DOI: http://dx.doi.org/10.26483/ijarcs.v16i1.7191

Volume 16, No. 1, January-February 2025

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

DESIGN OF AUGMENTED DIFFUSION MODEL FOR TEXT-TO-IMAGE REPRESENTATION

Ms. Subuhi Kashif Ansari
Research Scholar, School of Engineering & Technology,
Shri Venkateshwara University,
Gajraula, U.P., India.

Rakesh Kumar
Assistant Professor, School of Science and Technology,
Shri Venkateshwara University,
Gajraula, U.P., India.

Abstract: The exciting as well as demanding challenge of generating images from natural language commands is no small feat. This research provides a new architectural paradigm for Text to Images (T2I) approaches and demonstrate that a well-designed neural architecture may reach state-of-the-art (SOTA) presentation with only one generator and one discriminator, all in one step of training. This paper concludes with a call to action for researchers in the arena of T2I conversion, which has just begun to explore novel neural architectures. This work uses a Contrastive Language-Image Pretraining (CLIP) + Image approach to T2I generation, which optimises in the hidden space of a regular Generative Adversarial Network (GAN) to identify images with the highest possible semantic significance score given the input score as determined by the CLIP model. The CLIP+GAN technique is zero-shot, flexible, and doesn't need training, in contrast to conventional approaches that start from the beginning when training generative models to map T2I. With these essential methods, this research proposes an enhanced GAN that enhances the CLIP + GAN approach in this work. A CLIP score is made in this research which is more robust. Finally, this research can produce superior images with diverse objects, creative styles, backdrops as well as original counterfactual notions that do not occur in the training data of the GAN that this work utilise when supported by various input text. For the dataset created, the proposed methodology output images achieve top-level Inception Score (IS) as well as Frechet Inception Distance (FID) scores quantitatively, all without any further architectural design or training.

Keywords: T2I synthesis, Generative models, Multimodal learning, Dataset, IS, FID

1. INTRODUCTION

Using a text input, T2I generative models may generate several high-quality images representing concepts; these models have shown remarkable capability in picture synthesis, image translation, etc. [1]. It is possible to combine many different concepts into new arrangements and settings using T2I diffusion models (DM). Thoughts, unusual combinations, or structured concepts like hand palms remain elusive to them [2]. This paper talks about ODISE, a panoptic segmentation method that uses an open vocabulary and combines text-image DM that have already been trained with discriminative models [3]. Re-Imagen is a generative picture model that practices recovered data to yield lifelike images of any organism, no matter how rare or mysterious. Whenever a user inputs text, Re-Imagen searches an external multi-modal knowledge base for suitable (image, text) pairs. It then utilizes these references to create an image. Re-Imagen learns more about the entities described above's low-level visual features and high-level semantics via this retrieval procedure. Therefore, the creatures' visual appearance is more accurate [4].

A motion-generating framework called ReMoDiffuse has been developed. It is based on a DM and uses a retrieval mechanism to improve denoising. With its three primary layouts, ReMoDiffuse expands the adaptability and diversity of text-driven motion creation. To begin with, hybrid retrieval uses kinematic and semantic similarity to find relevant database references. (2) The semantic-modulated transformer fixes the difference between the planned motion sequence along with the retrieved samples by selectively taking in retrieval information. 3) By using the retrieval database more extensively during inference, Condition Mixture circumvents the scale sensitivity in classifier-free guiding [5]. This T2I DM has an extraordinary degree of photorealism as well as deep understanding of language. Imagen draws on the efficiency of big-transform language models for text accepting along with the power of DM for high-fidelity picture synthesis. Improving sample fidelity as well as image-text alignment is significantly more achieved by expanding the language model in Imagen compared to expanding the image DM [6]. Overall, the results show that wide-ranging big language models, such as T5, which are pretrained on text-only corpora, are amazingly effective at encoding text for image synthesis. It can get a better idea of how people and things interact by using a CLIP model to improve the pre-trained detector's human and object representations and an adapter-style tuning approach to get different representations from a frozen DM that are semantically related. To fix the problems with current HOI datasets, it suggests SynHOI, a fake dataset that is big, varied, and fair in terms of class; it has more than 140K HOI photos marked up with triplets [7]. Because of the limits of verbal explanation, written prompts alone may struggle to capture some aspects of artwork (including colour tone, composition, or brushwork). This is achieved via the provision of DreamStyler, a cutting-edge framework for creative picture synthesis that can integrate T2I synthesis and style transfer. By combining a multistage textual embedding with a context-aware text prompt, DreamStyler is able to generate images of exceptional quality [8]. Discussing the most fundamental visual depiction of human drawing using AIGC is a "backward" step. This work think creative sketching is best done in groups and place an emphasis on the originality of the drawings. Additionally, it solves the issue of "I can't sketch." It lets text dictate sketch ideas, allowing for an unfettered definition of uniqueness. Using a text-conditioned DM trained on picture pixel representations, a method for producing controlled drawings is detailed. The proposed technique, known as SketchDreamer, includes a differentiable rasterizer of Bézier curves that optimises an initial input in order to extract abstract semantic information from a pretrained DM [9].

Surveys will be used for the remaining tasks. Part II offers a synopsis of a number of recent and ongoing projects. Section III, define the proposed approach. The references are given after the results and analysis summary in Section IV.

2. LITERATURE REVIEW

Trabucco et al. [10] found these new visuals lack variety of significant semantic axes in the data. Existing augmentations are unable to alter the picture's high-level semantic features, such as animal species, in order to expand the data variety. To resolve the issue of data augmentation's lack of variety, parameterize image-to-image transformations using pre-trained T2I DM. Using a pre-made DM, the approach changes the semantics of photos and can generalize to new visual notions from a small number of tagged examples.

Zhong et al. [11], created the "Contrast-augmented DM with Fine-grained Sequence Alignment" (FSA-CDM) model, uses negative as well as positive samples that are different from each other inside the DM to make markup-to-image production work better. Practically speaking, in order to build trustworthy feature representations, you should provide a fine-grained cross-modal alignment module that looks at the arrangement comparison between the two modalities in depth.

Lu et al. [12] described to achieve their aim of learning an unseen style use only a limited amount of shots (less than 10) to fine-tune a pre-trained DM. The model is then taught to create high-quality photographs of any items in this style. To accomplish such very low-shot fine-tuning, a new toolset of tuning methods is used, including T2I customised data augmentations, content loss to aid content-style disentanglement, as well as sparse updating that focuses on just a few time steps. Diffusion Specialist is compatible with existing DM frameworks and may be readily customised using various methods.

Zhang et al. [13], explained the underlying DM for image synthesis is which stands alone, before moving on to a consideration of how circumstances or instruction improve learning. Based on this, it discusses the most current techniques to text-conditioned image synthesis (also called T2I). Along with that, it comes with two other apps: text-guided creative output and text-guided picture alteration.

Elsharif et al. [14]. detailed a method for improving the depiction of Arabic culture in T2I models via the use of prompt engineering. To accomplish the goal of prompt augmentation, this study provides a humble, method for rapid engineering that takes advantage of the domain knowledge of a SOTA language model, GPT. Using a GPT model and an approach called in-context learning, we can take a simple initial prompt and utilise it to produce several more detailed prompts about Arabic culture from other categories.

Bie et al. [15] described the employment of models that can read text input along with generate fine images from word descriptions as "text-to-image generation" (T2I). The autoregressive transformer and the GAN were crucial in the development of neural networks that could convert T2I. The DM is a well-liked kind of generative model used to produce graphics by meticulously adding noise and iterating through phases. Due to their outstanding results on picture synthesis, DM have been firmly established as the principal image decoder used by T2I models, catapulting T2I generation to the vanguard of machine learning (ML).

Chen et al. [16], created Geo Diffusion, lets pre-trained T2I DM give good detection data by changing different geometry conditions into text prompts on the fly. When compared to previous L2I algorithms, Geo Diffusion has the potential to accommodate more geometric situations, including camera views in a self-driving scenario, together with bounding boxes.

Xiao et al. [17] highlighted a simple method for utilizing the attention mechanism of T2I DM. It is possible to instantaneously achieve sentence-level semantic grounding without retraining or optimising inference time. The technique's performance is compared to current approaches in a poorly supervised semantic segmentation environment utilising Pascal VOC 2012 and Microsoft COCO 014.

Zhong et al. [18] offered a set consists of a fantastic picture, a challenging keyword-based task, and an understandable narrative prompt. To help SUR-adapter improve its reasoning and semantic understanding and create a high-quality textual semantic illustration for T2I generation, it aligns the complex prompts' semantic representation with the narrative prompts' semantic representation and gives it access to large language models (LLMs).

This work makes the following formal contributions:

- This research offers a new sentence interpolation technique that enables the generator to study a flat Conditional Space (CS),
- This research demonstrate how SOTA T2I models are generated using a contemporary residual neural architecture, which allows for one-stage training at the target image scale.
- Finally, this research thoroughly examines the characteristics of T2I models.

3. PROPOSED T2I MODEL

3.1 Pre-processing

As a collection of tokens, text may be seen. "Tokenization" describes the procedure by which these tokens are extracted; a token is analogous to a meaningful fragment of text. It could be a single word or a whole statement. After that, you may use stemming or lemmatization to standardise those tokens. Making use of the recovered tokens as model features is the next step. In a bag of words style, a basic counter feature. A massive vector of counters is used in lieu of each text. In a similar vein, n-grams may be added to maintain local ordering. It really improves the accuracy of text categorization. Typically, efficiency is improved when counters are replaced with Term Frequency - Inverse Document (TF-IDF) values [23].

3.2 Proposed Enhanced GAN (EGAN)

Compared to other well-known approaches, the proposed strategy is quite different. This work greatly streamlines the foundation for converting T2I. In order to streamline and quicken the training process, first abandon the multi-stage design in favour of a single-stage contemporary deep residual network. The second contribution is a sentence interpolation method that enhances results and enables image modification using arithmetic operations in CS. This technique enables the generator to learn a smooth CS, which is crucial for both performance and accuracy. The research concludes by showing that the proposed approach is superior to those that have relied only on the sentence vector. The proposed approach is thoroughly described in this section. Similar to other T2I synthesis approaches, these neural architectures employ multi-stage training using many networks. On the other hand, this option raises the computing costs and training complexity of these models. Based on this design, this work has a completely different approach. By combining an appropriate neural architecture with a straightforward phrase interpolation approach, this research displays that SOTA outcomes are possible. A single discriminator and generator are all that's needed for our method's one-stage training. Next, the neural architecture, sentence interpolation mechanism, and text encoder that make up the proposed solution is described.

This research follows the procedures set forth by Brock et al. [19], who presented BigGANDeep, the cutting-edge architecture for GANs. A more efficient and easier-to-train deep network is achieved by basing this architecture on the residual blocks with the bottleneck organization of He et al. [20]. In the generator and discriminator, BigGAN-Deep uses Spectral Normalisation and Non-local Blocks, similar to SAGAN. With the use of Conditional Batch Normalisation and the projection method of Miyato et al. [21], BigGAN-Deep finally incorporates conditioning information into the generator and discriminator, respectively. Using the ImageNet dataset in a supervised environment, BigGAN-Deep achieved a new SOTA result. Hence, it was made to be dependent on class labels. The architecture is expanded to manage the phrase vector since it uses dense embeddings to express class labels. In particular, the trainable class embeddings were substituted with the fixed phrase vectors. To prepare for projection conditioning, the discriminator linearly projects sentence vectors.

The construction of BatchNorm gains and biases in the generator entails the linear projection of sentence vectors joined with the noise vector z; biases are zero-centered and gains are one-centered. Utilising the sentence vectors that are fixed, compels the discriminator and generator to conform to the CS acquired by the DAMSM encoder. This leads to intriguing characteristics, including the generator's capacity to manage arithmetic operations in CS. It was initially intended for use in large-scale training when the BigGAN-Deep architecture was developed. Using a large batch size as well as training the models across numerous devices is how large-scale training is done. It is needed to make some more adjustments so the users can use this architecture on a smaller scale.

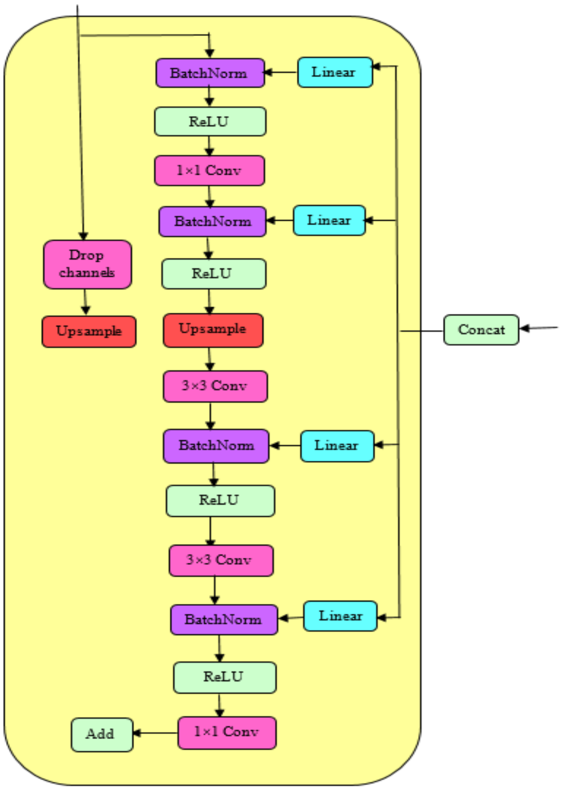

Fig.1. Proposed Architecture

A change from relu activation to leaky relu is made first. Because of the second adaptation, this aids in avoiding sparse gradients. Additionally, this work decreases the amount of parameters for both networks. By lowering the channel multiplier ch from 128 in the default BigGAN-Deep architecture to 96, it is able to decrease the amount of parameters in the generator along with discriminator. This reduction results in a decrease in discriminator parameters and a decrease in generator parameters. The training process is completed by immediately applying the target resolution of 256×256 pixels. Without using several generators and discriminators, no prior T2I approach could train directly at this resolution.

A combined image-text encoder, notably CLIP, contains of a couple of language encoder f_{txt} and image encoder f_{img}, which map a text T_x as well as an image I_g into a shared latent space on which their relevance can be calculated by cosine comparison.

$$S_{CLIP}\left(T_x, I_g\right) = \frac{\langle f_{txt}(T_x), f_{img}\left(I_g\right)\rangle}{\|f_{txt}(T_x)\| \cdot \|f_{img}\left(I_g\right)\|} \qquad (1)$$

The CLIP model was trained such that semantically connected sets of T_x and I_g have high comparison scores.

CLIP+GAN One can synthesis a T2I generator by combining a pre-trained GAN g and CLIP $\{f_{txt}, f_{img}\}$. Given an input text T_x , the research generate a realistic image I_g that is semantically related to T_x by optimizing the latent code ε such that the generated image $I_{img} = g(\varepsilon)$has maximum CLIP score $S_{CLIP}\left(T_x, I_g\right)$. Formally,

$$max_{\varepsilon} S_{CLIP}\left(T_x, g(\varepsilon)\right) \qquad (2)$$

This maximises the output image's semantic relevance to the input text while limiting it to the space of natural images. Adam is used to tackle the optimisation problem in references. As is common procedure when using BigGAN, we reduce z to the interval [-2, 2] [25]. The margins established in the discriminator loss function, the hinge loss is more stable than the WGAN loss, while it still operates similarly.The hinge loss for the discriminator may be expressed as:

$$V_D(\hat{G}, D) = E_{x,s\sim q_{info}}\left[min(0, -1 + D(x,s))\right] + E_{z\sim P_z, S\sim q_{info}}\left[min\left(0, -1 - D(\hat{G}(z,s), s)\right)\right], \qquad (3)$$

where x as well as s are real images along with their associated text vectors. For a specified phrase vector s and a random vector z, the generator produces a false image $\hat{G}(z,s)$. Keep in mind that we are not updating the generator's weights in this instance because of the hat in G. In a similar vein, the generator's loss function is:

$$V_G(G, \hat{D}) = E_{z\sim P_z, S\sim q_{info}}\left[\hat{D}(G(z,s), s)\right] \qquad (4)$$

where, the hat in D represents the discriminator's weights and is not being updated [24].

Algorithm: Enhanced T2I Generation using EGAN

Input: T_x, G, D and noise vector

Output: Generated Image I_g, Performance Metrics (IS, FID)

```
1. Start
2. Input: Text description $T_x$
3. Tokenization: Convert $T_x$ into tokens using NLP preprocessing
4. Feature Extraction:
    a. Apply stemming or lemmatization
    b. Convert tokens into vector representation (e.g., TF-IDF or Word2Vec)
    c. Generate fixed-length sentence embeddings
5. Initialize GAN Model:
    a. Load pre-trained BigGAN model $(G, D)$
    b. Modify architecture to use sentence embeddings as class embeddings
    c. Apply spectral normalization and conditional batch normalization
6. Generate Latent Space Vector:
    a. Sample noise vector $z$ from Gaussian distribution
    b. Concatenate $z$ with sentence vector embedding s
7. Train Generator and Discriminator:
    a. For each iteration:
        i. Generate image $G, D = G(z,s)$
        ii. Compute Discriminator Loss: $V_D$ (Equation 3)
        iii. Compute Generator Loss: $V_G$ (Equation 4)
        iv. Update weights of $G$ and $D$ using Adam optimizer
    b. Apply hinge loss to stabilize training
8. Sentence Interpolation:
    a. Modify embeddings for smooth latent space transition
    b. Adjust generated images by interpolating sentence vectors
9. Evaluate Performance:
    a. Compute Inception Score (IS)
    b. Compute Frechet Inception Distance (FID)
10. Compare results with baseline models
11. Output final generated image $I_g$
12. End
```

4. RESULTS

Here the outline of the experimental apparatus is explained. For this purpose, this research extensively tests on the most popular T2I datasets.

4.1 Datasets

This research's training data encompasses a variety of images and scenes, such as mist over verdant hills, broken plates scattered on the grass, cosmic love and attention, a time traveller amidst a crowd, life during the plague, a tranquil forest for meditation, AI, cosmic love and attention, a fiery sky, an icy pyramid, a solitary home in the woods, a mountain wedding, a lantern hanging from a tree in a foggy cemetery, a vivid dream, balloons over a city's ruins, the astronomer's tragic death, the tragic intimacy of an eternal conversation with oneself, demon fire, a snow pyramid, a solitary house in the woods, and an apple beside a fireplace. The graphic illustrates some of the results from the proposed technique.

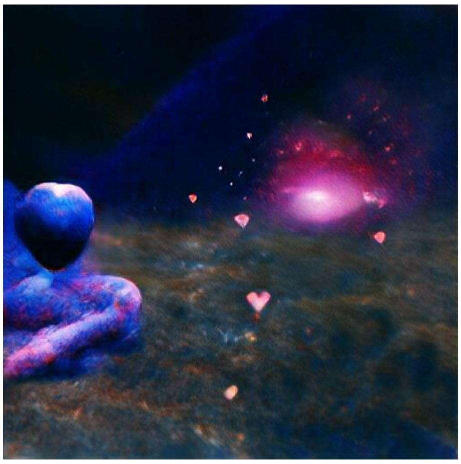

Fig.2. Cosmic love and attention

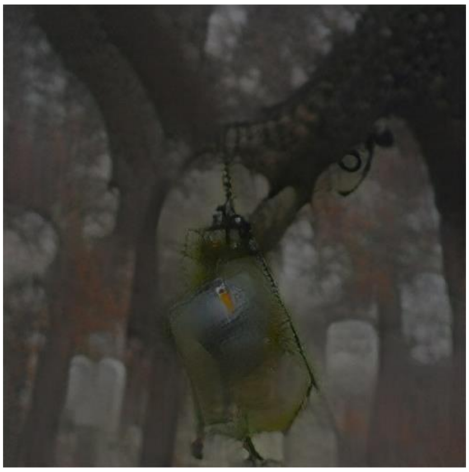

Fig.3. Lantern dangling from a tree in a foggy graveyard

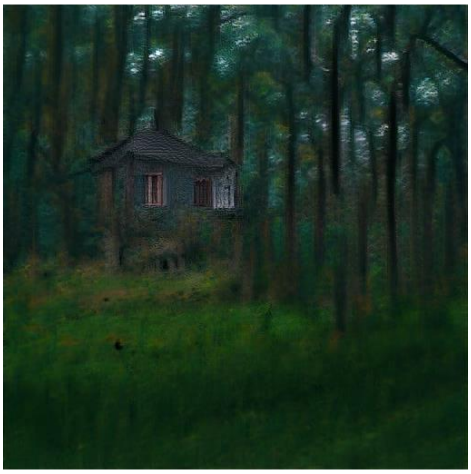

Fig.4. A lonely house in the woods

4.2 Evaluation

To test our approach, we make use of the IS as well as the FID, the two most popular measures for generative models. The greater the IS score, the more objectivity and diversity there is. This research employ the same Inception Networks that were used to assess prior work, calculate IS and also compare outcomes. The IS has the drawback of ignoring the statistics that are already present in the real data. The IS score of a generative model would be very high, even if its variety is modest, if it only produces a small number of high-quality samples for each class. Heusel et al. [22] created the FID to solve this issue. Since FID takes the training data statistics into account, it is feasible to assess if the generative model educated a distribution with comparable statistics. FID employs an Inception Network to calculate the activation features of the images in the training set and the produced images. Next, the Frechet distance is calculated using the images' features, both real along with false. A lower score indicates greater performance in FID, which measures the similarity between the output images' statistics and those in the training set.

4.3 Implementation Details

In PyTorch, this research employs the official pre-trained BigGAN model. This work has a batch size of 1000. With a learning rate of 0.07 along with no weight decay, Adam Optimizer is used for optimization. The implementation is based on the BigGAN-512 pretrained model.

Table 1. Quantitative Comparison of T2I Methods

Method	IS	FID
GAN-INT-CLS	2.88±0.04	-
GAWWN	3.60±0.07	-
StackGAN	3.70±0.04	55.28
StackGAN++	4.04±0.05	15.30
TAC-GAN	-	-
HDGAN	4.15±0.05	-
Ours	**4.23±0.05**	**11.17**

Quantitative data and the total count of discriminator as well as generator networks used in each study are shown in Table I. By achieving top performance in all metrics with only one discriminator and one generator in the overall architecture, it demonstrates that our approach is the preferable option. Across all datasets and metrics, it significantly beats all of the baseline techniques. Based on our approach, the dataset shows the most significant improvement. While compared to the strongest baseline with publicly accessible findings, it offers a relative improvement of around 7% IS and almost 300% FID. Our approach clearly produces much better results on the CUB dataset, which permits a decrease of around 24% in FID.

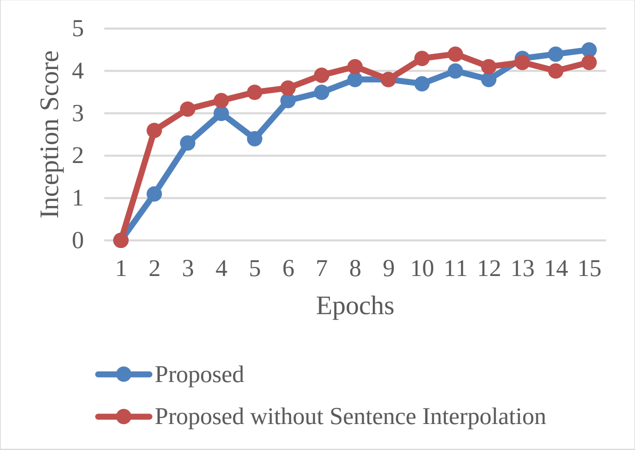

Fig. 5. Proposed model during IS training epochs without along with Sentence Interpolation

All of the IS values obtained during training are shown in Fig. 5. This finding provides further evidence that the proposed approach is important. There is little difference between the two outcomes in the first hundred epochs of training, but the difference becomes more apparent later on. Notably, the model without sentence interpolation only managed to achieve that mark twice during the training, but IS performance with it remained continuously greater than 4.00 after the 400th epoch [24].

5. CONCLUSION

In this study, we propose an original technique that changes the architectural paradigm of T2I conversion. Achieving SOTA presentation with an only step of training right at the target resolution is shown using an effective neural architecture. Not only do this work provide a more straightforward approach to T2I synthesis, but this research also charts a course for future work in this area. Our proposed solution EGAN utilizes CLIP-guided GAN to bring about SOTA T2I creation with exceptional quality. Our technique is zero-shot, training-free, and very customizable, making it ideal for users with limited computing resources or specific needs. It also outperforms classic training-based systems. Our novel approaches to Aug CLIP score, over-parameterized CLIP score, and composition generation are intriguing and applicable to various hidden CLIP optimization issues. This research investigates other approaches to using the sentence vector as a discriminator condition in further work. The projection conditioning was meant to labor with trainable class embeddings, so there is room for more research when the disorder is a fixed sentence vector. Attention modules and memory networks are components of more recent research that we would also want to study.

REFERENCES

[1] Bussell, Chris, Ahmed Ehab, Daniel Hartle-Ryan, and Timo Kapsalis, "Generative AI for Immersive Experiences: Integrating Text-to-Image Models in VR-Mediated Co-design Workflows," In International Conference on Human-Computer Interaction, pp. 380-388. Cham: Springer Nature Switzerland, 2023.

[2] Samuel, Dvir, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik, "It is all about where you start: Text-to-image generation with seed selection," arXiv preprint arXiv:2304.14530 ,2023.

[3] Xu, Jiarui, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2955-2966, 2023.

[4] Chen, Wenhu, Hexiang Hu, Chitwan Saharia, and William W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," arXiv preprint arXiv:2209.14491 ,2022.

[5] Zhang, Mingyuan, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu, "ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model," arXiv preprint arXiv:2304.01116, 2023.

[6] Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour et al., "Photorealistic text-to-image diffusion models with deep language understanding," Advances in Neural Information Processing Systems 35 pp: 36479-36494,2022.

[7] Yang, Jie, Bingliang Li, Fengyu Yang, Ailing Zeng, Lei Zhang, and Ruimao Zhang "Boosting Human-Object Interaction Detection with Text-to-Image Diffusion Model," arXiv preprint arXiv:2305.12252 ,2023.

[8] Ahn, Namhyuk, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong, "DreamStyler: Paint by Style Inversion with Text-to-Image Diffusion Models," arXiv preprint arXiv:2309.06933 ,2023.

[9] Qu, Zhiyu, Tao Xiang, and Yi-Zhe Song, "SketchDreamer: Interactive Text-Augmented Creative Sketch Ideation," arXiv preprint arXiv:2308.14191 ,2023.

[10] Trabucco, Brandon, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov, "Effective data augmentation with diffusion models," arXiv preprint arXiv:2302.07944 ,2023.

[11] Zhong, Guojin, Jin Yuan, Pan Wang, Kailun Yang, Weili Guan, and Zhiyong Li, "Contrast-augmented Diffusion Model with Fine-grained Sequence Alignment for Markup-to-Image Generation," arXiv preprint arXiv:2308.01147 ,2023.

[12] Lu, Haoming, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi, "Specialist Diffusion: Plug-and-Play Sample-Efficient Fine-Tuning of Text-to-Image Diffusion Models To Learn Any Unseen Style," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14267-14276, 2023.

[13] Zhang, Chenshuang, Chaoning Zhang, Mengchun Zhang, and In So Kweon, "Text-to-image diffusion model in generative ai: A survey," arXiv preprint arXiv:2303.07909 ,2023.

[14] Elsharif, Wala, James She, Preslav Nakov, and Simon Wong, "Enhancing Arabic Content Generation with Prompt Augmentation Using Integrated GPT and Text-to-Image Models," In Proceedings of the 2023 ACM International Conference on Interactive Media Experiences, pp. 276-288, 2023.

[15] Bie, Fengxiang, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu et al, "RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model," arXiv preprint arXiv:2309.00810 ,2023.

[16] Chen, Kai, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung, "Integrating Geometric Control into Text-to-Image Diffusion Models for High-Quality Detection Data Generation via Text Prompt," arXiv preprint arXiv:2306.04607 ,2023.

[17] Xiao, Changming, Qi Yang, Feng Zhou, and Changshui Zhang,"From Text to Mask: Localizing Entities Using the Attention of Text-to-Image Diffusion Models," arXiv preprint arXiv:2309.04109 ,2023.

[18] Zhong, Shanshan, Zhongzhan Huang, Wushao Wen, Jinghui Qin, and Liang Lin, "Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models," arXiv preprint arXiv:2305.05189 ,2023.

[19] Brock, J. Donahue, and K. Simonyan, “Large scale gan training forhigh fidelity natural image synthesis,” arXiv preprint arXiv:1809.11096,2018.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[21] T. Miyato and M. Koyama, “cgans with projection discriminator,” arXiv preprint arXiv:1802.05637, 2018.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in Advances in Neural Information Processing Systems, 2017, pp. 6626–6637.

[23] Basha, Syed Muzamil, J. M. Priyadharsheni, Sajeev Ram, and N. Ch SN Iyengar. "A Study on Natural Language Processing approaches for Text2Image using Machine Learning Algorithms."

[24] Souza, Douglas M., Jônatas Wehrmann, and Duncan D. Ruiz. "Efficient neural architecture for text-to-image synthesis." In 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2020.

[25] Liu, Xingchao, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. "Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization." arXiv preprint arXiv:2112.01573 ,2021.