Text to Image Generation using Stable Diffusion Model on Medical Images

Shubham Vatsal [1] **Rahul Meghwal** [1] **Harshit Mehta** [1]

Abstract

Text-to-image generation is a challenging problem in the field of artificial intelligence, as it involves generating visual data from linguistic input. In this paper, we investigate the use of the latest stable diffusion model for text-to-image generation. Stable diffusion models are a variant of diffusion models that have been shown to be more effective at generating high-quality images from text descriptions. In this work, we fine-tune a stable diffusion model on medical and multimodal imaging Radiology Objects in COntext (ROCO) dataset and generate images corresponding to their mapped textual inputs. Once the model has been fine-tuned, we evaluate its performance on a downstream classification task, manually analyze the images as well as calculate Frechet Inception Distance (FID). The discussed work apart from providing insight into the capabilities of the used stable diffusion model for text-to-image generation also publishes base pipeline code which can be easily used to fine-tune any other dataset.

1. Introduction

It is practically impossible to convince an astronaut to ride a horse in his or her space suit and then take their picture or it is impossible to make one's favourite basketball player ride a dinosaur. Is there a way we can visually perceive these imaginary pictures. Of course there have been many video editing tools from ages but it requires a lot of effort even using those tools to create such an unreal image. It is challenging to create such fictitious scenes, which requires synthesizing examples of specific subjects in novel settings such that they organically and flawlessly fit into the scene. But large deep learning based generative models successfully handle this challenge.

Figure 1. Quality of Images Generated Over the Years

The improvement of quality of generated images rendered by various generative models over the years can be seen in Fig 1.

Text-to-image generation can be defined as construction of images according to natural language descriptions in an automated environment. Descriptive language phrases are an intuitive and flexible way to describe visual notions for creating images when compared to other types of input such as sketches, audio and object masks. Learning from unstructured description and managing the differing statistical features between vision and language inputs are the primary challenges for text-to-image synthesis. It is a fundamental problem in many applications, such as art generation and multimedia content creation. Text-to-image models offer unprecedented freedom to guide creation through natural language. It has also made multimodal domain across vision and language as one of the most actively researched areas (Reed et al., 2016b;a; Zhu et al., 2018; Xu et al., 2015; Gan et al., 2017; Antol et al., 2015; Yang et al., 2016; Zhang & Dana, 2018; Zhang et al., 2018a).

Medical images are fundamentally different from natural images. It is quite easy for human beings to understand an image containing a bird or a car but it requires a lot of expertise to identify if there is any medical condition by looking at a chest x-ray. The language used to record pertinent information in medical data is distinct, with a limited but semantically rich vocabulary. Unsurprisingly, multi-modal models do not typically generalize well to the medical domain when trained on real image-text pairs. The lack of high-quality, annotated medical imaging datasets could be lessened by creating generative imaging models that accurately reflect medical concepts while offering compositional variation. With more medical imaging data, we can solve a number of issues currently being faced by the medical industry. The

[1] Courant Institute of Mathematical Science, New York University, New York, USA. Correspondence to: Shubham Vatsal <sv2128@nyu.edu>.

Proceedings of the 39^{th} *International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

machine learning models being presently used in the medical domain can be improved by retraining them with more new generated data. New data can lead to better diagnosis of rare medical condition and can compliment the doctors in making more accurate clinical decisions. In this work, we experiment with such automated generation of medical data conditioned on their natural language descriptions.

The rest of the paper is organised in the following way. Section 2 talks about the related work. We illustrate the working of the used stable diffusion model in section 3. Section 4 concentrates on the experiments we conducted and the corresponding results we achieved. The final section gives a summary of this paper.

2. Related Work

There have been many work done in the field of image generation conditioned on various factors using different kinds of generative models. (Zhu et al., 2017) talks about the image-to-image generation by modelling a distribution of possible outputs in a conditional generative modeling setting. They propose that the ambiguity of the mapping is distilled in a low-dimensional latent vector, which can be randomly sampled at test time. (Yan et al., 2016) investigates generation of images from visual attributes. The authors model an image as a composite of foreground and background and develop a layered generative model with disentangled latent variables that can be learned end-to-end using a variational auto-encoder. (Taigman et al., 2016) generates emojis from images using their proposed Domain Transfer Network which employs a compound loss function that includes a multi-class Generative Adversarial Network (GAN) loss, an f-constancy component, and a regularizing component. (Chen et al., 2017) uses conditional generative adversarial networks to generate images from audio. (Li et al., 2019a) introduces a word-level spatial and channel-wise attention-driven generator that can disentangle different visual attributes and allow the model to focus on generating and manipulating sub-regions corresponding to the most relevant words. They further discuss a word-level discriminator to provide fine-grained supervisory feedback by correlating words with image regions, facilitating training an effective generator which is able to manipulate specific visual attributes without affecting the generation of other content.

A few years back, GANs (Goodfellow et al., 2020) appeared as one of the most promising approaches for generative modeling. (Bao et al., 2017) presents variational generative adversarial networks, a general learning framework that combines a variational auto-encoder with a generative adversarial network, for synthesizing images in fine-grained categories, such as faces of a specific person or objects in a category. (Zhang et al., 2022) proposes double attention which simultaneously leverages the context of the local and the shifted windows, leading to improved generation quality. Lifelong GAN (Zhai et al., 2019) employs knowledge distillation to transfer learned knowledge from previous networks to the new network. This makes it possible for them to perform image-conditioned generation tasks in a lifelong learning setting.(Liao et al., 2022) talks about a novel framework Semantic-Spatial Aware GAN for synthesizing images from input text. The authors introduce a simple and effective Semantic-Spatial Aware block which learns semantic-adaptive transformation conditioned on text to effectively fuse text features and image features and learns a semantic mask in a weakly-supervised way that depends on the current text-image fusion process in order to guide the transformation spatially.

Diffusion models have become quite popular nowadays for beating GANs at the task of generative modelling. (Ho et al., 2022) proposes cascaded diffusion model comprising of a pipeline of multiple diffusion models that generate images of increasing resolution, beginning with a standard diffusion model at the lowest resolution, followed by one or more super-resolution diffusion models that successively upsample the image and add higher resolution details. (Saharia et al., 2022) discusses their simple implementation of image-to-image diffusion model and claim it outperforms strong GAN and regression baselines on all tasks, without task-specific hyper-parameter tuning, architecture customization, or any auxiliary loss or sophisticated new techniques needed. The authors further uncover the impact of an L2 vs. L1 loss in the denoising diffusion objective on sample diversity, and demonstrate the importance of self-attention in the neural architecture through empirical studies. (Gu et al., 2022) presents the vector quantized diffusion (VQ-Diffusion) model for text-to-image generation based on a vector quantized variational autoencoder (VQ-VAE) whose latent space is modeled by a conditional variant of Denoising Diffusion Probabilistic Model (DDPM). (Ruiz et al., 2022) leverages the semantic prior embedded in the model with a new autogenous class-specific prior preservation loss which enables synthesizing the subject in diverse scenes, poses, views, and lighting conditions that do not appear in the reference images. (Rombach et al., 2022) achieves state-of-the-art synthesis results on image data. It introduces cross-attention layers into the model architecture which allows diffusion models to become flexible generators for general conditioning inputs such as text or bounding boxes allowing high resolution synthesis in a convolutional manner.

There have been several prior work published on text-to-image generation using GANs and diffusion models. GANs came up with promising results on text-to-image generation (Zhang et al., 2017; 2018b). AttnGAN (Xu et al., 2018) proposes a multi-stage refinement framework to generate

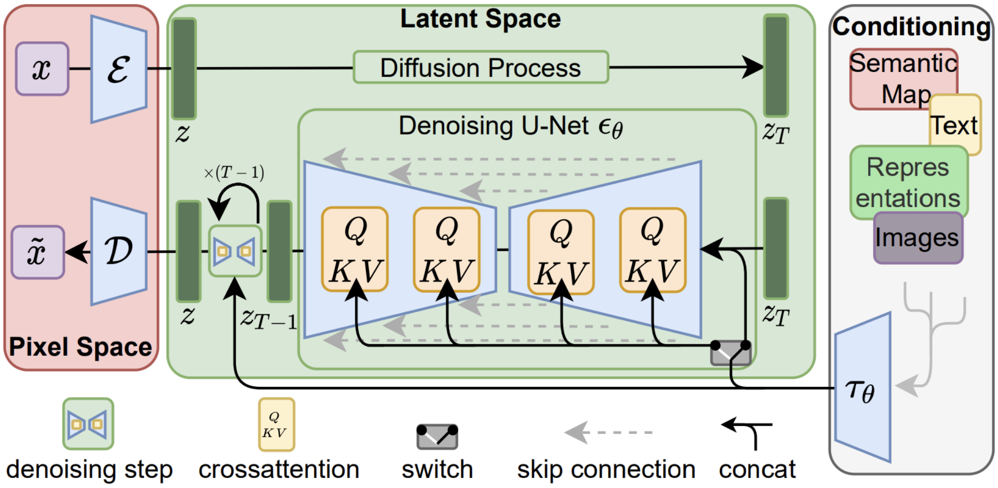

Figure 2. Stable Diffusion

fine-grained details by attending to relevant words in the description. Recent GAN-based approaches (Li et al., 2019b; Koh et al., 2021) propose object-driven, hierarchical approaches that explicitly model object instances within an image. (Balaji et al., 2022) trains an ensemble of text-to-image diffusion models specialized for different synthesis stages. (Nichol et al., 2021) explores diffusion models for the problem of text-conditional image synthesis and compare two different guidance strategies: CLIP guidance and classifier-free guidance.

Within the medical domain also, there has been substantial amount of image generation work. (Han et al., 2018) focuses on generating synthetic multi-sequence brain Magnetic Resonance (MR) images using GANs. (Frid-Adar et al., 2018) exploits GAN architectures for synthesizing high quality liver lesion ROIs. The work further discusses a novel scheme for liver lesion classification using CNN. (Iqbal & Ali, 2018) proposes a new GAN for Medical Imaging (MI-GAN) which generates synthetic medical images and their segmented masks, which can then be used for the application of supervised analysis of medical images. Particularly, the paper presents MI-GAN for synthesis of retinal images. (Guan et al., 2022) talks about a medical image augmentation method, namely, a texture-constrained multichannel progressive generative adversarial network (TMP-GAN). TMP-GAN uses joint training of multiple channels and an adversarial learning-based texture discrimination loss to further improve the fidelity of the synthesized images.

But there has been limited research revolving around to text-to-image generation for medical images.(Zeng et al., 2022) showcases a method based on attention mechanism and content preservation loss to improve image quality. Their model consists of three stages, and each stage generates feature map of different scale combined with attention features as the input of the next stage to optimise semantic consistency between image and text. (Chambon et al., 2022) develops a strategy to overcome the large natural-medical distributional shift by adapting a pre-trained latent diffusion model on a corpus of publicly available chest x-rays (CXR) and their corresponding radiology (text) reports. The paper further investigates the model's ability to generate high-fidelity, diverse synthetic CXR conditioned on text prompts. Along similar lines, in this work, we experiment with fine-tuning of a stable diffusion model for image generation conditioned on their corresponding text prompts.

3. Model Architecture & Datasets

In this section, we talk about the architecture of the stable diffusion model (Rombach et al., 2022) which we use for all our experiments. We further discuss the dataset which we use to conduct our experiments.

3.1. Model

A variation of the diffusion model (DM) known as the latent diffusion model (LDM) is used by stable diffusion. Diffusion models, which were first used in 2015, are trained with the goal of eradicating repeated applications of Gaussian noise on training images, which can be compared to a series of denoising autoencoders. The variational autoencoder (VAE), U-Net, and an optional text encoder make up stable diffusion. The VAE encoder condenses the image from pixel space to a more basic semantic meaning in a less dimensional latent space. Forward diffusion involves repeatedly applying Gaussian noise to the compressed latent representation. The output of forward diffusion is denoised backwards

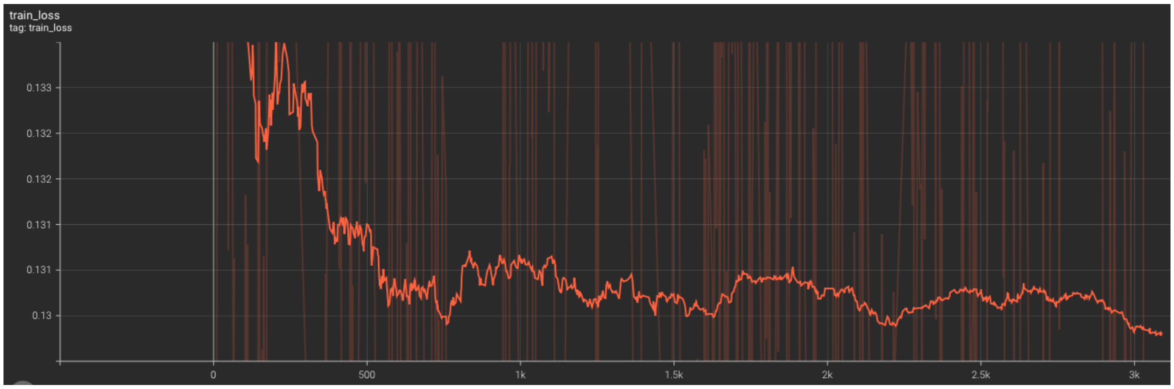

Figure 3. Training Loss Vs Epoch Graph

by the U-Net block, which is built from a ResNet backbone, to produce latent representation. The VAE decoder then transforms the representation back into pixel space to create the finished image. Flexible conditions can be applied to the denoising step based on a list of text, a picture, and other modalities. Through a cross-attention method, the encoded conditioning data is made available to denoising U-Nets. The fixed, pretrained CLIP ViT-L/14 text encoder converts text prompts to an embedding space for text conditioning. As a benefit of LDMs, researchers highlight to greater computing efficiency for training and generation. Fig 2 shows the visual depiction of conditioned LDMs either via concatenation or by more general cross-attention mechanism (Rombach et al., 2022). In contrast to a diffusion model a Latent Diffusion model works in latent space rather than in pixel space to minimize the loss in equation 1 which aims to learn the denoising conditioned on other inputs(text inputs for our problem). $\epsilon_\theta\left(z_t, t\right)$ is the donoising that needs to be learnt for a given step t and latent representation z_t at step t.

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon-\epsilon_\theta\left(z_t, t\right)\right\|_2^2\right] \qquad (1)$$

3.2. Datasets

We use the Radiology Objects in Context (ROCO) dataset (Pelka et al., 2018), which has over 81k radiology images from various medical imaging modalities, including computer tomography, ultrasound, x-ray, fluoroscopy, positron emission tomography, mammography, magnetic resonance imaging, and angiography. In ROCO, each image has a related description, keywords, UMLS Concept Unique Identifiers, and Semantic Type. To enhance prediction and classification performance, a series of 6k images, ranging from digital artwork to synthetic radiology figures, is made available outside of the classroom. It is possible to describe caption and keyword generation systems using ROCO, enabling multimodal representation for datasets lacking text representation. ROCO can be used to build systems with the objectives of semantic information tagging and picture structuring, which is advantageous and helpful for image and information retrieval. The test split for ROCO has 8177 images, each image is annotated with a text caption and the related concepts, this helps us in evaluating our model.

Table 1. Stable Diffusion Hyper-Parameters

Hyper-Parameters	**Values**
Epochs	10
Learning Rate	1e-05
Batch Size	16
Mixed Precision	bf16
Resolution	512
LR Scheduler	Constant
Max Grad Norm	1
Gradient Accumulation Steps	8
Loss	Mean Squared Error

4. Experiments

As discussed, we use the latest stable diffusion [1]model for fine-tuning on ROCO. The hyper-parameters used for fine-tuning can be found in Table 1. We do the fine-tuning on the training and validation splits of ROCO while generate images corresponding to the test split captions or text descriptions. The code for all the experiments has been released [2] . We generate both qualitative and quantitative performance results for our experiments. Details of which are discussed in the next section.

5. Results & Evaluation

The qualitative results can be seen in Fig 4. Quantitative results are evaluated using two methods. Firstly, we use a new

[1] https://github.com/CompVis/stable-diffusion

[2] https://github.com/shubham30vatsal/Text-To-Image-Generation

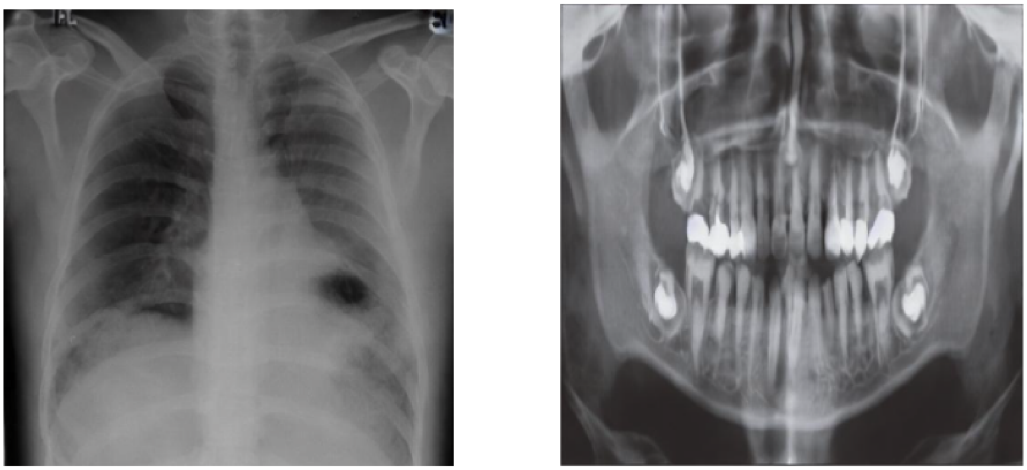

Figure 4. Qualitative Results for ROCO. The left image was generated using text prompt ”Radiograph of the chest revealing bilateral pleural effusion.”. Image on right was generated using text prompt ”Panoramic radiograph showing the absences of right permanent maxillary canine and all third molars.”

method inspired from GAN Test (Shmelkov et al., 2018) which aims to measure the semantic differences between the generated and test images using a multi-label classifier. Then we use FiD score(Seitzer, 2020) to measure the differences in the distribution of test images and generated images.

5.1. GAN Test

The concepts associated with each image in ROCO dataset seem to capture a semantic representation of the image, this led us to use the concept IDs for evaluating the diffusion model's performance. Inspired from GAN Test (Shmelkov et al., 2018), we perform a multi-label classification on the ROCO dataset to evaluate diffusion model's performance. We train a multi-label classification model on ROCO Dataset using only the top 50 most frequent UMLS Concept Unique Identifiers as labels for an image. We use modified DenseNet121 (Huang et al., 2017) (added one extra linear hidden layer with 256 units for the classifier sub-network) architecture for the multi-label classifier, hereafter just called as DenseNet. The classifier is trained on the ROCO training split and its performance is evaluated using the test split. Once trained we generate UMLS concept IDs for the images generated corresponding to the textual inputs from test split. The GAN Test results have been listed out in Table 3. The first two columns represent the ground truth labels and the predicted labels respectively. The next two columns give the weighted F1 and micro F1 scores. The first row in Table 3 gives us the performance of our modified DenseNet, where we compare predictions (Y Pred) from the classifier on ROCO test set images with the corresponding ROCO Test set concept IDs (Y True). The second row shows the performance of the model on the generated image set where we compare the predictions (Y Pred) on generated image (the images are generated using ROCO Test set captions) with the corresponding ROCO test set image's concept IDs. The third row shows results by comparing the predictions on ROCO Test images (Y True) with the predictions on generated image set (generated using the caption of ROCO test image). We believe this comparison overlooks the classification model's performance (first row of table 3) since here both **Y True** and **Y Pred** are predictions from the classifier i.e. they both come from a common understanding of image semantics.

Table 2. FID

Splits	FID
Train vs Test	4.1812
Train vs Generated Test	**23.3657**
Test vs Generated Test	**26.2831**

5.2. FiD Score

The results based on FID can be seen in Table 2. The lower the FID, the better are the quality of images generated. The first column in the table indicates the splits across which the FID is being measured. The low FID when one of the splits is our generated test set suggests that fine-tuning of the stable diffusion model on ROCO does a good job in generating images with high fidelity.

6. Conclusion

This paper goes through the recent development of diffusion models in the literature and specifically investigates the use of the state-of-the-art stable diffusion model for text-

Table 3. GAN Test

Y True	Y Pred	Weighted F1 Score	Micro F1 Score
Test	DenseNet (Test)	0.2788	0.3569
Test	**DenseNet (Generated Test)**	**0.2745**	**0.3621**
DenseNet (Test)	**DenseNet (Generated Test)**	**0.5214**	**0.5126**

to-image generation in the medical domain. Based on the mapped textual inputs, we fine-tune the model and generate high quality images for ROCO dataset. We evaluate the quality of generated images using the traditional FID and GAN Test downstream classification metrics. As we can see from both the qualitative and quantitative results, stable diffusion produces high quality images with respect to the dataset it is being fine-tuned even in case of the medical domain. FID showcases the efficiency of this new stable diffusion model. Moreover, the GAN Test shows that the performance of the modified DenseNet being similar on both the ground truth as well as the generated test set which further provides strong evidence about the capabilities of stable diffusion.

Acknowledgements

The authors would like to thank Professor Rajesh Ranganath and Adriel Saporta for their regular inputs and encouraging us to pursue this problem statement.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Bao, J., Chen, D., Wen, F., Li, H., and Hua, G. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pp. 2745–2754, 2017.

Chambon, P., Bluethgen, C., Delbrouck, J.-B., Van der Sluijs, R., Połacin, M., Chaves, J. M. Z., Abraham, T. M., Purohit, S., Langlotz, C. P., and Chaudhari, A. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

Chen, L., Srivastava, S., Duan, Z., and Xu, C. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 349–357, 2017.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., and Deng, L. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5630–5639, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

Guan, Q., Chen, Y., Wei, Z., Heidari, A. A., Hu, H., Yang, X.-H., Zheng, J., Zhou, Q., Chen, H., and Chen, F. Medical image augmentation for lesion detection using a texture-constrained multichannel progressive gan. *Computers in Biology and Medicine*, 145:105444, 2022.

Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., and Nakayama, H. Gan-based synthetic brain mr image generation. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 734–738. IEEE, 2018.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Iqbal, T. and Ali, H. Generative adversarial network for medical images (mi-gan). *Journal of medical systems*, 42 (11):1–11, 2018.

Koh, J. Y., Baldridge, J., Lee, H., and Yang, Y. Text-to-image generation grounded by fine-grained user attention.

In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 237–246, 2021.

Li, B., Qi, X., Lukasiewicz, T., and Torr, P. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019a.

Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., and Gao, J. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12174–12182, 2019b.

Liao, W., Hu, K., Yang, M. Y., and Rosenhahn, B. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18187–18196, 2022.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Pelka, O., Koitka, S., Rückert, J., Nensa, F., and Friedrich, C. M. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pp. 180–189. Springer, 2018.

Reed, S., Akata, Z., Lee, H., and Schiele, B. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 49–58, 2016a.

Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., and Lee, H. Learning what and where to draw. *Advances in neural information processing systems*, 29, 2016b.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.

Seitzer, M. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.2.1.

Shmelkov, K., Schmid, C., and Alahari, K. How good is my gan? In *Proceedings of the European conference on computer vision (ECCV)*, pp. 213–229, 2018.

Taigman, Y., Polyak, A., and Wolf, L. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018.

Yan, X., Yang, J., Sohn, K., and Lee, H. Attribute2image: Conditional image generation from visual attributes. In *European conference on computer vision*, pp. 776–791. Springer, 2016.

Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016.

Zeng, X., Huang, Z., Xu, L., and Xie, Y. Cp-gan: Meet the high requirements of diagnose report to medical image by content preservation. *IET Image Processing*, 16(1): 29–38, 2022.

Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M., and Mori, G. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2759–2768, 2019.

Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., and Guo, B. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11304–11314, 2022.

Zhang, H. and Dana, K. Multi-style generative network for real-time transfer. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., and Agrawal, A. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, 2018a.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018b.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.

Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., and Elgammal, A. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1004–1013, 2018.