GENERATING 4D PLANT MODELS FOR VIRTUAL REALITY ENVIRONMENTS USING THE INSTANT NEURAL GRAPHICS PRIMITIVES AND STABLE DIFFUSION MODEL

ANQI HU[1], NOBUYOSHI YABUKI[2] and TOMOHIRO FUKUDA[3]
[1,2,3] *Division of Sustainable Energy and Environmental Engineering, Osaka University Graduate School of Engineering.*
1anqihu1028@gmail.com, 0000-0002-8568-1030
[2]yabuki@see.eng.osaka-u.ac.jp, 0000-0002-2944-4540
[3]fukuda.tomohiro.see.eng@osaka-u.ac.jp, 0000-0002-4271-4445

Abstract. This paper addresses the challenge of enhancing realism in virtual reality (VR) environmental design, particularly by overcoming the limitations of traditional 3D plant modeling methods that fail to capture dynamic and temporal nuances across annual seasons. The approach integrates realistic, time-sensitive modifications into VR plant modeling. It employs a methodology where source images for instant neural graphics primitives (Instant-ngp) are preprocessed using a Stable Diffusion model optimized by a Low-Rank Adaptation (LoRA) focusing on tree structures. This preprocessing step enriches Instant-ngp's input data, enabling the creation of 4D plant models that exhibit both spatial detail and temporal dynamics, mirroring natural seasonal variations. Stable Diffusion and LoRA are applied beforehand to improve the realism of the generated models. Virtual source trees are utilized for testing and refining the approach, aiming to enhance the representation of plant models in VR environments. This research contributes to making VR simulations more immersive and realistic, with potential applications in virtual landscaping, urban planning, and therapeutic environments. The study acknowledges the initial nature of this research and the ongoing need for exploration to fully realize these applications' potential.

Keywords. Neural Radiance Fields (NeRF), Diffusion Models, 4D Plant Modeling, Virtual Reality (VR), Environmental Design.

1. Introduction

The field of environmental design has experienced noteworthy advancements, particularly in the realm of virtual reality (VR), wherein the precise representation of natural elements is essential for captivating and immersive experiences. Conventional plant modeling methodologies, which are typically insufficient in capturing the temporal and dynamic aspects of natural environments, encounter restrictions, particularly in interactive VR environments.

ACCELERATED DESIGN, Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2024, Volume 3, 411-420.

This paper innovatively combines artificial intelligence and neural rendering techniques to address these issues. Key to this method is the combination of instant neural graphics primitives (Instant-ngp) (Müller et al., 2022) with a secure diffusion model (Rombach et al., 2022), enhanced by additive networks and ControlNet (L. Zhang et al., 2023), and with a special focus on the LoRA model (Hu et al., 2021) for image refinement. This revolutionary fusion has the potential to revolutionize VR plant modeling by adding a temporal aspect. Instant-ngp plays a crucial role in its rapid and effective neural rendering, which can convert 2D images into intricate 3D plant scenes that are the foundation of high-quality VR simulations. The technology's capacity to render intricate models in real time is critical for VR applications where responsiveness and detail are the cornerstones of user immersion. The model's temporal aspect is determined by stable diffusion models that process 2D plant images, endowing the models with both temporal and seasonal changes. Such models expertly handle noise and yield an array of visual outcomes, realistically imitating fluctuating environmental conditions - features that traditional methods have historically failed to assimilate.

In addition, the innovation of this study is demonstrated by the data set of plants specifically trained for the LoRA model, a step that greatly improves the quality and fidelity of the model. The effect is further enhanced by the selection of an appropriate ControlNet integration into the extended network. Through this collaboration, 4D plant models can be generated - 3D structures that change over time and mimic the dynamics of real plants. This method enables the creation of accurate plant models that can be utilized in various applications, including future scenarios such as virtual landscaping, urban planning, and therapeutic VR environments.

This method surpasses conventional modeling techniques by offering a more engaging and interactive VR environment design experience. The resulting 4D plant models not only enhance the visual realism of VR environments but also introduce elements of time-based variation and interactivity that were previously unachievable in VR plant modeling, marking a noteworthy advancement in the field.

2. Related works

2.1. TRADITIONAL PLANT MODELING METHODS

While traditional plant modeling methods such as procedural modeling (Talton et al., 2011), mesh modeling (Pavllo et al., 2021), L-systems (Ruoxi Sun et al., 2009) and point cloud modeling (Bournez et al., 2017) have been fundamental in representing vegetation, they have significant limitations in VR environments. These techniques often fail to capture the organic randomness and dynamic aspects essential for immersive VR realism. The labor-intensive nature of mesh modeling, the complexity of L-systems, and the high cost and extensive data processing of point cloud modeling render them inadequate for dynamic, large-scale ecosystems in VR, creating a gap in achieving a realistic and interactive virtual representation of natural environments.

2.2. NEURAL RENDERING TECHNOLOGIES

Neural rendering technologies, including neural radiance fields (NeRF) (Mildenhall et al., 2020) and Instant-ngp, have greatly enhanced the creation of 3D scenes and models from 2D images by merging computer vision with deep learning techniques to improve

realism and efficiency. NeRF utilizes a deep learning model to scrutinize a set of 2D images taken from varying angles and constructs a volumetric scene representation (Mildenhall et al., 2020). The NeRF model predicts the color and density of light in space, which enables the creation of new scene perspectives beyond the original images (Lin et al., 2022). Instant-ngp, created by NVIDIA Research, significantly reduces both the training and inference times of models such as NeRF (Müller et al., 2022). Additionally, an interactive GUI with a VR mode provides the ability to view neurographic primitives through a VR headset, thus enabling real-time applications, and amplifying its suitability for tasks such as scene reconstruction (Instant Neural Graphics Primitives, 2022/2023).

However, despite these advancements, a significant research gap remains. The challenge lies not just in harnessing these technologies for static scene reproduction, but in dynamically integrating them within VR environments for plant modeling. Current applications of NeRF and Instant-ngp primarily focus on static scenes, lacking the capability to simulate the temporal and seasonal variations intrinsic to plant life. Plants in the real world exhibit complex behaviors: they grow, the color of their leaves changes with the seasons, etc. Capturing these dynamic changes in VR requires the technical capabilities to render these changes in real time.

2.3. STABLE DIFFUSION MODEL AND COMPARISON WITH GENERATIVE ADVERSARIAL NETWORKS (GANS)

Stable diffusion models excel in generating realistic and diverse images (Borji, 2023), providing an alternative to GANs (Goodfellow et al., 2014), which sometimes suffer from mode collapse and low diversity. While stable diffusion models excel in variety and complexity, the research gap lies in the effective application of these models, along with addition networks, LoRA, and ControlNets, for 4D plant modeling in VR.

LoRA is chosen for its ability to adapt large models efficiently without extensive retraining, making it ideal for iterative VR modeling. ControlNets, on the other hand, offer superior control over the generated outputs, crucial for the precise depiction of plant seasonal changes. This combination not only enhances image quality but also introduces temporal dynamics into the modeling process. This integration is pivotal for creating immersive VR environments, where interactive elements and temporal changes are key.

3. Proposed Method

This study introduces a system designed to create dynamic, 4D representations of plants within a VR environment, capable of simulating seasonal changes.

3.1. THE MAIN PROCESS OF THE SYSTEM

The process involves several key steps, as shown in Figure 1:

1. Video Capture: A comprehensive video of the intended plant is recorded to ensure full coverage from every angle. Plants, within the scope of video production, can be categorized as either real or virtual. This is achieved by filming over a full 360-degree rotation at a consistent speed.

2. Frame Extraction: The video is fragmented into frames using FFmpeg (FFmpeg

Developers, 2016). The count of frames obtained is reliant on the length of the video, intending to secure over 200 frames to facilitate comprehensive scrutiny.

3. Image Processing and Analysis: 3.1.a. Preliminary Analysis for 3D Reconstruction: The initial phase of processing involves crucial steps like feature extraction, feature matching, and camera pose estimation. These steps are foundational for creating a 3D reconstruction of the plant. While this can be achieved using various software tools, one such example is COLMAP (Schonberger & Frahm, 2016). However, it's important to note that these processes are not exclusive to COLMAP and can be effectively conducted using other structure from motion (SfM) software. The choice of tool depends on specific project requirements and available resources, allowing flexibility in approach while adhering to the underlying principles of image processing and analysis

3.1.b. Creating 3D Scenes with Instant-ngp: These frames, along with the data derived from COLMAP, are input into Instant-ngp. This step generates a realistic 3D rendition of the plant, which can be explored in VR, offering real-time interaction and the ability to modify various aspects like lighting and viewpoint.

3.2.a. Seasonal Transformation via Stable Diffusion: Separately, frames are input into Stable Diffusion to generate images of the plant in different seasonal states. Techniques like LoRA in the additional networks are utilized for accurate seasonal depiction, while ControlNet maintains the plant's structural integrity and orientation relative to the background.

3.2.b. Integrating Seasonal Changes in 3D Scenes: The seasonally transformed images, combined with the data from the COLMAP analysis, are then processed through Instant-ngp. This step creates a detailed 3D scene reflecting the plant in various seasons.

4. Switching Scenes User Interface: The system comes with a web-based user interface (UI) crafted for this study. It enables the user to seamlessly switch between various seasonal VR plant representations, enhancing their interactive experience.

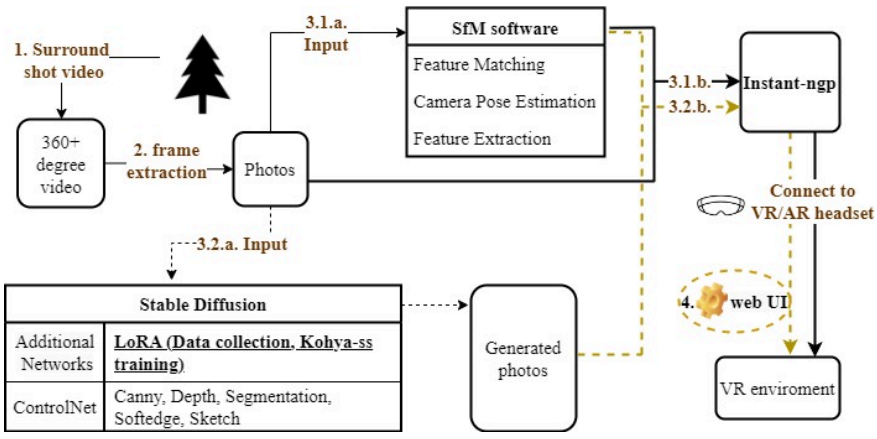

Figure 1. 4D generation system for plants in VR environments.

3.2. TRAINING FOR LORA

The system produces 4D plant models from 2D images, where temporal changes are integrated into 3D models through Stable Diffusion for seasonal transformations. The aim is to create desired outcomes that accurately reflect the text descriptions. LoRA's efficiency in training and low computational demands provide significant benefits. The first step is to prepare more than 50 images of plants, consistently styled and in moderate resolution, across seasons. These suitable images are adjusted to the standard 512x512 pixels due to their varying sizes. The BLIP neural network automatically labels these images, which are then stored in a text file. The Koyha_ss framework (bmaltais, 2022) is used for training due to its dataset compatibility and performance. The final model is selected based on analyzing the result plots from each LoRA model, with emphasis placed on their weights and effectiveness, rather than solely on loss results.

3.3. USAGE OF CONTROLNET IN STABLE DIFFUSION

ControlNet is an AI image generation plug-in that enhances standard image generation techniques by offering more accurate control.

The study employs three distinct models: the first emphasises edge detection and converts uploaded images to line drawings while preserving the composition in the newly generated image. The second model highlights depth and enables a more precise reproduction of the image's 3D structure. Finally, the system utilises a model to identify and categorise distinct parts of the image. By utilising these models, ControlNet can produce images that are abundant in detail and context.

4. Results

Software-generated plants were employed as targets to test the system's efficacy. Table 1 provides a comprehensive account of the equipment, software, and model versions utilised in this study, along with detailed parameters employed for the training of the LoRA model.

Table 1. Equipment software and parameters used.

PC spec for verifications: self-made		Parameters for LoRA training	
CPU	Intel Core i5-11400	LoRA type	Standard
GPU	NVIDIA GeForce RTX 3060 Ti	LR Scheduler	Cosine_with_re-starts
RAM	DIMM 16 GB	LR warmup (% of steps)	10
Analytical Model & Software		Optimizer	AdamW8bit
FFmpeg	Version: 5.1.2	Max resolution	512,512
COLMAP	Version: 3.7-win-dows-no-cuda	Network Rank and Alpha	128,128

Instant-ngp	RTX-3000-and-4000	Total steps	2000
Stable diffusion	Version ID: baf6946	Train batch size	1
Additional networks	Version ID: e9f3d62	Epoch	1
ControlNet	Version ID: 3011ff6	Regulatization factor	1
Kohya_ss	Version 2.0	Mixed_precision	fp16

Furthermore, the system is versatile enough to be used in both virtual and real environments. For the purposes of this study, virtual plants are used as a representative example. This approach is primarily due to the challenges associated with collecting comprehensive data from real plants across all seasons.

The selection of a summer maple for photographic data collection to train the LoRA algorithm was driven by its advantageous characteristics: the distinctive leaves and branching patterns of the summer maple provide clear, consistent data points ideal for algorithmic analysis. Its lush summer foliage ensures seasonal consistency, providing a rich data set. The tree's common presence and photogenic nature facilitate accessibility and high-quality image capture, critical for effective algorithm training. These contribute to a robust and diverse dataset, improving the accuracy and efficiency of the LoRA algorithm. Such data is imperative for generating precise and varied simulated models.

Figure 2 highlights the LoRA model's effectiveness within the Stable Diffusion framework, showcasing the target plant across different seasons—spring, autumn, late autumn, and winter. Table 2 lists all parameters applied, further illustrating the seasonal variations of plant imagery as discussed in section 3.2.b.

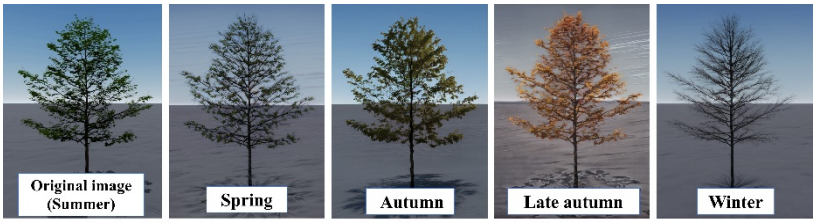

Figure 2. Stable Diffusion-generated plant photos for each season.

Table 2. Table of Parameters Used in the Stable Diffusion

Parameters			
Steps	26	Size	1920x1080
Denoising strength	0.33	Model hash	ca2e3bd9f9
CFG scale	10	Model	landscapeRealistic_v20WarmColor

Seed	3373133097	Sampler	DPM++ 2M SDE Karras
Spring Prompt	green leaf, bare tree, <lora:winter:0.2>		
Autumn Prompt	yellow leaf, <lora:maple_autu:0.6>		
LateAutumn Prompt	maple tree, <lora:maple_autu:0.6>		
Winter Prompt	bare tree, <lora:winter:0.6>		
Negative prompt	ground, background, trunks, branches, tree roots, small trees, people, other trees, classifier for paintings etc, background, building, trunks, branches, small trees, grass, mountains, lake, sloping land, kkw-Autumn		
ControlNet 0	Module: canny, Model: None, Weight: 1.2, Resize Mode: Resize and Fill, Low Vram: False, Processor Res: 512, Threshold A: 100, Threshold B: 200, Guidance Start: 0, Guidance End: 1, Pixel Perfect: False, Control Mode: ControlNet is more important		
ControlNet 1	Module: depth_midas, Model: None, Weight: 1, Resize Mode: Crop and Resize, Low Vram: False, Processor Res: 512, Guidance Start: 0, Guidance End: 1, Pixel Perfect: False, Control Mode: Balanced		
ControlNet 2	Module: seg_ofade20k, Model: None, Weight: 1.25, Resize Mode: Crop and Resize, Low Vram: False, Processor Res: 512, Guidance Start: 0, Guidance End: 1, Pixel Perfect: False, Control Mode: ControlNet is more important		

Figure 3 showcases results from three ControlNet models used in this study. The right image demonstrates the Canny model's ability to capture tree edges by converting images into line drawings, maintaining composition consistency. The center image shows the depth analysis model's accurate 3D tree structure replication via depth map extraction and layout reconstruction. This model also segments the image into tree and background pixels, as displayed in the right panel of Figure 3.

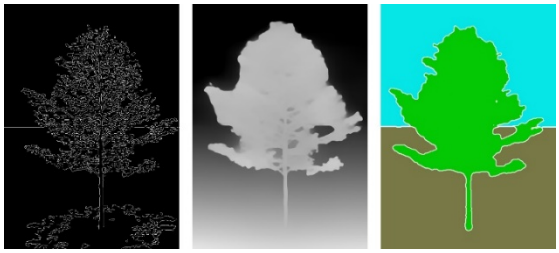

Figure 3. Comparative Results of ControlNet Models – Edge Detection, Depth Analysis, and Semantic Segmentation.

Table 3. Comparative analysis of image quality metrics.

	Spring	Autumn	Late autumn	Winter
PSNR	31.512	39.468	39.455	29.366
SSIM	0.952	0.965	0.848	0.733
LPIPS	0.457	0.398	0.432	0.589

The quality of these scenes is evaluated using peak signal to noise ratio (PSNR), structural similarity index (SSIM) (Hore & Ziou, 2010), and Learned Perceptual Image Patch Similarity (LPIPS) (R. Zhang et al., 2018) metrics, chosen for their detailed assessment over simpler methods. PSNR gauges perceptual errors, SSIM measures visual changes, and LPIPS evaluates nuanced perceptual differences via deep learning. The process starts with Instant-ngp converting 2D images into a 3D scene, focusing on light and color details. Reference frames set comparison standards, with Instant-ngp aligning frames to these benchmarks. PSNR identifies peak errors (higher values mean better quality), SSIM checks for visual similarities (closer to 1 is better), and LPIPS, using deep learning, prefers lower scores for closer original image resemblance. Random frame selection ensures unbiased evaluation across various scenes. Results show Instant-ngp's high-quality images, as illustrated in Table 3, with visual demonstrations of its capability to depict real seasonal changes in VR.

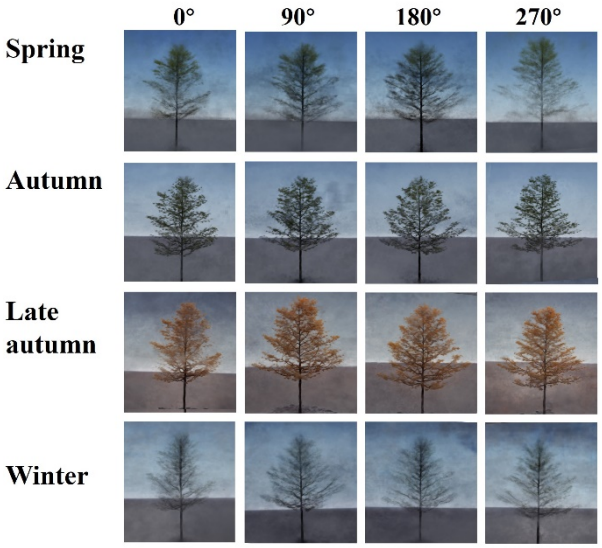

Figure 4. The state of the target plant in the VR environment in all seasons.

Furthermore, Figure 4 shows the target plants in Instant-ngp in different seasons as seen through a VR headset, further demonstrating the success of the system in creating real seasonal changes.

5. Discussion

The investigation into digital tree modeling within virtual environments has unveiled significant insights into the modeling process. Yet, it is recognized that such controlled conditions may not fully represent the complexities of natural environments, potentially impacting the applicability of these techniques to real trees. For example, the consistency of backgrounds that benefits the Instant-ngp process might not translate well to the varied backgrounds of real-world settings, potentially affecting model accuracy and effectiveness.

A notable advantage of employing a virtual original reference tree is its utility as a benchmark for evaluating the NeRF model. This approach allows for a direct comparison between the virtual reference and the NeRF models, offering a valuable method for assessing model fidelity. The current study's lack of this comparative analysis represents a limitation, suggesting future research should include detailed comparisons to enhance model validation and provide a more robust layer of validation for the modeling process.

While the system shows promise for 4D plant observation in VR, it faces some limitations, such as only being able to simulate a single plant at a time, which is not ideal when trying to simulate multiple plants at once in larger, more complex environments. These challenges highlight the need for future research that focuses on integrating semantic segmentation model enhancements to improve accuracy. In addition, there is a need to improve the accuracy of modeling plants in different seasons in VR environments to reduce uncertainty and increase the fidelity of seasonal transitions, thus enhancing the overall realism of the VR experience.

6. Conclusion

This research significantly improves 4D scene reduction and plant modeling in VR environments, by integrating NeRF with stable diffusion techniques to capture temporal changes in 3D scenes. The key achievement is the enhanced ability to depict temporal changes in 3D scenes, resulting in more realistic and detailed VR simulations of plants across different time periods. The study's findings demonstrate a significant improvement in the depth and authenticity of plant modeling in VR environments. This progress not only offers more immersive and accurate representations of plant behavior but also has implications for broader scientific research and enhancing user experiences in VR. The study contributes to the development of more advanced and realistic VR simulations, offering potential benefits in fields such as botany, ecology, and interactive education.

Acknowledgments

This work was supported by JST SPRING, Grant Number JPMJSP2138.

References

Bmaltais. (n.d.). Kohya's GUI [Python]. Retrieved 1 September 2023, from https://github.com/bmaltais/kohya_ss (Original work published 2022)

Bournez, E., Landes, T., Saudreau, M., Kastendeuch, P., & Najjar, G. (2017). FROM TLS POINT CLOUDS TO 3D MODELS OF TREES:A COMPARISON OF EXISTING ALGORITHMS FOR 3D TREE RECONSTRUCTION. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W3, 113–120. https://doi.org/10.5194/isprs-archives-XLII-2-W3-113-2017

FFmpeg Developers. (2016). http://ffmpeg.org/

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets.

Hore, A., & Ziou, D. (2010). Image Quality Metrics: PSNR vs. SSIM. 2010 20th International Conference on Pattern Recognition, 2366–2369. https://doi.org/10.1109/ICPR.2010.579

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models (arXiv:2106.09685). arXiv. http://arxiv.org/abs/2106.09685

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis (arXiv:2003.08934). arXiv. http://arxiv.org/abs/2003.08934

Müller, T., Evans, A., Schied, C., & Keller, A. (2022). Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. ACM Transactions on Graphics, 41(4), 1–15. https://doi.org/10.1145/3528223.3530127

Pavllo, D., Kohler, J., Hofmann, T., & Lucchi, A. (2021). Learning Generative Models of Textured 3D Meshes from Real-World Images (arXiv:2103.15627). arXiv. http://arxiv.org/abs/2103.15627

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

Ruoxi Sun, Jinyuan Jia, & Jaeger, M. (2009). Intelligent tree modeling based on L-system. 2009 IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design, 1096–1100. https://doi.org/10.1109/CAIDCD.2009.5375256

Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-Motion Revisited. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4104–4113. https://doi.org/10.1109/CVPR.2016.445

Talton, J. O., Lou, Y., Lesser, S., Duke, J., Měch, R., & Koltun, V. (2011). Metropolis procedural modeling. ACM Transactions on Graphics, 30(2), 1–14. https://doi.org/10.1145/1944846.1944851

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models (arXiv:2302.05543). arXiv. http://arxiv.org/abs/2302.05543

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 586–595. https://doi.org/10.1109/CVPR.2018.00068