Proceedings of CONF-MPCS 2024 Workshop: Quantum Machine Learning: Bridging Quantum Physics and Computational Simulations

DOI: 10.54254/2753-8818/52/2024CH0129

Performance Comparison of ControlNet Models Based on PONY in Complex Human Pose Image Generation

Qinyu Zeng

School of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China

162150121@nuaa.edu.cn

Abstract. Over the past two years, text-to-image diffusion models have advanced considerably. The PONY model, in particular, excels at generating high-quality anime character images from open-domain text descriptions. However, such text descriptions often lack the granularity needed for detailed control, especially in the context of complex human pose generation. To mitigate this limitation, recent research has introduced ControlNet to enhance the control capabilities of stable diffusion models. Nevertheless, the efficacy of a single model remains suboptimal for generating complex poses, highlighting the potential of combining multiple ControlNet models. This paper introduces the Depth+OpenPose methodology, a multi-ControlNet approach that enables simultaneous local control of depth maps and pose maps, in addition to other global controls. Distinct from single or other combined methods, Depth+OpenPose incorporates an additional conditional input. For addressing limb occlusion issues, depth maps provide positional relationships, while OpenPose captures facial expressions and hand poses, surpassing the performance of single models. Furthermore, Depth+OpenPose demonstrates superior speed and quality relative to other combinations. It is crucial to note that an excessive number of combinations can lead to too many conditional inputs, thereby reducing control efficacy. Through comprehensive quantitative and qualitative experimental comparisons, Depth+OpenPose proves its superiority in terms of speed, image quality, and versatility over existing methodologies.

Keywords: ControlNet, Stable Diffusion, Image Generation, Complex Human Posture, PONY.

1. Introduction

Since the stable diffusion (SD) [1-4] model has shown excellent performance in image generation and has attracted widespread attention, text-to-image diffusion models have become a popular choice for synthesizing high-quality images based on text input. From version 1.5 of the SD model to the stable diffusion XL [2] version (SDXL) of the SD model, the generation quality of the model has been improved. Among them, the most popular one in the past year is a fine-tuned parameter model of the SDXL version: the PONY model.

These models face challenges in understanding complex text and generating corresponding images [3,4], so it is necessary to enable more control models beyond text descriptions. The most popular solution in this regard is to use the ControlNet network [4], which allows the above models (such as the SD model) to use different local controls (such as depth maps, canny [4,5] maps, etc.) to generate images.

DOI: 10.54254/2753-8818/52/2024CH0129

This paper introduces the depth+openpose combination method, which shows the best results in generating images with complex human poses. By targeting the above problems, this paper proposes a method of combining the ControlNet model based on the PONY model. By combining the Depth and OpenPose models, this method can improve the image generation quality while ensuring the generation speed. Specifically, the Depth model is responsible for providing the depth information of the image [4,6], ensuring the edge control and detail consistency of the generated image. The OpenPose model provides the pose information of the character [4,7], making the generated image more natural and accurate in pose. Experimental results show that the Depth+OpenPose combined model can not only effectively avoid the defects of a single model when generating complex pose images, but also significantly improve the overall quality and generation speed of the image. This method is not only suitable for the generation of single-person poses, but also shows good adaptability in complex pose scenes of multiple people. The average generation time is 3 minutes, which meets the needs of practical applications. Through experimental verification, the method proposed in this paper has broad application prospects in design development, video production and other fields [8,9].

2. Method

2.1. Stable Diffusion、SDXL、PONY

Stable Diffusion is a variant of the diffusion model, which consists of three parts: VAE, U-Net, and text compiler [1,2]. VAE is trained to transform images into low-dimensional latent representations, add and remove Gaussian noise to the latent representation, and then decode the final denoising process into pixel space. In the forward diffusion process, Gaussian noise is iteratively applied to the compressed latent representation. Each denoising process is completed by the U-Net [1,4] architecture containing the ResNet backbone, and the latent representation is obtained by denoising in the reverse direction through forward diffusion. Finally, the VAE decoder converts the latent representation back to the pixel space to generate the output image.

The SDXL model adds a Refiner operation on the basis of SD [2]. In short, it can automatically optimize images, improve image quality and clarity, and reduce the need for manual intervention. The PONY model is a popular parameter fine-tuning model of the SDXL model.

2.2. ControlNet

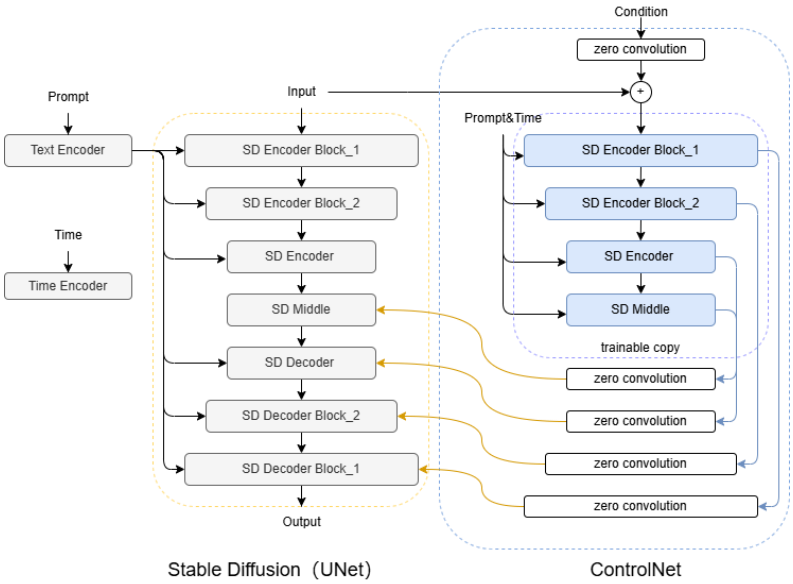

Figure 1. ControlNet Internal Architecture [4].

DOI: 10.54254/2753-8818/52/2024CH0129

ControlNet plays a guiding role in the denoising process of the SD model. As shown in Figure 1, all parameters in the UNet of StableDiffusion are locked and cloned into the trainable copy of the ControlNet side [4], and then the copy is trained using the external condition vector. ControlNet introduces a zero convolution layer and uses conditional information (such as depth map, canny map, posture map, etc.) to guide the UNet denoising process, thereby affecting the final image generation result of the SD model.

2.3. Principle of the method

In order to solve the problem of image conditional control generation under complex human postures based on the PONY model, this paper proposes a method of combining multiple ControlNet networks of Depth and Openpose for image generation under complex human postures. The whole method contains multiple modules as shown in Figure 2, including diffusion process, posture and depth preprocessing, feature fusion and conditional control. Different from the general text image, in addition to the text, image, semantic segmentation map, etc. in the Conditioning part on the far right as conditional input, the ControlNet model part is also added as a control condition to guide the denoising process of UNet. Moreover, in the conditional part of ControlNet, not only one processor is used to process the input original image, but the conditional input is obtained by using the preprocessors of openpose and depth respectively. Then, after being fused by an M processor and retaining the features according to the hyperparameters, it is ensured that the pixel blocks in the Denoising Process are the same size. Finally, it is sent to the zero convolution layer in the corresponding model of ControlNet to guide UNet denoising, and denoising is guided iteratively in the T time step. After multiple rounds of denoising, the final potential representation is obtained and the final image is obtained by entering the pixel space through the D processor.

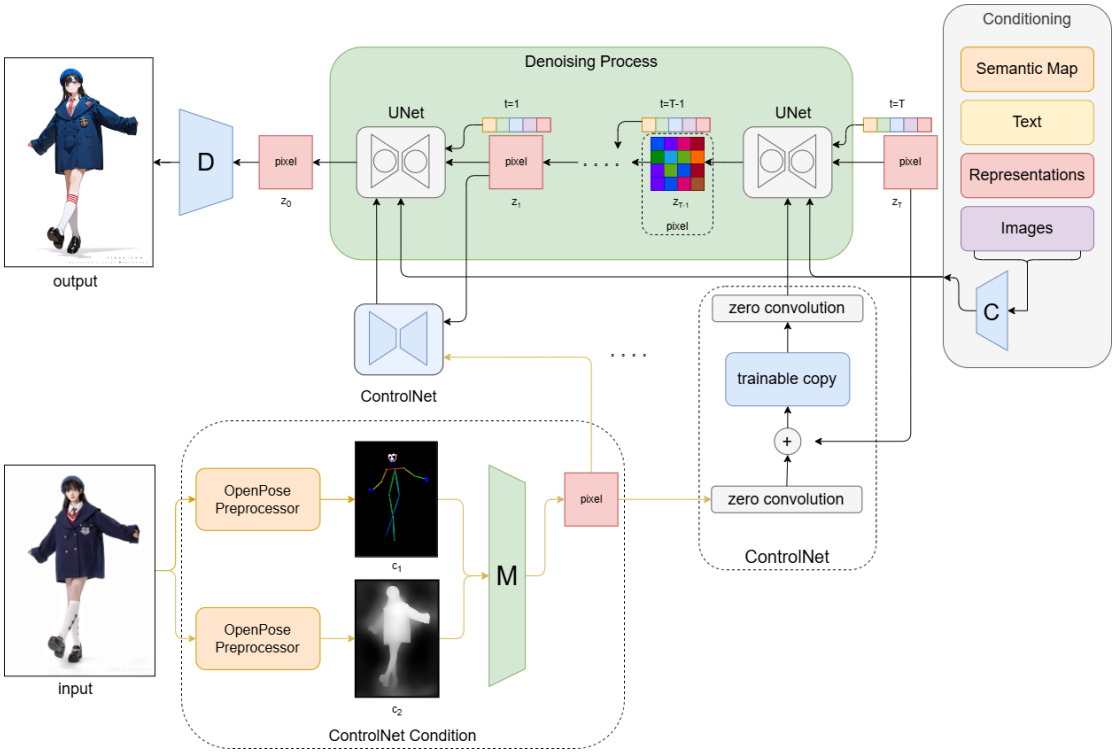

Figure 2. The overall framework of Depth+Openpose method.

DOI: 10.54254/2753-8818/52/2024CH0129

2.4. Evaluation Metrics

In order to quantitatively evaluate the advantages and disadvantages of the Depth+Openpose model compared with other models, the following evaluation indicators are given

2.4.1. Fréchet Inception Distance(FID). The below formula consists of two parts, r represents the real image, and g represents the generated image. In the first part, μ represents the mean of the distribution, and the first part is the square of the difference between the two means μ. In the second part, Σ represents the covariance, Tr represents the trace (the sum of the elements on the diagonal of the matrix), and the second part is the sum of the covariance matrices minus the trace of the product of the covariance matrices under the square root. FID calculates the distance between two distributions. The smaller the distance, the closer the generated distribution is to the real distribution [10-12].

$$FID = \|u_r - u_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_y\right)^{1/2}\right) \quad (1)$$

Expert evaluation. Since the quality of image generation is subjective and cannot be accurately judged by a single objective indicator, this paper also uses the questionnaire scores of 20 experts in the field of image design as indicators. The evaluation process is to provide the original image and multiple generated images of different models to the investigators, and let them give a score of 0-10 based on their own judgment. Finally, the average score of the generated images of each model is taken as the score.

3. Experiment

3.1. data set

The dataset used has two categories: simple half-body images and complex full-body images. The experimental data is shown in Figure 3.

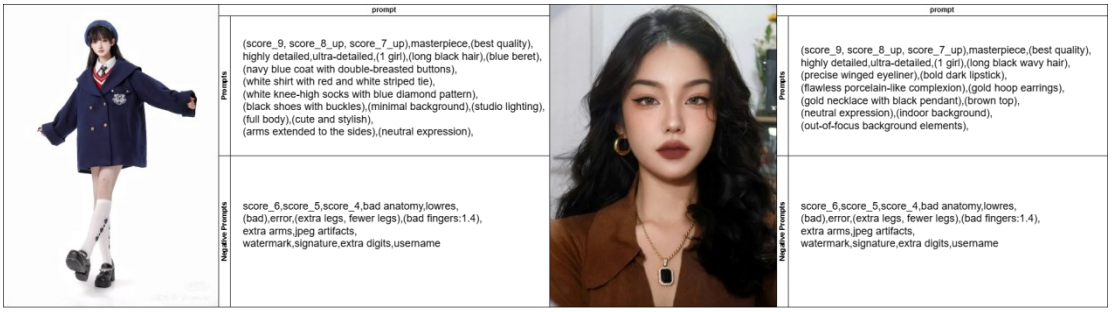

Figure 3. (Left) A full-body image of a complex person and its prompt, (right) a half-body image of a simple person and its prompt. Score_9, score_8_up and other similar prompt words are fixed semantic representations of the PONY model, and the rest are image features.

3.2. Experimental parameter design and hardware equipment environment

The specific experimental parameter design and equipment environment are shown in Table 1

Table 1. Experimental parameters and equipment environment.

Steps	Sampler	CFG Scale	Size	Model hash	Denosing strength	Clip skip	Graphics	Cuda
20	DPM++ 2M Karras	7	original	ac17f32d24	0.75	2	Nvidia 2060super	12.5.51

DOI: 10.54254/2753-8818/52/2024CH0129

3.3. Experimental Results

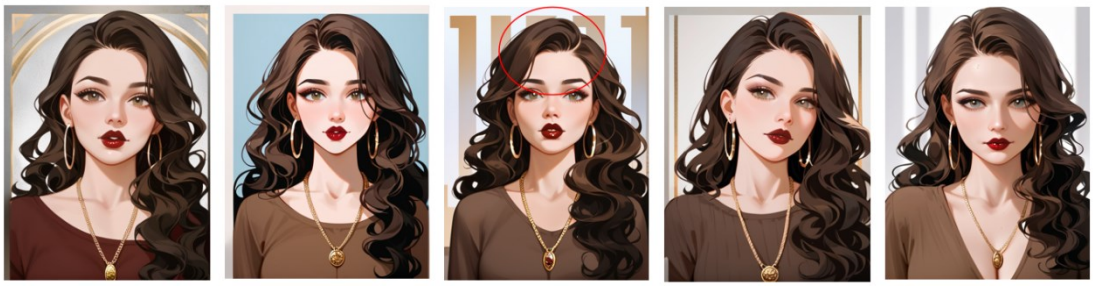

Figure 4. Simple figure half-length image generation effect of canny, depth, openpose, scribble, hed_softedge models.

Figure 5. Complex character full body image generation effect of canny, depth, openpose, scribble, hed_softedge models.

3.3.1. Simple half-length figure. The control generation effects of a simple half-body image under different controlnet models are shown in Figure 4: from left to right, they are the generation effects of the canny model, depth model, openpose model, scribble model, and hed_softedge model.
As can be seen from Figure 4, all models have good performance in terms of generation effect, except that the openpose model has a reversed hairstyle and some fine-tuning of facial expressions, but most of them can meet the requirements. Basically, for a simple half-body image, the SDXL version of controlnet can basically achieve good image generation.
These guidelines, written in the style of a submission, show the best layout for your paper using Microsoft Word. If you don't wish to use the Word template provided, please use the following page setup measurements.

3.3.2. Complex character poses. In general, the single model has different levels of defects in complex image generation, such as missing legs, missing clothes, loss of the original style of clothes, incorrect recognition of leg positions, etc. The generation effect of a single model is shown in Figure 5.

3.3.3. Depth+openpose method optimization. For complex human postures, this paper uses a combination of multiple controlnet networks of depth+openpose to achieve better quality results than a single model. The combination and comparison results can be seen in Figure 6. The upper left corner shows the generation effect of the OpenPose single model and the posture map, the lower left corner shows the generation effect of the Depth single model and the depth map, the middle is the combined generation effect, and the right is the original image. In the generation, the weights of depth and openpose are 0.5 respectively. In terms of guidance, denoising is coordinated with U-net from beginning to end.

DOI: 10.54254/2753-8818/52/2024CH0129

Figure 6. Depth+OpenPose method generation effect.Top left: OpenPose, bottom left: Depth, middle: generated, right: original.

3.4. Method versatility test

The proposed depth+openpose combination method is applied to various complex postures, even in the case of multiple people. It can be found that its performance is good and the generation speed is ideal, with an average time of 3 minutes. The generation effect is shown in Figure 7.

Figure 7. Other tests of Depth+OpenPose method.

3.5. Indicator evaluation

The Depth+OpenPose model excels in expert ratings while maintaining a good balance in both FID and generation time. Although some models perform well in certain aspects, such as the Depth model having the lowest FID and the Canny model having the shortest generation time, they fall short in other areas. Therefore, the Depth+OpenPose model is the best overall performer because it strikes the optimal balance between image quality and generation efficiency. This conclusion is drawn from Table 2.

Table 2. FID, Expert Score, and Generation Time of images generated by each Controlnet model.

Model	FID	Expert Score	Generation Time
Canny	114.741	6.36	**1min26s**
Depth	**71.333**	7.24	2min33s
Hed_Softedge	73.483	7.57	5min9s
Depth+OpenPose	84.788	**8.93**	3min17s
OpenPose	149.327	7.1	1min49s

DOI: 10.54254/2753-8818/52/2024CH0129

Table 2. (continued).

Scribble	132.011	6.8	1min37s
Hed_Softedge+OpenPose	119.577	8.07	18min15s
Canny+Depth	105.003	7.61	4min8s
Hed_Softedge+Depth+OpenPose	92.837	7.63	32min52s

4. Discussion

4.1. Analysis of single model image generation results

The generation effect of a single model is shown in Figures 5 and 8. From the generation effect of Figure 8 (divided into expression portrayal effect, limb overlap effect, and clothing consistency), under a single canny model, the most prominent problem is that the area with little difference in light and dark cannot be well obtained, resulting in incomplete generation of the overlapping part of the legs. In addition, in the animation style, the linear features of real people are maintained, and the face cannot restore the style of the model well. The Canny edge detection method mainly captures the edge information in the image. If the input edge image is of poor quality or inaccurate, the generated image may be affected. For example, if the edge detection result contains too much noise or false detection, the generated image may be distorted or contain unnecessary details. For complex character postures, the edge information may not be sufficient to fully describe all the details of the posture. Edge detection only provides the contour information of the image and lacks contextual information (such as color, texture, etc.), which may cause the loss of some details when generating images, especially when the complex texture and color of the original image need to be retained. The image generated by the depth map is much better than the canny model in the overlapping part of the limbs, but it is not good in facial expressions and clothing details. There are holes in the socks and the size of the face is not coordinated. The image generated by the posture map is very detailed in the depiction of facial expressions, but the clothing details are randomly generated according to the prompt words. The images generated by the graffiti pictures also rely on the prompt words in the details of the clothes. The clothes will have holes, and the generated legs are too thin and not coordinated. The soft edge generation has problems with the overlapping position of the legs, and it is impossible to identify the front and back positions of the legs. And the generation speed is very slow which is in Table 2.

4.2. Analysis of image generation results of combined model

Since generating complex human pose images requires more human pose information, the paper consider combining different information together. For example, the canny+depth model mentioned here uses not only the depth map but also the linear hard edge map to guide the generation. The effect is better than that of a single model, but it still has the defects of the canny model. As shown in Figure 9, there will be conflicts in recognizing real faces to anime styles. Compared with a single softedge model, the openpose model provides poses, which can obtain richer facial expressions and accurate pose performances. However, the generation speed will be slowed down by using two complex networks at the same time, and it will take more than 30 minutes to generate. This is a fatal defect for assembly line production, so it cannot be considered a good method. The combined openpose+depth model achieves perfect coordination between the generation effect and the generation speed. It can be seen that not only the depth map can well control the edge of the generated image to control the thickness of the legs and the consistency of some parts, but also the two models can control the position relationship of the legs, and the completeness of the clothes can also maintain a balance. The generation speed is also very fast. In addition, it is wrong to say that the more combined information, the better. Experiments have shown that even with multiple experiments using different weights, good results cannot be achieved. After analysis, it is because the original purpose of using controlnet is to not use a lot of dimensional information. If the information dimension is too high, a graph similar to the effect of a graph generated

DOI: 10.54254/2753-8818/52/2024CH0129

by a raw image will be produced, as shown in Figure 9, which means that the effect of controlnet on single simple information control will be lost.

Figure 8. Analysis of the defects of the single model in terms of facial expression portrayal, limb overlapping, and clothing consistency.

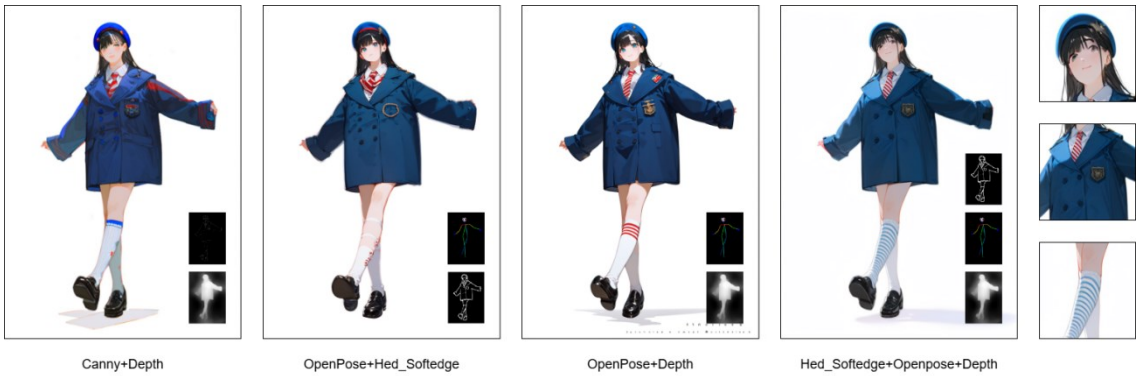

Figure 9. Analysis of composite model generation effect.

5. Conclusion

OpenPose+Depth has generalization and stability when controlling the PONY model to change the style of an image, especially when facing complex character poses and relationships. It can ensure that the image has no obvious errors and the speed is considerable. It can be well used in various fields such as design development and video production for pipeline generation. Secondly, when facing a simple half-length portrait, a single controlNet model can be used, while a combination of depth+openpose can be used for complex character pose images. At the same time, it is not recommended to use multiple controlNet models at the same time, which will cause the controlNet to lose the force of a single control condition, thus becoming a process similar to the image generation process.

In the future, it can be combined with other models to produce smooth videos and animations, which will require less manual modification, greatly reducing labor costs and time costs. With the corresponding lora model or clothing map, specific styles and clothes can be added to meet specific needs.

DOI: 10.54254/2753-8818/52/2024CH0129

References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10684-10695.

[2] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.

[3] Zhao, S., Chen, D., Chen, Y. C., Bao, J., Hao, S., Yuan, L., & Wong, K. Y. K. (2024). Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems, 36.

[4] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836-3847.

[5] Rong, W., Li, Z., Zhang, W., & Sun, L. (2014). An improved CANNY edge detection algorithm. In 2014 IEEE international conference on mechatronics and automation. 577-582.

[6] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10371-10381.

[7] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7291-7299.

[8] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022). Video diffusion models. Advances in Neural Information Processing Systems, 35, 8633-8646.

[9] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[10] Barratt, S., & Sharma, R. (2018). A note on the inception score. arXiv preprint arXiv:1801.01973.

[11] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.

[12] Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., & Ma, W. Y. (2019). Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6609-6618.