

# Abstract Art Interpretation Using ControlNet

Rishabh Srivastava  
Columbia University  
New York  
rs4489@columbia.edu

Addrish Roy  
Columbia University  
New York  
ar4613@columbia.edu

## Abstract

*Our study delves into the fusion of abstract art interpretation and text-to-image synthesis, addressing the challenge of achieving precise spatial control over image composition solely through textual prompts. Leveraging the capabilities of ControlNet, we empower users with finer control over the synthesis process, enabling enhanced manipulation of synthesized imagery. Inspired by the minimalist forms found in abstract artworks, we introduce a novel condition crafted from geometric primitives such as triangles.*

## 1. Introduction

Visual inspiration often strikes unexpectedly, prompting a desire to immortalize fleeting mental imagery into tangible forms. With the advent of text-to-image diffusion models [14], such aspirations have become increasingly attainable through the simple act of textual description. However, the inherent limitations of these models in providing precise spatial control over image composition persist, presenting challenges in accurately conveying complex layouts, poses, shapes, and forms solely through textual prompts.

While existing methodologies primarily rely on textual prompts to guide image generation, the integration of additional image conditions has emerged as a promising approach to empower users with finer control over the synthesis process. One major breakthrough in this domain was the ControlNet architecture [21]. ControlNet revolutionized the field by enabling the integration of diverse conditioning inputs with Stable Diffusion, ranging from edge maps and human pose skeletons to segmentation maps and depth information. By treating these additional images as conditions for the image generation process, ControlNet paved the way for enhanced spatial control and nuanced manipulation of synthesized imagery.

Building upon the foundation laid by the ControlNet architecture introduced in the previous year, our project focuses on training ControlNet with a new condition. Inspired by the simplicity and versatility of geometric shapes in ab-

stract artworks, our new condition is crafted from geometric primitives like triangles. These primitives are meticulously arranged to approximate the original image in the training dataset, yet imbue it with the essential qualities of abstract artistry.

The concept stems from the recognition that abstract art often relies solely on shapes to convey its message, emphasizing the power of minimalist forms in eliciting diverse interpretations. Such artistry, characterized by its reliance on geometric elements, serves as the foundation for our new condition image. Moreover, the subjective nature of human imagination further fuels this endeavor, as diverse textual prompts can yield a myriad of representations from the same abstract artwork. Thus, our pursuit is grounded in the rich tapestry of abstract expressionism, where geometric shapes serve as the catalyst for exploring the boundless realm of artistic interpretation and creativity.

In summary, our project presents a pioneering exploration into the intersection of abstract art interpretation and text-to-image synthesis. We aim to train a novel condition and conduct qualitative assessments of the outcomes, condition fidelity, and the overall quality of the generated images.

## 2. Related Work

Diffusion models belong to a category of probabilistic generative models. They operate by gradually altering data through noise injection and subsequently learn the ability to reverse this process for generating samples. Sohl-Dickstein and colleagues pioneered Image Diffusion Models [16], a concept subsequently adapted to image generation [3]. Latent Diffusion Models (LDM) [14] streamline computation by conducting diffusion steps in the latent image space [4]. Text-to-image diffusion models, utilizing pre-trained language models such as CLIP [11], excel in generating images by translating textual inputs into latent vectors. Stable Diffusion [20], a large-scale implementation of latent diffusion, is notable in the field.

Controlling Image Diffusion Models empowers users with the ability to personalize, customize, and generate images tailored to specific tasks. Methods for guiding image

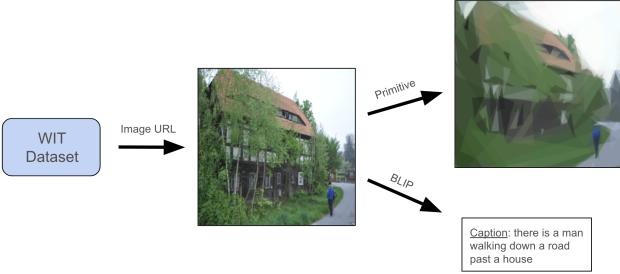


Figure 1. Representation of our dataset generation pipeline.

generation through text focus on adjusting prompts, manipulating CLIP features, and modifying cross-attention mechanisms [12]. Approaches like MakeAScene [7] encode segmentation masks into tokens to control images, while GLIGEN [10] adapts attention layers of diffusion models to learn new parameters for grounded generation. Personalization techniques such as Textual Inversion [8] and DreamBooth [15] fine-tune image diffusion models using user-provided example images to customize content in generated images. Tools for prompt-based image editing [1] offer practical ways to manipulate images with textual prompts.

In image-to-image translation, an image from one domain is transformed into an image from another domain. Conditional Generative Adversarial Networks (GANs) [2] and transformers [5] have demonstrated their capacity to learn mappings between disparate image domains. Manipulating pre-trained GANs allows for addressing specific image-to-image tasks; for example, StyleGANs can be controlled using additional encoders [13].

Unlike image diffusion models that only rely on text prompts for image generation, ControlNet [21] extends the capabilities of pre-trained large diffusion models to incorporate additional semantic maps, such as edge maps, segmentation maps, key points, shape normals, and depth cues. We delve further into the ControlNet architecture in Section 3.2.

### 3. Methodology

#### 3.1. Dataset Preparation

The dataset utilized in this study was meticulously curated from a subset of the Wikipedia-based Image Text (WIT) Dataset [17], a colossal multimodal multilingual repository comprising an astounding 37.6 million image-text examples.

To construct our dataset, we utilized the image URLs sourced from the WIT dataset to download the original images. These images served as the target images when we trained our ControlNet model. While the WIT dataset offers image descriptions, their specificity and granularity were not ideally suited for training our model. Thus, cap-

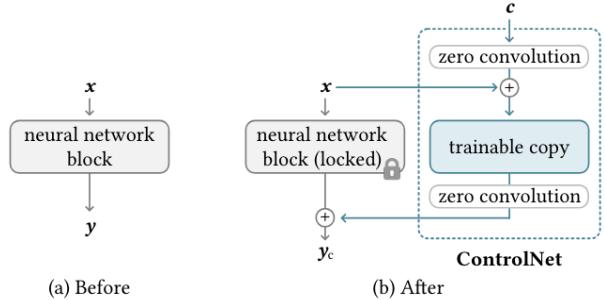


Figure 2. To incorporate a ControlNet into the block illustrated in (a), the original block is locked and a trainable duplicate is generated as the first step. These two blocks are then linked using zero convolution layers, specifically employing  $1 \times 1$  convolutions with weight and bias parameters set to 0. Here, the conditioning vector  $c$  represents additional information we aim to integrate into the network, as depicted in (b).

tions for the images were generated using Hugging Face’s implementation of the Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BLIP) model [9].

In tandem with obtaining images and captions, control images were derived from the downloaded images using the innovative Primitive software [6]. Employing an iterative methodology, Primitive progressively refines a target image by adding geometric shapes to minimize the disparity between the target and the drawn image. This iterative refinement process continues until a visually appealing yet abstract representation is achieved, imbuing the dataset with a nuanced understanding of artistic abstraction and creativity. The current dataset uses 50 triangles to approximate the target image.

In summary, our dataset [18] comprises 14,279 pairs of control and target images, each accompanied by meticulously crafted captions. By leveraging the inherent variability and complexity within the dataset, ControlNet can be trained to adeptly manipulate stable diffusion images based on newly introduced conditions.

#### 3.2. ControlNet Architecture

Fig. 2 above helps to visualize the setup for ControlNet where we can see that ControlNet integrates additional conditions into the main neural network block (for example, it can be resnet block, multi-head attention block, transformer block, etc.). Consider the trained neural network block to be represented by  $F(\cdot; \Theta)$  where the  $\Theta$  refers to the parameters of the block. Let,  $x$  &  $y$  denote the input & output of this block. Then,

$$y = F(x; \Theta) \quad (1)$$

To inject a ControlNet into this block, we lock the parameters  $\Theta$  of the pre-trained original block and duplicate the

block to create a trainable copy with parameters  $\Theta_c$ . The input to the trainable copy is a conditioning vector,  $c$  which comes from the outside. Let  $Z(\cdot; \cdot)$  denote the zero convolution layers that connect the trainable copy to the locked model. Two instances of such zero convolutions are used with parameters,  $\Theta_{z1}$  &  $\Theta_{z2}$  respectively. Then, the output of the ControlNet can be represented by

$$y_c = F(x; \Theta) + Z(F(x + Z(c; \Theta_{z1}); \Theta_c); \Theta_{z2}) \quad (2)$$

After the 1st training step, the above equation (2) reduces to the equation below:

$$\mathbf{y}_c = \mathbf{y} \quad (3)$$

Thus, when the training starts harmful noise cannot impact the hidden states of the network block in the trainable copy. Additionally, given that  $Z(c; \Theta_{z1}) = 0$  and the trainable copy also takes the input image  $x$ , it remains fully operational and preserves the capabilities of the extensive, pre-trained model. This enables it to function as a robust foundation for continued learning.

### 3.3. Training

Once the dataset is prepared, the next step is to load the pre-trained model. The model architecture is defined in a YAML configuration file, encapsulating its structure and parameters. Loading the pre-trained model weights initializes the model with prior knowledge learned from extensive training on large datasets. This initialization kickstarts the model's ability to generate meaningful images based on textual input.

With the dataset and model in place, the training process is initialized. This involves setting up a DataLoader to efficiently feed batches of data into the model during training. The DataLoader handles tasks such as batching, shuffling, and parallel data loading, optimizing the training workflow.

The model is then trained to learn to generate images by iteratively processing image-text pairs and optimizing its parameters to minimize the discrepancy between generated and ground truth images like in Stable Diffusion. During each iteration, the model adjusts its internal representations, gradually improving its ability to translate textual prompts into realistic images. The use of GPU acceleration speeds up this process significantly, allowing for efficient exploration of the high-dimensional parameter space.

### 3.4. Inference

Once the ControlNet parameters are trained, the model becomes capable of generating new images. Inference becomes straightforward by inputting a sample abstract image along with a prompt. While inference can technically function without prompts, it is important to note that our model's performance without prompts may be sub-optimal. This limitation arises from the fact that our model has not been extensively trained on datasets lacking prompts.

## 4. Experimental Results

We applied the ControlNet architecture on Stable Diffusion and trained the resultant model using 14000 images from our dataset. The model was trained on an Nvidia T4 GPU, and a batch size of 2 was employed to ensure compliance with the memory constraints of the GCP Compute Engine instance. The code can be found at [19].

As stated in the ControlNet paper [21], we also observed the sudden convergence phenomenon (Fig. 3). As a result of zero convolutions, ControlNet consistently generates high-quality images throughout the training process. However, at a specific stage in training, the model undergoes a sudden learning event wherein it begins to adhere closely to the input condition, which is noted as the sudden convergence phenomenon.

Fig. 4 illustrates the results generated by our trained model for a given test image. This same image can be perceived in multiple ways, as evidenced by the diverse prompts showcased. Each prompt breathes life into its own distinct imaginative space, showcasing the richness and depth of creative interpretation.

## 5. Discussion

Successfully training a ControlNet model on a novel control image concept marks a significant achievement in our research endeavor. To facilitate this, we meticulously curated a custom dataset tailored to our specific requirements.

During the training process, our ControlNet exhibited remarkable adaptability and proficiency, effectively learning to generate diverse images in accordance with the provided prompts. Notably, it adeptly preserved the approximated object locations, underscoring its ability to maintain spatial consistency while exploring the creative possibilities inherent in the input conditions.

However, a notable limitation was observed: while the trained model exhibited proficiency in determining object placements and aligning with the desired prompt descriptions and input control image, it struggled to accurately replicate colors sometimes. This shortfall likely stems from the need for further training. Regrettably, due to constraints on compute resources, we were unable to conduct more extensive training sessions to address this deficiency.

## 6. Conclusion

In conclusion, our exploration delves into the confluence of abstract art interpretation and text-to-image synthesis, addressing the intrinsic challenges of achieving precise spatial control over image composition solely through textual prompts. By extending the capabilities of pre-trained large diffusion models, ControlNet empowers users with finer

Step	1650	3628	7002	<b>7361</b>	8861
Prompt	there is a building with a tower on top of it	purple flowers with yellow centers in a field of grass	there is a man running in a marathon race in front of a crowd	there are many fish swimming in the water together	there is a green bug that is on a white surface
Control Image					
Target Image					
Generated Image					

Figure 3. Sudden convergence phenomenon: ControlNet consistently produces high-quality images throughout training, with a marked instance (e.g., bolded step 7361) where it abruptly starts intersecting with the control image. Although the intersection of the generated image with the target image increases at step 7361, the colors do not match completely.

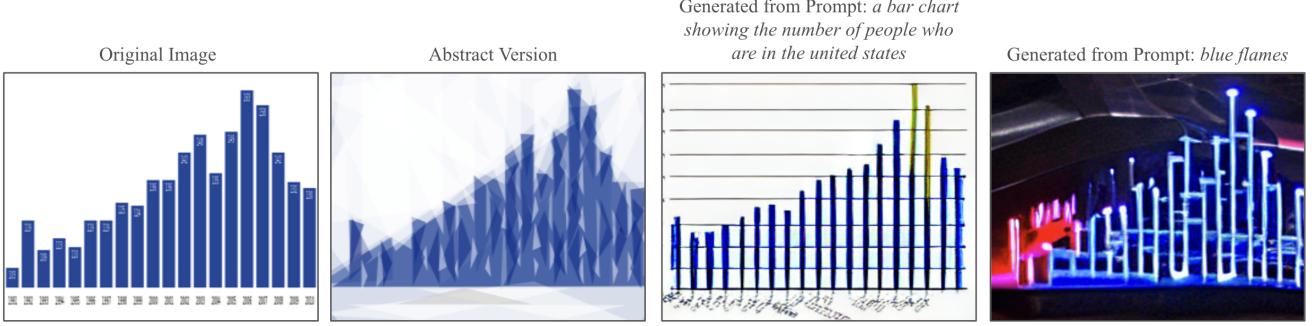


Figure 4. An example of how the same abstract image can be interpreted differently.

control over the synthesis process, laying the groundwork for enhanced spatial manipulation of synthesized imagery.

Drawing inspiration from the minimalist forms found in abstract artworks, our project introduces a novel condition crafted from geometric primitives, such as triangles. Rooted in the recognition of abstract art’s reliance on shapes to convey its message, our endeavor harnesses the subjective nature of human imagination, where diverse textual prompts yield myriad representations from the same abstract artwork.

Despite notable proficiency in preserving object locations and aligning with desired prompts, our trained ControlNet model exhibited limitations in accurately replicating colors, likely necessitating further training to address this deficiency, hampered by compute resource constraints.

In our future work, we aim to enhance our dataset by incorporating a wider variety of geometric shapes and increasing the number of shapes used to approximate different images. We also aim to explore evaluating the synthesized images quantitatively.

## References

- [1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2023. 2
- [2] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, jun 2018. 2
- [3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. 1
- [4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021. 1
- [5] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, jun 2021. 2
- [6] Michael Fogleman. Primitive. <https://github.com/fogleman/primitive>, 2016. 2
- [7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, page 89–106, 2022. 2
- [8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2
- [9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 2
- [10] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023. 2
- [11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 1
- [12] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2
- [13] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2287–2296, 2021. 2
- [14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 1
- [15] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2023. 2
- [16] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. 1
- [17] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’21*. ACM, July 2021. 2
- [18] Rishabh Srivastava. geometricshapes14k. <https://www.kaggle.com/datasets/rishabhsrivastava66/images-made-up-of-geometric-shapes-controlnet/data>, 2024. 2
- [19] Rishabh Srivastava and Addrish Roy. Abstract-art-interpretation-using-controlnet. <https://github.com/RishabhS66/Abstract-Art-Interpretation-Using-ControlNet>, 2024. 3
- [20] Stability. Stable diffusion v1.5 model card, <https://huggingface.co/runwayml/stable-diffusion-v1-5>, 2022. 1
- [21] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 1, 2, 3