Assessing Synthetic Data Quality and Model Generalization for Planetary Imagery

Clara Salditt[1,2,*], Karan Molaverdikhani[1,2,3] and Barbara Ercolano[1,2,3]

[1]Ludwig-Maximilians-Universität München (Munich University), Geschwister-Scholl-Platz 1, D-80539 München
[2]Universitäts-Sternwarte, Scheinerstr 1, D-81679 München, Germany
[3]Exzellenzcluster 'Origins', Boltzmannstr 2, D-85748 Garching, Germany

Abstract

AI is set to play a crucial role in the future of space missions, enabling autonomous rover navigation, landing procedures, and terrain analysis. For these systems to perform reliably, they must be trained on large volumes of high-quality, task-specific data. However, in space science, data is often limited due to the high costs and power demands of transmission, and more critically it is not fully controllable. Synthetic data offers a promising solution by being both controllable and significantly more cost- and time-efficient. Yet, for synthetic data to genuinely enhance model performance, its quality must be rigorously evaluated. This work addresses that challenge by assessing the quality of synthetic data generated with StyleGAN2-ADA, trained on HiRISE imagery. An evaluation pipeline was developed to analyze the data using a range of established metrics. At the same time, it examines the reliability and relevance of these metrics themselves. The findings reveal a perceptual mismatch between model-based feature extractors and human judgment, raising concerns about the trustworthiness of current evaluation practices.

Keywords

Data Quality, Gen AI, Model Generalization in Planetary Imagery, Synthetic Images Quality

1. Introduction

AI applications in space science, such as surface classification models, autonomous navigation and landing of rovers, and terrain analysis have become increasingly relevant in recent years [1, 2, 3]. These systems not only enable real-time decision-making and increase the likelihood of mission success, but also broaden the scope of space missions by reducing the need for human intervention in every decision. However, as with any AI application, training such models requires large amounts of high-quality data to ensure reliable performance. In space science and planetary exploration, data acquisition is inherently limited, due to constraints such as costs, data security and the energy-intensive nature of data transmission. Moreover, there is minimal control over image content in terms of angles, weather conditions, or lighting.

Synthetic data may help mitigate these challenges. It offers a cost- and time-efficient method to generate data that can be tailored to address gaps and biases in real-world datasets, making it a valuable complement to fully real datasets. Still, the effectiveness of synthetic data hinges on its quality. Therefore, evaluating the quality of the data is fundamental to building reliable, high-performing AI systems and helps trace the root causes of arising problems or failures. With the rise of generative AI many evaluation metrics were introduced, these metrics typically rely on features extracted by convolutional neural networks (CNNs) or vision transformers (ViTs), assuming that such models generalize well to novel domains. Yet recent studies have highlighted biases and a lack of robustness in these feature extractors when applied to unfamiliar domains [4, 5, 6].

Workshop on AI-driven Data Engineering and Reusability for Earth and Space Sciences (DARES'25), co-located with the 28th European Conference on Artificial Intelligence (ECAI 2025), Bologna, Italy, October 25, 2025

*Corresponding author.
clara.salditt@web.de (C. Salditt); karan.Molaverdikhani@colorado.edu (K. Molaverdikhani); ercolano@usm.lmu.de (B. Ercolano)
https://github.com/ClaraSalditt/Metrics_on_Mars.git (C. Salditt)
0009-0006-9730-4773 (C. Salditt); 0000-0002-0502-0428 (K. Molaverdikhani); 0000-0001-7868-2740 (B. Ercolano)

Planetary imagery is particularly relevant in this context, as widely used feature extractors like Inception, CLIP, and DINO have been primarily trained on Earth-based, object-centric images, making planetary data distinctly out-of-distribution [7]. This work contributes to the understanding of this issue by evaluating the robustness of such models and the metrics used to assess synthetic data. This work proposes an evaluation pipeline that examines a range of metrics concurrently and present benchmark results for StyleGAN2-ADA trained on HiRISE imagery. The findings suggest a fundamental perception gap between the representations of backbone models and human judgment.

2. Dataset and Preprocessing

Data from the HiRISE (High Resolution Imaging Science Experiment) catalog was used for this project [8]. HiRISE is onboard the Mars Reconnaissance Orbiter, which has been orbiting Mars since 2006 at an altitude of approximately 250 to 316 kilometers. It captures images with a resolution of around 0.3 meters per pixel [9]. For training, the "JPEG IRB color no map" images were downloaded via web scraping using BeautifulSoup [10]. These images are not true-color RGB; instead, a min-max stretch is applied to each color band to enhance visual contrast [11]. After visual inspection, several common image artifacts were identified, including corrupted upper borders and vertical black or blue columns. Therefore, pre-processing began with the removal of the upper border, followed by the detection and removal of black and blue columns. To detect blue columns, scripts were first tested on a small, manually curated subset of images in order to fine-tune the hue range in the HSV color space. A hue range of [85, 95] (on OpenCV's 0–179 scale) was found to be most effective. In addition to these specific artifacts, some images exhibited more severe quality issues. A custom script was developed to detect extreme hue values and flag such images for manual review. These flagged images were then sorted into "keep" and "remove" sets based on visual inspection. After curation, the valid images were cropped into square tiles of size 512 × 512 pixels. The final dataset used for training StyleGAN2-ADA comprised 173,603 tiles generated from 24,609 original images. In contrast, the test dataset consisted of 50,000 tiles cropped from 19,998 original images.

3. Experiments

The metrics used in this analysis can be grouped into two categories. Distribution-based metrics evaluate the entire datasets by extracting feature maps and treating them as samples from underlying distributions. These metrics then compute some form of statistical distance or divergence between the distributions. Pairwise image similarity metrics, on the other hand, operate at the level of individual image pairs, comparing them either directly in pixel space or in an embedded feature space.

- Distribution-based metrics: FID [12], KID-poly [13], KID-rbf, CMMD [14], Precision and Recall [15], PPL [16], ISC [17]
- Pairwise image similarity metrics: MS-SSIM (pixel-based) [18], PSNR (pixel-based), LPIPS-Alex [16], LPIPS-VGG, DreamSIM [19]

To ensure a consistent evaluation benchmark, a dataset of 50,000 generated images was created using a modified script based on `generate.py` from the StyleGAN2-ADA repository. Truncation was set to $\psi = 1.0$, and the noise mode was fixed to `const`. For reproducibility, the dataset was generated using a fixed random seed.

3.1. Distribution based metrics

Table 1 shows the results of distribution-based metrics evaluated on the final StyleGAN2-ADA model. KID-poly, KID-rbf, and ISC were calculated using InceptionV3 embeddings with the Pytorch- Fidelity implementation [20]. CMMD was computed with the PyTorch implementation from [21]. FID was

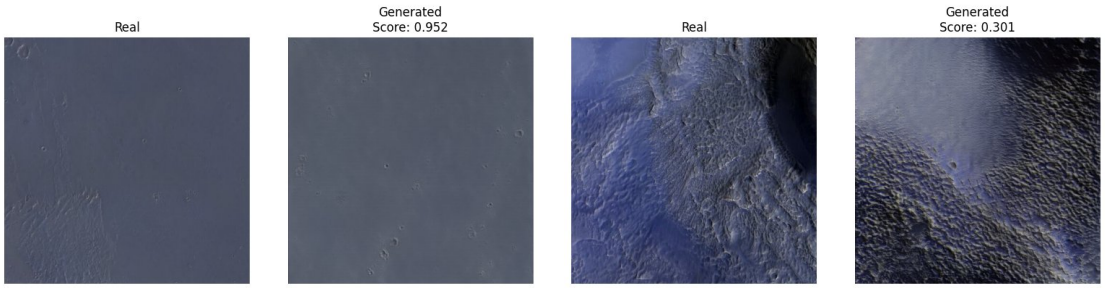
Metric	Score
FID (fid50k)	9.94359
FID (fidelity)	11.669854
KID (fidelity poly)	0.003367
KID (fidelity rbf)	0.00329
ISC (fidelity isc)	5.3313
CMMD	2.172
PPL (pplzend)	42.2750
PPL (pplwend)	25.6877
PPL (pplzfull)	42.6170
PPL (pplwfull)	25.1294
Precision	6.6
Recall	0.59

Table 1
Distribution based metrics for the final StyleGAN2-ADA-PyTorch models on the generated dataset.

evaluated both with Pytorch-Fidelity and NVIDIA's original FID50k implementation, the latter using image samples generated directly from the network rather than from a pre-generated dataset.

Notably FID from Fidelity is consistently higher compared to the FID from StyleGAN-ADA's implementation. These differences may be due to variations in the image generation procedure. Although both implementations use the InceptionV3 network, Fidelity uses a PyTorch version, while NVIDIA's original FID50k implementation relies on TensorFlow.

While no standardized benchmarks exist for synthetic planetary surface data, it is useful to contextualize the reported values using results from more established datasets. For example, FID values for StyleGAN2 on well-known datasets typically range from 5.3 to 12.96 (see Table 1 in [22]). KID provides an even more favorable comparison, the values reported in Figure 11 of [23] range between 0.15 and 3.36, depending on the training dataset. In contrast, Inception Scores between 8.55 and 10.02 are reported in the same paper, which the models trained in this thesis do not reach. Regarding PPL in w-space, values for StyleGAN2 have been shown to lie between 125 and 802 (see Table 2 in [24]), which is considerably worse that the performance of StyleGAN2-ADA in this thesis. In the CMMD benchmarks reported by [21], scores span roughly 0.55 to 1.14 for a variety of non–style-based generative models. StyleGAN2-ADA, however, falls outside this range.

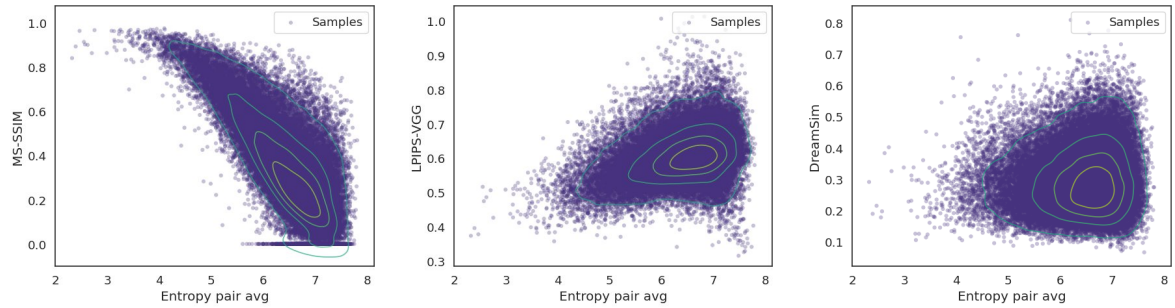


(a) good scoring pair (b) bad scoring pair

Figure 1: Examples of nearest-neighbor matched images pairs by CLIP of generated sample and test data sample with their corresponding MS-SSIM scores.

3.2. Pairwise Image Similarity Metrics

Pairwise Image Similarity Metrics were evaluated with three distinct matching schemes. First, each generated image was paired with a randomly selected real image from the test set. As reference random



(a) MS-SSIM (b) LPIPS-VGG (c) DreamSIM

Figure 2: Metric scores plotted against Image Entropy averaged over three color channels and image pair

matching was also performed for real-to-real images pairs and generated-to-generated. Second, the same k-nearest neighbor algorithm used for precision and recall was applied to match every generated image to its closest real counterpart in the test dataset on the base of the image embedding extracted from DINOv2-ViT-B/14, CLIP-ViT-B/32 and InceptionV3. Third, the nearest-neighbor matching was repeated, but this time using the training set instead of the test set. For all three schemes, pairwise Euclidean distances were computed in the feature spaces of DINOv2-ViT-B/14, CLIP-ViT-B/32 and Inceptionv3 to find the nearest neighbor.

The qualitative behavior of all metrics was the same. The average score improved when replacing random paring with nearest-neighbor matching and even more, albeit modest, when nearest-neighbors where drawn from train dataset. Interestingly, the choice of feature extractor has as much impact on the mean score as switching from random to either of the nearest-neighbor pairings. This is highlighted by an overlap analysis that examines the agreement between extractors: Fewer than 1 % of the generated images are matched to a crop originating from the same real image, regardless of the extractor pair. Among the three models, DINO and Inception agree the most, yet even in that case only 0.7 % of matches point to the same original image.

For both nearest-neighbor schemes in all metrics looked at, CLIP yields the lowest average similarity score, followed by Inception and DINO changing position. A visual examination of the image matches showed that the highest- scoring pairs look visually similar but often lack large-scale structure, whereas mid-range pairs typically show broader structural elements (see figure 1). Paradoxically, some lower-scoring pairs do not appear perceptually less similar than certain higher-scoring ones, suggesting that the metrics are disapproving complex scenes. To probe the observation that high scores are only archived by pairs lacking structure, mean Shannon entropy was computed for each color channel (RBG) of every image, averaged over the pair, and plotted against the corresponding metric score. Figure 2 shows the results plotted for (a) MS-SSIM as a pixel-based metric, (b) LPIPS-VGG as feature based metric, and (c) DreamSIM which weights were trained to align with human perception [19]. For MS-SSIM a clear bias is evident, pairs with low entropy tend to achieve the best MS-SSIM scores. This bias reduces for LPIPS-VGG where both the best and worst results are achieved by high-Entropy pairs. The best scoring pairs with high entropy show only very small repetitive structures no complex large scenes. For the DreamSIM this trend vanishes. The corresponding Pearson matrix shown in Figure 3 confirms a pronounced negative correlation $\rho \approx$-0.75 for MS-SSIM and $\rho \approx$-0.80 for PSNR, meaning that image pairs with higher Shannon entropy are systematically assigned lower similarity scores. By contrast, LPIPS (AlexNet and VGG backbones) exhibits only a weak positive correlation with entropy ($\rho \approx$0.30), while DreamSim is virtually independent of entropy.

4. Conclusion and Outlook

This work has demonstrated that the fidelity and diversity of the generated images are comparable to benchmarks from popular datasets, even though the training dataset used here was significantly

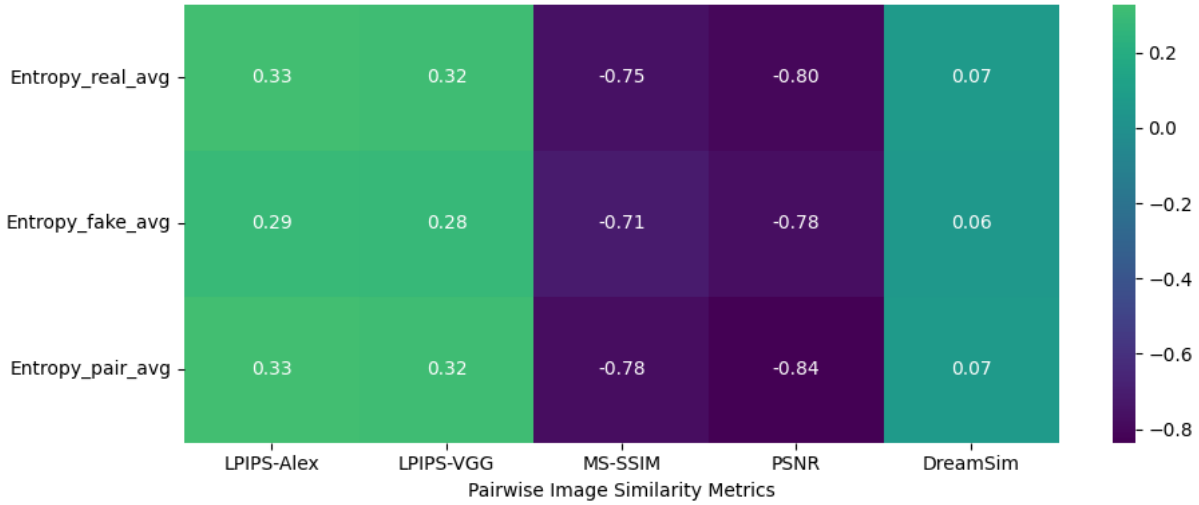


Figure 3: Pearson coefficient Heatmap of Image Entropy for average of real and generated (fake) images and the Entropy of the image pair average with different Pairwise Image Similarity Metrics.

smaller. However, direct comparisons across entirely different datasets are not necessarily reliable, as metric scores can vary greatly due to inherent dataset characteristics rather than the actual quality of generated images.

Furthermore it was shown that merely looking at the numerical scores can mask underlying flaws in the metrics themselves that arise on the particular dataset being looked at. For example, MS-SSIM and PSNR showed a strong correlation with image entropy, while LPIPS displayed this effect to a lesser extent. This indicates limitations in these metrics' ability to recognize perceptual similarity, especially in images with larger-scale structures. This bias could either be an inherit property of the metric or it could be caused by flawed nearest neighbor matches and therefor be an problem in the backbone model. This is plausible since the nearest neighbor matches created doubts if the embedding match human judgment of similarity, pointing towards a lack of generalization in the feature underlying extraction models. To really proof there a perceptually better matches, a comprehensive human judged study would have to be carried out.

Quality control for synthetic data generation therefore still poses a challenge. Fine-tuning these models to align with human judgment could play an essential role in making all feature-based metrics more reliable and interpretable. There are two primary approaches to achieving this alignment: either fine-tuning on a curated dataset specific to the context and application, or employing a more holistic approach, such as that presented in [25], which generally aims to match internal structure to human cognition. However, fine-tuning on a specific dataset typically comes at the cost of generalization and relies heavily on the availability of such curated data. This poses a significant challenge for planetary science, as there are neither extensively labeled (besides datasets for crater detection) nor otherwise curated datasets suitable for model training. Furthermore it is particularly difficult because the nature of these environments themselves is not yet fully understood, making it harder to definitively judge if generated images accurately resemble reality. While a more holistic approach might offer greater benefit, it too relies on potentially biased, often object-based, datasets. Future work should assess the performance of such fine-tuned models within this unique planetary domain.

Therefore, until a sufficiently trustworthy and universal metric is established, anyone generating, and especially using, synthetic data should meticulously examine it rather than solely relying on single scores like FID, IS, KID. The details of such an examination are highly domain- and application-specific. Some domains may have pre-defined statistics the generated images should match, or simulated data that they can be compared with. But visual assessments even on small portions of data can highlight underlying deficiencies, in the sense that the metric does not align with human perception.

Acknowledgments

This research was supported by the Excellence Cluster ORIGINS which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC-2094 - 390783311.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar, spelling check and paraphrase and reword. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] H. Viggh, S. Loughran, Y. Rachlin, R. Allen, J. Ruprecht, Training deep learning spacecraft component detection algorithms using synthetic image data, in: 2023 IEEE Aerospace Conference, 2023, pp. 1–13. doi:10.1109/AERO55745.2023.10115578.

[2] P. Suwinski, A. Liesch, B. Liu, F. Schnitzer, T. Kohlsmann, K. Janschek, 2d and 3d data generation and workflow for ai-based navigation on unstructured planetary surfaces, in: AIAA SCITECH 2024 Forum, 2024. URL: https://arc.aiaa.org/doi/abs/10.2514/6.2024-1744. doi:10.2514/6.2024-1744. arXiv:https://arc.aiaa.org/doi/pdf/10.2514/6.2024-1744.

[3] A. Escalante Lopez, P. Ghiglino, M. Sanjurjo-Rivo, Applying machine learning techniques for optical relative navigation in planetary missions, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 1–11. doi:10.1109/TGRS.2024.3374454.

[4] W. Tu, W. Deng, T. Gedeon, Toward a holistic evaluation of robustness in CLIP models, 2024. URL: http://arxiv.org/abs/2410.01534. doi:10.48550/arXiv.2410.01534. arXiv:2410.01534 [cs], version: 1.

[5] D. Torpey, R. Klein, On the robustness of self-supervised representations for multi-view object classification 161 (2022) 82–89. URL: https://www.sciencedirect.com/science/article/pii/S0167865522002276. doi:10.1016/j.patrec.2022.07.016.

[6] M. Baharoon, W. Qureshi, J. Ouyang, Y. Xu, A. Aljouie, W. Peng, Evaluating general purpose vision foundation models for medical image analysis: An experimental study of DINOv2 on radiology benchmarks, 2024. URL: http://arxiv.org/abs/2312.02366. doi:10.48550/arXiv.2312.02366. arXiv:2312.02366 [cs], version: 4.

[7] X. Li, C. Wen, Y. Hu, N. Zhou, RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision 124 (2023) 103497. URL: https://www.sciencedirect.com/science/article/pii/S1569843223003217. doi:10.1016/j.jag.2023.103497.

[8] H. Team, Hirise image catalog, 2006–present. URL: https://www.uahirise.org/, accessed: 2025-05-16.

[9] A. S. McEwen, E. M. Eliason, J. W. Bergstrom, N. T. Bridges, C. J. Hansen, W. A. Delamere, J. A. Grant, V. C. Gulick, K. E. Herkenhoff, L. Keszthelyi, R. L. Kirk, M. T. Mellon, S. W. Squyres, N. Thomas, C. M. Weitz, Mars reconnaissance orbiter's high resolution imaging science experiment (HiRISE) 112 (2007). URL: https://onlinelibrary.wiley.com/doi/abs/10.1029/2005JE002605, publisher: John Wiley & Sons, Ltd.

[10] L. Richardson, Beautiful soup documentation, April (2007).

[11] A. S. McEwen, M. E. Banks, N. Baugh, K. Becker, A. Boyd, J. W. Bergstrom, R. A. Beyer, E. Bortolini, N. T. Bridges, S. Byrne, B. Castalia, F. C. Chuang, L. S. Crumpler, I. Daubar, A. K. Davatzes, D. G. Deardorff, A. DeJong, W. Alan Delamere, E. N. Dobrea, C. M. Dundas, E. M. Eliason, Y. Espinoza, A. Fennema, K. E. Fishbaugh, T. Forrester, P. E. Geissler, J. A. Grant, J. L. Griffes, J. P. Grotzinger, V. C. Gulick, C. J. Hansen, K. E. Herkenhoff, R. Heyd, W. L. Jaeger, D. Jones, B. Kanefsky, L. Keszthelyi, R. King, R. L. Kirk, K. J. Kolb, J. Lasco, A. Lefort, R. Leis, K. W. Lewis, S. Martinez-Alonso, S. Mattson,

G. McArthur, M. T. Mellon, J. M. Metz, M. P. Milazzo, R. E. Milliken, T. Motazedian, C. H. Okubo, A. Ortiz, A. J. Philippoff, J. Plassmann, A. Polit, P. S. Russell, C. Schaller, M. L. Searls, T. Spriggs, S. W. Squyres, S. Tarr, N. Thomas, B. J. Thomson, L. L. Tornabene, C. Van Houten, C. Verba, C. M. Weitz, J. J. Wray, The high resolution imaging science experiment (HiRISE) during MRO's primary science phase (PSP) 205 (2010) 2–37. URL: https://www.sciencedirect.com/science/article/pii/S0019103509001808. doi:10.1016/j.icarus.2009.04.023.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL: http://arxiv.org/abs/1706.08500. doi:10.48550/arXiv.1706.08500. arXiv:1706.08500 [cs].

[13] M. Bińkowski, D. J. Sutherland, M. Arbel, A. Gretton, Demystifying MMD GANs, 2021. URL: http://arxiv.org/abs/1801.01401. doi:10.48550/arXiv.1801.01401. arXiv:1801.01401 [stat].

[14] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, S. Kumar, Rethinking FID: Towards a better evaluation metric for image generation, 2024. URL: http://arxiv.org/abs/2401.09603. doi:10.48550/arXiv.2401.09603. arXiv:2401.09603 [cs].

[15] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, T. Aila, Improved precision and recall metric for assessing generative models, 2019. URL: http://arxiv.org/abs/1904.06991. doi:10.48550/arXiv.1904.06991. arXiv:1904.06991 [stat].

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL: http://arxiv.org/abs/1801.03924. doi:10.48550/arXiv.1801.03924. arXiv:1801.03924 [cs].

[17] S. Barratt, R. Sharma, A note on the inception score, 2018. URL: http://arxiv.org/abs/1801.01973. doi:10.48550/arXiv.1801.01973. arXiv:1801.01973 [stat].

[18] M. Abdel-Salam Nasr, M. F. AlRahmawy, A. S. Tolba, Multi-scale structural similarity index for motion detection 29 (2017) 399–409. URL: https://www.sciencedirect.com/science/article/pii/S1319157816300088. doi:10.1016/j.jksuci.2016.02.004.

[19] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, P. Isola, DreamSim: Learning new dimensions of human visual similarity using synthetic data, 2023. URL: http://arxiv.org/abs/2306.09344. doi:10.48550/arXiv.2306.09344. arXiv:2306.09344 [cs].

[20] A. Obukhov, M. Seitzer, P.-W. Wu, S. Zhydenko, J. Kyl, E. Y.-J. Lin, High-fidelity performance metrics for generative models in pytorch, 2020. URL: https://github.com/toshas/torch-fidelity. doi:10.5281/zenodo.4957738, version: 0.3.0, DOI: 10.5281/zenodo.4957738.

[21] S. Paul, Cmmd: Clip-based maximum mean discrepancy metric in pytorch, 2024. URL: https://github.com/sayakpaul/cmmd-pytorch, accessed: 2025-06-08.

[22] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, J. Lehtinen, The role of ImageNet classes in fréchet inception distance, 2023. URL: http://arxiv.org/abs/2203.06026. doi:10.48550/arXiv.2203.06026. arXiv:2203.06026 [cs].

[23] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila, Training generative adversarial networks with limited data, 2020. URL: http://arxiv.org/abs/2006.06676. doi:10.48550/arXiv.2006.06676. arXiv:2006.06676 [cs].

[24] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of StyleGAN, 2020. URL: http://arxiv.org/abs/1912.04958. doi:10.48550/arXiv.1912.04958. arXiv:1912.04958 [cs].

[25] L. Muttenthaler, K. Greff, F. Born, B. Spitzer, S. Kornblith, M. C. Mozer, K.-R. Müller, T. Unterthiner, A. K. Lampinen, Aligning machine and human visual representations across abstraction levels, 2024. URL: http://arxiv.org/abs/2409.06509. doi:10.48550/arXiv.2409.06509. arXiv:2409.06509 [cs].

A. Online Resources

The sources for the code used:

- GitHub