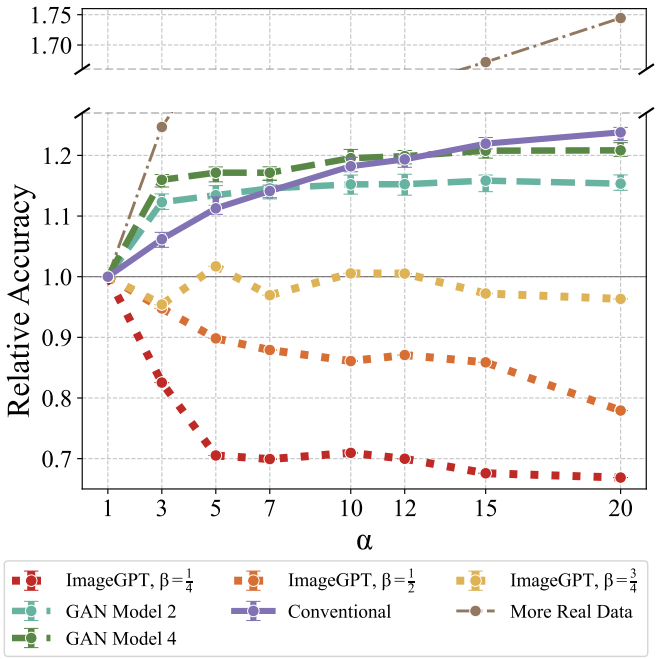


Fig. 7: Comparison of relative accuracy for all original and synthetic datasets. Error bars show median and interquartile range across 5 repeats.

The dynamics for the accuracy of Model 2 (best FID) are more surprising, since maximal accuracy was attained for a smaller amount of synthetic data with $\alpha \approx 7$, and after which accuracy was approximately constant for larger α. This suggests that there is some useful information gained from using this GAN, though not as much as can be gained from the other GAN or conventional augmentation.

V. Discussion and Conclusion

In this paper, we compared the performance of different classes of synthetic data for image data augmentation during classifier training and explored the relationship between classification accuracy and FID as utility and fidelity measures for assessing our synthetic data.

A. Evaluation Methods

Our comparisons between classification accuracy and FID for conventional and ImageGPT synthetic data showed that, while there is a link between utility and fidelity, they were not strongly correlated in our study. For ImageGPT synthetic data, lower FID generally indicated higher accuracy. However, for conventional data, an n-shaped relationship emerged, with optimal accuracy achieved at a moderate amount of transformation and an intermediate FID value. These patterns held for any constant value of α, so are functions of the data variation/dissimilarity parameters γ and β only. We can therefore see that conclusions about one synthetic data type need not carry to another.

Throughout the study, classification accuracy was highly dependent on the amount of synthetic data. For the top-performing models (conventional SDG and GAN Model 4), accuracy improved until it plateaued (see in Figs. 3b and 7). However, models such as the small-prompt ($\beta = \frac{1}{4}$) ImageGPT and GAN Model 2, showed optimal accuracy with lower α, meaning less augmentation. Despite these exceptions, the best-performing models consistently benefited from larger amounts of synthetic data. Conversely, as seen in Fig. 6, FID was largely invariant to varying α. This is a desirable behaviour for a fidelity measure, so that it can well assess the difference between two datasets without dependency on the size of the datasets, but it further limits its ability to predict a utility measure such as classification accuracy, since we have seen that utility depends on α.

FID, based on the Inception-v3 model, may be biased toward datasets like ImageNet [35]. Replacing Inception-v3 with a domain-specific classifier [36] or adopting class-conditional approaches [37] could improve robustness.

Finally, Jayasumana *et al.* [36] introduced CMMD, a GAN evaluation metric based on CLIP embeddings and maximum mean discrepancy with a Gaussian RBF kernel. Comparing CMMD to FID, alongside utility measures like classification accuracy, could be a valuable direction for future research.

B. ImageGPT vs Conventional

We compared ImageGPT and GANs as examples of newer generative models with conventional image transformations. We assessed different SDGs with the relative accuracy of a classifier trained on synthetic-augmented data against a baseline of the classifier trained on real-only data. We found that ImageGPT was able to produce visually useful synthetic data. However, all synthetic data produced by ImageGPT was detrimental to the overall classifier performance. Only the ImageGPT generative synthetic datasets generated using larger prompt size β and larger augmentation ratio α were able to achieve similar accuracy to original data only. Compared to classifier models trained with conventional augmentation, the ImageGPT synthetic data performed worse throughout. While a useful and interesting novel method, these results do not suggest that it is a better technique for augmentation that the conventional transformation-based techniques, which are also much faster to compute: informal experimentation found conventional image transformation to be at least 2000 times faster than ImageGPT data generation.

While ImageGPT's prompt-extrapolation approach adds new information, our results show it was not beneficial for image classification. Since ImageGPT was trained on the ImageNet dataset rather than CIFAR-10, and is not conditioned on class labels, it generates images without specific class information. As a result, for small and medium prompt sets ($\beta = \frac{1}{4}$ and $\frac{1}{2}$), there is likely too much 'semantic loss' (feature-level loss, rather than pixel-level) in the image generation. Since ImageNet has a much larger domain space than CIFAR-10, images may resemble ImageNet classes but not CIFAR-10, and in some cases, smaller prompts may lead to images resembling other CIFAR-10 classes. For larger prompt sets ($\beta = \frac{3}{4}$), the synthetic images were closer to real data, but the lack of diversity in generated images resulted in no improvement in classifier accuracy.

C. GAN vs Conventional

Although the GAN models generated images that were visually different having higher FID scores and thus lower fidelity, augmentation with GAN generated data improved the relative accuracy and Model 4 (maximum utility) achieved similar accuracy to conventional data augmentation. This demonstrates that very different augmentation methods can achieve similar classification accuracy, despite varying FID scores or levels of diversity. However, as discussed previously, the accuracy of the GANs increased only up to augmentation ratios up to $\alpha \approx 10$, further highlighting the advantage of conventional methods, which allow more images to be generated with greater control over augmentation intensity (γ).

Despite the complexity and higher resource demands of training GANs, these results suggest that they are still useful. Furthermore, these conventional data augmentations are only applicable in the image domain, as they use geometrical transformations of the image data. While GANs were originally designed for the purposes of image generation, there are numerous examples of generalisations to different types of data [16], [38]. Our results, while focused on the image domain, suggest that GANs could be a promising candidates for data augmentation in other domains.

This study showed that GANs are promising for synthetic data generation but are limited by computational constraints, requiring further testing and hyperparameter tuning to better understand their utility. While GAN augmentation achieved good classification accuracy, the high FID scores suggest further investigation into the relationship between these metrics. Applying this evaluation framework may also help assess other deep learning models.

Chen *et al.* [6] noted that ImageGPT's transformer-based architecture, faces challenges in generating large images. Training ImageGPT on high-dimensional pixel data is computationally intensive, so pre-trained models are only available for 32×32 images, such as CIFAR-10. This limited our study to small images which may not fully capture the potential of this type of architecture to generate highly realistic images.

REFERENCES

[1] J. Jordon, A. Wilson, and M. van der Schaar, "Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods," *CoRR*, vol. abs/2012.04580, 2020.

[2] O. Mendelevitch and M. D. Lesh, "Fidelity and Privacy of Synthetic Medical Data," *ArXiv*, 2021.

[3] R. McKenna, G. Miklau, and D. Sheldon, "Winning the NIST Contest: A scalable and general approach to differentially private synthetic data," *Journal of Privacy and Confidentiality*, vol. 11, no. 3, Dec. 2021.

[4] C. Shorten and T. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, Jul. 2019.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *Adv. Neural Inf. Process. Syst.*, vol. 3, Jun. 2014.

[6] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning*, PMLR, 2020, pp. 1691–1703.

[7] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification," *Neurocomputing*, 2018.

[8] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "GANs for Medical Image Synthesis: An Empirical Study," *Journal of Imaging*, vol. 9, no. 3, 2023.

[9] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller, *Synthetic Data – what, why and how?* 2022. arXiv: 2205.03257 [cs.LG].

[10] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. of the 31st NeurIPS*, 2017, pp. 6629–6640.

[11] A. Krizhevsky and G. Hinton, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, Tech. Rep. 0, 2009.

[12] Sid2412, *CIFAR-10 CNN Model 85.97% accuracy*, Accessed: 17th Dec, 2024, 2020. [Online]. Available: https://www.kaggle.com/code/sid2412.

[13] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved Techniques for Training GANs," in *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[14] M. Fréchet, "Sur quelques points de calcul fonctionnel," *Rendiconti del Circolo Mathematico di Palermo*, 1906.

[15] H. Alt and M. Godau, "Computing the Fréchet Distance between Two Polygonal Curves," *Int. J. Comput. Geometry Appl.*, vol. 5, pp. 75–91, Mar. 1992.

[16] A. Borji, "Pros and Cons of GAN Evaluation Measures: New Developments," *CVIU*, vol. 215, Jan. 2022.

[17] G. Parmar, R. Zhang, and J.-Y. Zhu, "On Aliased Resizing and Surprising Subtleties in GAN Evaluation," in *2022 IEEE CVPR*, Jun. 2022, pp. 11 400–11 410.

[18] M. J. Chong and D. A. Forsyth, "Effectively Unbiased FID and Inception Score and Where to Find Them," *2020 IEEE CVPR*, pp. 6069–6078, 2019.

[19] P. Simard, D. Steinkraus, and J. Platt, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," in *Seventh ICDAR Proc.*, 2003.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84, 2017.

[21] TorchVision Maintainers and Contributors, *TorchVision: PyTorch's Computer Vision library*, Nov. 2016.

[22] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of captions," 2020. arXiv: 2006.11807 [cs.CV].

[23] P. Dutta and K. Kunal, "Implementation of Image Generative models using ImageGPT," in *2023 ICCAMS*, vol. 1, 2023, pp. 1–6.

[24] A. Borji, "Pros and cons of GAN evaluation measures," *Computer Vision and Image Understanding*, 2019.

[25] Z. Ding, S. Jiang, and J. Zhao, "Take a close look at mode collapse and vanishing gradient in GAN," in *IEEE ICETCI*, 2022.

[26] Y. Kossale, M. Airaj, and A. Darouichi, "Mode Collapse in Generative Adversarial Networks: An Overview," in *8th ICOA*, 2022.

[27] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *ArXiv*, vol. abs/1411.1784, 2014.

[28] Z. Ahmad, Z. u. A. Jaffri, M. Chen, and S. Bao, "Understanding GANs: fundamentals, variants, training challenges, applications, and open problems," *Multimedia Tools and Applications*, 2024.

[29] Z. Zhao, Z. Zhang, T. Chen, S. Singh, and H. Zhang, *Image Augmentations for GAN Training*, 2020. arXiv: 2006.02595 [cs.LG].

[30] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," in *4th ICLR*, 2016.

[31] A. Lechner, *PyTorch-GAN*, Accessed: 17th Dec, 2024, 2018. [Online]. Available: https://github.com/alexlechner/PyTorch-GAN.

[32] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proc. of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016.

[33] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs Created Equal? A Large-Scale Study," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[34] S. v. Steenkiste, K. Kurach, J. Schmidhuber, and S. Gelly, "Investigating object compositionality in Generative Adversarial Networks," *Neural Netw.*, 2020.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *2016 IEEE CVPR*,

[36] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, "Rethinking FID: Towards a Better Evaluation Metric for Image Generation," in *2024 IEEE CVPR*, Jun. 2024, pp. 9307–9315.

[37] Y. Benny, T. Galanti, S. Benaim, and L. Wolf, "Evaluation Metrics for Conditional Image Generation," *IJCV*, vol. 129, no. 5, pp. 1712–1731, Mar. 2021.

[38] P. Kotha, V. Janardhan Babu, and S. Ankam, "Generative Adversarial Networks: A Comprehensive Review," in *Proc. of Fifth ICCCT*, 2024, pp. 105–114.