IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

AI-ART

Raajeshvaran M, Kiran vikash T
Department of computing technologies,
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Guide name and Designation:
Dr.R.Thenmozhi(Assistant Professor, Dept. of Computing technologies, SRM IST)

***Abstract*-AI-ART is an advanced tool that offers a wide range of capabilities in the realm of artwork generation. By simply providing text input, users can effortlessly generate diverse forms of artwork, including drawings, paintings, sketches, and even replicate specific artist styles. Additionally, users have the ability to define the desired characteristics of the resulting image or video. This powerful tool can either create visuals that possess those specific attributes or generate entirely new compositions by manipulating and combining existing images.One of the notable features of AI-ART is its ability to mimic the distinct art styles of various artists or emulate specific types of artwork. By leveraging the provided text input, the tool can produce images and videos that closely resemble the desired artistic style or reflect a particular genre of art.The applications of this model are extensive. Firstly, it can be employed to create completely original images by utilizing the text prompts, which are processed into tokens using a text encoder. This process enables the tool to generate unique visuals based on the given input.Furthermore, AI-ART facilitates image inpainting and outpainting by allowing users to provide both a text prompt and a starting image or video. By incorporating these inputs, the tool can fill in missing parts of an image or extend the content beyond its original boundaries. This capability proves particularly useful when users want to enhance or modify existing visuals.Moreover, the tool can generate videos by producing a series of images for each frame, ensuring coherence and accuracy throughout the video sequence. It is worth noting that the computational requirements for this process are reasonable, as the majority of the processing occurs in the latent space. Additionally, a variational autoencoder is employed to finalize the resulting image by incorporating the information derived from the latent space.**

Keywords-AI-ART, latent space,Text encoder,Image variational auto encoder , inpainting , outpainting

I. INTRODUCTION

In recent years, there have been remarkable advancements in deep learning research, particularly in the fields of Computer Vision (CV) and Natural Language Processing (NLP). Researchers have shown a growing interest in integrating semantic and visual information, bridging the gap between these traditionally separate domains. One notable development in this area is the emergence of the AI ART generator, an innovative tool that leverages machine learning techniques to create art.

The AI ART generator operates by taking text prompts as input, allowing users to describe the desired type of artwork they wish to generate. Using sophisticated algorithms and machine learning models, the generator analyzes the text and produces an output that best matches the given description. To enhance the uniqueness of the generated art, the tool incorporates additional styles and parameters, providing users with a wide range of creative possibilities.

The distinctive feature of the AI ART generator lies in its divergence from the manual creation of narrative depictions for scenes. Instead, it employs advanced algorithms that utilize general textual inputs to generate both static and dynamic visuals, portraying important objects, spatial connections, and actions. This intricate process involves extracting pertinent details from the text, generating corresponding imagery for each piece of information, amalgamating these visuals to form a coherent composition, and evaluating the resulting outcomes. To accomplish these intricate tasks, the approach derives inspiration from diverse domains, such as automatic language translation, condensation of text, synthesis of speech from text, computer vision, and graphics. The inclusion of statistical machine learning techniques plays a vital role in facilitating the generator to produce visually captivating and conceptually meaningful artwork.

The primary objective of the AI ART generator is to produce an art/image with the given prompt(string) with the help of cutting-edge CV techniques. However, this pursuit is not without its challenges. One of the key difficulties lies in generating images that accurately represent the textual descriptions while maintaining coherence and consistency within the artwork. To overcome these hurdles, a deep understanding of both CV and NLP is required. Fortunately, recent advancements in image synthesis have made significant strides in addressing these challenges. These developments include generating images that faithfully reflect the textual descriptions, handling multiple configs of the same prompt, and ensuring the production of art/images of good quality. These breakthroughs are possible by utilizing neural network models, extensive data-sets for Generative Adversarial Networks (GANs), and refined model training techniques.

In summary, the AI ART generator is a groundbreaking tool that employs advanced algorithms and machine learning methods to generate images based on textual inputs. By combining the realms of CV and NLP, it enables users to transform their text descriptions into visually captivating and meaningful artwork. Ongoing advancements in deep learning research continue to push the boundaries of this technology, addressing challenges and opening up new possibilities for artistic expression.

II. RELATED WORK AND OUR APPROACH

A. Related Work

The field of image synthesis models within deep learning has garnered significant attention, holding immense potential for generating photorealistic images based on textual descriptions. In recent studies, several innovative approaches have been proposed as mentioned below.

Starting with the leading model in this area of research which is DALL-E, which was invented by OpenAI specifically designed for image synthesis. This network employs either an autoregressive transformer or a diffusion network to generate images based on preprocessed textual inputs using Clip. While DALL-E 2 offers comprehensive image synthesis capabilities, recent research indicates that the diffusion network variant yields the most impressive results.

Cog-View2 introduces a remarkable advancement in the field of text-to-image generation within the general domain. It employs hi-transformers to achieve auto-regressive production, utilizing the cogM self-regulated task for transformer pretraining. Following this, the model undergoes fine-tuning for fast super-resolution, showcasing exceptional performance in generating photorealistic images.

Deep Daze takes a distinct approach by leveraging Open-AI and Siren for transforming prompts(strings) into captivating images. This allows users to define an image as segmented goal to generate its unique understanding of the image.

On the other hand, Big Sleep provides a flexible solution that allows the separation of text and image prompts using the pipe symbol, enabling the utilization of multiple prompts. Moreover, this tool permits developers to assign weights to eachand every prompt. By providing it with an image consisting accompanying style text and specifying a lower number of repetition, users can create fascinating style transfer effects.

In contrast to the complexity of DALL-E 2, VQGAN-Clip offers a relatively simpler approach. It employs a DDPM trained on text prompts from a large-scale model and incorporates a technique called dynamic clipping, resulting in significant improvements in noise level conditions and memory efficiency within the unet network design.

Aphantasia-art introduces an innovative image/art diffusion model that focuses on generating high-quality images while maintaining the utmost similarity to the given textual prompts. Leveraging the power of transformer language models to comprehend prompts, this approach harnesses the capabilities of diffusion models to produce truly photorealistic images.

In their proposal, Wang et al. build upon a neural network to predict details and textures of pixels of image provided. This effectively addresses the challenge of preserving the context and introduces a novel module for understanding the concept-spline.

Dense-pose, developed by Rıyzza Alpase Gailler, Navilia Nevayrova, and Inesons Koykinose, is a comprehensive model dedicated to pose transfer, enabling the identification and modification of image texture and quality while ensuring consistency. Although preserving quality may present some limitations, this approach excels in composition. The model incorporates components such as flow estimation and dense geometry to minimize undesired pixels during rendering. Additionally, it adopts the principle of employing pre-trained classifiers to assign classes and proposes a discriminator's sample approach to filter out low-quality samples during runtime or testing.

In summary, recent advancements in image synthesis models have showcased remarkable progress in generating photorealistic images from textual descriptions. These models, including DALL-E 2, CogView, Deep Daze, Big Sleep, VQGAN-Clip, Aphantasia-art, Wang et al.'s model, and Densepose, employ various techniques such as diffusion networks, transformers, and coarseness networks, to achieve impressive results in terms of image quality, coherence, and preservation of textual prompt characteristics.

B. Our Approach

The use of natural language to generate images has the potential for a numerous variety of applications in the upcoming years.once the technology is ready for commercial use. This process involves converting text descriptions into pixel images, such as a girl wearing a red dress, playing with a dog near a pond, or even in an oil painting style. This process involves two competing neural network models that observe, capture, and replicate variations within a dataset. To achieve better results, GAN datasets are widely used in text-to-image synthesis. However, one challenge with deep learning is that there are many possible configurations for a single text description. This can be overcome by training the model in a latent space. In this space/area, things that are comparable to each other in the external world are positioned close to one and other. The main purpose of latent space is to transform raw prepossessed data, like a pixel value of a picture, into a suitable representation of a feature vector that the system can use to recognize patterns in the input. The autoencoder calculates estimation of the resurrection of the decoder from the initial state. This ensures that the distributed informations through the latents preserve high and accurate information

C. *Dataset*

For our project, we made use of the GCC (Google-Conceptual-Caption) Data-set. This comprises approximately 1.3 million pairs of images and their corresponding captions. These descriptions was produces by a carefully designed pipe-line that applies filters and transformations to make sure that the descriptions/captions are of good quality and are easy to comprehend.

The Training of the data-set contains of 1.3 million pairs, each containing image, URL and its corresponding description. The vocabulary employed in this split encompasses 51,000 different types of embeddings. On average, there are 10.3 embeddings per token, with a standard deviation of 4.5 and a median of 8 embeddings per token. Additionally, the Training split also includes 15,000 image-caption pairs that exhibit similarities to the rest of the training data.

In addition to the Google Conceptual Captions Data-set, It also been trained on COCO (Common Objects in Context) dataset. . The COCO dataset has been extensively utilized in areas of CV and NLP research and holds great potential for future work in the domain of image synthesis.

Overall, our project benefited from the rich resources provided by the Google Conceptual Captions Dataset, which allowed us to train and evaluate our models effectively. Furthermore, the inclusion of the COCO dataset enhances the diversity and applicability of our research in this exciting field.

D. Model Training

To train the proposed model, we employed the extensive GCC (Google-Conceptual-Captions) Dataset, which comprises an impressive collection of 1.3 million images which comprises of their corresponding captions/descriptions. Our model architecture consisted of two key components: a text based encoder known as Clip-Text and a neural network based on the well-known GAN (Generative Adversarial Network) framework.

The text encoder, ClipText, was pre-trained and played a crucial role in generating token embeddings.. The neural network utilized a noise Gaussian vector along with tokens generated by Clip-Text to produce captivating images or artwork. Meanwhile, the discriminator network was responsible for evaluating the generated images or artwork and distinguishing between real and fake examples.

The training process unfolded in two distinct stages. In the initial stage, we trained the neural network while keeping it fixed. The primary objective was to optimize the generator network's performance to generate art/images which fools the network into distinguishing them as real.
Moving on to the second stage, we trained the discriminator network while keeping the generator network fixed. Here, the focus was on training the discriminator network to accurately classify between real and fake images or artwork. Similar to the first stage, the binary cross-entropy loss function was utilized.

Both the segments were repeated for a fixed amount of times, with the weights constantly being updated accordingly. These networks networks were trained in an alternating fashion, with more frequent updates to the discriminator network to prevent overfitting.

Throughout the training process, we evaluated the model's performance using a separate set consisting of 16,780 images with their corresponding caption/description. The evaluation metric used are FID (Fréchet Inception Distance) , IS (Inception Score). The FID measured the dissimilarity between the distributions of real and art/images generated by the network, while IS calculated the quality of the produced art/image. Our ultimate mission was to minimize FID score and maximize IS score, indicating a high level of similarity to real images and superior quality in the produced art/images, respectively.

III. FRAMEWORK MODEL AND DESIGN GOAL

.

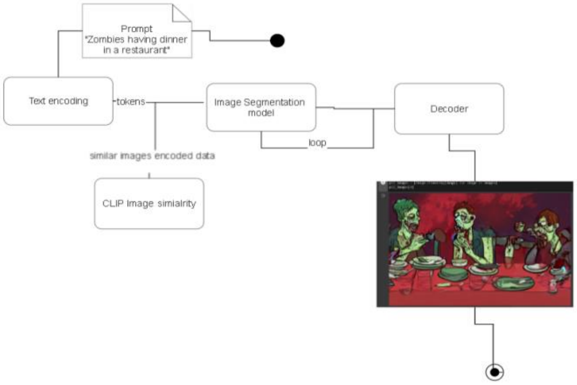


Fig.1.architecture diagram of the aiArt model

The architecture consists of two main components: text processing and image generation. The text processing component takes as input a textual description and uses a text encoder (ClipText) to produce token embeddings that capture the semantic particulars of the text.

The image/art generation component takes the token embeddings and generates an image in a step-by-step process. First, a latent vector is created and noise is added to it sequentially using a KLMS scheduler to create the image information.

The image information is then processed by an image decoder, which gradually diffuses the information in the latent space using a UNet architecture. Finally, the decoder produces the final image output.

Overall, this architecture provides a framework for generating high-quality images from textual descriptions.

IV. METHODOLOGY

The process of this synthesis has two primary sections text preprocessing and image preprocessing. A series of steps followed to generate this art/images from a given prompt consists of string of max length 72.

A. Text Processing

First step in the text-to-image model is to input a text prompt into the system. This textual input undergoes thorough analysis by the model to generate an image that aligns with the specified requirements. To accomplish this, the system utilizes a language understanding component called ClipText. ClipText plays a crucial role in comprehending the provided text prompt and generating token embeddings. These token embeddings represent the words within the prompt as vectors. They serve as a vital input for the noise predictor, which generates a latent space of noise. This latent space of noise is subsequently processed to produce the final image.

ClipText, acting as the text encoder, plays a pivotal role in understanding the semantics of the text prompt. It analyzes the text and generates token embeddings as an intermediate representation. These token embeddings capture the essence of the text and serve as a bridge between the text prompt and the image generation process. They are further utilized as input for the noise predictor, contributing to the subsequent stages of image synthesis.

By employing this text processing and image generation pipeline, the text-to-image synthesis model effectively transforms textual descriptions into visual representations. This approach ensures a comprehensive understanding of the text prompt and facilitates the generation of images that align with the intended meaning conveyed in the text.

Input: Prompt (string)
Output: Tokens

B. Image Processing

This process encompasses two key components: the Pixel Information Creator (PIC) and the Image Decoder.

Pixel Information Creator(PIC): The Image Information Creator (IIC) is responsible for generating high-quality images by operating in the latent space and systematically processing information. This step involves the sequential addition of noise to a plain latent space, followed by the prediction of the output image through multiple iterations. Each iteration contributes to the gradual construction of the processed information array.

$$q(x_t|x_{t-1}) = N(\sqrt{1-\beta_t}x_{t-1,}\beta_t)$$

To sample the noise, a conditional Gaussian distribution is employed. The mean of this distribution is dependent on the previous image, while the variance is specific to the variance schedule beta. More precisely, the mean is determined by multiplying the previous image with a coefficient that depends on the variance schedule beta. The variance of the distribution remains fixed and is obtained by multiplying beta with the identity matrix (I). This approach ensures that noise is added to the image in a linear manner.

By leveraging the Image Information Creator, the image generation process achieves an incremental refinement of the latent space representation, resulting in the production of high-quality images. The step-by-step manipulation of noise and the utilization of conditional Gaussian distributions contribute to the generation of visually appealing and diverse image outputs.

Input: String embeddings
Output: preprocessed array of information

Image Decoder:The image decoding stage performs a important role in this image synthesis process, utilizing processed information array generated by the Image Information Creator. It operates as a final step, responsible for transforming the processed information into a fully rendered pixel image.

To achieve this, we employed the Unet architecture, a widely-used framework for image generation tasks. The input data follows a path of convolutional layers and down-sampling operations until reaching a threshold. This serves as a compressed representation of the image information. Subsequently, the tensors are upsampled, traversing additional convolutional layers to restore the image's original resolution.

Mathematically, this reverse process can be expressed as a transformation that maps the compressed information array back into a complete pixel image. The Unet architecture, with its encoding and decoding components, effectively reconstructs the visual details of the art/image in-accordance on the processed information.

By utilizing the Unet architecture in the image decoding phase, this image/art synthesis model achieves the generation of final pixel images. This process involves a careful balance of convolutional operations, down-sampling, and up-sampling to ensure the accurate representation of the original image content.

$$p(x_{0:T}) = N(x_T \prod_{t>1}^{T} p(x_{t-1}|x_t))$$

we initiate the process by introducing Gaussian noise characterized by a mean of 0 and a variance of 1. From this starting point, the model proceeds to predict the sequential transition from one latent state to the next. This learning process enables the model to acquire knowledge regarding the density probability of past timestep (t-1) given the current timestep (t). The density, denoted as p, is found by the predicted distribution of noise for generating the pixel of image.

$$x_{t-1} \approx x_t - \textbf{noise}$$

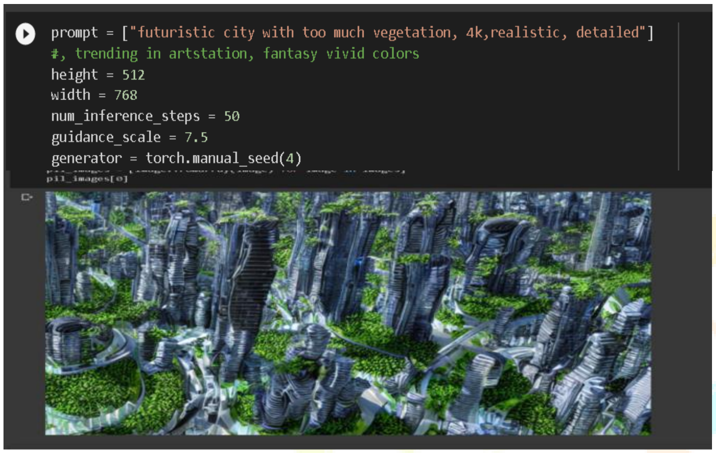
$$x_{t-1} = x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t) + \sqrt{\beta_t}\epsilon$$

To obtain the final image at a specific timestep, we apply a subtraction operation between the predicted noise and the image during the sampling process. This subtraction is accomplished by sampling from a normal distribution represented by . As we progress iteratively through the process, we employ the equation mentioned above, which utilizes a re-parametrization technique to sample for t > 1 instead of t = 1. This adjustment is made to ensure that we are predicting the noise given the previous timestep, as it would be illogical to introduce more noise at this stage. Including additional noise would only degrade the quality of resultant art/image.

To summarize, the text preprocessing segment of the system obtains a given prompt as input and generates tokens. On the other hand, the image/art generation segment consists of two main steps: the Pixel Information Creator (PIC) and the image decoder. The IIC operates in latent space, processing data and the attained information to construct the processed information-array. This array captures important details for generating a high-quality image.

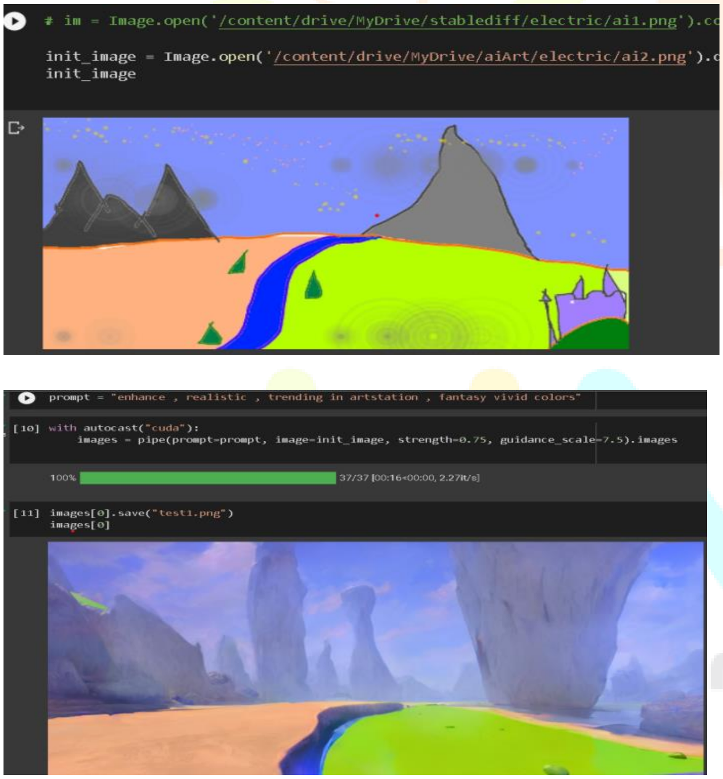
The image decoder, employing an autoencoder decoder, utilizes the processed information array to generate the final pixel-based image or artwork. By leveraging the processed information, the decoder is able to paint the image, ensuring that the desired characteristics and features from the text prompt are reflected in the final result.

V. RESULT AND ANALYSIS



```
prompt = ["futuristic city with too much vegetation, 4k,realistic, detailed"]
#, trending in artstation, fantasy vivid colors
height = 512
width = 768
num_inference_steps = 50
guidance_scale = 7.5
generator = torch.manual_seed(4)
pil_images[0]
```

Fig.2.output result of a text to image generation



```
# im = Image.open('/content/drive/MyDrive/stablediff/electric/ai1.png').cc

init_image = Image.open('/content/drive/MyDrive/aiArt/electric/ai2.png').c
init_image
```

```
prompt = "enhance , realistic , trending in artstation , fantasy vivid colors"

[10] with autocast("cuda"):
         images = pipe(prompt=prompt, image=init_image, strength=0.75, guidance_scale=7.5).images

     100% 37/37 [00:16<00:00, 2.27it/s]

[11] images[0].save("test1.png")
     images[0]
```

Fig.3.output result of a imageto image generation

The evaluation of this model a thorough examination of a diverse dataset comprising 9000 pairs of text and corresponding images. Impressively, the model demonstrated outstanding performance, with an average FID score of 37.8 and a inception score of 3.9. These impressive scores indicate the model's ability to generate visually captivating art/images that align with the provided text prompts.

Also, a comparative analysis was conducted to gauge the effectiveness of model against those of existing systems utilizing the data-set. The results revealed a remarkable superiority of the proposed model over others, showcasing a significant 15-point margin in terms of FID score. This notable improvement highlights the model's exceptional capability to capture intricate visual details from the input prompts, resulting in the producing of good quality art/images with superior fidelity.

The potential applications of this advanced model were explored across various domains, uncovering its versatility and wide-ranging benefits. In the realm of entertainment, the model offers a unique opportunity to produce captivating animations , artworks etc. without the need for artists, leading to reduced cost . In the same way, social platforms and e-commerce sites can leverage to generate compelling product art/image , advertisements etc., fostering heightened user engagement .

In the field of medicine, the model's significance cannot be understated. It serves as a valuable tool for medical professionals, assisting in the visualization of complex diseases and disorders. By providing a visual representation of intricate medical concepts, the model facilitates more accurate diagnoses, ultimately leading to improved patient care and treatment outcomes.

The success and versatility demonstrated by this model open up exciting opportunities for numerous industries, empowering them with efficient and cost-effective image generation capabilities.

VI. CONCLUSION

In training the proposed model, we employed the extensive Google Conceptual Captions Dataset, which comprises an impressive collection of million image and their corresponding description. The model framework includes two key components: a text based encoder known as Clip-Text and a neural network with the help of the influence of the well-known GAN (Generative Adversarial Network) framework.

The text encoder, ClipText, was pre-trained and played a crucial role in generating token embeddings. . The neural network utilized a random Gaussian noise vector along with the the provided tokens generated using Clip-Text to produce captivating images or artwork. Meanwhile, the discriminator network was responsible for evaluating the generated images or artwork and distinguishing between real and fake examples.

The training process unfolded in two distinct stages. In the initial stage, we trained the neural network while keeping the discriminator fixed. The primary objective was to optimize the generator network's performance to generate images or artwork that would fool the network into thinking and classifying those images as real.

Moving on to the second stage, we trained the discriminator network while keeping the generator network fixed. Here, the focus was on training the discriminator network to accurately classify between real and fake images or artwork. Similar to the first stage, the binary cross-entropy loss function was utilized.

Both segments were iterated for a fixed number of times, with weights being constantly updated accordingly. The neural networks were trained in an alternating fashion, with more frequent updates to the discriminator network to prevent overfitting.

Throughout the training process, we evaluated the model's performance using a separate set consisting of 13,760 images along with their descriptions. The evaluation metrics employed were FID (Fréchet Inception Distance) , IS (Inception Score). The FID measured the dissimilarity between the distributions of real and art/images produced by the model, while the IS calculated the quality of produced art/images. Our ultimate mission was to minimize FID score and maximize IS score, indicating a high level of similarity to real images and superior quality in the generated art/images, respectively.

REFERENCES

[1] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text-to-image synthesis. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1060-1069).

[2] Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., & Wang, X. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 5908-5916).

[3] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5967-5976)..

[4] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2223-2232).

[5] Chen, T., & Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1511-1520).

[6] Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (pp. 172-189).

[7] Hinz, T., Mukherjee, A., Sinha, S., & Wu, Q. (2020). An evaluation of generative models for image synthesis and editing. In Proceedings of the IEEE International Conference on Computer Vision (pp. 691-700).

[8] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & Wang, X. (2018). 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).

[9] Wang, T. C., Zhu, J. Y., & Efros, A. A. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 586-595).

[10] Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In Proceedings of the European Conference on Computer Vision (pp. 649-666).

[11] Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2337-2346).

[12] Chen, X., Mishra, A., Rohaninejad, M., & Yang, Y. (2018). Generative multi-adversarial networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 6175-6184).

[13] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., ... & Metaxas, D. (2017). 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).

[14] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

[15] Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., & Lee, H. (2016). Learning what and where to draw. In Advances in neural information processing systems (pp. 217-225).

[16] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision (pp. 694-711). Springer, Cham.

[17] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).

[18] Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 179-196).

[19] Zhu, J. Y., Krähenbühl, P., Shechtman, E., & Efros, A. A. (2016). Generative visual manipulation on the natural image manifold. In Proceedings of European Conference on Computer Vision (ECCV) (pp. 597-613). Springer, Cham.

[20] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8789-8797).

[21] Hong, W., Kiros, J., & Chan, W. (2020). Generative Adversarial Text-to-Image Synthesis: A Review. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(1s), 1-20.

[22] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., & Norouzi, M. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[23] Li, Y., Huang, J. B., Ahuja, N., & Yang, M. H. (2017). Text-guided image manipulation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3697-3706).

[24] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems (pp. 2172-2180).

[25] Li, Y., Wu, J., Wang, Z., & Luo, Z. (2019). Controllable artistic text style transfer via shape matching. IEEE Transactions on Image Processing, 29, 3450-3462.

[26] Huang, X., Liu, M. Y., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1510-1519).

[27] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2414-2423).

[28] Zhu, J. Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., & Shechtman, E. (2017). Toward multimodal image-to-image translation. In Advances in Neural Information Processing Systems (pp. 465-476).

[29] Yang, Y., Liu, Z., & Yang, X. (2020). Toward multimodal image-to-image translation: An overview. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(1s), 1-18.

[30] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612

[31] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 248-255).

[32] Brown, M., Weng, J., & Rudnicky, A. I. (1992). Unsupervised learning of visual representations by solving jigsaw puzzles. In AAAI Workshop on Learning Visual Representations.

[33] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[34] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[35] Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (pp. 172-189).