

Fréchet Inception Distance: Reliability and Robustness in Evaluating Generated Images

Cody Fang

Thesis presented for:
Advanced Studies in Computing (COMP30013)

School of Computing and Information Systems
The University of Melbourne

October 2023

Abstract

In this paper, we present evidence that suggests inconsistencies found in the widely used Fétchet Inception Distance (FID score) which claims to evaluate the realism of model-generated images. More specifically, there are instances where the FID score irrationally improves when there is no ostensible difference in image quality. Our experiments subject 10000 real test images to a variety of image transformations, and compare the resulting FID scores with the baseline (calculated against 10000 training samples). These testing and training images are sampled from MNIST, CIFAR-100, and CelebFaces data sets, and we include transformations such as gamma adjustments, Gaussian blurring, and more. We then perform the same experiments across 10000 test images generated by diffusion models (DDPM and LDM) trained on CIFAR and CelebFaces data sets. Transformations that enhance the FID score are further applied in combination to test for compounding effects. Overall, we find significant improvements in the FID score in response to gamma, sharpness, and saturation adjustments. Our results furthermore show that the most FID-efficient adjustments are dependent on the diffusion model and its training data set. Ultimately, a combination of these transformations is able to reduce the FID score by up to 20%, although the effects are subadditive. Our results and code can be found in Github.

Declaration

I certify that:

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.
- the thesis is 2842 words in length (excluding text in images, table, bibliographies and appendices).

Signed: Cody Fang

Date: October 28th, 2023

Acknowledgement

I would like to extend my gratitude to Dr. Kris Ehinger, the supervisor of this work, for her guidance and insightful feedback in the research process.

Contents

1	Introduction	4
1.1	Background	4
2	Related Work	5
3	Experiments	6
3.1	Methodology	6
3.2	Image Transformations	7
4	Results	8
4.1	Experiment 1: Real Test Images	9
4.2	Experiment 2: Diffusion Generated Images	10
4.3	Compounding Effects	12
5	Conclusion and Future Works	12
A	Experimental Results	15
A.1	All Transformations - Real Test Images	15
A.2	All Transformations - Generated Test Images	18

List of Figures

1	FID score comparison	5
2	Model generated outputs	6
3	Sample display of transformations	8
4	FID plots of the CIFAR test images under a various transformations	9
5	FID curves of all generated image sets under gamma adjustment.	10
6	FID curves of all generated image sets under saturation adjustment	11
7	FID curves of all generated image sets under sharpness adjustment	11
8	Sample displays of mixed transformation	13
A1	FID curves of transformations on real test images	15
A2	FID plots of transformations on real test images	16
A3	FID curves of transformations on generated test images	18
A4	FID plots of transformations on generated test images	19

List of Tables

1	All transformations tested, their parameters and respective increment step sizes.	7
2	Image sections transformed ($x_1 : x_2, y_1 : y_2$)	8
3	FID optimal transformation parameters.	10
4	FID scores of mixed transformation	12
A1	Miscellaneous transformations with corresponding FID (real images).	17
A2	Miscellaneous transformations with corresponding FID (generated images).	20

1 Introduction

The Fréchet Inception Distance score (FID) is an evaluation metric that quantifies the realism (or similarity) of model generated images in relation to the model’s training data set. Being a significant improvement to its predecessors, the FID score’s convincing results in distinguishing between images of “good quality” and “bad quality” have brought about its widespread usage in tuning and comparing complex generative models such as Generative Adversarial Networks and diffusion models.

However, the FID score is not without flaws. It has been the subject of scrutiny and investigations in literature due to its lack of robustness. These phenomena can be categorised into two cases, the first being where seemingly innocuous changes to test images have a significant influence over the FID, and the second is where two image sets with observable quality differences yield similar FID scores. Nevertheless, it has remained the standard and authority in model evaluation.

These weaknesses could have practical implications since FID has the potential to mislead. For example in model selection, the FID could unreasonably penalise realistic models simply because the model generates specific features that the metric is particularly sensitive toward, and instead favour models that learn patterns that do not necessarily correlate with perceived realism.

In this work, we investigate the reliability and robustness of the FID in evaluating transformed images. We improve upon the previous experiments by expanding the varieties of models and images tested, as well as by attempting to measure the severity of the FID’s weaknesses. We propose the following **research question**: Are there ways to artificially boost the FID score that does not improve the realism of generated images? If so, how significant is the effect?

1.1 Background

As proposed by Martin Heusel et al. in [1], the FID is a comparison between generated samples and real world samples. Specifically, these images are converted to latent feature representations within the second last layer of the Inception-v3 Neural Network (trained on the ImageNet data set) in order to obtain vision relevant features. These are aggregated as multi-variable (2048 corresponding to each latent feature) Gaussian distributions with estimated parameters. The FID score is formulated as the Fréchet distance between these distributions. Here a lower FID score indicates more realistic images because there are fewer differences between the real and generated samples. For two estimated distributions $X_r \sim N(\boldsymbol{\mu}_r, \Sigma_r)$ and $X_g \sim N(\boldsymbol{\mu}_g, \Sigma_g)$:

$$\text{FID}[(\boldsymbol{\mu}_g, \Sigma_g), (\boldsymbol{\mu}_r, \Sigma_r)] = \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_r\|^2 + \text{Tr}(\Sigma_g + \Sigma_r - 2(\Sigma_g \Sigma_r)^{\frac{1}{2}}) \quad (1)$$



Figure 1: Different image qualities and their corresponding FID scores. Images on the left look more distorted than ones on the right. **Left:** FID 45 and **Right:** FID 13 [1]

2 Related Work

Fréchet Inception Distance. In the original proposal paper [1], FID’s robustness in evaluating certain modified images was also demonstrated. These disturbances included blurring, blackening, and image contamination, performed on images in the CelebFaces Attributes data set. However, since its conception, FID has been criticised for its weaknesses in producing evaluations of model generated images consistent with human judgement.

The procedure to obtain the FID score has been dissected and examined by related works. [2] finds perception bias in the ImageNet Inception v3 Neural Network latent space used for feature encoding in the FID, resulting in generated images yielding similar FID scores despite visible quality differences (example of FastGAN in comparison to StyleGAN2). Another discovery is that FID does not account for the bias instilled by finite sample sizes, despite the original author’s recommendation of using a minimum of 10000 or more sample images in FID calculation[3][1].

Other prominent weaknesses discovered of the FID include the unreasonably harsh evaluations of compressed and resized images, despite these transformations resulting in no visible differences in comparison to the baseline[4]. Other experiments display that the realism ranking of adjusted images using the FID seems to contradict human perception[5]. Similarly, we aim to test transformations that are not yet investigated, on a variety of data sets including some currently not covered by literature in order to better understand the depth and breadth of these weaknesses.

Diffusion Generative Models. There are many image generation models that the FID evaluates, and the current state-of-the-art methods involve diffusion models. Although the idea of applying techniques in thermodynamics to deep learning models was proposed in 2015, the Probabilistic Diffusion Model (DDPM) yielded promising results only recently due to the adoption of Markov chain as well as variance inference approximations in model training[6].

However, despite its effectiveness, each image requires one thousand generation steps. In light of this, new algorithms and methods are introduced with the goal of improving efficiency. Firstly, [7] introduced the technique of predicting noise entirely in latent feature space instead of the high-dimensional pixel space (LDM - Latent Diffusion Model). Another method eliminates the necessity of Markov processes in the generation process, allowing the algorithm to “skip” steps [8]. We use outputs from some of these models in our investigation (Figure 2).



Figure 2: Sample model generated outputs. **Left:** DDPM (CIFAR), **Middle:** DDPM (Celeb-A), **Right:** LDM(Celeb-A)

3 Experiments

3.1 Methodology

The objective of our experiments is to perform a variety of transformations to images, both real and generated (from diffusion models). In general, we calculate the FID scores of both the original and transformed images based on a common set of real images (will be referred to as the “training set”) so that the scores are comparable. The difference observed will indicate FID’s sensitivity towards certain transformations.

In experiment 1, we first randomly select 10000 samples from the training partition of each data set as the training set. Then, we select another 10000 samples from the test partition to eliminate the overlapping of images. The training samples will be considered as “real” images and the test samples (which we transform) as “generated” images in the context of FID calculation. The following data sets below are included in experimentation to ensure sufficient coverage in different types of images:

- MNIST (handwritten numerical digits)
- CIFAR-100 (collection of animals and objects)
- CelebFaces Attributes (images of human faces)

This procedure is performed again in experiment 2, but with 10000 model generated images used for testing. A variety of models were tested to investigate the model specific behaviours of FID. We included generated outputs from the following models:

- DDPM (trained on both CIFAR-10 (32×32) and CelebFaces (256×256))
- Latent Diffusion Model (trained on CelebFaces (256×256))

3.2 Image Transformations

In this experiment, we focus on edge-enhancing and colour-shifting transformations such as gamma, saturation, and sharpness adjustments due to the limited research and results in these areas available in the current literature. Additionally, some other transformations such as "salt and pepper" noise and Gaussian blurring, have been included for which there are previous results [1]. Each transformation is tested with a range of parameters if applicable. All transformations selected have a variety of implementations available. We use the Torchvision Library in Python to conduct these experiments.

Transformation	Parameters	Increment
Gamma	0.2 to 3.0	0.2
Saturation	0.2 to 3.0	0.2
Sharpness	0.2 to 3.0	0.2
Hue	-0.5 to 0.5	0.125
Invert Image	—	—
Rotations	90° to 270°	90 degrees
Sectional Rotations	180°	—
Crop and Swap	—	—
Gaussian Blur	0.5 to 1.4	0.3 (2 dimensions)
Salt and Pepper Noise	0% to 10%	1%

Table 1: All transformations tested, their parameters and respective increment step sizes.

We provide descriptions of some notable parameters in Table 1. The gamma adjustment is a non-linear shift of image intensities of every RGB entry. Transformed intensity I' is given in terms of the original intensity I with parameter γ is defined as $I' = 255 \times \left(\frac{I}{255}\right)^\gamma$ [9]. The saturation parameter determines the weight given to the given image when overlaid with the gray-scale image. The magnitude of this parameter value corresponds to the degree of colour intensity of the picture. The sharpness parameter behaves almost identically, only the overlay image is instead produced via convolution of a 3 by 3 Laplacian sharpening kernel.

The parameter for Gaussian blur is the standard deviations σ used to determine the amount of blur [9]. Here, the kernel standard deviations can vary in each direction (horizontal and vertical) independently to produce different blurring effects (we include combinations of the range in our experiment). Finally, Table 2 includes the selected image sections changed by spatial transformations. These include rotating the test images, rotating a cropped section of test images upside-down, as well as cropping two identically sized sections of the images and swapping them with each other.

Data Set	Rotated Section	Swapped Sections
MNIST	(9:19, 9:19)	(5:10, 10:20) and (20:25, 15:25)
CIFAR (All)	(10:22, 10:22)	(5:10, 10:20) and (20:25, 15:25)
Celeb-A (All)	(40:140, 90:160)	(40:65, 100:130) and (125:150, 100:130)

Table 2: Image sections transformed ($x_1 : x_2$, $y_1 : y_2$)

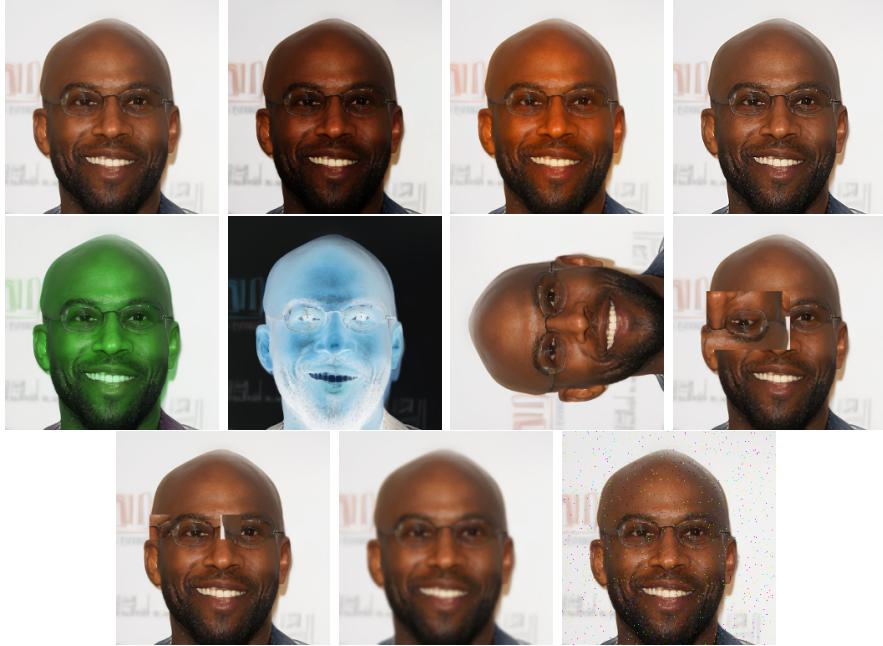


Figure 3: Sample display of transformations, with original (top left) and in 1 presentation order.

Ultimately, we identify transformations that result in more favourable FID scores and apply these in combination to examine compounding effects based on each data set/model. We utilise a heuristic approach in ultimately determining the final boosted FID score to report by (1) applying a 1-dimensional grid search to find optimal parameter combinations, and (2) assuming that transformations are commutative, and yield identical FID regardless of order. These allow for significantly reduced computation intensity while maintaining approximately accurate results.

4 Results

In this experiment, we show that while the FID behaves largely according to expectation in response to transformations, a number of transformations significantly improve the FID of model-generated images.

4.1 Experiment 1: Real Test Images

Across test images from MNIST, CIFAR-100, and Celeb-A data sets, all transformations resulted in a higher FID score compared to the baseline for each data set; these images were rated as less realistic by the FID due to these transformations. All but a few transformations yielded FID higher than 10. Furthermore, higher levels of distortion generally result in higher FID. This is expected, especially for transformations such as Gaussian blurring and “salt and pepper” noise which have been investigated previously. We also verified that the FID correctly responds to spatial disturbances. Figure 4 demonstrates a pattern that can be observed throughout the FID curves from this section, where the minimum FID coincides occurs at the original image.

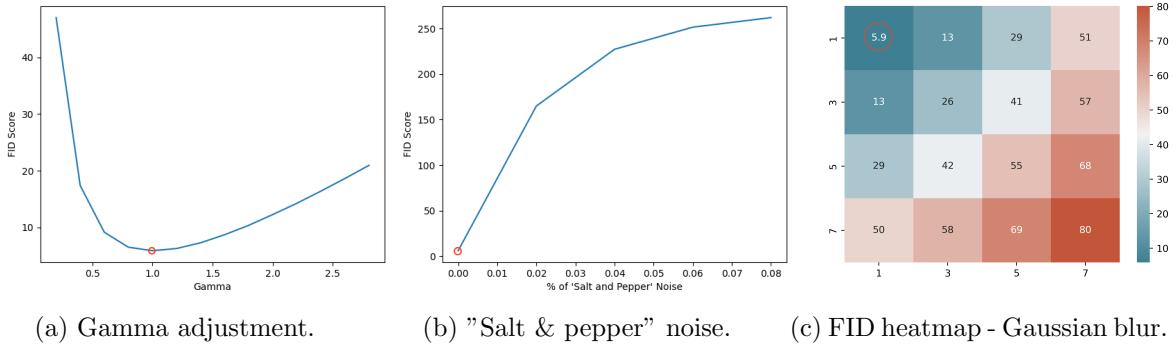


Figure 4: FID plots of the CIFAR test images under various transformations. The X and Y axes of heat map represent blurring levels in their respective dimensions. We calculate Gaussian standard deviation with $\sigma = 0.15 \times \text{level} + 0.35$. Circle indicates original FID.

MNIST (1.522 baseline FID). MNIST images demonstrated high degrees of fluctuations in FID in response to the tested transformations. Inverting MNIST images yielded the highest score of 224.668. Although inverting MNIST images preserves the shapes of handwritten digits, we expect the FID, which measures similarity, to detect that black digits on a white background is a different colouring compared to the MNIST training set. On the other hand, the FID changes comparatively less under gamma and sharpness adjustments. While rotating these images by 180 degrees yields an impressively low score of 6.616, this can be attributed to the rotational symmetries of many digits (“0”, “1”, “6” and “9”). Therefore, this particular result falls within reasonable expectations as well.

CIFAR-100 (5.892 baseline FID), Celeb-A (3.542 baseline FID). Across the colour image data sets, image inversion once again produced the largest FID scores (62.262 and 118.498) respectively. In addition, the metric is sensitive towards the upside-down rotation of Celeb-A images (97.688) because all training images do not have rotations in contrast to CIFAR where some objects appear in different orientations. Comparatively, both sharpness and saturation have the least impact on FID. We also observe that in transformations such as blurring and to an extent sharpness and saturation also, the FID scores evaluated for Celeb-A test images seem to be affected less. However, all Celeb-A images have higher dimensions

than the 32 by 32 standard of CIFAR. These particular transformations could be less effective in distorting the perception of the Inception Neural Network due to increased the volume of pixels.

4.2 Experiment 2: Diffusion Generated Images

Overall, most transformations exhibit similar patterns on diffusion model generated images as when they are applied to real test images. For example, image inversions again led to the highest FID, and all transformations except for gamma, sharpness, and saturation adjustments resulted in increased FID scores (deemed less realistic). In applying gamma, sharpness, and saturation adjustments to the generated test images, we observe FID improvements for a small range of distortion factors and select the parameters that yield minimum FID for further investigation (Table 3).

Model	Gamma	Saturation	Sharpness
DDPM (CIFAR)	0.8	0.8	2.4
DDPM (Celeb-A)	1.4	0.4	1.2
LDM (Celeb-A)	1.2	0.6	0.4

Table 3: FID optimal transformation parameters.

For the models trained on Celeb-A, the FID is a minimum at γ values around 1.2 to 1.4; the images are slightly darkened. In contrast, FID favours generated images of objects that are slightly brighter. The improvements are relatively small but still significant, with differences of approximately 1 point (Figure 5). Potential reasons for FID’s different behaviours on CIFAR

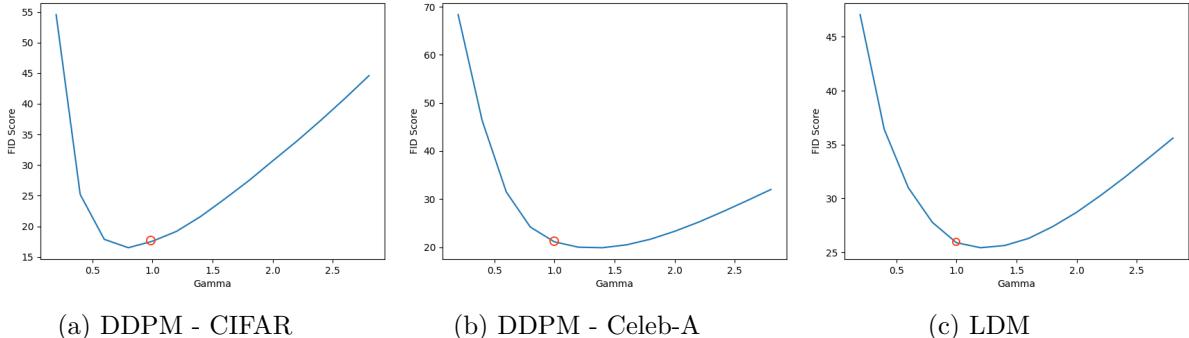


Figure 5: FID curves of all generated image sets under gamma adjustment.

trained and Celeb-A trained models could be related to the Inception neural network’s unique feature encoding of human faces (the network output does not contain a “face” category). While the diffusion models could be simply generating images of the human face that are too bright, an examination of the outputs in comparison to the Celeb-A data set suggests that there are no noticeable differences in perceived brightness overall (Figure 2). Therefore, one would not expect improvements from gamma adjustments alone.

Similarly, saturation adjustments also improve the FID for distortion factors anywhere between 0.4 to 0.8, with approximately the same degree of improvement as gamma adjustments (Figure 6). Arguably, the images are of lesser quality after the transformation because of the faded

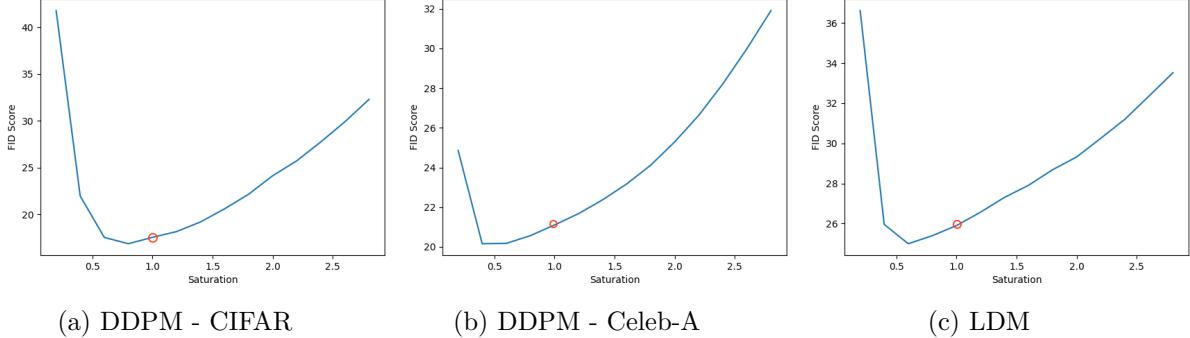


Figure 6: FID curves of all generated image sets under saturation adjustment.

colours. However, the faded colouring seems to be the only type of transformation that improves FID consistently across all test images from all models and data sets, suggesting that this is a general weakness in the FID. We conjecture that saturation adjustments uniformly shift the mean of the Gaussian distributions of the feature space, reducing the Fréchet distance.

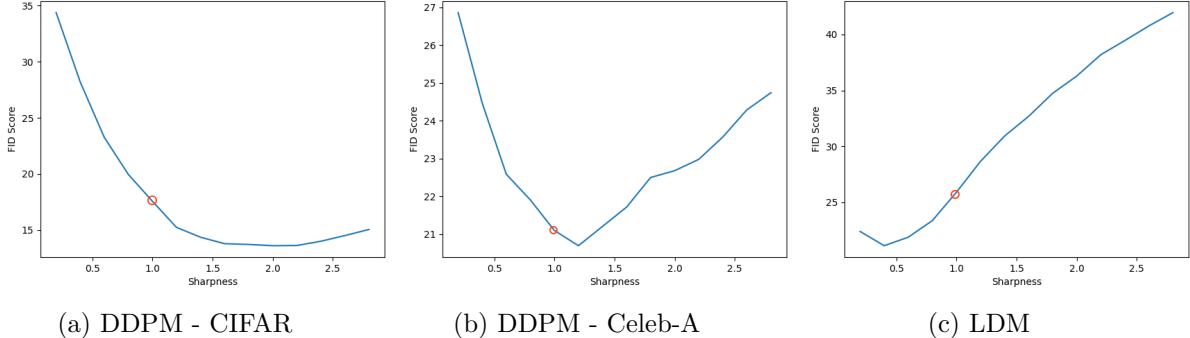


Figure 7: FID curves of all generated image sets under sharpness adjustment.

Finally, the greatest FID improvement comes from image sharpening, with the largest FID reduction of 4 (Figure 7). Both DDPM generated test image sets benefit from greater sharpness, whereas LDM images reduce substantial FID improvements when the edges are blurred. Notably, the generated CIFAR images appear to achieve the minimum FID consistently as the sharpness mixing parameter increases. This could be because the CIFAR images are low resolution, and enhanced edges assist Inception in perceiving spatial features. In contrast, DDPM (Celeb-A) images converged to a minimum FID around the parameter value of 1.2, whereas LDM (Celeb-A) has a minimum of 0.4. We speculate that LDM produces pictures with edge-dense malformed regions, and by blurring edges, these can be effectively "disguised"

from the Inception model. Yet, humans are still able to recognise these features so it provides little benefit in terms of overall realism (Figure 8, 2).

4.3 Compounding Effects

Transformation	DDPM (CIFAR)	DDPM (Celeb-A)	LDM (Celeb-A)
Baseline	17.562	21.0984	25.8999
Gamma	16.50699	19.8723	25.4313
Saturation	16.8806	20.1566	24.9979
Sharpness	13.991	20.6958	21.1268
Gamma + Saturation	16.1701	18.9286	24.63
Saturation + Sharpness	14.7277	19.6132	20.72
Gamma + Sharpness	13.9172	19.2284	20.6503
All	15.5441	18.1613	20.6537

Table 4: FID scores of mixed transformation. Lower FID indicates more realistic images

Overall, across all models, we see significant reductions in FID when gamma, sharpness, and saturation transformations are applied in combination compared to the baseline (parameters of each transform are based on previously found minimums). Some but not all combinations resulted in better FID scores compared to individual transformations (Table 4). For example, DDPM (Celeb-A) test images with gamma and saturation adjustments produced superior FID in comparison to either of these transformations alone. We report the maximum observed FID improvement of 5.25 (20% of the baseline), which comes from tests involving the LDM test images.

Nevertheless, the effects of each transformation are clearly subadditive and not independent from each other. Applying them together does not always result in the optimal FID (Table 4). For instance, in testing DDPM (CIFAR) test images, we find that adding saturation adjustments on top of gamma and sharpness adjustments increases the FID by 1.6 instead of providing improvements. Similarly, LDM test images demonstrate no clear FID difference between the applications of only gamma and sharpness adjustments in comparison to when all three are applied. While the first example is reasonable, the image quality clearly changes in the latter example, and one would expect the FID to change significantly as well (Figure 8. The inconsistency of the FID in this regard can be a cause for concern in evaluating model outputs.

5 Conclusion and Future Works

In our experiments, we found that although the FID responds as expected towards most transformations, gamma, saturation, and sharpness adjustments can artificially improve the FID evaluation of generated test images across all models and data sets (each with specific parameter values). We have identified the largest FID decrease of 5.25, 20 percent of the baseline FID, which illustrates the severity of the effect. Because these transformations do not

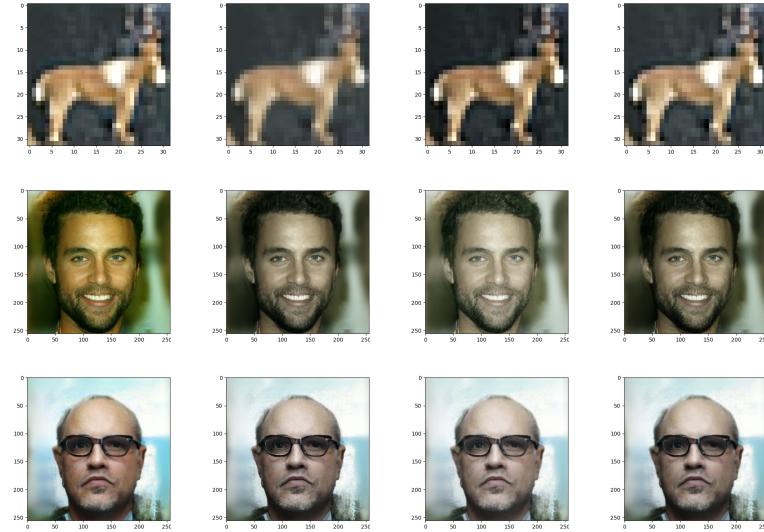


Figure 8: Sample displays of mixed transformation. **Row 1:** DDPM (CIFAR), **Row 2:** DDPM(Celeb-A), **Row 3:** LDM(Celeb-A). **Columns from left to right:** gamma + sharpness, gamma + saturation, saturation + sharpness, all three.

improve the perceived realism of the images to the same extent suggested by the FID reduction, this result suggests that the FID alone is misleading in evaluating generative models. Models that simply replicate patterns to appease the FID, by producing sharper and brighter images for instance, risk replacing more suitable models that produce more realistic images. Hence, corroborating FID evaluation results with human perception or other well-known metrics such as the Inception score can mitigate these issues of the FID to a certain extent.

We propose further investigation on the different perception behaviours of the Inception-v3 Neural Network in response to transformed images. This would provide an in-depth understanding of the problems within the FID architecture, and therefore methods of addressing the outlined issues associated with the FID. In addition, we recognise that every person will have a unique understanding of image realism. Thus, an examination of how closely all available evaluation metrics coincide with human judgement could be beneficial in determining more suitable evaluation metrics.

References

- [1] Martin Heusel et al. “GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium”. In: *CoRR* abs/1706.08500 (2017). arXiv: 1706 . 08500. URL: <http://arxiv.org/abs/1706.08500>.
- [2] Tuomas Kynkänniemi et al. *The Role of ImageNet Classes in Fréchet Inception Distance*. 2023. arXiv: 2203.06026 [cs.CV].
- [3] Min Jin Chong and David A. Forsyth. “Effectively Unbiased FID and Inception Score and where to find them”. In: *CoRR* abs/1911.07023 (2019). arXiv: 1911 . 07023. URL: <http://arxiv.org/abs/1911.07023>.
- [4] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. “On Buggy Resizing Libraries and Surprising Subtleties in FID Calculation”. In: *CoRR* abs/2104.11222 (2021). arXiv: 2104 . 11222. URL: <https://arxiv.org/abs/2104.11222>.
- [5] Shaohui Liu et al. “An Improved Evaluation Framework for Generative Adversarial Networks”. In: *CoRR* abs/1803.07474 (2018). arXiv: 1803.07474. URL: <http://arxiv.org/abs/1803.07474>.
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models”. In: *CoRR* abs/2006.11239 (2020). arXiv: 2006 . 11239. URL: <https://arxiv.org/abs/2006.11239>.
- [7] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *CoRR* abs/2112.10752 (2021). arXiv: 2112 . 10752. URL: <https://arxiv.org/abs/2112.10752>.
- [8] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising Diffusion Implicit Models”. In: *CoRR* abs/2010.02502 (2020). arXiv: 2010 . 02502. URL: <https://arxiv.org/abs/2010.02502>.
- [9] TorchVision maintainers and contributors. *TorchVision: PyTorch’s Computer Vision library*. <https://github.com/pytorch/vision>. 2016.

A Experimental Results

A.1 All Transformations - Real Test Images

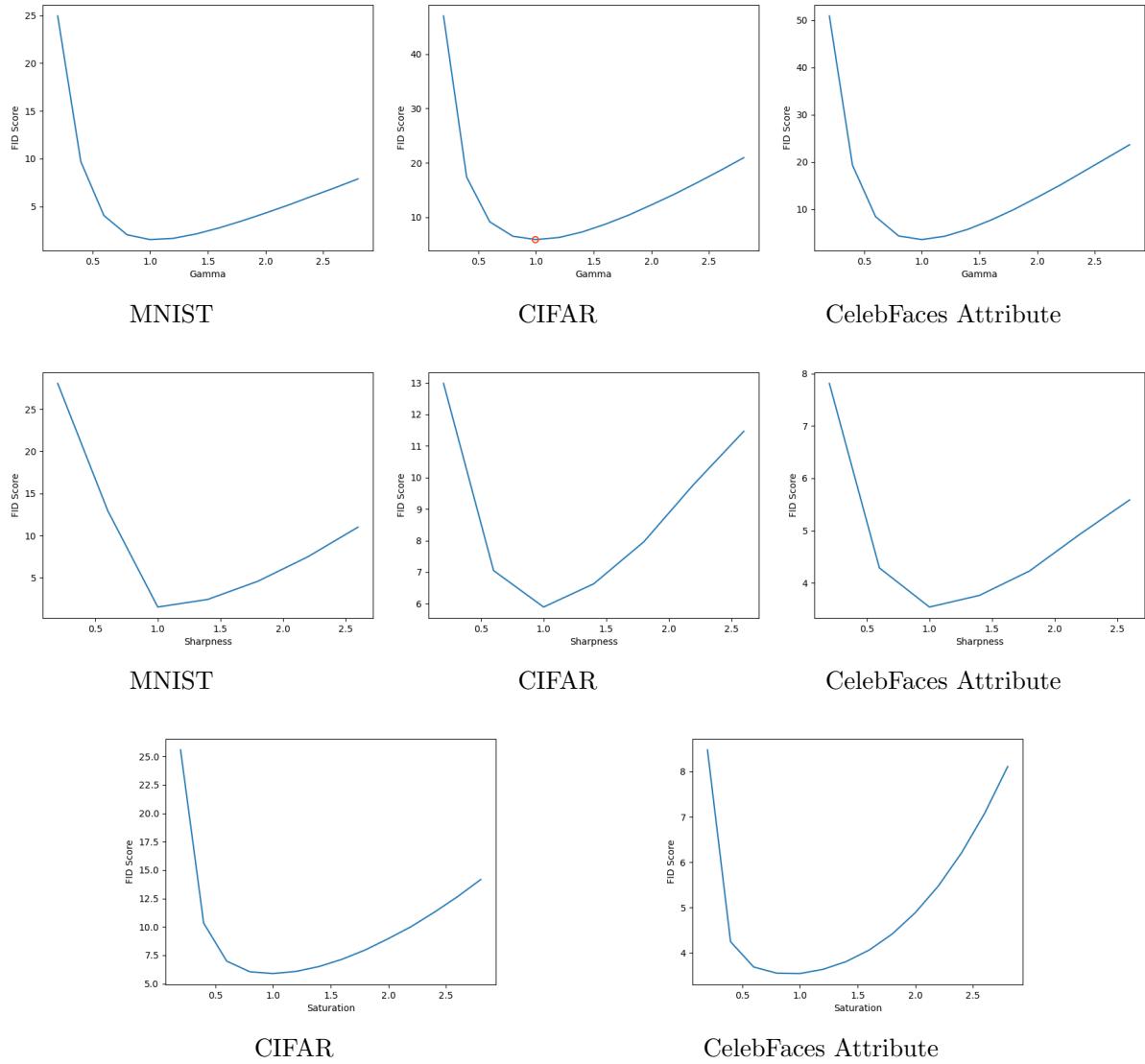
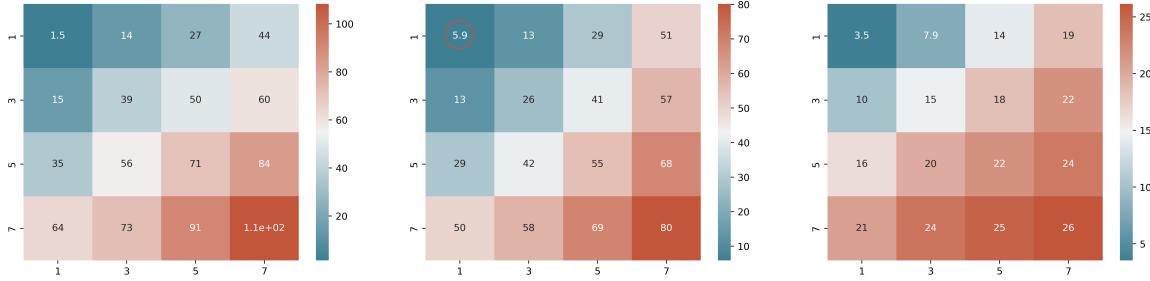


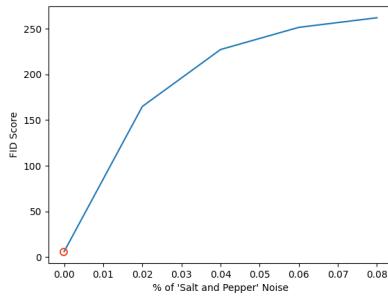
Figure A1: FID curves of transformations on real test images. (**Row 1** - gamma adjustment, **Row 2** - sharpness adjustment, **Row 3** - saturation adjustment)



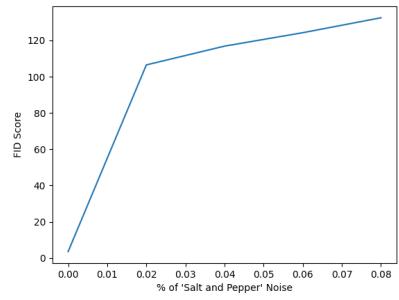
MNIST

CIFAR

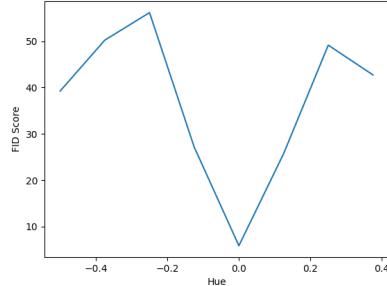
CelebFaces Attribute



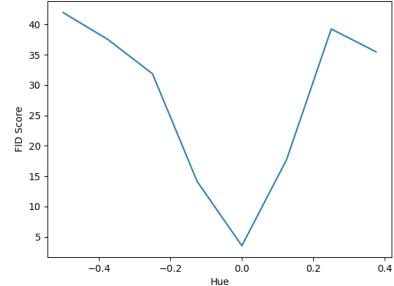
CIFAR



CelebFaces Attribute



CIFAR



CelebFaces Attribute

Figure A2: FID plots of transformations on real test images. (**Row 1** - Gaussian blur, **Row 2** - "salt and pepper" noise, **Row 3** - hue adjustment)

Transformation	MNIST	CIFAR-100	CelebFaces
Colour Inversion	224.6677	62.2618	118.4976
Sectional Rotation	41.2843	16.0380	25.1445
Crop and Swap	32.6173	16.3801	24.816
Rotation 90°	24.5572	26.3609	—
Rotation 180°	6.6159	19.507	97.6885
Rotation 270°	24.6562	26.2615	—

Table A1: Miscellaneous transformations with corresponding FID (real images).

A.2 All Transformations - Generated Test Images

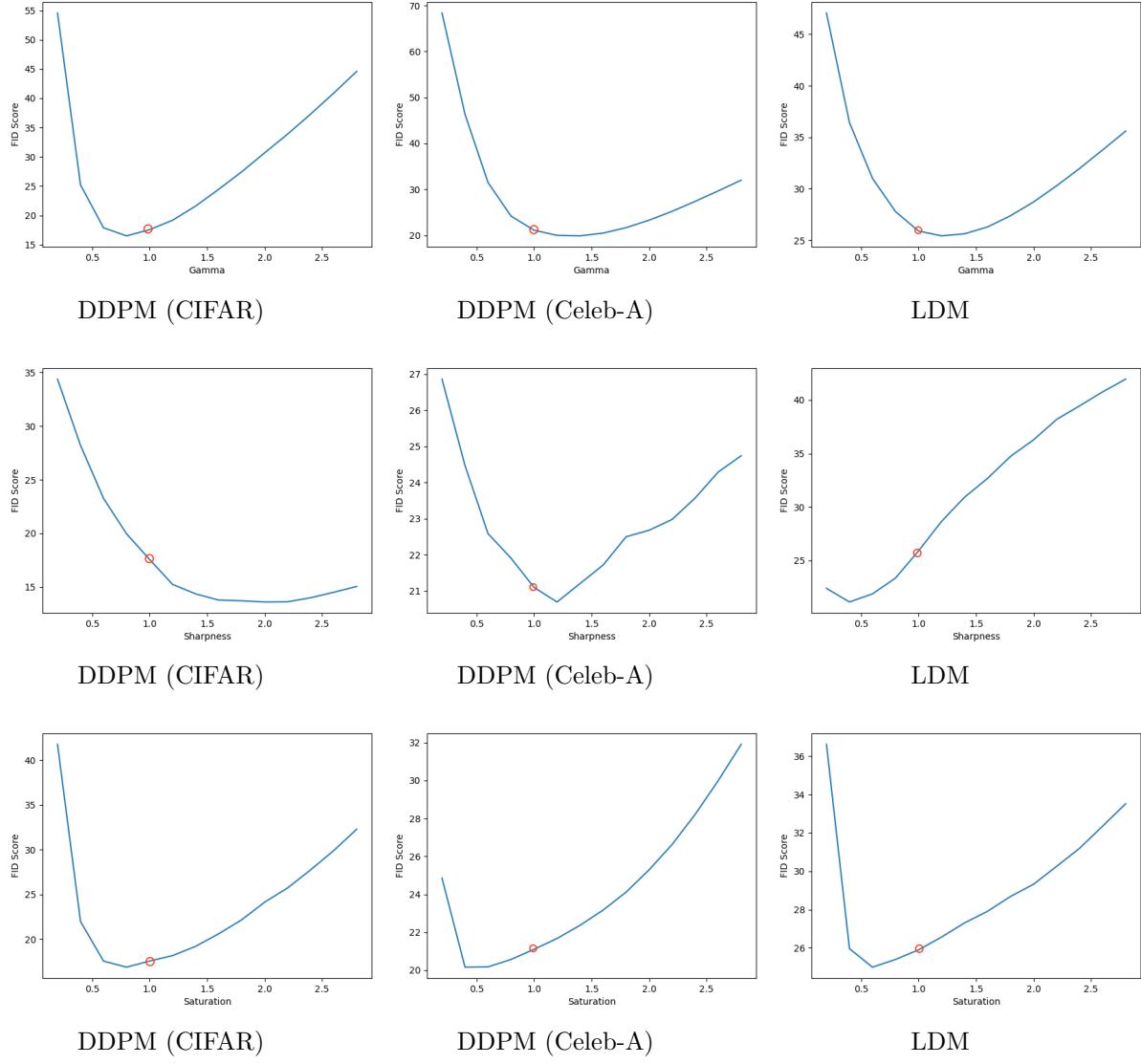
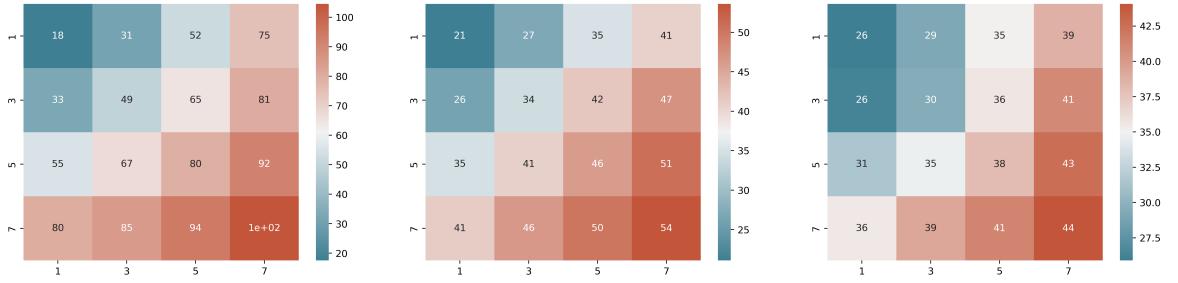


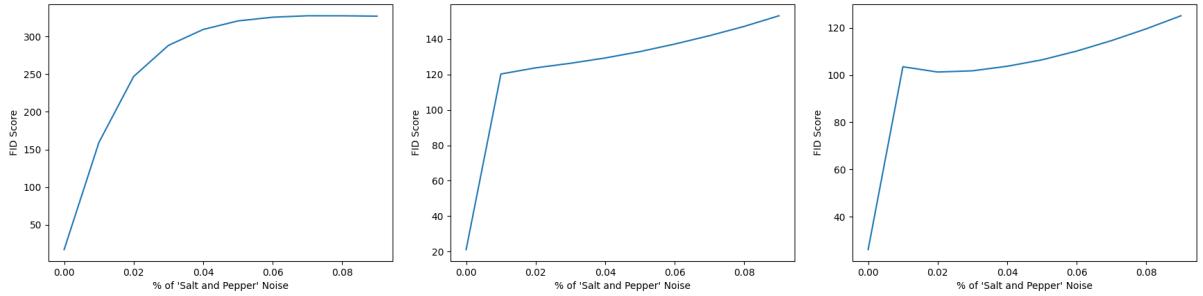
Figure A3: FID curves of transformations on generated test images. (**Row 1** - gamma adjustment, **Row 2** - sharpness adjustment, **Row 3** - saturation adjustment)



DDPM (CIFAR)

DDPM (Celeb-A)

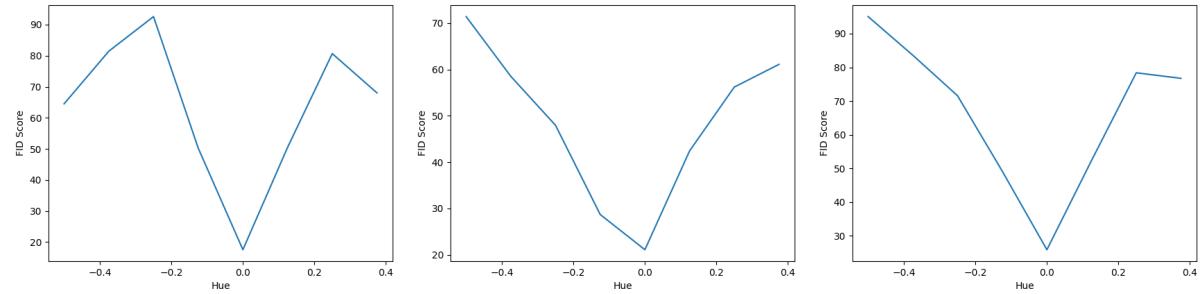
LDM



DDPM (CIFAR)

DDPM (Celeb-A)

LDM



DDPM (CIFAR)

DDPM (Celeb-A)

LDM

Figure A4: FID plots of transformations on generated test images. (**Row 1** - Gaussian blur, **Row 2** - "salt and pepper" noise, **Row 3** - hue adjustment)

Transformation	DDPM (CIFAR)	DDPM (Celeb-A)	LDM (Celeb-A)
Colour Inversion	90.7103	167.7517	167.2434
Sectional Rotation	36.169	88.9273	112.1213
Crop and Swap	32.3748	81.6819	111.95
Rotation 90°	66.4796	127.4804	121.3738
Rotation 180°	47.1745	98.2941	107.9098
Rotation 270°	66.4604	126.9873	120.6286

Table A2: Miscellaneous transformations with corresponding FID (generated images).