Facing the Truth: Generating Faces using Stable Diffusion and Implicit Neural Representation

Jason Uwaeze[1]

[1]**ju6@rice.edu**

https://colab.research.google.com/drive/1vCLc5Hd3SPIky$_BqXmBTvgOvh5Gpr7YQ?usp = sharing$

INTRODUCTION

In this project, I applied Stable Diffusion, one of the most widely used public diffusion models, for both image generation and inpainting, and implicit neural representations to fit images of the legendary Lionel Messi (Molaei et al. (2023); Sitzmann et al. (2020)).

1 DIFFUSION MODELS

Figure 1. Example images generated by Stability-AIs text-to-image stable diffusion model. The prompts used to generate these images are as followed: (1) A humanoid robot doing yoga in houston texas, (2) Jason Statham angrily reading a children's book on a plane, and (3) Tom Brady interviewing President Barack Obama on a podcast. Demonstrate that prompts excluding celebrities generated most satisfactory images.

I explored various creative prompts on Stability-AIs text-to-image diffusion model. The prompts used in this experiment include: (1) A humanoid robot doing yoga in houston texas, (2) Jason Statham angrily reading a children's book on a plane, and (3) Tom Brady interviewing President Barack Obama on a podcast. I found that prompts that didn't include names of celebrities outperformed prompts that did (see fig. 1). To validate this hypothesis, I tested the following prompts: (1) Allyson Felix running on a track, (2) Steph Curry playing basketball, and (3) Tom Brady playing football. With these prompts the model continued to struggle with generating faces of celebrities but did a good job of generating photos individual body parts like arms and legs (see fig. 2). I noticed that text was difficult for the model to generate. In figure 2 the model failed to generate the texts Stephen Curry's jersey, which should state "Golden State Warriors".

Figure 2. Example images generated by Stability-AIs text-to-image stable diffusion model. The prompts used to generate these images are as followed: (1) Allyson Felix running on a track, (2) Steph Curry playing basketball, and (3) Tom Brady playing football.

I performed two experiments with varying num_inference_steps and found the more steps resulted in higher fidelity images, but too many steps resulted lower fidelity images. In the first experiment, I tested 20, 40, 80, 160, and 320 inference steps with "Angry clown hiding in storm drain" as the prompt (see fig. 3). I found that 40 inference steps generated the image with the highest fidelity, but 160 inference steps was closest to the prompt. In the second experiment, I tested 20, 40, 80, 160, 320 with "A chessboard

where the pieces are represented by miniature cities" as the input prompt. I found that 40 inference steps also generated the image with highest fidelity but 20 inference steps more accurately address the prompt.

Figure 3. Example images generated by Stability-AIs text-to-image stable diffusion model with varying num_inference_steps and "Angry clown hiding in storm drain" as the input prompt. *Top Left:* 20 inference steps, *Top right:* 40 inference steps, *bottom Left:* 80 inference steps, *bottom right:* 160 inference steps

1.1 Short Questions

1. Which prompt are effective for generating images, and which ones lead to less pleasing results? Why?

Prompts including celebrities, like president Barack Obama and Tom Brady, leads to unfavorable results like distorted faces. This could be cause there are many details or local features found in celebrity faces, and the diffusion model is trained to model global or overall structure. The stable diffusion model also struggles with generating detailed appendages like arms and legs which are often a result of prompts that include people or animals. However, I found that the diffusion model does a great job generating backgrounds or landscapes. **Poor people and appendage generation could be a result of ill-defined prompts, such that the model does before with more or less specific prompts**. Great landscape generation could be because there are less details or less complicated patterns needed to complete the task, unlike people with eyes, ears, posture and many more.

2. How does the model's performance vary with different numbers of inference steps?

In the example of 'A chessboard where the pieces are represented by miniature cities', the diffusion model did a poor job generating the chessboard, but progressively got better as the number infer-

Figure 4. Example images generated by Stability-AIs text-to-image stable diffusion model with varying num_inference_steps and "Angry clown hiding in storm drain" as the input prompt. *Top Left:* 20 inference steps, *Top right:* 40 inference steps, *middle Left:* 80 inference steps, *middle right:* 160 inference steps, *bottom:* 320 inference steps

ences steps increase from 20 to 320. All generated outputs failed to include cities as chess pieces or simulate a realistic chess games (e.g. all outputs had an unreasonable amount of pons).

Figure 5. Input and PixelLib created mask for image inpainting.

2 IMAGE INPAINTING

I performed image inpainting with Stability-AIs stable diffusion model for image generation, and PixelLib's instance segmentation model for creating input annotations. The PixelLibs segmentation model takes an image path and an output folder path as input. PixelLib creates and save multiple masks of all instances of objects and people segmented from the original input image. I would then choose the mask of my choice to perform image inpainting (see fig. 5).

https://github.com/prateekralhan/Instance-Segmentation-using-PixelLib

For this experiment I chose to replace Lionel Messi in famous image of him celebrating. Figure 5 shows the input and corresponding mask I used to complete the inpainting. I tested various 100, 200, 300, 500 inference steps and found 300 produce the most appealing results, while 200 results were the most frightening (see fig. 6). The 200 and 300 results can be view in the second and third rows of figure 7. This experiment also showed that the diffusion model struggle to produce quality images when given "A person with a backpack" as its input prompt. I further tested "A soccer player" and "A basketball player" as input prompts, producing considerably better results (see fig 8).

I explored the use of guidance scale as an input parameter for image inpainting, and found that a scale of 8.5 produce some of the most appealing results (see fig 9). I also noticed that the diffusion model struggle to generate images of basketball player, generating cartoon like images. This could be because in the model has yet to see a basketball player making the same pose as Lionel Messi, especially since basketball players either running while dribbling a basketball or run while guarding another player.

3 IMPLICIT NEURAL REPRESENTATIONS

3.1 Short Questions

i. According to the SIREN paper, what does the periodic activation contribute to the representational capabilities?

Sitzmann et al, 2020, discusses the universal function approximation properties of the period activation for their proposed neural network, with an emphasis representation of signal derivatives. Signal derivatives contain information of how pixel values change in the image. The periodic activation essential allows for better representations of signal derivatives. Better derivative representation results in better edge detection, feature extraction, and overall image quality. newline

ii. How do the SIREN models perform with different numbers of hidden_layers? Why?

I noticed a higher PSNR and lower loss with more hidden layers. With more hidden layers the SIREN

Figure 6. Image inpainting with 300 num_inference_steps. The prompts used to generate images is as follows. *Top:* "A person running", *Middle:* "An astronaut", *Bottom:* "A person with a backpack".

model was able to grenerate higher quality images, or images with less noise.

iii. How do the SIREN models perform with different numbers of hidden_features? Why?

Visually I noticed less noise with less hidden features, such that 128 hidden features outperformed 256 and 512 hidden features. This could because with too many hidden features, the model is overfitting the data.

iv. How could implicit representations be used? Explain some examples of the possible applications.

Implicit representations can be used to reconstruct, segment, register, and generate novel medical images because of INR's resolution agnostic nature, memory efficiency, and ability to avoid locality biases (Molaei et al. (2023)).

REFERENCES

Molaei, A., Aminimehr, A., Tavakoli, A., Kazerouni, A., Azad, B., Azad, R., and Merhof, D. (2023). Implicit neural representation in medical imaging: A comparative survey. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2381–2391.

Sitzmann, V., Martel, J. N., Bergman, A. W., Lindell, D. B., and Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*.

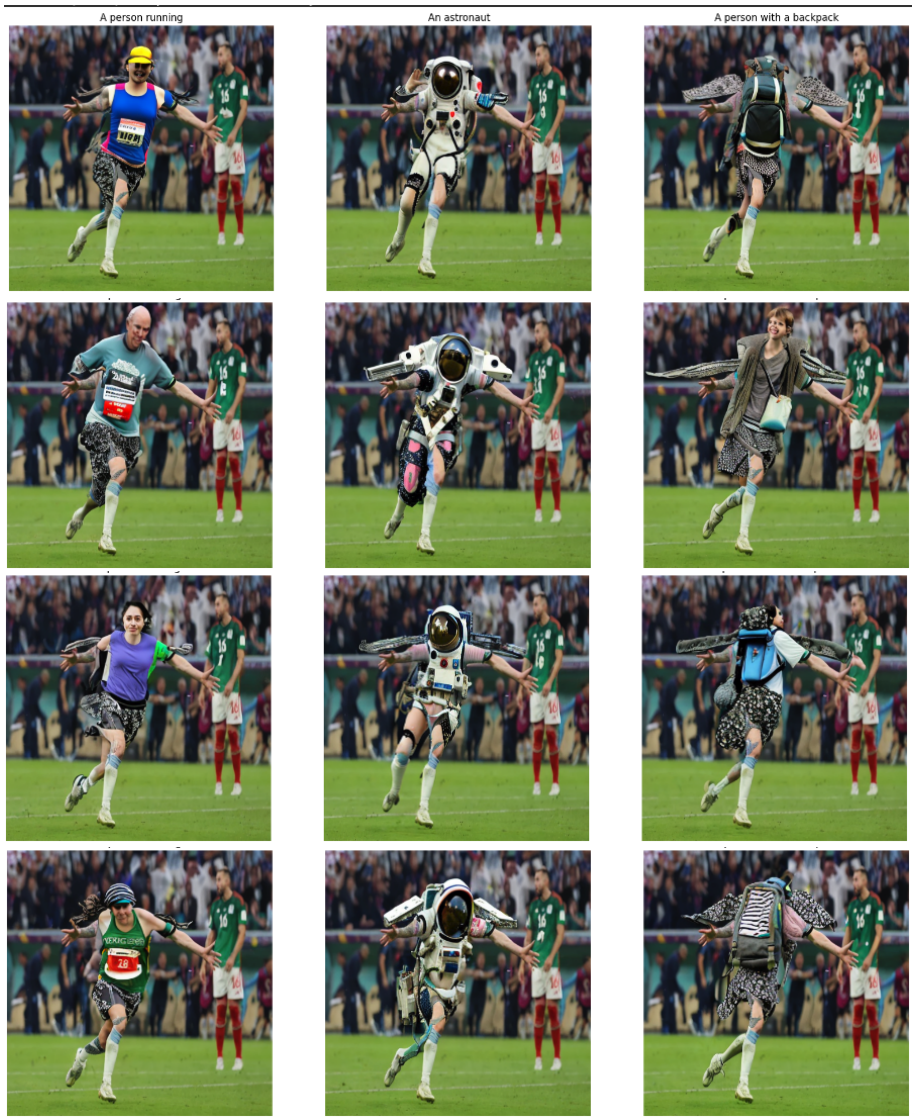

Figure 7. Image inpainting with various num_inference_steps, where the rows from top to bottom corresponding to 100, 200, 300, 500 inference steps, respectively. The prompts used to generate images is as follows. *Left Column:* "A person running", *Middle Column:* "An astronaut", *Right Column:* "A person with a backpack".

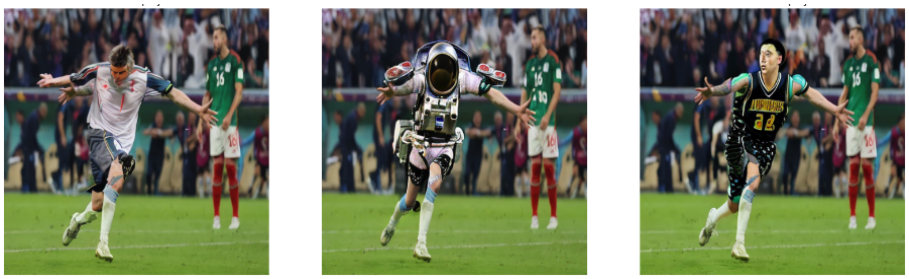

Figure 8. Image inpainting with 500 inference steps. The prompts used to generate images is as follows. *Left Column:* "A soccer player", *Middle Column:* "An astronaut", *Right Column:* "A basketball player".

Figure 9. Image inpainting with guidance scale set to 8.5. The prompts used to generate images is as follows. *Left Column:* "A soccer player", *Middle Column:* "An astronaut", *Right Column:* "A basketball player".

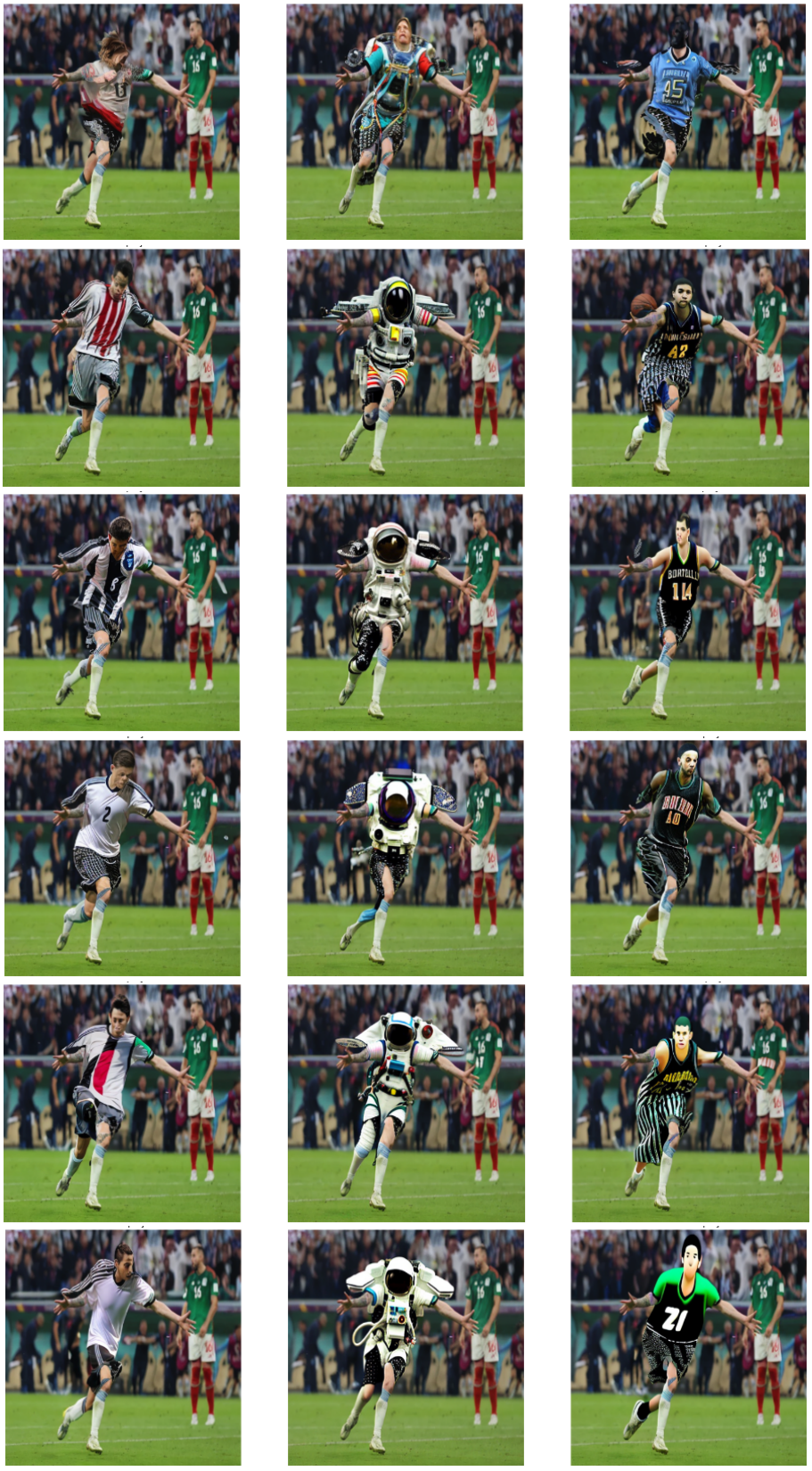

Figure 10. Image inpainting with various guidance scale values. The rows correspond to 3, 7.5, 8.5, 10, 12, and 15 guidance scale values. The prompts used to generate images is as follows. *Left Column:* "A soccer player", *Middle Column:* "An astronaut", *Right Column:* "A basketball player".