SaveFace: Controlling Face in Image Diffusion Models

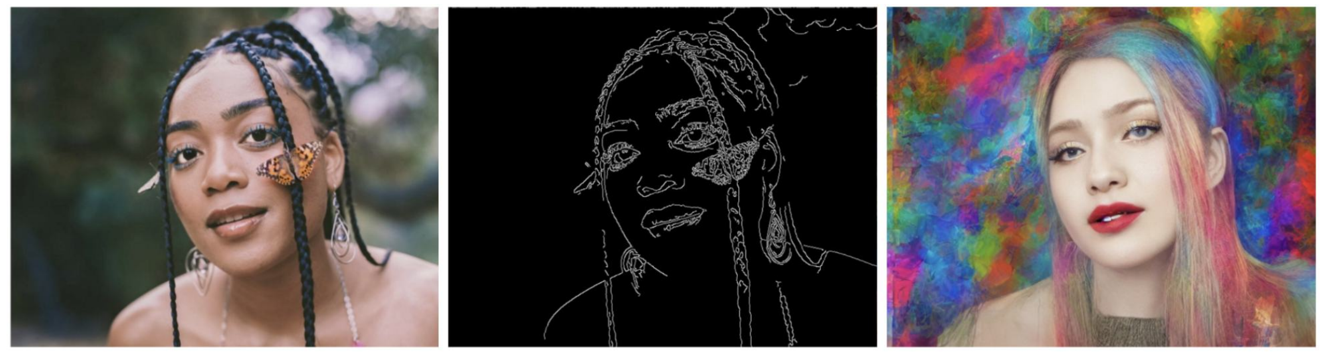
Ashna Khetan, Isabel Sieh, Laya Iyer | {ashnak, isabelrs, laya}@stanford.edu

CS231N Spring ‘24

Motivation

Text-to-image diffusion models perform remarkably on image generation tasks, even faces, and have great design potential. But generating graphics with a photo input tends to **over manipulate the image.** They also contain social biases that harm **underrepresented** genders and ethnicity (a study found that SD representation of “White/Man” was ~50%) that result in homogenized facial features, losing aspects like skin-tone.

E.x. **ControlNet Canny generated** images based on the left-most image (Order of images: `Input, Canny edge condition, and output image`)



In this project, we introduce **SaveFace**, a ControlNet model with a custom control that aims to preserve facial structure and race. Overall, we aim to *bridge gaps created by biases due to lack of representation in the data used for other diffusion models.*

Goals

- Diffuse a person’s environment, background, accessories, etc while maintaining one thing **constant: their face**
- Focus on creating a **control** for ControlNet that preserves race and facial features → **improving** upon the **biased** performance of other controls for ControlNet such as Canny ControlNet

Dataset

The **dataset** that we used to train and evaluate our models were a combination of the following datasets:

- UTKFaces → contain **20k** images of faces, labeled with various details
- Labeled Faces in the Wild (LFW) → **13k** images for facial recog tasks
- Human Faces (Kaggle) → **7k** web-scraped human faces
- 10k AI-generated (10k) → **10k**

5k images separated from the combined dataset for quantitative analysis

References

L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 2, 4

G. AI. Mediapipe face landmarker. https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker, 2023. Accessed: 2023-06-06. 2

C. Team. Controlnetmediapipeface. https://github.com/ControlNet/ControlNetMediaPipeFace, 2023. 2

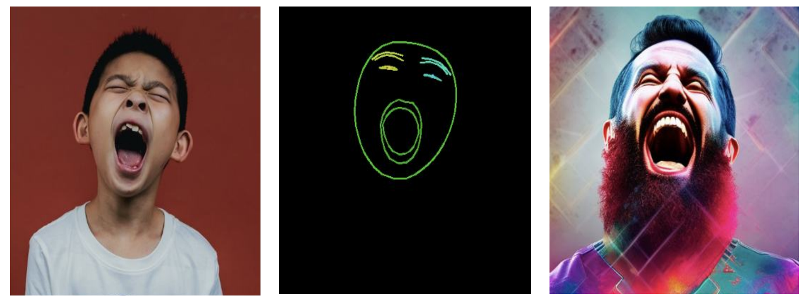
T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 2

Benchmark Models

ControlNet: Finetuning on large diffusion models is challenging due to limited data for specific conditions, causing overfitting or catastrophic forgetting. ControlNet addresses this by locking the parameters of the large model, while allowing encoding models to be trained to learn diverse conditional controls. It can controls SD with a conditioning image.

ControlNet-Canny:
Canny edges are successful in largely maintaining the face pose and general features but can drastically change aspects like ethnic features, or hairstyle.

ControlNet-MediaPipeFace:
Maintains face pose as its conditional mask but does not retain the original image’s identity

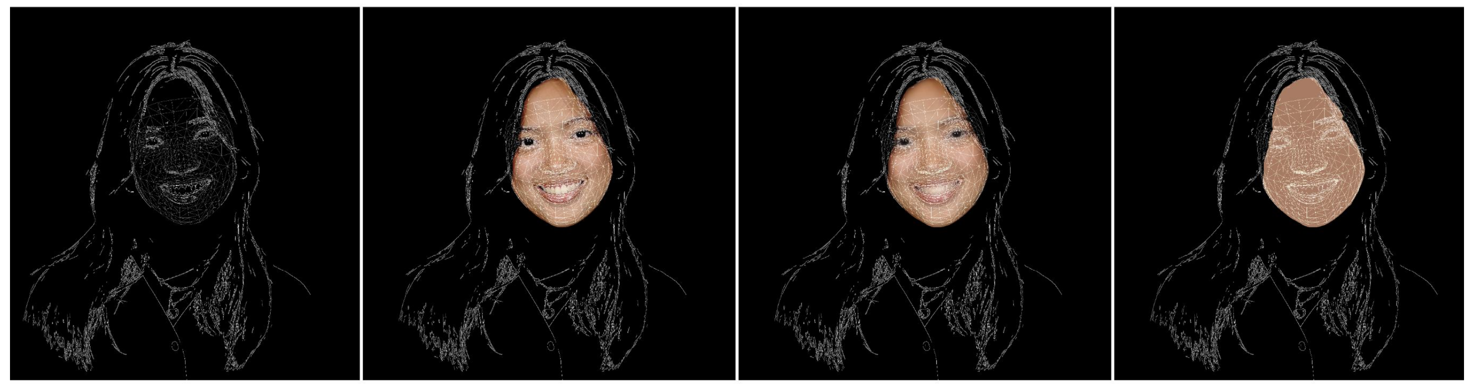


InstructPix2Pix
A conditional diffusion model that combines GPT-3 and SD in order to edit images based on human text instructions.

“Replace the fruits with cake”

SaveFace: A ControlNet Model

To train our variation of ControlNet, first, we had to **develop a mask** that would be passed in as our training image. For this we used multiple techniques such as **Facial Landmarks** (Mesh), Facial Landmarks + **AlphaMask**, Facial Landmarks + **Gaussian Blur**, and Facial Landmarks + **Single-Color** as shown in the image below:



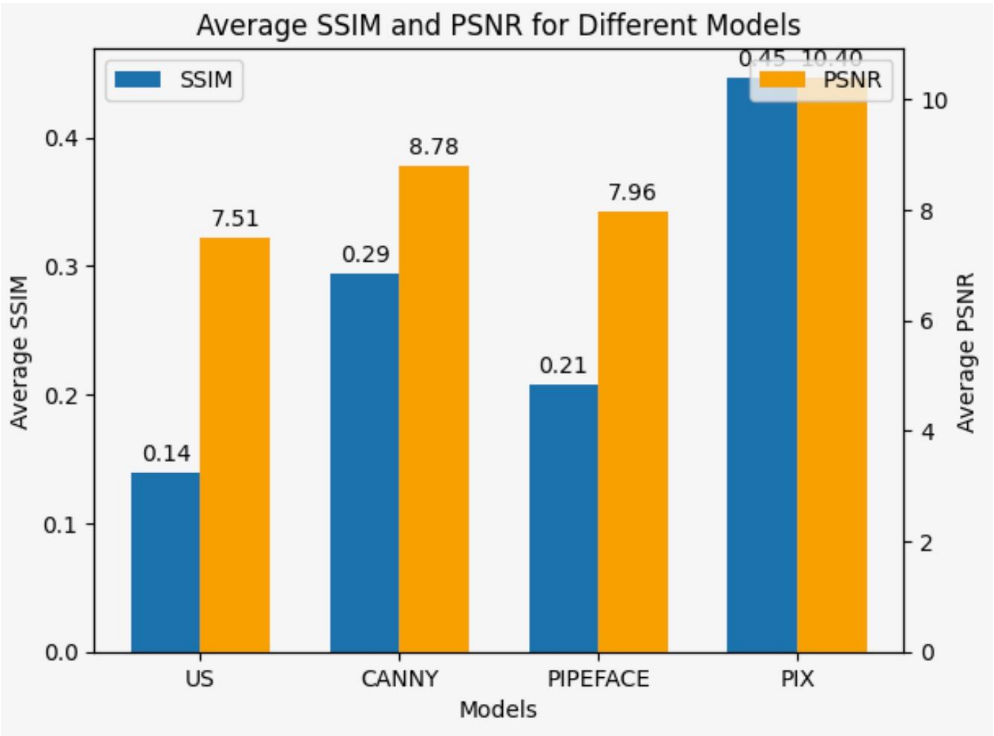
Then we train using ControlNet’s architecture which reuses the encoding layers and pre-training from a given diffusion model and learns a specific set of conditional controls. The complete ControlNet then computes:

$$y_c = F(x;\Theta) + Z(F(x + Z(c;\Theta_{z1});\Theta_c);\Theta_{z2}) \quad (1)$$

where y_c is the output of the ControlNet block, Z is the zero-convolution layer, $F(x;\ \Theta)$ is a trained neural block, and c is the condition. The loss function of ControlNet is as follows:

$$L \ = \mathbb{E}_{z_0,t,c_t,c_f,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t,t,c_t,c_f)\|_2^2\right] \quad (2)$$

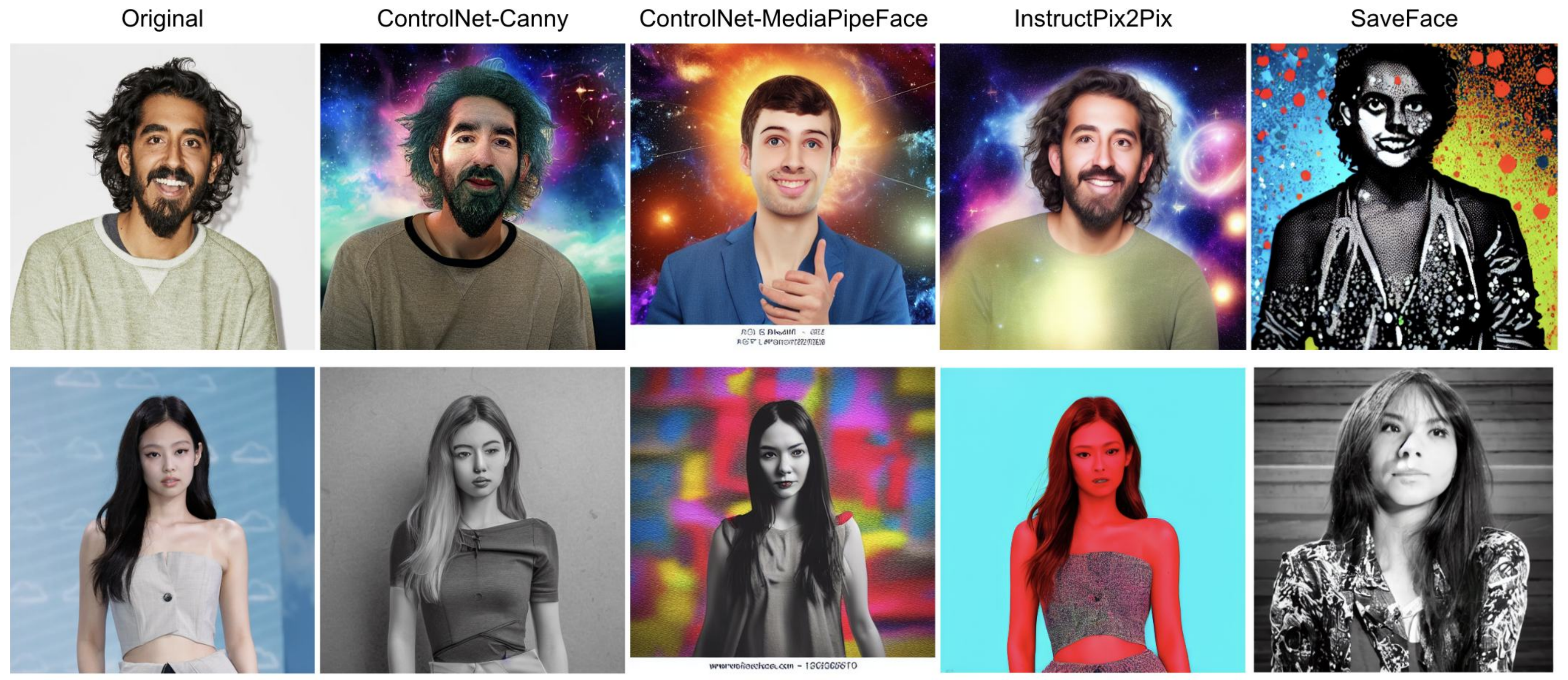
Results



SSIM: Structural Similarity Index (Range: -1 to 1) and tells us how structurally similar two images are
PSNR: Higher Peak Signal-to-Noise Ratio means we have better quality images
Result Quality: Qualitative measure of how the result appears to a viewer
Face Fidelity: Qualitative measure of how much a face is preserved

Method	Result Quality	Face Fidelity
ControlNet-Canny	1.8	2.8
ControlNet-MediaPipeFace	2.6	3.2
InstructPix2Pix	2.3	1.3
SaveFace	**3.4**	**3.1**

Analysis



- Facial pose, made up of the human’s angle, the way they look at the camera, etc, was captured well by our model → mesh landmarks provided valuable information **just not enough valuable information.**
- Diffusion models undergo the **Sudden Convergence Phenomenon**

Future Work

- Vary the **mask type**, the **base model**, and the **caption type** given in training
- Use **ControlNet-Canny** as the base model to **train** on top of instead of Vanilla ControlNet to see how that affects performance
- Experiment with captions to describe the individual’s face in more detail