Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/132/2024.20706

The Effect of Dataset Imbalance on the Performance of Image-to-Cartoon Generative Adversarial Networks

Yuqi Wang[1*], Zihan Cai[2], Qinghong Zhang[3]

[1] Benjamin Franklin International School, Barcelona, 08017, Spain, yqw.yuqiwang@gmail.com
[2] Anhui University, Anhui, 230039, China, czh1412@outlook.com
[3] Shanghai electricity of power university, Shanghai, 200090, China, youkuxiaozhang@hotmail.com

*corresponding author

Abstract. This report investigates the impact of dataset imbalance on AI-powered image-to-anime style transfer, focusing on the AnimeGANv3 model. Despite the common perception of AI as free of human bias, we highlight that machine learning systems inherently reflect societal prejudices through their training data. The anime art style, popular worldwide but limited in its representation of diverse ethnicities and cultures, serves as a case study for this phenomenon. We analysed AnimeGANv3's training datasets and compared its performance on over- and underrepresented image classes using quantitative and qualitative metrics. Results demonstrate that users from minority groups likely experience inferior outcomes due to dataset imbalance. The study emphasises the need for transparent and responsible dataset curation for machine learning systems to ensure ethical AI development and improved model performance across all user groups.

Keywords: generative adversarial networks, dataset imbalance, image stylization, AI ethics, image-to-image translation.

1. Introduction

The conventional anthropomorphisation of AI depicts machine learning programs as unfeeling algorithms free of human bias. While this presentation of AI is convenient for entertainment and media, the notion that AI is absolutely objective can become a dangerous presumption as machine learning systems become ever more prevalent in daily life [1]. Starting from the most basic levels of a model's training process, human biases are imbued into the model during dataset creation.[2][3] Regardless of whether the insertion of bias is intentional, models are trained on man-made data and, by extension, propagate the prejudices and inequalities present in society. This is not a reason to discredit AI systems, rather, these biases must receive appropriate recognition and understanding so that technological advances can be made with integrity.

The style of Japanese animation commonly referred to as 'anime' has gained popularity worldwide within the past decades. Although anime has a worldwide fanbase, it predominantly depicts young Asian or Caucasian characters in asiatic settings, rarely representing other ethnicities or cultures. This homogeneity creates challenges for image-to-anime conversion models, as their training datasets lack

DOI: 10.54254/2755-2721/132/2024.20706

diversity and fail to reflect the broad spectrum of users worldwide.

In recent years, machine learning based image-to-image translation methods such as generative adversarial networks (GANs) have achieved successful artistic style transfer results[4][5][6][7]. However, anime style transfer presents a unique challenge. While artistic style transfer focuses on applying painting textures and colours to target images, anime style transfer aims to generate images that closely resemble hand-drawn anime artwork[8]. Although some GAN models have demonstrated the ability to generate results similar to the work of anime artists, the generated images are often degraded by issues such as noise, artefacts, excessive blur, and semantic structure mismatch. These problems appear more frequently for images of scenes and individuals that are underrepresented in anime content.

To better understand the above problem, we studied the effect of dataset imbalance on the AnimeGANv3 model, the most recent iteration in the AnimeGAN series known for anime style transfer. We analysed its training datasets and compared model performance on overrepresented and underrepresented image classes. The results were evaluated qualitatively through visual observation and quantitatively using the FID (Fréchet Inception Distance) and KID (Kernel Inception Distance) metrics. FID and KID are widely used metrics in the field of generative models. In the context of anime style
transfer, these metrics are suitable because they effectively capture the nuances of visual similarity between the real and generated images. Comparison of generated images from overrepresented and underrepresented classes reveals that dataset curation significantly impacts model performance, with users from minority classes likely experiencing inferior results.

Transparency and responsible curation of balanced training datasets is key to ethical machine learning development. Current approaches to address dataset imbalance include resampling, augmentative oversampling, and synthetic data generation. These methods can improve dataset balance by providing more representative samples and reducing bias. However, it is ultimately crucial for AI developers to actively seek diverse data sources and collaborations with underrepresented communities to ensure comprehensive representation in training datasets both for the sake of improved model performance and ethical AI development.

2. Related Work

2.1. Double tail generative adversarial network

The model used in our study, AnimeGANv3, is described in the paper “A Novel Double-Tail The model used in our study, AnimeGANv3, is described in the paper “A Novel Double-Tail Generative Adversarial Network for Fast Photo Animation” by Liu et al. In this paper, the authors propose a double output tail approach to produce high quality anime style images. Architecturally, significant emphasis was placed on the careful selection of normalisation and loss functions. This included the use of an improved grayscale loss to prevent the colour of the anime image from interfering with texture learning, and a revised colour reconstruction loss that utilises the Lab colour space to achieve colour accuracy more closely aligned with human vision [8].

In contrast, their discussion of dataset-related work is brief. The paper mentions that the model was trained on landscape images from CycleGAN's photo data and an anime dataset consisting of frames captured from anime films. However, it fails to acknowledge that the anime image data does not exclusively feature landscapes. Instead, the images are random screenshots that include character faces, empty walls, buildings, miscellaneous objects, and everything in between. It could be suggested that the innovations in model architecture were limited by the dataset curation, and a dataset selection process matching the level of effort put into the model's structural design holds the potential to greatly improve performance.

DOI: 10.54254/2755-2721/132/2024.20706

2.2. Imbalance problems in computer vision tasks

The development of any computer vision model begins with collecting images and data, followed by any preprocessing and pattern recognition to carry out the intended task [9]. However, if the collected images are significantly imbalanced or insufficient, the desired outcome may become unachievable regardless of the quality of the model itself. In "A survey on generative adversarial networks for imbalance problems in computer vision tasks", Sampath et al. describe the impact of inter-class dataset imbalance, noting that classifiers developed with such imbalanced datasets are prone to predicting minority classes as rare events, often treating them as outliers or noise, which leads to the misclassification of these minority classes [9].

Some other negative effects of dataset imbalance include the discriminator potentially becoming too strong due to the imbalances, which results in poor gradients for the generator concerning the minority classes. Consequently, GANs trained on such imbalanced datasets are prone to mode collapse, where the generator produces limited or repetitive outputs. As a result, the generator may focus predominantly on generating samples from the majority class, neglecting the minority class due to its dominance in the dataset [9]. In order to address this issue, several solutions to synthetically adjust to the training data such as resampling, augmentative oversampling, semi-supervised learning, and cost sensitive learning are explored. While each method can be successful in specific situations, synthetic images created using traditional data level approaches may not be truly representative of the training set and methods involving the duplication of data, such as resampling, will not contribute productively to improve classifier performance.

This survey provides a comprehensive, multifaceted overview on the issue of dataset imbalance in computer vision tasks, however, it covers a wide spectrum of content at a technical, abstract level, leaving room for case-specific application for more intuitive understanding.

2.3. Bias in data driven AI systems

Given the extensive influence of AI-guided algorithms on government, corporate, and other institutional decisions, it is crucial to understand and study biases in AI to prevent AI-based decision-making from amplifying pre-existing biases or creating new ones [9][10]. The issues of understanding, mitigating, and accounting for bias are addressed in the survey "Bias in data-driven artificial intelligence systems" by Ntoutsi et al. In machine learning, bias refers to the assumptions made by a specific model. With imbalanced datasets, this bias manifests as an "inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair" [10].

Statistical machine learning inferences require that the training data be representative of the data on which it is applied. However, data collection often suffers from biases that lead to the over- or under representation of certain groups, particularly because many datasets are not created with the rigour of a statistical study, such as those used for casual, aesthetic applications like image-to-anime conversion. In these cases, human biases can enter AI systems and may be further amplified by the complex sociotechnical cycle.

The survey suggests that these biases can be mitigated through pre-processing, in-processing, and post-processing methods. Pre-processing techniques focus on the primary source of bias to produce "balanced" datasets for training; in-processing solutions address the classification problem by explicitly incorporating the model's discrimination behaviour in the objective function through regularisation or constraints; and post-processing methods include white-box and black-box approaches that modify the model or its predictions after training [4]. In our study, we will explore the pre-processing approach by analysing the correlation between dataset diversity and model fairness.

3. Methodology

The AnimeGANv3 model generator is formed by two output tails: a support tail for coarse-grained anime style images with specific high frequency noise and artefacts and a main tail for denoising and removing artefacts from the support tail outputs. Linearly adaptive denormalization (LADE), which can obtain the global data distribution of all channels and use this distribution to guide instance

DOI: 10.54254/2755-2721/132/2024.20706

normalisation, is used to solve artefacts in generated anime images. The model employs various loss functions, including a region smoothing loss function, which is used to weaken the texture details of the generated images to achieve anime effects with abstract details, and a fine-grained revision loss function, which is used to eliminate artefacts and noise in the generated anime style image while preserving clear

edges. Additionally, a lightweight attention module was used to reduce the size of the generator, resulting in AnimeGANv3 needing only 1.02 million parameters in the inference phase. The model was trained on unpaired data in an unsupervised manner, with an initial pre-training phase for the generator to accelerate convergence and enhance training stability [8].

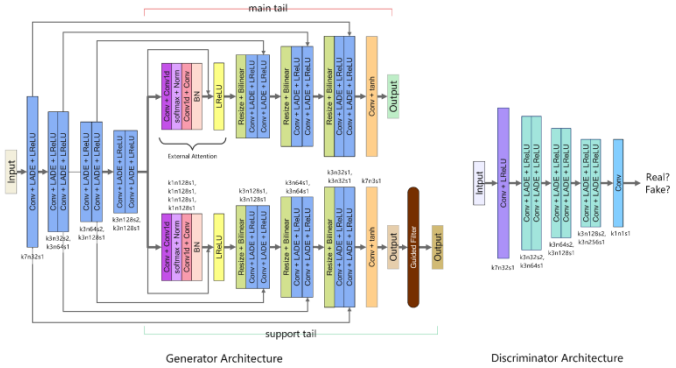

Figure 1. AnimeGANv3 model architecture

From the previous paragraph, it is evident that the developers of AnimeGANv3 prioritised architectural innovations to maximise model performance. Our experiment aims to investigate how error propagation through the model's layers, originating from an imbalanced training dataset, may constrain the model's potential despite its innovative architecture. We performed inference with two pretrained models on two different image sets: one provided by the project repository, similar to the training dataset,

and one we created ourselves with handpicked, underrepresented images, each containing 60 photos. Each image set was processed five times, and an average was calculated to ensure accuracy in our results and mitigate the impact of random error. This approach will help us understand the impact of data quality on model performance, independent of architectural advancements.

The AnimeGANv3 GitHub repository offers two pretrained models: a Shinkai model trained on the artworks of director Makoto Shinkai, and a Hayao model trained on the artworks of director Hayao Miyazaki. Both models were trained on 256x256 unpaired real-world and anime images and tested on real-world photos. For the anime datasets, around 1500 screenshots were collected from the films of the respective directors to best capture their artistic styles [11]. Although these films have garnered international acclaim from fans worldwide, they primarily feature young Japanese characters within the Japanese countryside. Regarding the inference datasets, the first is the official inference image set from the GitHub repository. This set features images highly similar to the anime images used during training, depicting Asian individuals in Asian settings. The second set was curated by us after analysing the model outputs to identify weak points corresponding to certain classes of images, including individuals with dark skin tones, elderly individuals, and Western metropolitan settings.

Our experimental results were evaluated quantitatively using the FID (Fréchet Inception Distance) and KID (Kernel Inception Distance) metrics. FID measures the similarity between the feature representations of real images and generated images with lower FID scores indicating that the model successfully captures the characteristics represented in real anime photos. Likewise, KID is another metric that compares the feature representations of real and generated images, specifically using kernel methods to estimate the distributional discrepancy. Similar to FID, lower KID scores suggest that the distribution of features in the generated images aligns well with the distribution in real cartoon images.

DOI: 10.54254/2755-2721/132/2024.20706

This indicates that the model successfully captures both local and global features that are characteristic of animated styles.

Qualitatively, we evaluated the visual impact of changes in the aforementioned quantitative metrics by assessing each image for two key aspects: (1) successful anime-style adaptation, characterised by the presence of distinctive anime stylistic elements and absence of discrepancies such as cracked lines; and

(2) fidelity to the original image, measured by the preservation of essential characteristics like facial structure and key identifying features.

The combination of quantitative and qualitative metrics allow us to perform a comprehensive study into the impact of dataset imbalance on image-to-cartoon style transfer models.

4. Experimental Results

The following experiments were conducted on Google Colab using a single Intel(R) Xeon(R) CPU @ 2.20GHz with two logical processors. Quantitative results are shown in Table 1 and qualitative results are shown in Figures 2 and 3. For the following evaluation, 'overrepresented image' is abbreviated as OI and 'underrepresented image' as UI.

Table 1. Quantitative Comparison of Model Performance

	FID	KID
AnimeGANv3 training results	40.77	0.42
Shinkai inference with OI	54.06	1.78
Hayao inference with OI	61.44	2.10
Shinkai inference with UI	93.66	6.22
Hayao inference with UI	119.10	11.80

In terms of baseline performance, the original paper reports an FID of 40.77 for AnimeGANv3. Our results, shown in Table 1, demonstrate higher FID scores even with the official dataset, this is likely due to difference in dataset size and variations in implementation. For performance on underrepresented data, there's a significant increase in both FID and KID scores when using the dataset of underrepresented photos. The FID score increased by 73% for the Shinkai variant and 94% for the Hayao variant, while the KID score increased by 249% and 462% respectively. Although the Shinkai model performs slightly better than the Hayao model on both datasets, both show similar patterns of degradation. This suggests that the issue of dataset imbalance affects different stylization approaches similarly.

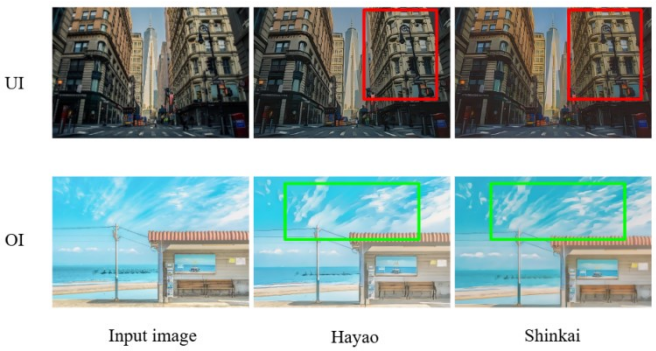

Figure 2. Qualitative Comparison of Model Performance on OI and UI Landscapes

DOI: 10.54254/2755-2721/132/2024.20706

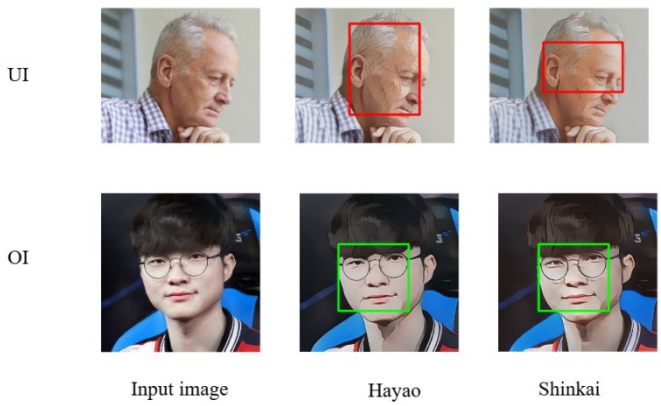

Figure 3. Qualitative Comparison of Model Performance on OI and UI Faces

In Figure 2, the model successfully stylizes the OI of the Japanese seaside train stop. This is most evident in the clouds, which are transformed into anime-like paint strokes after processing. The rest of the image also reflects this success, as the train stop adopts a simple outline with smoothed colours, similar to its depiction in an anime film. In contrast, the UI of Manhattan shows minimal changes after processing. Although the colours are segmented and smoothed, no further stylization is apparent beyond this rudimentary colour blocking. For the Shinkai model output, both images exhibit changes in brightness and saturation, suggesting that the model attempts to apply the colour palette used in Shinkai films. However, there remains a significant disparity in stylization between the OI and UI, highlighting the restrictive nature of an imbalance dataset on model performance.

In Figure 3, the model effectively simplifies and stylizes the facial features in the processed OI to fit an anime aesthetic. However, the facial features in the processed UI are flawed and lack proper stylization. Specifically, the wrinkles on the elderly man's face become cracked lines, and his pale hair appears jagged and unnatural. This discrepancy arises because the model's training data lacks sufficient representation of features commonly found in older individuals. Consequently, the model's feature space is not well-adapted to handle these characteristics, leading to a failure in preserving the defining attributes and authenticity of the original image. This highlights the model's limitation in generalising across different age groups and its difficulty in accurately stylizing underrepresented features.

Our findings align with broader concerns in AI about bias and representation in training data. Models trained on biased datasets (e.g. predominantly young Asian faces) may perform poorly or produce biased results when applied to a global, diverse user base. This highlights the need for careful consideration of dataset composition in AI development, especially for applications intended for diverse user groups.

5. Discussion and Future Work

In this report, we examined the effect of dataset imbalance on the performance of anime style transfer GAN models by comparing inference results from overrepresented and underrepresented images. Due to hardware limitations, we were unable to perform the full training process and were restricted to using the pretrained model for inference. Ability to train the model would produce more nuanced and conclusive results as we originally aimed to track the differences between overrepresented and underrepresented images at each stage of image processing. This would involve monitoring the training process for the main and support branches by observing loss functions, gradients, and the output images at various stages of processing to study how the quality evolves. Additionally, data from training the model would also allow for more accurate comparison with the results presented in the original AnimeGANv3 paper.

This issue of dataset imbalance is especially prevalent in image-to-anime style transfer models due to the monocultural nature of anime content. The impact of imbalance datasets on biased model performance demonstrated in our study can potentially marginalise minority users and perpetuate

DOI: 10.54254/2755-2721/132/2024.20706

societal inequalities. This issue extends beyond aesthetic applications in image-to-cartoon translation, affecting critical fields such as facial recognition, medical diagnosis, and automated decision-making systems[10]. It is therefore demonstrated that dataset imbalance is an obstacle that must be addressed not only to improve model performance but also to ensure the ethical development of AI technologies that fairly serve diverse populations.

References

[1] "What is ai ethics?," IBM, https://www.ibm.com/topics/ai-ethics (accessed Jul. 21, 2024).
[2] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, "A deeper look at dataset bias," Domain Adaptation in Computer Vision Applications, pp. 37–55, 2017. doi:10.1007/978-3-319-58347-1_2
[3] Simone Fabbrizzi et al., "A survey on bias in visual datasets," Computer Vision and Image Understanding, https://www.sciencedirect.com/science/article/abs/pii/S1077314222001308 (accessed Jul. 21, 2024).
[4] R. Wu, X. Gu, X. Tao, X. Shen, Y. Tai, and J. Jia, "Landmark assisted CycleGAN for cartoon face generation," CoRR, vol.abs/1907.01424, 2019.
[5] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative Adversarial Networks for Photo Cartoonization," Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit. (CVPR2018), pp.9465–9474, Dec. 2018.
[6] J. Chen, G. Liu, and X. Chen, "AnimeGAN: A Novel Lightweight GAN for Photo Animation," Commun. Comput. Info. Sci. (ISICA2019), pp.242–256, 2020.
[7] X. Wang and J. Yu, "Learning to Cartoonize Using White-Box Cartoon Representations," Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit. (CVPR2020), pp.8087–8096, 2020.
[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017. doi:10.1109/cvpr.2017.632
[9] G. LIU, X. CHEN, and Z. GAO, "A novel double-tail generative adversarial network for Fast Photo Animation," IEICE Transactions on Information and Systems, vol. E107.D, no. 1, pp. 72–82, Jan. 2024. doi:10.1587/transinf.2023edp7061
[10] V. Sampath, I. Maurtua, J. J. Aguilar Martín, and A. Gutierrez, "A survey on generative adversarial networks for imbalance problems in computer vision tasks," Journal of Big Data, vol. 8, no. 1, Jan. 2021. doi:10.1186/s40537-021-00414-0
[11] E. Ntoutsi et al., "Bias in data‐driven Artificial Intelligence Systems—an introductory survey," WIREs Data Mining and Knowledge Discovery, vol. 10, no. 3, Feb. 2020. doi:10.1002/widm.1356