PoolImagen: Text-to-Image Diffusion Models with An Efficient Transformer without Attention

Hyeeun Ku, Minhyeok Lee*
School of Electrical and Electronics Engineering
Chung-Ang University
Seoul 06974, Republic of Korea
{hae10202, mlee}@cau.ac.kr

Kanghyeok Ko, Sun Jae Baek
Department of Intelligent Semiconductor Engineering
Chung-Ang University
Seoul 06974, Republic of Korea
{dogworld12, baechu}@cau.ac.kr

Abstract— **Recent advancements in the field of image generation models have been particularly notable for diffusion models. Imagen has the most remarkable image generation capabilities among these models, particularly at high resolutions. However, Imagen comes with limitations, as creating high-quality results requires considerable computational resources and lengthy training times. To address these limitations, we propose PoolImagen, a novel and improved variant of Imagen that combines high performance with low computational costs. PoolImagen introduces various improvements to overcome the constraints of Imagen. Notably, we adopted the idea, first propose in MetaFormer, which suggests replacing the attention module with a pooling structure in the transformer architecture of Imagen to mitigate the issues related to increased training costs and computational complexity. Additionally, considering the influence of text encoder size on text-to-image transformation quality, we incorporate the large language models (e.g. flan-t5-xxl), an extension of the t5 model that offers more parameters and refined text processing capabilities. With a well-trained transformer, PoolImagen achieves image generation with consistent performance and significantly accelerated training velocities. In experiments based on bird image datasets, PoolImagen demonstrates improved performance in terms of Fréchet Inception Distance (FID) and training time. In the case of the bird datasets, PoolImagen exhibits an approximately 11.29% improvement in FID compared to Imagen, while training time is reduced by 2.25 times. In addition, we conducted additional experiments to evaluate ability of PoolImagen to represent domain-specific features in generated images. These findings emphasize the potential of PoolImagen as a powerful tool for rapidly generating text-to-image outputs and suggest promising directions for enhancing the future performance of diffusion models.**

Keywords—text-to-image synthesis, diffusion models, image generation, transformer

I. INTRODUCTION

Various forms of multimodal learning have gained significant attention recently, particularly emphasizing text-to-image synthesis [1, 2] and image-text contrastive learning [3, 4]. Diffusion-based text-to-image models [5-7] have demonstrated remarkable progress in generating realistic content using text prompts. These models have revolutionized the research landscape and garnered extensive interest due to their innovative image generation [8] and editing capabilities [9] . They have a profound impact on content creation [10], image synthesis , video synthesis [11], inpainting [12], and super-resolution [2], among others. However, this impact is accompanied by a significant increase in computational demands required to execute these models.

Imagen [13], a text-to-image diffusion model aimed at high-realism and high-resolution, enhances quality through a larger text encoder. However, it has been observed that its effectiveness comes at the cost of extensive computations and computational demand, leading to prolonged training periods. In this context, Imagen incorporates the attention mechanism from the transformer structure within the enhanced U-Net architecture [14].

In the context of contemporary transformers [15] have commonly employed attention mechanisms [16, 17] . In recent research endeavors have ventured into challenging the conventional approaches to attention. To overcome this limitation, MetaFormer [18] have been introduced, explored a novel avenue by replacing the attention mechanism with an exceedingly simplistic non-parametric operator known as pooling in the token mixer component [19]. Remarkably, this substitution demonstrated competitive performance in image classification tasks [20]. In other words, it has been revealed that the main driving force behind the effectiveness of the transformer does not solely rely on attention. Instead, it is the combined performance of various other components that plays a crucial role [21].

In this paper, we present PoolImagen, a diffusion text-to-image model that uses pooling within a transformer architecture [18] to resolve the difficulties associated with diffusion. We focus on faster training and better image production to achieve this. The simple computing procedure of pooling compresses data and extracts crucial visual features [22]. By incorporating pooling into the U-net architecture, we aimed to accelerate the learning process. This modification allowed us to expedite the training process without sacrificing image generation performance. Pooling for the attention mechanism balanced learning speed and performance. Thus, the model learns faster and produces images that are superior.

In addition, we achieved this through alterations in the U-Net architecture and the incorporation of a generic large language model [23, 24] as the text encoder. The text encoder applied to PoolImagen is the Flan t5-xxl [25], which, in contrast to the conventional t5 text encoder [24], offers superior performance for text-based tasks. Flan t5-xxl is an extension of the t5 model that has been trained with additional data and time, resulting in more advanced text processing capabilities than the standard t5 model.

Furthermore, PoolImagen was evaluated using a datasets: the well-known, bird dataset, CUB [26]. We compared the

Fréchet Inception Distance (FID) [27] scores and training times for various image sizes in both datasets. As a result, it achieves performance in maintaining the quality of generated images while also shortening the training time.

In summary, the main contributions of this study are as follows: 1) In the U-net structure, we used pooling instead of attention mechanism to improve the learning speed. By maintaining the performance, we were able to increase the learning speed. 2) We managed to accomplish text-to-image synthesis by utilizing a new text encoder, Flan t5-xxl [25], which provides sophisticated text processing. 3) Our objective is to demonstrate experimentally the viability of PoolImagen for accelerating learning and generating realistic images.

II. Related Works

A. Diffusion Probabilistic Model

The diffusion probability model was initially introduced in [6-8] . Its successful implementation in the field of image generation was initially observed for small-scale images, but its performance improved significantly for relatively larger images as demonstrated in [28]. This diffusion model's architecture has continued to evolve, incorporating significant advances in learning and sampling methodologies, such as Denoising Diffusion Probabilistic Model (DDPM) [5], Denoising Diffusion Implicit Model (DDIM) [29], and Score-Based Diffusion [30].

The image diffusion technique is often implemented by directly utilizing pixel color information from the training data [30]. In such instances, researchers examine computational resource conservation solutions, especially for high-resolution images [31]. Usually, these strategies rely on neural network architectures such as U-net. To optimize computational resources for training diffusion models, "Latent Images" formed the basis for the Latent Diffusion Model (LDM) [2], which has now been extended to Stable Diffusion [2] to improve learning.

B. Text-to-Image Diffusion

The diffusion model can be effectively trained with conditioning input channels, allowing for the generation of conditional images [7, 8]. Recent applications of the diffusion model in text-to-image synthesis have attracted considerable attention, especially for their innovative synthesis capabilities.

Typically, this is accomplished by translating textual inputs into latent vectors using pre-trained language models such as CLIP [3]. For instance, Glide [7] is a text-guided diffusion model that facilitates image generation and modification. Disco Diffusion [32], on the other hand, is a CLIP-guided implementation designed to handle text prompts. Stable Diffusion [2] is a large-scale implementation of latent diffusion developed to generate text-to-images. Imagen [13], on the other hand, utilizes a text-to-image structure that avoids the use of latent images and diffuses pixels directly using a pyramidal structure.

Imagen, a Diffusion Text-to-image model [13], combines large-scale transformer language models for text understanding with diffusion models for high-resolution image production. Imagen competes with DALL-E 2 [8] and outperforms it in evaluations, especially in terms of high-resolution image generation. This benefit comes from the revelation that pretraining on text-only corpora with large-scale language models like T5 [24] successfully encodes text.

In Imagen [13], boosting the language model improves sample quality and image-text alignment more than increasing the diffusion model.work also presents an efficient method for training the diffusion model using pre-trained text encodings and large language models.

C. Pooling for Poolformer

The transformer architecture has demonstrated remarkable performance in various computer vision tasks , with the key to this success often attributed to the attention mechanism . As a result, ongoing research has focused on how to effectively enhance this attention mechanism. By introducing periodic shift methods, models such as Swin Transformer [33] have attempted to enhance the attention structure. Similarly, models such as ResMLP [34] and MLP-Mixer [35] have achieved high performance by substituting attention modules with simplified spatial MLPs . While these models consistently deliver exceptional performance, the adoption of the transformer architecture in a large-scale model incurs a significant increase in training costs and computational requirements.

This sthdy introduces the MetaFormer [18], a general architecture without attention structure restrictions, to ocercome these difficulties. The MetaFormer provides flexible token mixer component swaps without requiring transformer attention structure. To demonstrate that attention is not the core of the transformer architecture, but that superior performance arises from the combination of various modules, the paper conducts experiments by replacing the attention module with an extremely simple pooling structure.

The results of these experiments indicate that the PoolFormer [18] model continues to deliver superior results even when attention is replaced by pooling. This model utilizes spatial pooling operators to divide input images into smaller fragments before combining the extracted features from each fragment to represent the entire image. It highlights the importance of spatial pooling operators and shows how to attain good performance with less parameters and computing effort. Using these findings as a foundation, this paper proposes PoolImagen, which maintains efficacy amid faster training speeds.

III. METHOD

Our goal is to develop a diffusion model-based image synthesis framework with improved U-Net structure and learning speed. Present Imagen uses self-attention in U-Net transformer module. Accordingly, many researchers thought the transformer's great performance was due to the attention-based token mixer module, but spatial MLP [34, 35] replaced it and it still performed well. Even though attention and attention-based token aggregator modules have been widely studied, self-attention and spatial MLP are computationally costly. To address these issues, in this paper, the Poolformer framework was introduced in PoolImagen. As illustrated in Figure 1, PoolImagen consists of a text encoder that maps text to a sequence of embeddings and a cascade of conditional diffusion models that map these embeddings. In the following subsections, we describe each of these components in detail.

A. Pretrained text Encoder

Text-to-image models require robust semantic text encoders capable of comprehending the complexity and compositionality of diverse natural language inputs. Text encoders trained on paired image-text data are standard in recent text-to-image models; they can be trained from scratch [1, 7] or on image-text data , such as CLIP [8], that has been pretrained. The image-text training objectives imply that these text encoders are capable of encoding visually expressive and meaningful representations that are especially pertinent to the text-to-image generation task. In addition, large language models can be used to encode text for text-to-image generation. Recent advancements in large language models (e.g., BERT [23], GPT [36] , and T5 [24]) have resulted in substantially improved textual comprehension and generative capacities. Language models are trained on a text-only corpus that is substantially larger than paired image-text data, exposing them to a text distribution that is extremely rich and extensive. Furthermore, these models have significantly larger model sizes compared to the text encoders used in existing image-text models [3, 4]. (e.g., PaLM [37] has 540B parameters, while CoCa [38] has a 1B parameter text encoder).

Thus, it is natural to investigate text encoders for the text-to-image task. In our model, PoolImagen, we utilized the pre-trained text encoder Flan-t5-XXL [25]. To streamline our approach for greater efficiency, we have frozen the weights of these text encoders. This approach allows offline embedding computation and low computational overhead and memory use during text-to-image model training. We found that text encoder size improves text-to-image conversion. Consequently, we chose the flan-t5-XXL encoder is an extension of the t5 model that conforms well to text and provides more sophisticated text processing capabilities than the standard t5 encoder.

B. Efficient model arichitecture

The structure in general is illustrated in figure 1 (a) below. Through a text encoder, the entered caption is expressed numerically, and this expression is used as a condition to generate the result value z. The image generator utilizes a diffusion model, leveraging embeddings and sample noise from z_t to proceed with training, ultimately generating images based on the input text.

The architecture of the model is built upon U-Net. For denoising, the position value was included in the encoding to conduct conditioning at each time step. Encoding is generated at each timestep and applied to U-Net, where images, time, and text are all encoded, facilitating conditioning.

We enhanced the architecture of the model for rapid learning. Our model is simpler and converges faster than Imagen by using a more efficient transformer block. The U-Net architecture in PoolImagen, as shown in Figure 2 (b), consists of the following components: Conv layers in the encoder section, 4 down-sampling blocks, 2 ResNet blocks, and Transformer blocks in the middle layer, and 4 up-sampling blocks and Conv layers in the decoder section.

In a typical U-Net, down-sampling block, the down-sampling operation occurs after the convolutions, and in an up-sampling block, the up-sampling operation occurs prior the convolution. Figure 2 (c) illustrates the down-sampling and up-sampling blocks of U-Net, respectively. We employed an efficient U-Net that flipped the order of convolution for down-sampling and up-sampling blocks to boost forward pass speed and avoid performance deterioration. ResNet blocks have 8 grouped convolution layers and are used depending on parameters.

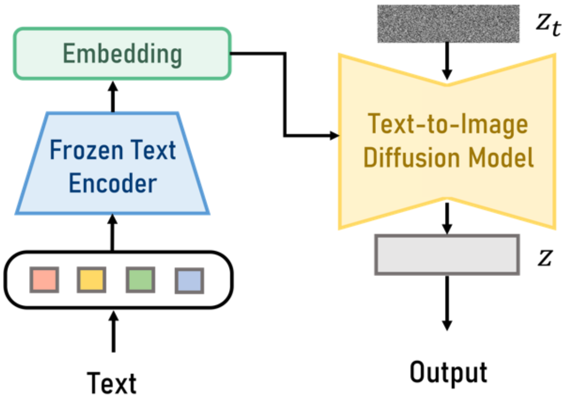

Fig 1. Structure of the proposed PoolImagen. PoolImagen uses a frozen test encoder to encode the input text into text embeddings. A conditional diffusion model with inputs of noise vector and embedding generates an image.

Furthermore, in PoolImagen, we replace with pooling instead of self-attention in the token mixer section of the transformer structure. This modification enables us to maintain performance while employing fewer parameters and computations, which has a positive effect on training velocity.

C. Training specifications

To compare the efficacy of the text-to-image diffusion model [13], we employed the same learning procedure as Imagen but with several adjustments in PoolImagen. During the training of PoolImagen, we utilized the Adam optimizer [39] with a learning rate set to $1e-4$ and conducted training for 4000 epochs.

U-Net used an 8-channel model with a 64x64 input image. Each down- and up-sampling layer had twice as many channels due to dimensional scaling. For each layer, ResNet [40] blocks were set to (1, 2, 4).

In the diffusion model, we used the L2 loss [41] as the loss function and set the number of timesteps to 200. During image generation, the model employs text and other conditional information, and the conditional dropout probability is specified as 0.1. The model is trained using a batch size of 16 on one A6000 GPU and one RTX 3090 GPU with 48GB and 24GB of memory, respectively. The Python version used was 3.8.15, and the Pytorch version was 2.0.1 with CUDA 11.7 and 1.12.1 with CUDA 11.3, respectively.

Imagen, the baseline model, has a structure that enhances resolution by employing the image generated by the diffusion model and the text encoding value as conditions. However, in this study, we focused on generating images without increasing resolution due to considerations related to time and computational costs. To achieve this, we applied the same training details as the baseline model.

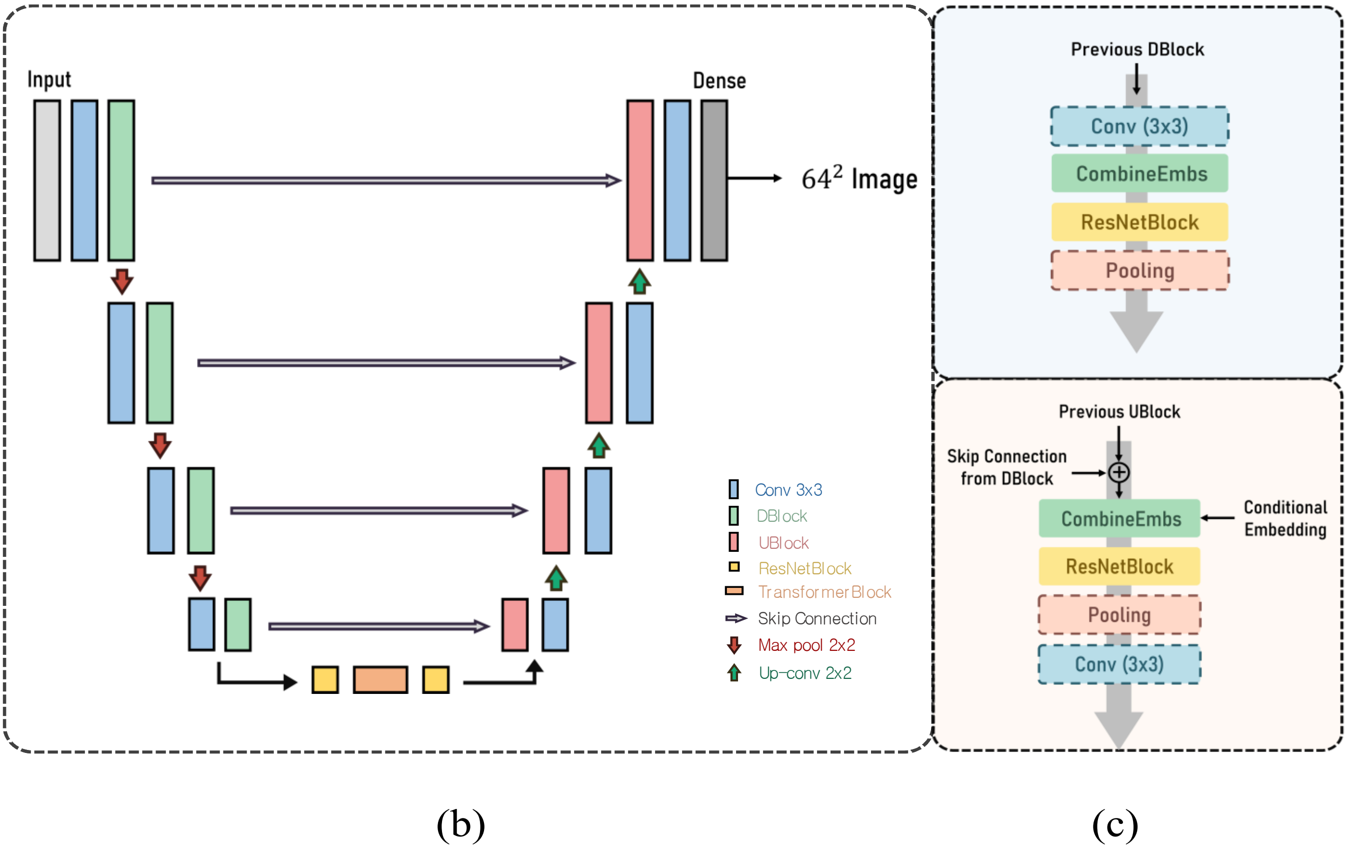

Fig 2. This is the improved U-Net structure used in the diffusion model. (b) represents the overall U-Net structure, which is a conditional diffusion model that maps text embeddings into 64x64 images. (c) shows DBlock and UBlock of efficient U-Net, respectively. The dashed blocks for the down-sampling block (DBlock) and up-sampling block (UBlock) are optional components.

IV. EXPERIMENTS

In this section, we evaluate the performance of PoolImagen on a datasets: Caltech-UCSD Birds-200 (CUB) [26]. CUB datasets comprises 11,788 images representing 200 different bird species. Each image is annotated with attributes describing the color and shape of the bird. We utilized these annotations to create multiple-sentence text captions that provide textual descriptions for each image.

For our experiments, we generated 64x64 pixel images for all datasets and evaluated their quality utilizing the Fréchet Inception Distance (FID) [27] along various dimensions. In addition, we examined the relationship between the training speed and the FID scores of the synthetic images.

A. Baseline models

Since the proposed model is a modification of Imagen [13], Imagen was chosen as the baseline for the experiments. In order to demonstrate that the improvements made in PoolImagen were not caused by unrelated changes, two important criteria were considered. Firstly, Imagen was trained using the google/t5-base-xxl [24] text encoder, which comprises approximately 18 million parameters. In contrast, the google/flan-t5-xxl [25] used in PoolImagen has approximately 25 million parameters. This emphasizes the significance of employing a large text encoder when generating text-based images. Second, Imagen utilized self-attention, a commonly used component in the transformer architecture. PoolImagen, with its modified transformer structure utilizing pooling, maintained performance while addressing the issue of computational complexity. This choice of the base model aimed to emphasize the significance of addressing computational complexity while obtaining performance gains through pooling.

B. Quantitative evaluation

We evaluate the Fréchet Inception Distance (FID) [27] between randomly sampled real images and generated images, which is one of the conventional methodologies to assess generative models. We provide two evaluations to compare the generated quality and training time for each dataset and each caption. We evaluate randomly conditioned images with various image dimensions and datasets by employing FID.

Figure 3 indicates that PoolImagen performs similarly to Imagen in FID relative to training time, but with much shorter training times. We demonstrate that the use of a pooling structure in PoolImagen improves training time performance compared to Imagen. For example, in the case of generating 16-dimensional CUB bird images, PoolImagen shows an FID value of 181.24, which is approximately 11.29% different from the baseline model, while reducing the training time to 8 days, which is 2.25 times quicker. Comparing training durations, the three models are approximately 1.68 times, 2.25 times, and 2.8 times, respectively, faster. Moreover, FID and training time perform better in 8-dimensions. In conclusion, despite the fact that FID may imply relatively inferior performance on the CUB dataset, performance of PoolImagen is successfully proven when compared to training time.

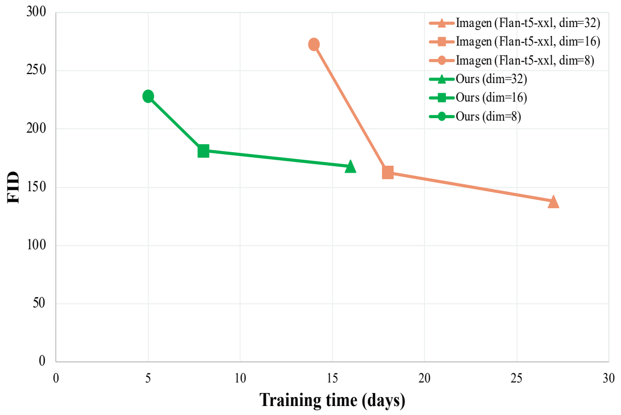

Fig 3. Quantitative comparison with FID-Training time with CUB dataset. This is a comparison of the FID-Training time of the base model and our proposed model, PoolImagen, using CUB dataset. The results are shown according to the size of each dimension.(i.e., 32, 16, 8)

C. Evaluation with generated Images

In the experiments, we trained PoolImagen using the CUB datasets. The architecture of PoolImagen was kept the same while specific details were altered to evaluate its efficacy. To enhance the image quality, we set the timesteps to 100. As depicted in Figure 3, we experimentally demonstrated the image generation capability of PoolImagen, which demonstrated relatively superior FID results. In addition, Figure 4 represents 16-dimensional images derived from the CUB dataset, effectively capturing the features of the CUB training set. For instance on the right, visually matching images to the text such as "this large waterbird has a black crown and nape, long black bill, black wings, neck, chest, and belly, with various shades of yellow" were observed. This proves PoolImagen is capable of overcoming timing constraints without losing image quality. In conclusion, we effectively demonstrated that the model can represent visual features and generate images rapidly.

V. Conclusion

The study introduces PoolImagen, a refined variant of Imagen, a diffusion model in image generation. While Imagen has faced challenges with computational burden and prolonged training times, PoolImagen addresses these issues without compromising performance. Inspired by the MetaFormer model, PoolImagen incorporates a pooling mechanism in place of the attention module, reducing computational complexity and accelerating training times. This pooling operation efficiently extracts features, striking a balance between speed and performance.

Text encoding capabilities are enhanced by integrating Flan t5-xxl, an advanced version of the t5 model, improving text-to-image transformations. Empirical evaluations on the CUB bird dataset demonstrate compelling results. PoolImagen not only maintains but also improves Fréchet Inception Distance (FID) scores, outperforming Imagen by approximately 11.29% in FID while reducing training time by a factor of 2.25.

In summary, PoolImagen represents a significant advancement in text-to-image synthesis, addressing computational and temporal limitations of Imagen. Architectural modifications and improved text encoding contribute to superior performance. Future research should explore further optimizations, recognizing PoolImagen's potential as an efficient tool for text-to-image synthesis and paving the way for refining and expanding diffusion models in this field.

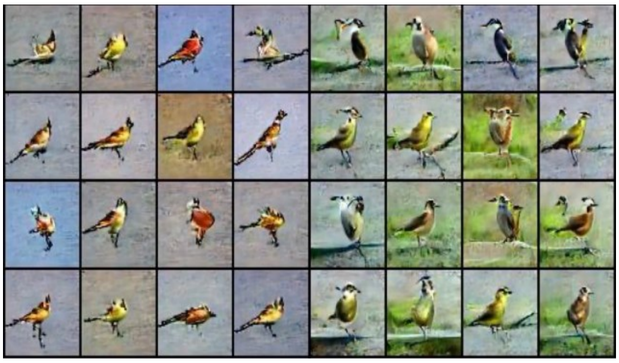

Fig 4. Generated 16×16 dimension images by PoolImagen using Cub dataset.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00251528), and Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0020967, Advanced Training Program for Smart Sensor Engineers)

References

[1] A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*, 2021: PMLR, pp. 8821-8831.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684-10695.

[3] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021: PMLR, pp. 8748-8763.

[4] C. Jia *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*, 2021: PMLR, pp. 4904-4916.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems,* vol. 33, pp. 6840-6851, 2020.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems,* vol. 34, pp. 8780-8794, 2021.

[7] A. Nichol *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided difmifusion models," *arXiv preprint arXiv:2112.10741,* 2021.

[8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image

generation with clip latents," *arXiv preprint arXiv:2204.06125,* vol. 1, no. 2, p. 3, 2022.

[9] T.-J. Fu, X. E. Wang, and W. Y. Wang, "Language-driven image style transfer," *arXiv preprint arXiv,* vol. 2106.00178, 2021.

[10] E. Spolaore and R. Wacziarg, "The diffusion of development," *The Quarterly journal of economics,* vol. 124, no. 2, pp. 469-529, 2009.

[11] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," *arXiv preprint arXiv:2302.03011,* 2023.

[12] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11461-11471.

[13] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems,* vol. 35, pp. 36479-36494, 2022.

[14] C. Saharia *et al.*, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1-10.

[15] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[16] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[17] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems,* vol. 34, pp. 15908-15919, 2021.

[18] W. Yu *et al.*, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10819-10829.

[19] Y. Li *et al.*, "Efficientformer: Vision transformers at mobilenet speed," *Advances in Neural Information Processing Systems,* vol. 35, pp. 12934-12949, 2022.

[20] H. Pashler, J. C. Johnston, and E. Ruthruff, "Attention and performance," *Annual review of psychology,* vol. 52, no. 1, pp. 629-651, 2001.

[21] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*, 2021: PMLR, pp. 8162-8171.

[22] A. Zafar *et al.*, "A comparison of pooling methods for convolutional neural networks," *Applied Sciences,* vol. 12, no. 17, p. 8643, 2022.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[24] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research,* vol. 21, no. 1, pp. 5485-5551, 2020.

[25] J. Wei *et al.*, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652,* 2021.

[26] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[27] 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," *Advances in neural information processing systems,* vol. 30, 2017.

[28] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096,* 2018.

[29] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502,* 2020.

[30] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," *Advances in Neural Information Processing Systems,* vol. 34, pp. 11287-11302, 2021.

[31] G. Huang, D. Chen, T. Li, F. Wu, L. Van Der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," *arXiv preprint arXiv:1703.09844,* 2017.

[32] T. Wang *et al.*, "DisCo: Disentangled Control for Referring Human Dance Generation in Real World," *arXiv preprint arXiv:2307.00040,* 2023.

[33] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012-10022.

[34] H. Touvron *et al.*, "Resmlp: Feedforward networks for image classification with data-efficient training," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 4, pp. 5314-5321, 2022.

[35] I. O. Tolstikhin *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Advances in neural information processing systems,* vol. 34, pp. 24261-24272, 2021.

[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog,* vol. 1, no. 8, p. 9, 2019.

[37] J. Kaplan *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361,* 2020.

[38] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917,* 2022.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.

[40] 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.

[41] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging,* vol. 3, no. 1, pp. 47-57, 2016.