ControlNet for X-Ray Image Generation

Brian Rapanan
bprap25@mit.edu

John Readlinger
jred@mit.edu

James Moore
jmoore1@mit.edu

Abstract

Currently, medical doctors in the US and abroad are limited by laws like the Health Insurance Portability and Accountability Act (HIPAA) [9], which places a tight restriction on what information they can and cannot share with their patients. In the context of X-Rays, doctors are unable to show their patients third party examples of the same injury as a result. In addition, it may be difficult to obtain x-rays of certain orientations due to a patients' inability to achieve such a position. This paper explores the application of ControlNet integrated with stable diffusion to generate high-quality X-Ray images from conditional input. These inputs follow a general line sketch similar to those generated by the OpenPose library and textual descriptions specifying body parts. After finetuning our ControlNet for this specific task our best model achieved a mean MSE score of 101.4 over a test set. Our system is thus able to generate high quality X-Rays conditioned on the user's input, allowing for the user to freely control any generated x-ray. Implementations of image preprocessing, the model, and experiments can be found here https://github.com/jamesmoore24/CV-Project.

1. Introduction

The medical field is heavily regulated by laws and regulations, with the aforementioned HIPAA being one of the most cited regulations. While legislation like this is incredibly important to ensure the privacy of individuals' medical records, it can often be frustrating for doctors to navigate this complex set of rules. For example, HIPAA makes it virtually impossible to share outsider X-Rays with a patient, even if such data is anonymized or simply for reference. Moreover, for those patients receiving X-Rays and for doctors who want a certain angle of an X-Ray image, this can result in excruciating pain for the patient or an insufficient image, potentially leading to misdiagnosis. The diffusion model we propose here aims to bridge this gap in an ethical manner, allowing for the generation of new X-Ray images to either use as reference or to augment the diagnostic process from X-Ray images.

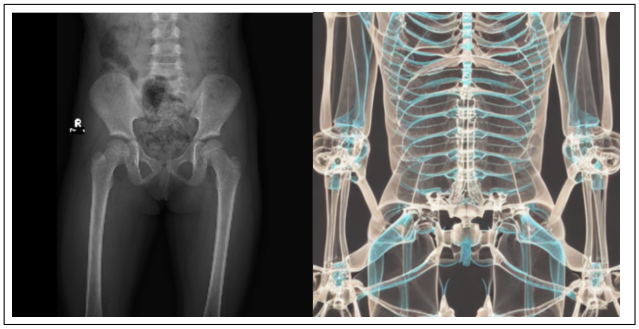

Figure 1. Real x-ray of hips and legs and vanilla stable diffusion example output with prompt "x-ray of hips and legs from front perspective"

Deep learning has made significant progress over the last decade including significant efforts in image generation. However as shown in the figure, state-of-the-art stable diffusion models struggle in generating consistent realistic x-rays. Notice how the x-ray is both convoluted and not able to produce the colors associated with realistic x-rays. This may confuse patients and is currently not applicable in a professional medical setting. By utilizing the latest developments in CNN architecture, attention mechanisms and transformer models, stable diffusion is able to understand and manipulate image features by denoising a pattern of random noise through a learned reverse diffusion process. ControlNet helps to guide this diffusion process based on a set of input constraints provided by a user. In this case, the constraints are a sketch of the X-Ray image that we want to generate and text that represents what we want to see in the image specifically. By specifying a certain part of the image we can leverage attention mechanisms to effectively focus on a part of an image that we want to change or create. We can apply this process to create realistic images of X-rays of certain body parts and characteristics (fractured, not fractured) that we describe which can be used in a medical context for patient diagnosis and care.

To evaluate the relative performance of our models, we plan to evaluate using MSE, precision, recall and f1 scores.

MSE takes our generated image set against a set of ground truth images and calculates the squared error be-

tween the pixel values, which is a good measure of image similarity. Precision, recall, and F1 are primarily concerned with the high level overlap of the generated and ground truth images, subject to a certain threshold. A number of experiments are designed around varying text prompts and noisy input images to test the robustness of the ControlNet with respect to these performance metrics. The implementation of our image preprocessing pipeline, ControlNet, and experiments is available at https://github.com/jamesmoore24/CV-Project.

2. Related Works

Although stable diffusion and ControlNet architectures are relatively new, there is a corpus of literature that deals specifically with x-ray generation. One such paper titled "Conditional Diffusion Model for X-Ray Segmentation Data Generation" [5] uses ControlNet to specify extra conditions that the diffusion model should follow, such as edge maps, depth maps, segmentation masks, or CLIP image embeddings. However, this research deals with generating x-rays specifically for patient with pulmonary nodules and not generalizable to other types of x-rays.

Other papers have looked at the specific architecture of the diffusion model to make it more effective at generating medical images. In "A New Chapter for Medical Image Generation: The Stable Diffusion Method" [7] they propose a lightweight architecture for diffusion models specifically designed for medical image generation.

To handle pre-processing our images we have considered this paper "Pre-processing methods in chest X-ray image classification" [6] which details approaches like histogram equalization, gaussian blur, bilateral filter, and adaptive masking to our x-ray dataset for better image generation.

When it came to evaluation metrics, we experimented with Inception Score (IS) and Frechet Inception Distance (FID), but ultimately decided on the familiar MSE metric, along with F1, Precision, and Recall, after consulting the paper "Creating Image Datasets in Agricultural Environments using DALL.E: Generative AIPowered Large Language Model" [8]. While the topic of the paper was far from medical imagery, the results section details the use of MSE to evaluate both Text-to-Image and Image-to-Image (variation) generated outputs. Along a similar vein, our project combines both text and imagery to generate an x-ray image.

We encountered the problem of "catastrophic neglect" in the paper "Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models" [4] which details the phenomenon where diffusion models fail to generate one or more of the subjects from an input prompt. The writers solved this issue through Generative Semantic Nursing (GSN), where the generative process is interrupted to re-excite the activations of all the tokens in the prompt, encouraging the model to generate all subjects. However, this solution proved to be computationally expensive, so we have devised another solution to the semantic problem. Since doctors already vary their terminology, we perform synonym swaps, which is very computationally inexpensive, to test the performance of synonymous inputs.

3. Methods

3.1. Dataset and Transformations

To train our ControlNet for X-Ray generation, we use the FracAtlas dataset [1], an open-source X-Ray dataset for use in computer vision tasks such as fracture classification, segmentation, and more. The dataset contains 4,083 examples labeled by medical professionals and makesense.ai, with over 900 of these images containing fractures. We use these X-Rays as our 'ground truth' values and generate ControlMaps similar to those generated by OpenPose [3]. Controlmaps in this style are extremely effective when used with a controlnet for stable diffusion, as they provide the controlnet with the proper context under which to generate the image. Moreover, these simple sketches consisting of only lines can be easily recreated by hand, which is the intended use when applied directly in the medical field, as doctors would be sketching these maps by hand to feed as a prompt into the model.

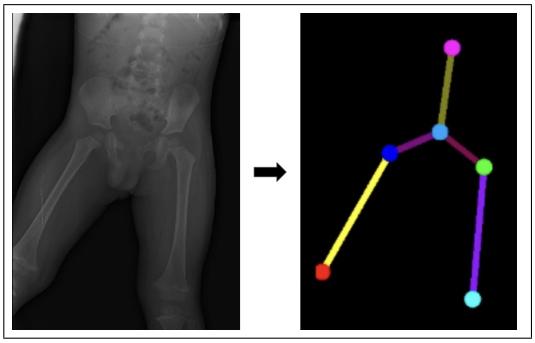

Figure 2. Example of pipeline from x-ray to pseudo-OpenPose sketch for input into ControlNet model

We modify OpenPose's methods with our own keypoint labeling to generate maps, as the OpenPose library relies on keypoints not present in X-Ray imaging. A sample controlmap from our image preprocessing pipeline can be found in 2. We can then use these alongside our text prompts to generate high-resolution X-Rays from an unbounded number of angles for use by medical professionals.

3.2. Generating Text Prompts from Data

The FracAtlas dataset has labels on fractured vs non-fractured as an indicator variable, but does not have plain text annotations in the form of prompts for our ControlNet.

To generate these prompts, we streamline this workflow by defining a grammar that is then filled in with the body part and the desired view (frontal, lateral, etc). Moreover, baked into our model is a set of keywords to avoid, ensuring our prompts are semantically similar and well-defined for the X-Ray generation task at hand. This is especially important given the findings in [2], which emphasizes the importance of semantic relevance in Stable Diffusion prompting. These text inputs are then augmented via synonym swap for a later experiment.

3.3. Model Architecture, Fine Tuning

Our high-level workflow of a forward pass in the model involves a number of inputs to the ControlNet as outlined in 3. First, we have a controlmap of the desired orientation of the output X-Ray (as generated from our image preprocessing pipeline). Finally, we include a text prompt such as "x-ray of hips and legs from a frontal view", or some other variant denoting the body part and desired image orientation. These input features are fed into the fine tuned ControlNet, which can generate the proper X-Ray image in the orientation defined by the controlmap and the text prompt.

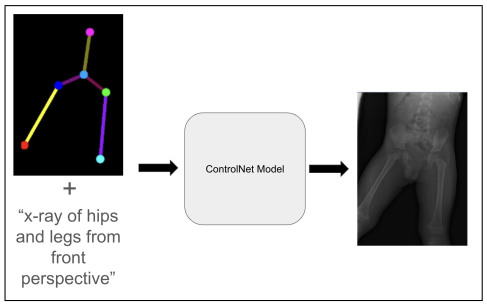

Figure 3. High-Level Overview of Workflow

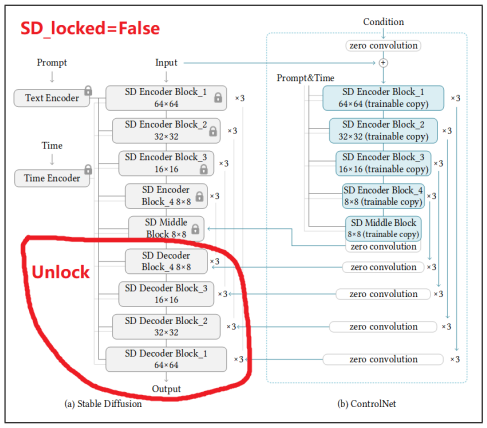

Figure 4. Modifications to ControlNet Architecture

To accomplish this generation process, we rely on the ControlNet architecture proposed in [10]. Specifically, the ControlNet is desirable for image generation tasks such as ours due to a number of reasons, such as its adaptability, vast resources on which it was pretrained, and deep architecture consisting of convolutional layers and an encoder-decoder pair. Specifically, the encoder-decoder layers allow for the model to be finetuned well for specific tasks and conditions while also incorporating much needed visual attention for feature detection, especially with our OpenPose controlmaps; the convolutional layers suit the model well for learning spatial relationships in our input images. After adjusting the original architecture as outlined in [10] and finetuning to fit the specific demands of our X-Ray dataset, the controlnet can effectively generate the detailed X-Rays discussed earlier. To train our controlnet, we feed sample controlmaps into the model with the accompanying text prompt, and the model then compares the output X-Ray to a corpus of the ground truth FracAtlas X-Rays, and the loss is computed.

To modify ControlNet, we unlock certain decoder blocks in the Stable Diffusion model as shown in 5. Doing so allows us to directly train these weights, which will help in generating better results for our X-Ray specific task, given that X-Ray images are much different than other image types, and thus will require maximal attention layers and blocks. Moreover, after some experimentation with the model, we settle on the following hyperparameters:

- Learning Rate = 1e-5
- Batch Size = 4
- only_mid_control = False

The remaining hyperparameters are experimented with later, each with important goals in mind.

4. Experiments

With our models, we aim to rigorously test the accuracy, quality, robustness, and generalizability of the images. To do so, we conducted 3 distinct experiments, primarily concerned with altering the image input, textual input and model architecture to test its effect on the quality of generated images.

To diversify our conditional image maps we added several different noise layers. The thought process here is doctors may incorrectly sketch a pose map, so a model of this type would need to handle noisy images well in order to perform well in the wild. To accomplish this, we lean on the torchvision.transforms library, which offers a suite of image transformations such as Gaussian blurring, motion blurring, reflections, image sharpening, and more. By training a model on these images with random amounts of noise, the resulting trained model should generalize better to a wider array of inputs and lead to better metrics and more flexible capabilities.

Next, we manipulated the text input and trained a ControlNet on this augmented dataset. Doctors are subject to using varying terminology when describing the desired output, even when referencing the same thing in theory. To test this, we can rely on synonym swaps provided by the NLPAUG library, which uses WordNet to randomly swap synonyms into sentences and train a model on this data. This should allow for the model to be more flexible in the range of inputs it can handle in testing, leading to better outputs.

We expect our model to perform best with varied input text and images so our first experiment is changing the architecture of our ControlNet but setting the SD_locked parameter to false. This opens up the SD decoder block in the architecture for training on the input data. We then generated X-Rays from our test dataset and measured the MSE, Precision, Recall, and F1 metrics to test the abilities of the SD block.

Using the best performing model from experiment one as a baseline our second experiment involves training a model on inputs that are not blurred. We predict that this model will not generalize to the test set well since the test set is made up of both blurry and regular images which will lead to decreased performance when compared to the baseline model.

Our final experiment will involve training a model on a singular standard text input. Again we predict that this model will not generalize to the test set well since the test set is made up of variable text prompts which will lead to decreased performance when compared to the baseline model.

5. Results and Figures

5.1. Quantitative Analysis

Below are the results of the experiments outlined above, along with accompanying figures and discussion.

	MSE Score	Precision	Recall	F1 Score
sd_locked = True	104.2017	0.3333	0.3333	0.3333
sd_locked = False	101.9814	0.3333	0.3333	0.3333

Table 1. Metrics for sd_lock = True and sd_lock = False

For our ControlNet with sd_locked = True, we achieved an MSE of 104.2017, a Recall of 0.33, a Precision of 0.33, and an F1 score of 0.33. We see that the MSE on the sd_locked = False is lower at 101.98, with identical precisions and recalls. While the performance boost is small, the additional layer allows for greater accuracy in the form of MSE, which decreases by about 3%.

These results imply that including controlmap inputs that have noise such as motion blur could help generalize our model to new inputs. For our ControlNet trained with images that have random motion blurs applied to them, we see that the performance on the test set outperforms an equal model with no noise added. Specifically, the MSE is lower here for the noisy inputs and each of Precision, Recall, and F1 score are all higher than those of the controlnet trained on solely non-blurred, noiseless images.

	MSE Score	Precision	Recall	F1 Score
Motion blurred inputs	101.9814	0.3333	0.3333	0.3333
Non-blurred inputs	104.8624	0.2666	0.2666	0.2666

Table 2. Metrics for motion blurred vs non-motion blurred inputs

	MSE Score	Precision	Recall	F1 Score
Variable text inputs	101.9814	0.3333	0.3333	0.3333
Singular standard text input	101.4314	0.3333	0.3333	0.3333

Table 3. Metrics for variable text input vs singular standard text input format

These results imply that training a controlnet on a singular textual input versus multiple textual inputs with synonym swapping does not significantly affect the quality of the model. This could be due to the type of problem of x-ray generation as we are generated very specific images so the textual input may not matter apart from a few key words like ”hip” and ”leg” which are not changed during synonym swap.

5.2. Qualitative Analysis

For a thorough analysis of our model, we also perform qualitative analysis by showing example outputs from our model under different conditions. As shown by the figure the largest noticeable difference in output came when we set sd_locked to false instead of true. By doing this we were able to fine tune the decoder blocks of the SD architecture. Before setting sd_lock to false we were noticing that most images weren't following a similar structure and that the model was unable to learn specific parts of our input like how bones connected to each other or even how different joints like knees and hips look. As a result sd_locked = True returned results that looked like a contorted amalgamation of different bones which followed the pose map somewhat. Also observe how the model was unable to distinguish the x-ray content from the black background which resulted in a grey monochrome output.

In contrast, sd_locked = False was able to correctly identify different regions of the body and link them together in a cohesive way. However, it often struggled to correctly orient the image to the control map and sometimes the resulting figure would generate multiple body parts unrelated to the control map. The model also frequently exhibited a ”feature duplication” issue where multiple body parts like spines or hips were generated close to each other and overlapped resulting in a incoherent x-ray. Despite it's drawbacks the sd_locked = False model shows promise in generating accurate and useful x-rays conditioned on certain control maps.

Both experiments concerned with constraining the input prompt to a single prompt and not including blurred control maps both did not significantly affect the output of the image.

6. Discussion

Overall, we see promising behaviors from our model. Our initial hypothesis that activating the sd_lock block for training was

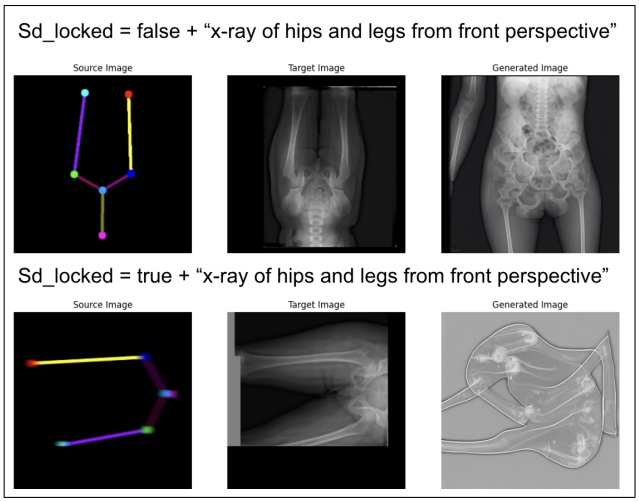

Figure 5. Sample output for given control map and prompt ”x-ray of hips and legs from front perspective” for baseline model sd_locked = False and sd_locked = True. Notice how sd_locked = False exhibits ”feature duplication” phenomenon of the hip.

correct, as we see an improvement in MSE and a slight qualitative improvement in the generated images. The additional attention to the input features allows for images with less overall error in comparison to the ground truth, and images that overall resemble a real X-ray. Moreover, making our controlmaps subject to noise as opposed to the noiseless alternative resulted in a better MSE. On the other hand, we observed that our methods for text augmentation did little to affect performance, indicating the need for additional inquiry into this field.

Still, the metrics and sample images from our model indicate that there is more progress to be made. Making certain architecture changes such as adding cross-attention and increasing the size of the dataset could allow for a better model allowing for more accurate diagnostics. Overall, the findings of this paper are optimistic for continuing this line of research in X-Ray generation using ControlNet for Stable Diffusion. The positive results obtained by our experiments on the additional Stable Diffusion attention blocks, augmented text data, and data augmented with noise can serve as building blocks for creation of a state of the art X-Ray generation model to be deployed for diagnostic purposes.

7. Conclusion

In this project, we set out to research the capabilities of ControlNet for X-Ray image generation with a few key architecture and data-related modifications and experiments. Specifically, our choices were to test the capabilities of the supplemental Stable Diffusion attention blocks, test the effects of variable input text data, and to examine the effect of adding noise to input controlmaps to measure each of their effects on model generalizability and accuracy. Our findings indicate that our methods relating to model architecture tuning and image noise application make a positive effect on the quality of X-Ray generations, while our methods for creating variable text inputs make a neutral effect, indicating the need for further research in this domain.

8. Division of Work and Experiments

Our group made every effort to make the division of work as fair as possible, as we took the massive amount of work to be done for the project and assigned tasks to each person based on their strengths.

James Moore spearheaded a lot of the software setup of the ControlNet. This required Colab Pro+ given the size of the model, and required proper environment and dependency setup. Once he had trained the model, he led the baseline experiment with no data corruption.

Brian Rapanan led most of the transformation work on the data, namely broadcasting features to constant dimensions, adding noise layers using torchvision.transforms, and iterated the design of our controlmap generation method. Brian then led an experiment on the noisy image data.

Jack Readlinger aided the team’s efforts in annotating data and generating/modifying text prompts. Jack made use of the NLPAUG library to later add synonym swaps to these prompts for his experiment surrounding variable text prompts.

References

[1] Iftekharul Abedeen, Md. Ashiqur Rahman, Fatema Zohra Prottyasha, Tasnim Ahmed, Tareque Mohmud Chowdhury, and Swakkhar Shatabda. FracAtlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs. *Scientific Data*, 10(1), aug 2023. 2

[2] Anh Bui, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Removing undesirable concepts in text-to-image generative models with learnable prompts, 2024. 3

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4), jul 2023. 2

[5] Zehao Fang. Conditional diffusion model for x-ray segmentation data generation. *Journal of Artificial Intelligence Practice*, pages 7–10, 2024. 2

[6] Agata Giełczyk, Anna Marciniak, Martyna Tarczewska, and Zbigniew Lutowski. Pre-processing methods in chest x-ray image classification. *PLoS One*, 17(4):e0265949, Apr. 2022. 2

[7] Loc X. Nguyen, Pyae Sone Aung, Huy Q. Le, Seong-Bae Park, and Choong Seon Hong. A new chapter for medical image generation: The stable diffusion method. In *2023 International Conference on Information Networking (ICOIN)*, pages 483–486, 2023. 2

[8] Ranjan Sapkota, Dawood Ahmed, and Manoj Karkee. Creating image datasets in agricultural environments using dall.e: Generative ai-powered large language model, 2024. 2

[9] U.S. Department of Health and Human Services. Standards for privacy of individually identifiable health information. Code of Federal Regulations, title 45, sec. 164.510(b)(3), 2003. Accessed: 2024-05-14. 1

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3