Enhancing ControlNet Performance in Image Synthesis using Edge and Color Information

Daniel Fleming

Paper ID

Abstract

Stable Diffusion models are powerful tools for image generation, but they can struggle to maintain image structure during the synthesis process. This work proposes a novel approach that leverages ControlNets to enhance the model's ability to generate images by leveraging both edge and color information. I achieve this by training the ControlNet on a dataset of colored edge maps. These edge maps are pre-generated from real images, preserving the original color information within the detected edges. By conditioning the ControlNet on these edge maps during training, we essentially teach it to learn and utilize both structural cues from the edges and color information embedded within them. This learned knowledge can then be applied during image synthesis, enabling the ControlNet to generate new images that are faithful to the underlying structure and color characteristics.

1 Introduction

Great progress has been made in image synthesis using deep-learning models in the past few years. Stable Diffusion models, in particular, have emerged as powerful tools for creating new images from latent codes or noise inputs. Most of these models, however, are plagued with the shortcoming of maintaining the structural integrity of the target image during the synthesis process. This can lead to blurry or distorted outputs that lack fidelity to the original image, particularly regarding the intricate details of object boundaries and shapes. One promising direction for overcoming this limitation is ControlNets. ControlNets belong to the family of generative models and focus on the conditional synthesis of images. These signals contain more information that guides the model in generating images in line with the features it desires.

This paper suggests a new direction to enhance the performance of ControlNet in the image synthesis task that focuses on the model's capability to capture and represent image structure. In this work, I use colored edge maps to harness this power. By incorporating color information within these edges, I can create a better representation that goes beyond simple structural cues. I hypothesize that if I train a ControlNet on a dataset of colored edge maps, it will obtain the ability to learn to use the color information coupled with the structural information to generate new images that are not only faithful to the overall structure of the target image but also retain the original color characteristics. My goal is to evaluate whether this approach using colored edge maps leads to a significant improvement in maintaining structural integrity and color integrity compared to the baseline Stable Diffusion model. I will evaluate the effectiveness of our approach using two key quantitative measures. I will assess structural similarity using Structural Similarity Index Measure (SSIM), and color fidelity using Color Mean Squared Error.

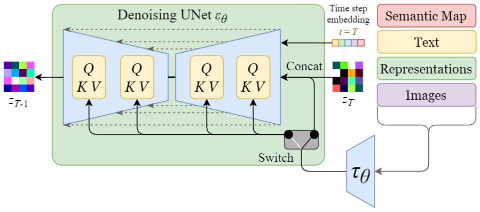
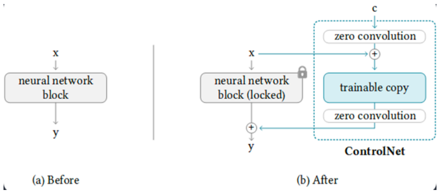


Figure 1: Stable Diffusion Process



(a) Before

(b) After

Figure 2: Visualization of ControlNet

2 Related Work

2.1 Stable Diffusion

Stable Diffusion is a recent advancement in text-to-image synthesis, generating photorealistic images from text-based descriptions [1]. It does so quite efficiently compared to previous models, which makes it runnable on common GPUs and thereby democratizes access to this powerful technology. Figure 1 depicts the Stable Diffusion process.

2.2 ControlNet

ControlNet is an architecture for neural networks designed to augment models like Stable Diffusion with spatial conditioning controls [2]. The model uses the pre-trained encoding layers of any existing diffusion model and then adds a trainable module, architected to receive any kind of conditioning input like depth maps, edge information, or human poses. This formalism provides a fine-grained process of control for image generation, maintaining the strength of the diffusion model underneath.

3 Data

3.1 Data Description

The CelebA dataset is a large-scale collection of celebrity face images [4]. It contains over 200,000 images that includes people from diverse ethnicities and a wide range of poses and expressions. The breadth of this dataset allows us to evaluate our method's ability to capture both structural details and color information.

4 Methodologies

4.1 Hypothesis

I hypothesize that color information, embedded within the edge maps used to condition the ControlNet, will allow the Stable Diffusion model to learn to use the color information in conjunction with structural information to generate new images that are faithful to the overall structure of the target image and also retain the original color characteristics.

4.2 Data Preprocessing

The CelebA dataset is pre-processed to generate a corresponding dataset of colored edge maps. I employ the Canny edge detection algorithm [3] to extract the structural information from the images. To preserve color information, I don't discard the original image data after edge detection. Instead, I use the edge map as a mask, assigning the original pixel color values to the corresponding edge pixels and setting the remaining pixels to black. This results in colored edge maps that retain both structural cues and color information.

4.3 Model Architecture

I use the existing Stable Diffusion [1] model architecture and integrate a ControlNet [2] module. The ControlNet receives the pre-processed colored edge maps as conditioning input alongside the text prompt during the image generation process.

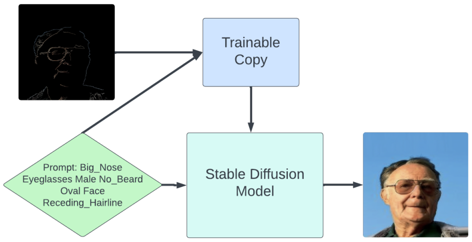

Trainable Copy
Prompt: Big_Nose Eyeglasses Male No_Beard Oval Face Receding_Hairline
Stable Diffusion Model

Figure 3: Enhanced Structural Fidelity in Stable Diffusion (ControlNet + Colored Edges)

Figure 3 llustrates the proposed method for enhancing structural fidelity in Stable Diffusion images using ControlNets conditioned on colored edge maps.

4.4 Evaluation Metrics

To evaluate the effectiveness of the approach I am using, I decided to use two quantitative benchmarks: Structural Similarity Index Measure (SSIM) [5] and Color Mean Squared Error (MSE)[6] scores achieved by this approach against a baseline Stable Diffusion model. This comparison will reveal to what extent ControlNet conditioned on colored edge maps influences the structural fidelity and color presrvation of its generated images. The SSIM metric ranges from -1 (perfect dissimilarity) to 1 (perfect similarity), while MSE typically falls within a non-negative range from 0 (same image) to positive infinity.

4.4.1 SSIM

This metric assesses the perceived structural similarity between the generated image and the ground truth image, considering luminance, contrast, and structure. SSIM [5] is a well-established metric in image quality assessment, and its focus on these aspects aligns with my goal of improving structural fidelity in the generated images.

4.4.2 Color MSE

This metric measures the average squared difference between corresponding pixels in the generated and ground truth images across all color channels (RGB). Since our approach incorporates color information in the edge maps, Color MSE [6] serves as a suitable metric to evaluate how well the generated images preserve the original color characteristics.

4.5 Training

The Controlnet requires a dataset with the following schema: (original image, conditioned image, prompt). I trained the Stable Diffusion model integrated with the ControlNet module this where the original image was a high resolution image from the CelebA dataset, the conditioned image was a colored edge map derived from the corresponding original image using the Canny edge detection algorithm [3], and the prompt was the list of facial attributes each image had which was also provided by the CelebA dataset. I used an Adam optimizer for training since it's a common choice for deep learning models [7]. I utilized a batch size of 4, used a learning rate of 1e-5, and trained the model for a maximum of 5 epochs on a single RTX 6000 Ada GPU with 32-bit precision on the Runpod platform. A subset of 5000 of the 200,000 images were used as the training set. PyTorch served as the deep learning framework for model implementation and training.

5 Results

I evaluated the effectiveness of ControlNet conditioning on the preservation of structural fidelity and color characteristics in image generation. The model's performance was assessed using the Structural Similarity Index Measure (SSIM)[5] and Color Mean Squared Error (MSE)[6] on a validation set of 1000 images. I compared the scores achieved by the model with ControlNet against a baseline model consisting of Stable Diffusion without ControlNet conditioning. Table 1 contains the respective SSID and MSE scores for both models.

Table 1: Comparison of ControlNet and Stable Diffusion on SSIM and MSE

Metric	**ControlNet**	**SD**
SSIM	0.5429 ± 0.0942	0.5188 ± 0.0231
MSE	0.1146 ± 0.0439	$0{,}1029 \pm 0{,}0403$

6 Evaluation

I can't definitively say that ControlNet is better than the baseline, just by looking at the SSIM and MSE scores. The calculated value for the similarity index is 0.5429 for ControlNet and 0.5188 for the baseline. A paired-sample t-test [9] shows that there is no significant difference (p-value > 0.05) between the SSID of ControlNet and the baseline. With MSE, the values do hint at possibly lower color accuracy for ControlNet than for the Baseline (0.1146 versus 0.1029), but the paired-sample t-test [9] again did not show any statistically significant difference (p-value > 0.05).

7 Discussion

Following the evaluation, this calls for a further discussion going into the potential reasons for the results obtained, leading to inconclusive results, and an outline of future directions of investigation.Though SSIM and MSE are useful in offering insights into the structural similarity and color accuracy of images, these signify just a minimal aspect of quality in images. These metrics sometimes can't bear the full load in the evaluation of the images because they dont capture the full spectrum of human perception. To gain a deeper understanding of ControlNet, future work could involve an ablation study [8] to gain a deeper understanding of ControlNet's influence.

8 Conclusion

This study investigated ControlNet conditioning for better image fidelity. While ControlNet did have a higher mean SSIM compared to the baseline, the difference was within the standard deviation. Inconclusive results from SSIM and MSE show limitations in just relying on these metrics. Further exploration with human studies and other metrics is needed for a more comprehensive understanding of ControlNet's impact on image quality.

9 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 (pp. 10684-10695).

[2] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv preprint arXiv:2302.05543.

[3] Canny, J. (1986). A Computational Approach To Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(6), 679-698.

[4] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. Retrieved from `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`

[5] 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.

[6] Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. IEEE Signal Processing Magazine, 26(1), 98-117.

[7] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[8] Meyes, R., Lu, M., Waubert de Puiseau, C., & Meisen, T. (2019). Ablation Studies in Artificial Neural Networks. Retrieved from `https://arxiv.org/abs/1901.08644`.

[9] Gosset, W. S. (1908). "The Probable Error of a Mean." Biometrika, 6(1), 1-25.

[10] Nilsson, J., & Akenine-Möller, T. (2020). Understanding SSIM. arXiv preprint arXiv:2006.13846. Available at: `https://arxiv.org/abs/2006.13846`