

# Enhancing Text-to-Image Synthesis using Data-Centric Approaches

Thivagarasarma Karthick

*Department of Computer Engineering  
University of Sri Jayewardenepura  
Nugegoda, Sri Lanka  
en93899@sjp.ac.lk*

Jathurshan Pathamarasa

*Department of Computer Engineering  
University of Sri Jayewardenepura  
Nugegoda, Sri Lanka  
en93827@sjp.ac.lk*

Sachchithananthan Gowreeshan

*Department of Computer Engineering  
University of Sri Jayewardenepura  
Nugegoda, Sri Lanka  
en93817@sjp.ac.lk*

Ishaqu Mohamed Infas

*Department of Computer Engineering  
University of Sri Jayewardenepura  
Nugegoda, Sri Lanka  
en93896@sjp.ac.lk*

Bhagya Nathali Silva

*Department of Information Technology  
Sri Lanka Institute of Information Technology  
Malabe, Sri Lanka  
nathali.s@slit.lk*

Dilani Ranaweera

*Department of Computer Engineering  
University of Sri Jayewardenepura  
Nugegoda, Sri Lanka  
dilani.ranaweera@sjp.ac.lk*

**Abstract**—In the quickly changing area of generative AI, text-to-image generation models are being noticed a lot because they can make detailed and varied images from textual explanations. Even though these models have potential to transform many things, they suffer from training times that are too long - this is caused by large amounts of data needed for training and also needing much computational power. This study looks at methods focusing on data to improve how efficiently we train text-to-image generation models. We utilize Curriculum Learning and Active Learning to make the data selection better. This helps us reduce training times while enhancing image quality. We assess our method thoroughly on three main benchmark datasets called Oxford-102, CUB-200, and MS COCO. Every dataset offers different difficulties in terms of complexity and variety of images found in them.

Through experiments, we find out that our model-agnostic method greatly decreases training duration and enhances the quality of images in most cases. Particularly, models like GAN-INT-CLS, StackGAN, AttnGAN and VQ-Diffusion show faster convergence and better performance when trained with our improved data subsets.

**Index Terms**—Text-to-Image synthesis, Curriculum learning, Active learning, Coreset selection

## I. INTRODUCTION

Text-to-image synthesis is a growing field in artificial intelligence that allows the creation of images from text descriptions. This technology has many uses, including making content, entertainment, design and accessibility. The progress made in this area recently has been shown by models like DALL-E [1] and Imagen [2]. But, these developments have also shown some key problems linked to data like dealing with data accuracy, expensive training costs and the quality as well as variety of datasets.

A big problem with text-to-image synthesis is that datasets often have incorrect labels. In a study by Northcutt et al. [3], they note that mistakes in labeling can cause models to not work as well and create misunderstandings about what the computer makes. The incorrect answers accentuate the impor-

tance of accurate and dependable labeling methods to uphold the truthfulness of training data. Furthermore, the considerable expenses related with training cutting-edge models are another significant problem. For illustration, DALL-E was trained on 250 million images using 1,024 x 16 GB NVIDIA V100 GPUs; Imagen needed 860 million images and it was trained with 256 x TPU v4 chips [1, 2]. The instances show the big amount of computational and financial resources required, underlining the necessity for training methods that can achieve similar or better performance with less resources.

Also, there is a second hurdle created by the shortages and prejudices in the available real-world data. In many cases, real-world data does not have enough labeling and shows biases that can affect model training and performance. According to [4], Generated images can be affected by biases present in the training dataset, causing the model to produce less diverse or representative outputs [4]. In addition, the absence of rich labeling limits the ability of models to comprehend and correctly generate complex scenes.

Taking into account these difficulties, our study is focused on investigating if using a Data-Centric AI (DCAI) method can result in statistically noticeable decrease in training time while at the same time improving quality and variety of generated samples. Through coreset selection, we plan to find and use the most informative parts of data sets. This could help reduce how much data is needed for training without sacrificing performance. We expect that this method will greatly decrease demands for computational resources and time.

Moreover, while we use DCAI methods for careful choosing and organizing of data, it can help to reduce the problems caused by label inaccuracies. This makes sure that the training data is of good quality and aids in effective learning for models. If a DCAI approach is followed, we suggest that similar or even better results can be obtained with just a portion of the data and resources used at present. This could result in major reduction of expenses related to training.

The study also concentrates on dealing with issues related to data bias and scarcity. A more balanced and complete set of information has potential to improve model's capacity in creating different types of samples with high quality.

Moreover, it is anticipated that the inclusion of methods such as active learning and curriculum learning will enhance the training process. Here, active learning enables the model to concentrate on difficult samples, thus enhancing its ability to learn effectively. In contrast, curriculum learning slowly introduces difficulty into the model's training procedure. This helps improve generalization and performance of the model.

The subsequent sections of this article are structured as follows: Section II offers a summary of related work, introducing knowledge from prior researches about data-centered methods for text-to-image generation. Section III delves deeper into proposed methodology. Section IV demonstrates procedure applied in the study, which includes the datasets used, selection of models and effectiveness of applying data-centric approaches. Section V shows results from experiments that have been done; it contains comparisons on quality and diversity as well as observations made during analysis along with insights gained from them. Lastly, in Section VI, the article is concluded by summarizing the main contributions of this study and pointing out its implications, recognizing restrictions, as well as discussion on future works.

## II. RELATED WORKS

Handling unlabeled data and using low amounts of data for training are challenges that recent text-to-image synthesis advancements have attempted to tackle through different proposed solutions. These approaches mainly concentrate on enhancing model performance within such restrictions, along with the efficient management of computational resources.

[5] presented a self-supervised method for tackling problems in scenarios with little data. Their strategy showed great improvement in low-data conditions, but it also demanded more computational resources. [6] used unlabeled data by adding a pretrained CLIP model into the process. This helped in making better use of data that did not have labels, but it also demanded lots of computer resources.

In a different study, [7] used many discriminators to create powerful and general images. This method enhanced the quality of produced images but didn't completely fix the issue of data scarcity. In [8], they used core-set selection method for conditional GANs that are based on inception embeddings. This brought about a significant decrease in the requirement for training resources, time and data; showing it as an encouraging technique for generating images conditionally.

[9] employed curriculum learning for low-data training, also using it on conditional GANs. They achieved quicker convergence and enhanced model steadiness by progressively raising the difficulty level of the training data. However, this technique makes an assumption about the existence of enough training data, which is not always true. [2] discussed the difficulties related to using a large amount of labeled data for text-to-image creation. They stressed that the model's

effectiveness is greatly influenced by how good and correct the labeled content is, highlighting problems like not having enough data and managing scalability.

In general, these studies highlight how crucial it is to develop new methods that can handle both limited and unlabeled data efficiently, all while minimizing the demand for computational resources. Our study builds upon these understandings by suggesting a Data-Centric AI technique for improving text-to-image synthesis. This method tackles issues related to data accuracy, reducing training costs, and improving the quality and diversity of generated samples.

## III. METHODOLOGY

In this section, we describe the methodologies used in our approach to reduce the training time for text-to-image generation models. We use a combination of three techniques: Curriculum Learning, Core Set Selection, and Active Learning.

### A. Curriculum learning

Curriculum learning [10] is a training strategy where the initializing point of the model is the easiest examples first and it progresses to more challenging examples. This emulates the human learning process and should make models converge faster and reach better performance. The dataset is partitioned into easy and hard data using a KeyBERT [11] model, from the text descriptions for each image. Specifically, data points with more than a threshold of 5 keywords per caption are considered hard data while the rest are regarded as easy data. This starts the learning process from easier examples and gradually increases the complexity.

### B. Coreset selection

Coreset selection [12] is the process of selecting a subset of the dataset optimal for representing the whole data distribution. This procedure minimally affects the training time of a model from a dataset and only incurs a small performance drop because it selects the most informative data points. We used greedy approach [13] for coreset selection. Firstly, we select one embedding randomly. After that, following embeddings are selected at each step based on the distance to the selected ones, so that the chosen are diverse. subset. This process iterates further until 10% of the easy data and 50% of the hard are selected. This ensures that the coreset is both representative and manages to be small in size, hence making the training faster.

### C. Active learning

Active learning [14] is whereby the model has an active process that queries the most informative data points to include in the training set. Technically, it is an iterative process to learn more smartly by eventually focusing only on the hardest examples that it currently most struggles with. First, train the Text-to-Image generation model on 10% of the easy data for 50 epochs. Use the trained model to predict 50% of the hard data. Among those predicted data, choose the instances which

the generator loss is highest, as they are the most difficult examples for the model. Append this selected hard instances to the initial training data. This augmented dataset is trained again for an extra 20 epochs. We repeat this predict, select and retrain process gradually adding more challenging instances to the training set until we have trained the model for a total of 200 epochs. By combining these three approaches we can significantly reduce the amount of time taken to train the model while actually improving its performance on text-to-image generation.

#### IV. EXPERIMENT

At the time of this writing, there had not been any work specifically on reducing the time required to train a text-to-image generation models, and therefore we did not make any comparison of our approach with other baselines.

##### A. Dataset

- 1) Oxford-102 Flowers dataset [15]: It contains 8,189 images of flowers in 102 categories. Each image is annotated with 10 different kinds of text descriptions. The descriptions of the text describe the flowers on different aspects that help generate images accurately and diversely.
- 2) CUB-200 Birds dataset [16]: It contains 8,855 training and 2,933 test images belonging to 200 bird species. Each image in this dataset is annotated with 10 detailed text descriptions thus it is densely annotated for the generation of high-fidelity bird images.
- 3) MSCOCO [17]: The dataset comprises 82K images for training and 40K images for validation. Each image is annotated with five descriptions. Unlike the CUB-200 and Oxford-102 datasets MS COCO dataset images contain multiple objects and a variety of backgrounds, which will make the text-to-image generation model more comprehensive and dynamic during training.

##### B. Experimental Settings

We used the same hyperparameters as in [18–21]. However we set the batch size as 8 for our training strategy as the sample population is less in the initial phase. Indeed it is generally desirable for better batch generalization to set a larger batch size, such as 64, for Text-to-Image generation models, to capture a bigger portion of the dataset with many different factors in it. diverse data points. This allows a smaller batch size to be effective, as noted in [8]. All the experiments were run on Python 3.10 and PyTorch 2.1.0 on Kaggle with an NVIDIA Tesla P100 GPU 16 GB.

##### C. Models

We selected the models from the pioneering papers in GANs:

- 1) GAN-INT-CLS [22]: This model architecture is an initial frame for generating images from text descriptions. This model forms the basis for evaluating improvements in text-to-image synthesis techniques.

- 2) StackGAN [19]: This architecture produces images that are of better quality through a two-stage generation. Low-resolution, and later, is filtered in the second stage to create high-resolution images. This is done especially to enhance the details for more fidelity of the image.
- 3) AttnGAN [20]: The attention mechanism is proposed for image generation through the refinement iterations. During the process, it significantly enhances the alignment and coherence between the generated image and its textual description.
- 4) VQ-Diffusion [21]: It exploits a vector quantized variational autoencoder combined with a diffusion process to generate high-quality images from text. By learning a discrete latent space and refining the generated images through a denoising diffusion process, it achieves superior image quality and diversity.

#### V. RESULTS AND DISCUSSION

In this paper, we compared four different text-to-image generation models: GAN-INT-CLS, StackGAN, AttnGAN, and VQ-Diffusion, using different data selection methods with respect to their efficiency in generating high-quality images. In particular, we have compared data selection methods such as full-dataset training, random sampling, text-based coreset selection, and image-based coreset selection.

The results as shown in Table I reveal that text-based coreset selection consistently delivers high performance while reducing training time. For example, the GAN-INT-CLS model got a high Inception Score when trained with text-based coreset. This method also helped to lessen training time considerably in comparison to using complete dataset as reference point. So, it exposes choosing coreset based on text is good at improving image quality and making training more efficient. Other GAN models like StackGAN, AttnGAN and VQ-Diffusion also showed similar advancements. This indicates the generalizability of this approach across different architectures.

The coreset selection method demonstrated a clear improvement when compared to random sampling. Over most variances, coreset selection—either based on text or images-based, was superior to random sampling on the basis of Inception Score. It means that the performance of the algorithm in drastically selecting an informative, smaller group of data points is way much better than choosing randomly chosen samples. For instance, StackGAN and AttnGAN models expressed better Inception Scores as they were trained using coreset data. This emphasizes how efficient this method is for making use of limited training resources.

The dataset we selected greatly impacted the model’s performance. Normally, when using the full dataset for training, Inception Scores were highest overall. But this took much longer time. The VQ-Diffusion model showed that just 10% of data with text-based coreset method had more Inception Score than random sampling - and it also significantly cut down on total training hours needed.

TABLE I: Comparison of data selection methods

(a) GAN-INT-CLS on Flowers and Birds dataset

Data Selection Method	Flowers (Inception Score / Time for 200 Epochs)	Birds (Inception Score / Time for 200 Epochs)
Full dataset	2.238 / 17.2 hours	4.179 / 21.1 hours
10% random data	1.523 / 2.1 hours	2.816 / 2.8 hours
10% text-based coreset	1.827 / 2.2 hours	3.239 / 3.4 hours
10% image-based coreset	1.502 / 2.4 hours	2.824 / 3.7 hours
10% image and text-based coreset	1.542 / 2.5 hours	2.974 / 3.8 hours

(b) StackGAN on Flowers dataset

Data Selection Method	Inception Score / Time taken for 200 Epochs
Full dataset	4.284 / 32.6 hours
10% random data	3.191 / 3.9 hours
10% text-based coreset	3.713 / 4.1 hours
10% image-based coreset	3.201 / 4.3 hours
10% image and text-based coreset	3.293 / 4.4 hours

(c) AttnGAN on Flowers dataset

Data Selection Method	Inception Score / Time taken for 200 Epochs
Full dataset	8.103 / 38.3 hours
10% random data	4.958 / 4.3 hours
10% text-based coreset	5.825 / 4.6 hours
10% image-based coreset (DAMSM)	4.921 / 4.9 hours

(d) VQ-Diffusion-S on Flowers dataset

Data Selection Method	Inception Score / Time taken for 100 Epochs
Full dataset	35.195 / 9.2 hours
10% random data	23.936 / 0.7 hours
10% text-based coreset	27.692 / 0.7 hours
10% image-based coreset (VQ-VAE)	24.914 / 1.2 hours

By concentrating on the Flowers and Birds datasets, we were able to carry out experiments that were easier to handle and understand, giving us clear understanding on how well certain methods for selecting data work.

Moreover, we made the decision to not test the Birds dataset with models other than GAN-INT-CLS. This choice was guided by our result that the text-based coreset method worked well in both Flowers and Birds datasets for GAN-INT-CLS. Therefore, conducting additional tests with Birds dataset for other models could be seen as unnecessary since core findings have already shown superiority of text-based coreset approach.

The results for each dataset using combinations of active and curriculum learning methods along with the full dataset as shown in Table II, give interesting observations about how effective it is to combine curriculum learning, coreset selection and active learning methods with the full dataset across different GAN models. We have tried these experiments on three datasets: Oxford-102 Flowers, Caltech-UCSD Birds and MS COCO.

Regarding the improvement in Inception Score for the Oxford-102 Flowers dataset, we observe that adding curriculum learning, coreset selection and active learning together resulted in clear enhancements. GAN-INT-CLS showed a significant boost with this combination - its score of Inception

was 2.5357 compared to just 2.2391 when using curriculum learning plus coreset selection alone; this shows how much active learning can help improve both image quality as well as efficiency by reducing data requirement (GAN-INT-CLS). Likewise, StackGAN also got a lot of advantage from combined method with the best Inception Score of 4.3924 and less training time. The largest improvement was seen in AttnGAN where the Inception Score rose from 8.1034 to 8.4361 showing that advanced training plan improved its capacity for creating top-level pictures better. VQ-Diffusion-S also showed greater Inception Score, indicating the method's performance on various models.

For the Caltech-UCSD Birds dataset, a significant enhancement was achieved by using curriculum learning, coreset selection and active learning. The improvement was especially noticeable for GAN-INT-CLS and StackGAN models. The Inception Score of GAN-INT-CLS went up from 4.1793 to 4.7924 with this method combination while StackGAN reached its highest score at 5.3683 percentiles. Even though AttnGAN showed only a small boost in performance, the joint approach still played a part towards making training more effective. VQ-Diffusion-S showed notable progress in its Inception Score, increasing from 48.5793 to 50.3960. This means that the joined approaches work well when dealing with big sets of data too.

TABLE II: Results for each dataset using different training strategies

(a) Flowers Dataset (IS / Time (Hrs) / #Data Points)

Model	Full Dataset	Curriculum + Coreset	Curriculum + Coreset + Active Learning
GAN-INT-CLS	2.2379 / 17.4 / 7,169	2.2391 / 3.2 / 2,545	2.5357 / 4.2 / 2,381
StackGAN	4.2837 / 32.3 / 7,169	4.3483 / 4.6 / 2,545	4.3924 / 5.1 / 2,357
AttnGAN	8.1034 / 38.7 / 7,169	6.8341 / 4.5 / 2,545	8.4361 / 7.6 / 2,320
VQ-Diffusion-S	35.1954 / 9.3 / 7,169	34.0157 / 1.7 / 2,545	35.753 / 2.2 / 2,034

(b) Birds Dataset (IS / Time (Hrs) / #Data Points)

Model	Full Dataset	Curriculum + Coreset	Curriculum + Coreset + Active Learning
GAN-INT-CLS	4.1793 / 21.5 / 8,855	3.6461 / 3.4 / 3,713	4.7924 / 5.6 / 3,292
StackGAN	4.7422 / 34.6 / 8,855	4.2368 / 4.3 / 3,713	5.3683 / 5.7 / 3,427
AttnGAN	8.2862 / 41.6 / 8,855	6.8452 / 5.4 / 3,713	8.2839 / 7.4 / 3,266
VQ-Diffusion-S	48.5793 / 12.6 / 8,855	45.5346 / 2.3 / 3,713	50.3960 / 4.4 / 3,278

(c) MS COCO Dataset (IS / Time (Hrs) / #Data Points)

Model	Full Dataset	Curriculum + Coreset + Active Learning
AttnGAN	26.1685 / 302.4 / 82,783	25.8941 / 54.6 / 28,143
VQ-Diffusion-S	79.2398 / 81.1 / 82,783	81.4546 / 29.6 / 29,643

We didn't carry out evaluations on COCO dataset for GAN-INT-CLS and StackGAN because of the massive amount of computation needed by such a large dataset. Our previous experiments with Flowers and Birds datasets demonstrated that combining curriculum learning along with coreset selection and active learning enhanced Inception Scores greatly, cutting down on training duration considerably. For this reason, we concentrated our testing efforts on these methods with models that had already shown excellent performance when working with smaller datasets. In the case of AttnGAN, the Inception Score decreased a little from 26.1685 when using full dataset to 25.8941 with combined method. But this came with significant reductions in training time and data needs. VQ-Diffusion-S: This method showed a small improvement in Inception Score, rising from 79.2398 to 81.4546. It proves that combining the strategy can improve performance even with big datasets.

The table presented in Table III the results from different GAN models: GAN-INT-CLS, StackGAN, AttnGAN and VQ-Diffusion. We compare their performance under two separate training methods - one on full dataset and another using our suggested method with coreset selection, curriculum learning plus active learning. The rows in this table have captions that describe distinct bird species while the cells show images created by these models.

Observations for the GAN-INT-CLS model are that the images created by it, when we use the full dataset, show clear pictures of described birds. For example, it can be seen in a image "a large and gray bird, and a yellow curved beak" which has good visual coherence matching with its caption. The images generated by this model using our way of training have similar quality to those created from complete dataset; there is no big loss in detail or accuracy when showing bird descriptions visually.

StackGAN creates good images, full of details and colors

when it's trained on the complete dataset. For instance, the image for "a large white and gray bird with a long yellow beak" is very clear and detailed. With our method, we maintain this quality in generated images. The bird's features are clear and the colors still strong, showing that our training method is working well.

































AttnGAN does an outstanding job, especially with intricate explanations such as "it is a bird that appears completely black; it has a small head in proportion to its body and possesses a small black bill which is slightly curved downward". The pictures are clear, and the feature representation improvement by attention mechanism is evident. When trained using our method, the images are almost identical to those created from full dataset. The attention mechanism keeps on functioning well, showing that our training approach does not affect the model's capacity to grasp intricate characteristics.

When VQ-Diffusion is trained with the full dataset, it exhibits excellent performance in producing high-resolution images. The image for "a brown bird with a black rectrices, the bill is black and curved, the head is small compared to its" shows great detail and realism. Our method also produces similar quality of images as those made from complete dataset. The high-resolution and detail is maintained, showing that our coreset selection, curriculum learning, and active learning strategy are successful in preserving the model's performance.

The outcomes show that our training plan, which merges coreset selection, curriculum learning and active learning is very successful. The pictures made by models trained using our method are almost identical to those trained on all data sets - this means it works in lowering necessary data size but not lowering image quality.

Coreset selection is useful for selecting a smaller group of data that represents the bigger set, making sure those important features needed to train are maintained. Curriculum learning, by step-by-step making the training examples more complex,

TABLE III: Comparison of generated images from different models

Caption	GAN-INT-CLS		StackGAN		AttnGAN		VQ-Diffusion	
	Full	Ours	Full	Ours	Full	Ours	Full	Ours
a large and gray bird, and a yellow curved beak								
a large white and gray bird with a long yellow beak								
a completely black bird, it has a small head for its body, and small black bill with a downward curve								
a brown bird with a black rectrices, the bill is black and curved, the head is small compared to its								

helps models learn better and perform well on complex tasks like generating images from scratch. Active learning makes certain that the most informative examples are chosen for training, improving the efficiency of model's learning process.

In general, our method lessens the amount of computational resources and time needed for training. At the same time, it sustains or improves image quality in most cases. This technique can be quite useful in training GANs when there's a lot of data available but not many computational resources.

## VI. CONCLUSION

This study proves that through having a schedule of data preprocessing, the training of text-to-image generation models can be made more efficient. With the help of Curriculum Learning and Active Learning methods we have reinforced the data selection process and minimized the training times. Also, that process enhanced the quality of the produced images. Therefore, the meta-analysis of our study across the Oxford-102, CUB-200, and MS COCO databases establishes that the identified techniques are equally efficient and robust in various and challenging data settings.

In the experimental results, training time has been cut and there is a general improvement of the images quality in most of the cases. This means that not only are our methods beneficial in terms of training in terms of time, but they also do not negatively affect, or in some cases positively affect, the quality of the models. These observations seem to represent the essence of what data-oriented approaches are capable of achieving to the overall generative AI field, in the way that will open new horizons for its practical and widespread application.

More complex work can involve applying these techniques to other generative models and other data-centric ways of improving the training methodologies. Thus, the further development of the generative AI will require the implementation of such innovative solutions that will remove current restrictions and expand a potential of the text-to-image generation in the future.

## REFERENCES

- [1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021. [Online]. Available: <https://arxiv.org/abs/2102.12092>
- [2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022. [Online]. Available: <https://arxiv.org/abs/2205.11487>
- [3] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," 2021. [Online]. Available: <https://arxiv.org/abs/2103.14749>
- [4] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach," in *Conference on Empirical Methods in Natural Language Processing*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:202537756>
- [5] Y. Tan, C.-P. Lee, M. Neo, K. Lim, and J. Lim, "Enhanced text-to-image synthesis with self-supervision," *IEEE Access*, vol. PP, pp. 1–1, 01 2023.
- [6] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "Lafite: Towards language-free training for text-to-image generation," 2022. [Online]. Available: <https://arxiv.org/abs/2111.13792>
- [7] Z. Zhang, Y. Zhang, W. Yu, J. Lu, L. Nie, G. He, N. Jiang, Y. Fan, and Z. Yang, *Text to Image Synthesis Based on Multiple Discrimination*, 09 2019, pp. 578–589.
- [8] S. Sinha, H. Zhang, A. Goyal, Y. Bengio, H. Larochelle, and A. Odena, "Small-gan: Speeding up gan training using core-sets," 2019. [Online]. Available: <https://arxiv.org/abs/1910.13540>
- [9] P. Soviany, C. Ardei, R. T. Ionescu, and M. Leordeanu, "Image difficulty curriculum for generative adversarial networks (cugan)," 2019. [Online].

Available: <https://arxiv.org/abs/1910.08967>

- [10] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum learning: A survey," *International Journal of Computer Vision*, vol. 130, pp. 1526 – 1565, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:231709290>
- [11] M. Grootendorst, "Keybert: Minimal keyword extraction with bert." 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.4461265>
- [12] B. Mirzasoleiman, J. A. Bilmes, and J. Leskovec, "Data sketching for faster training of machine learning models," *ArXiv*, vol. abs/1906.01827, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:174799238>
- [13] Y. Li, Y. Shen, and L. Chen, "Camel: Managing data for efficient stream learning," in *Proceedings of the 2022 International Conference on Management of Data*, ser. SIGMOD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1271–1285. [Online]. Available: <https://doi.org/10.1145/3514221.3517836>
- [14] P. Ren, Y. Xiao, X. Chang, P.-Y. B. Huang, Z. Li, X. Chen, and X. Wang, "A survey of deep active learning," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1 – 40, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:221397441>
- [15] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.
- [16] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.
- [17] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014. [Online]. Available: <https://api.semanticscholar.org/CorpusID:14113767>
- [18] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016. [Online]. Available: <https://arxiv.org/abs/1605.05396>
- [19] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5908–5916, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:1277217>
- [20] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:8858625>
- [21] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," 2022. [Online]. Available: <https://arxiv.org/abs/2111.14822>
- [22] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:1563370>