Probabilistic Image Inpainting: exploring conditioning and biases

Jiwoong Jeong (1223841892), Dipanshu Singh (1220267958)

March 11, 2025

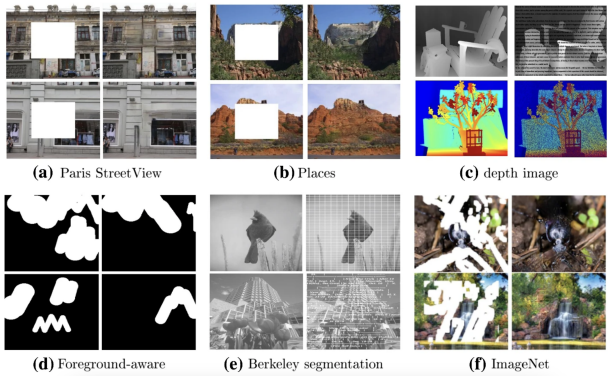

Figure 1: Various examples of inpainting tasks and datasets

Abstract

In this project, we aimed to explore probabilistic image inpainting through a series of progressively challenging goals categorized into Baseline, Medium, and Stretch levels. Our work successfully achieved both the Baseline and Medium goals, by implementing two models (PDE and LaMa) and on two different conditions (masks and noise) and evaluated them on the metrics outlined in our proposals and looked for any biases, correlations, or disparities. We found out through our experiments that both methods are fairly robust to different initial conditions but the PDE based method showed some bias in inpainting, specifically in images with noise with resepect to MSE.

1 Introduction and Motivation

Image inpainting (inpainting) is generally defined as any method that "paints in" missing parts of the image. These methods can range from simple pixel interpolation methods to more complex deep learning methods with probabilistic reasoning. No matter the method, the end goal of inpainting is to reconstruct damaged, or defective, portions of an image that "looks" reasonable. This is relevant and interesting in terms of the course because both simple and complex inpainting methods like Convolutional Neural Networks (CNNs) Li et al. [2021], Generative Adversarial Networks (GANs) Goodfellow et al. [2020], and diffusion-based Ho et al. [2020] methods all use some form of probabilistic methods to estimate a "most probable" image/pixels of the missing areas. These methods take into account either the unmasked areas or a learned probabilistic distribution (conditional or unconditional) into account when generating an inpainted image. As such inpainting is an interesting problem because outside of an experimental setting as there are no "ground truths" to compare the outputs to. This raises some interesting questions about how these models were trained and if there are any hidden biases within the probabilistic reasoning process of these models Jam et al. [2021]Lugmayr et al. [2022]. This is important as these hidden biases could result in biased and unfair sentencing as demonstrated in the COMPAS recidivism risk-scoring model Dressel and Farid [2018].

Our goal in this project was to perform inference on some of these inpainting methods and interrogate them with different initial conditions (masks and noise distributions) to explore and evaluate any strengths, weaknesses, or biases that might be hidden in these models. We did not train any models as training any generative models would take a long time (diffusion models), are notoriously difficult to balance (GANs), or would result in poor quality (VAEs) without significant training and testing. Our technical contributions include exploring: 1) the effect of different masks, 2) different initial noise distributions and 3) seeing if there are any biases based on protected attributes that may be seen in the inpainting models.

2 Background

Image inpainting, broadly speaking, can be thought of as missing data imputation e.g. some statistical process that replaces missing (pixel) values with estimated (inpainted) values based on the available (non-masked) information. The most basic method of imputation

replaces missing values with some summary statistic (generally mean or median). Essentially, inpainting can be framed as a problem of estimating the conditional probability distribution of the missing values given the observed ones. Some more complex approaches include k-nearest neighbors (KNN) and Gaussian Mixture Models (GMMs) - essentially clustering or more complex machine learning models that learn the relationship between some latent noise to the image through adversarial training (GANs), encoding and decoding (VAEs), or noising and denoising steps (diffusion).

3 Technical Contribution

For our project we tested two models: a PDE based inpainting method, PyInpaint, and a residual learning based inpainting method, LaMa on the CelebA dataset with different initial conditions to see if there were any interesting observations we could see from the types of masks and noise conditions. While we would have liked to train inpainting models that drew from different latent noise distributions and compare their performances, due to the time and resource constraints we could not. These initial conditions include different masks (line, triangle, circle, and rectangle) and different noises applied to the initial images (Gaussian, gamma, exponential, and uniform).

First we implemented the inpainting of random masks. Each different mask was generated randomly and applied to the image to get four different inpainted images for the same original image. Then we applied random noise equally to the images, applied the mask, then inpainted them. We made sure to keep track of the mapping between the original, noisy, mask, and inpainted images so that at the end we could also check the performance of our models on different labeled attributes in the CelebA dataset. To evaluate our model outputs, we used both traditional image metrics (Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Squared Error (MSE)) as well as more complex metrics of image quality (Frechet Inception Distance (FID), Kernel Inception Distance (KID), and Learned Perceptual Image Patch Similarity (LPIPS)).

We evaluated the mentioned metrics for all of our experiments with different initial conditions as well as evaluating them based on the attributes of the images in CelebA. We chose the attributes: Male, Young, and Pale Skin as sex, age, and skin tone are attributes that are not explicitly protected (attributes that should not be discriminated against in the workplace or in our case, model performance) but a close surrogate to them.

Table 1: Overall results of PDE inpainting on different masking shapes.

	Line	Triangle	Circle	Rectangle	Overall
SSIM (↑)	0.988±0.023	0.94±0.055	0.950±0.041	0.888±0.083	0.942±0.066
MSE (↓)	24.201±261.036	193.603±388.112	184.201±343.411	511.232±746.100	228.458±504.760
PSNR (↑)	41.699±5.044	30.329±7.569	30.029±7.433	24.103±5.727	31.537±9.126
FID (↓)	13.917	29.962	21.985	49.792	23.890
KID (↓)	0.007±0.000	0.012±0.000	0.008±0.000	0.019±0.001	0.009±0.001
LPIPS (↓)	0.546±0.071	0.574±0.071	0.563±0.070	0.596±0.073	0.571±0.073

Table 2: Overall results of LaMa inpainting on different masking shapes.

	Line	Triangle	Circle	Rectangle	Overall
SSIM (↑)	0.998±0.002	0.962±0.041	0.968±0.030	0.917±0.076	0.961±0.054
MSE (↓)	1.615±1.768	78.481±121.809	80.259±131.882	243.890±420.338	101.061±245.025
PSNR (↑)	47.912±4.086	34.725±8.348	34.898±8.881	28.134±6.720	36.417±10.202
FID (↓)	1.49	3.131	2.988	6.009	2.014
KID (↓)	0.001±0.000	0.001±0.000	0.001±0.000	0.002±0.000	0.001±0.000
LPIPS (↓)	0.558±0.069	0.559±0.071	0.559±0.069	0.566±0.071	0.560±0.072

3.1 Results

The overall results of the PDE inpainting 1 and LaMa inpainting 2 show that the both methods worked best on line masks. This makes sense because line masks have the thinnest and smallest area to fill and more ”known” pixels to estimate the missing values off of. Conversely, we can see that the rectangle masks performed the worst as it had the largest area to fill and the least information for the estimation to work off of. This can be seen in the selected samples of PDE and LaMa inpainting in figures 2 and 7.

In the second series of experiments, where we added noise to the input image, we can see that the both methods perform similarly to the non-noisy images as seen in 3 and 4. It is interesting that the MSE actually improves with the different noise added to the inputs, which is in contrast to what we would expect since the PDE method seems to interpolate the nearest pixel and not add noise as seen in figure 2. This is in contrast to the results of the noisy LaMa images where all metrics perform poorer that might have to do with the Fourier convolutions that LaMa uses. The FID, KID, and LPIPS metrics all suffer when compared to the normal images because they are metrics that are pretrained on natural images without noise.

Table 3: Overall results of PDE inpainting different masking shapes with added noise on the original image.

	Line	Triangle	Circle	Rectangle	Overall
SSIM (↑)	0.994±0.003	0.932±0.060	0.941±0.044	0.872±0.091	0.935±0.073
MSE (↓)	8.116±7.246	182.310±237.821	188.590±226.154	477.696±575.802	214.399±372.075
PSNR (↑)	39.676±2.223	29.066±6.117	29.043±6.480	23.663±4.711	30.356±7.772
FID (↓)	39.67	71.031	59.232	108.285	62.008
KID (↓)	0.012±0.002	0.022±0.003	0.018±0.002	0.043±0.007	0.021±0.002
LPIPS (↓)	0.720±0.084	0.757±0.089	0.750±0.086	0.787±0.089	0.756±0.099

Table 4: Overall results of LaMa inpainting different masking shapes with added noise on the original image.

	Line	Triangle	Circle	Rectangle	Overall
SSIM (↑)	0.991±0.004	0.928±0.063	0.938±0.044	0.862±0.104	0.930±0.079
MSE (↓)	11.444±5.945	149.311±182.372	139.465±157.483	374.033±495.354	168.563±304.792
PSNR (↑)	38.136±2.330	29.389±5.501	29.393±5.349	24.719±4.621	30.409±6.705
FID (↓)	2.795	6.414	5.148	12.726	4.457
KID (↓)	0.001±0.000	0.003±0.001	0.002±0.000	0.007±0.001	0.003±0.000
LPIPS (↓)	0.747±0.088	0.760±0.089	0.751±0.089	0.769±0.089	0.756±0.092

Table 5: PDE bias analysis

Shape	Attribute	SSIM (↑)	MSE (↓)	PSNR (↑)
Line	Male/Female	-0.005	**0.430**	0.006
	Young/Old	0.004	**-2.151**	0.006
	PaleSkin/DarkSkin	-0.001	**-2.360**	0.014
Triangle	Male/Female	-0.006	0.101	-0.010
	Young/Old	0.004	-0.126	0.007
	PaleSkin/DarkSkin	-0.009	**0.202**	-0.040
Circle	Male/Female	-0.004	0.054	0.015
	Young/Old	0.007	**-0.263**	0.003
	PaleSkin/DarkSkin	0.003	-0.139	0.025
Rectangle	Male/Female	-0.004	0.082	-0.019
	Young/Old	0.010	-0.075	0.020
	PaleSkin/DarkSkin	0.011	-0.126	-0.028
Overall	Male/Female	-0.005	0.091	0.000
	Young/Old	0.006	-0.161	0.008
	PaleSkin/DarkSkin	0.001	-0.058	-0.004

Table 6: LaMa bias analysis

Shape	Attribute	SSIM (↑)	MSE (↓)	PSNR (↑)
Line	Male/Female	0.000	-0.068	0.007
	Young/Old	0.000	0.060	-0.005
	PaleSkin/DarkSkin	0.000	0.107	-0.003
Triangle	Male/Female	0.000	0.029	0.001
	Young/Old	0.003	-0.070	0.008
	PaleSkin/DarkSkin	0.010	**-0.269**	0.038
Circle	Male/Female	0.004	-0.003	0.025
	Young/Old	-0.005	0.149	-0.036
	PaleSkin/DarkSkin	-0.002	0.289	-0.036
Rectangle	Male/Female	0.003	0.029	-0.007
	Young/Old	-0.003	0.145	-0.008
	PaleSkin/DarkSkin	0.001	0.191	0.010
Overall	Male/Female	0.002	0.022	0.007
	Young/Old	-0.001	0.105	-0.010
	PaleSkin/DarkSkin	0.002	0.153	0.002

Table 7: PDE bias analysis with noisy images

Shape	Attribute	SSIM (↑)	MSE (↓)	PSNR (↑)
Line	Male/Female	0.000	0.069	0.002
	Young/Old	0.001	**-0.267**	0.012
	PaleSkin/DarkSkin	0.002	**-0.411**	0.026
Triangle	Male/Female	0.006	-0.036	0.029
	Young/Old	0.016	-0.107	0.015
	PaleSkin/DarkSkin	0.024	**-1.063**	0.043
Circle	Male/Female	-0.009	0.098	-0.026
	Young/Old	0.005	0.188	-0.032
	PaleSkin/DarkSkin	0.011	**0.467**	-0.100
Rectangle	Male/Female	-0.015	0.187	-0.027
	Young/Old	0.011	**-0.200**	0.001
	PaleSkin/DarkSkin	-0.004	-0.037	-0.070
Overall	Male/Female	-0.004	0.125	-0.004
	Young/Old	0.008	-0.009	0.000
	PaleSkin/DarkSkin	0.008	0.050	-0.015

Table 8: LaMa bias analysis with noisy images

Shape	Attribute	SSIM (↑)	MSE (↓)	PSNR (↑)
Line	Male/Female	0.000	-0.040	0.006
	Young/Old	0.000	0.013	-0.001
	PaleSkin/DarkSkin	0.000	-0.094	0.011
Triangle	Male/Female	-0.001	-0.009	0.002
	Young/Old	0.005	-0.057	0.007
	PaleSkin/DarkSkin	0.017	**-0.250**	0.044
Circle	Male/Female	0.004	-0.048	0.016
	Young/Old	-0.005	0.151	-0.022
	PaleSkin/DarkSkin	-0.002	0.193	-0.016
Rectangle	Male/Female	-0.003	0.015	-0.006
	Young/Old	-0.002	0.114	-0.006
	PaleSkin/DarkSkin	0.008	0.172	0.005
Overall	Male/Female	0.000	-0.004	0.005
	Young/Old	-0.001	0.084	-0.005
	PaleSkin/DarkSkin	0.006	0.105	0.012

In terms of bias, we can see that in 5 and 6 that both methods do not have much bias at all, which makes sense as a pixel-wise PDE inpainter will not have any learned distribution outside of the image itself and ideally a Fourier transformed model in the frequency domain won't have bias as well. The MSE values in the PDE bias tables show that the line images have the most bias (bias being under 80 percent or over 120 percent of the baseline) but this is most likely due to the normalization applied to the images that brightened the original darker image in these samples. However, with noisy images, we can see in 3 that the PDE method has more bias in MSE for the circle and rectangle masks especially in the young/old category while the LaMa method doesn't increase bias at all. We also ran experiments to see if there is any difference in performance of all metrics based on the noise applied to the images but saw that there was no significant difference between the different noise (gaussian, gamma, uniform, exponential) added.

4 Related Work

There are lots of related work that explores the potential biases in generative networks. These papers conclude that these off the shelf models are biased significantly, especially on gender, race, and even facial expressions Zhou et al. [2024]Currie et al. [2024]. Specifically, Zhou et. al Zhou et al. [2024] tested Midjourney, Stable Diffusion, and DALL·E 2 by asking it to generate images of people with specific occupations and saw that there was a large disparity in represented gender across all occupations. Furthermore, they showed that there is a huge bias in the represented race of people in these occupations, though interestingly Stable Diffusion seemed to represent Asians more than white. However, as of writing this, there were no papers addressing any possible

Figure 2: Various examples of inpainting tasks. The left column shows the original image, PDE based inpainted images in the middle column, and the LaMa based inpainting on the right column.

Figure 3: Various examples of inpainting tasks with noise. The left column shows the original image, PDE based inpainted images in the middle column, and the LaMa based inpainting on the right column.

biases specifically on inpainting methods and different initial conditions in which the image is to be inpainted. While we did not find any huge biases in our experiments (perhaps because we didn't get to use complex enough models), it is not far-fetched to assume that the biases in generative models will be present in inpainting models.

5 Conclusion

For our baseline goal, we started looking for multiple models to work with and implement. We started with PyInpaint, based on partial differential equations (PDEs) in a graph framework. Its simplicity made it an excellent starting point in understanding the modifications and experimenting with different setups. Then, we tried multiple GAN-based inpainting models (region-wise-inpainting, generatitve-inpainting, pluralistic image completation) but did not include them in this project as it was difficult to modify for our initial purposes without extensive modifications to the code (though in hindsight, since we went with initial conditions of the input image, these models could have been implemented in our final experiments). Finally, we tested the LaMa model in the simple-lama-inpainting repository, which seeks to use Fourier convolutions for resolution-robust inpainting of big masked areas. All of these models had their own strengths and perspectives, and thus an overarching framework for the evaluation of different methodologies of image inpainting was developed.

Building on the baseline to achieve our medium goals, we finalized on implementing two inpainting models: PyInpaint (PDE) and LaMa. Selecting CelebA as our image dataset, the images were first used as it and then conditioned on different noise distributions, including Gaussian, uniform, exponential, and gamma.

We evaluated the outputs using reconstruction metrics such as PSNR, SSIM, MSE, FID, KID, and LPIPS. This allowed us to identify some correlations between specific conditions (adding masks, noise, and both) and inpainting behavior, uncovering potential biases and dependencies that merit further investigation. Unfortunately due to resource and time constraints, we were unable to complete the tasks outlined in the Stretch goals.

5.1 Limitations

Some limitations in our project were computational resources and time. Although we had access to the SOL cluster, since we didn't train models and just implemented and interrogated them, we chose to use Google Collab resources. Due to the space in Google Collab and time limitations (especially for running the PDE method - it sometimes took over 10 seconds per image based on the mask size) , we performed our experiments on about 1% of the whole CelebA dataset. Another limitation came from not having enough time to fully explore the model's code. LaMa outputs generated padded images of (224, 184, 3) vs the original CelebA dataset's size of (217,178,3) so initially the results of the LaMa looked horrible with SSIMs in the range of 50%. Once we cropped the image to the correct dimensions (we assumed the padding of LaMa to pad to the bottom and to the right - based on the observed artifacts), the results made more sense and had values we expected. However, if we had more time, maybe we could have explored the padding and generated outputs without padding for a strict comparison.

5.2 Future Work

In the future, we would aim to address the challenges outlined in the Stretch goals. This includes implementing additional inpainting models, such as GAN-based methods like GAN-image-inpainting, and to train them with different latent noise distributions to conduct a more comprehensive comparative study. We would also explore text-prompted inpainting to introduce semantic guidance into the reconstruction process and investigate techniques to debias the outputs systematically.

References

Geoffrey M Currie, K Elizabeth Hawk, and Eric M Rohren. Generative artificial intelligence biases, limitations and risks in nuclear medicine: An argument for appropriate use framework and recommendations. In *Seminars in Nuclear Medicine*. Elsevier, 2024.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jireh Jam, Connah Kendrick, Kevin Walker, Vincent Drouard, Jison Gee-Sern Hsu, and Moi Hoon Yap. A comprehensive review of past and present image inpainting methods. *Computer vision and image understanding*, 203:103147, 2021.

Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33 (12):6999–7019, 2021.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

Mi Zhou, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim, and Kannan Srinivasan. Bias in generative ai. *arXiv preprint arXiv:2403.02726*, 2024.

6 Supplementary Materials

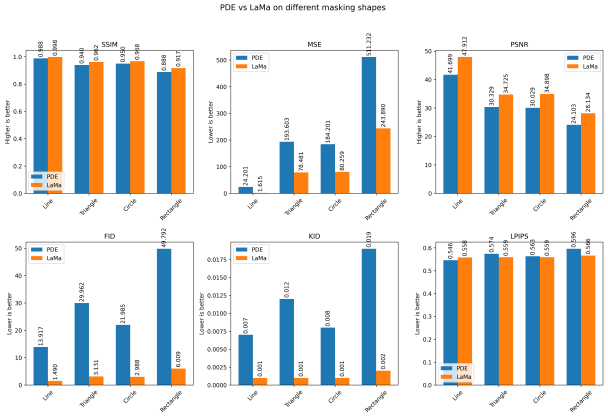

Figure 4: Comparison of the two models and their metrics on the original image.

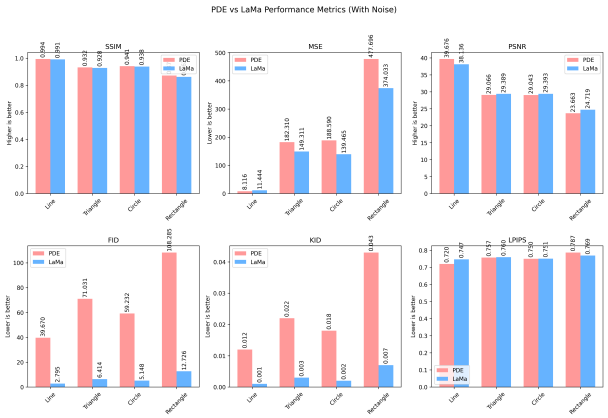

Figure 5: Comparison of the two models and their metrics on the noisy image.

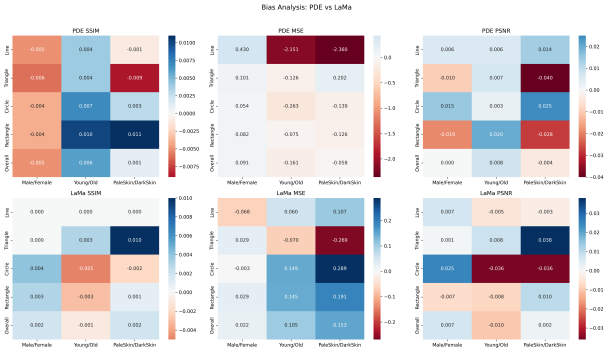

Figure 6: Comparison of the two models and their metrics on any potential bias on original images.

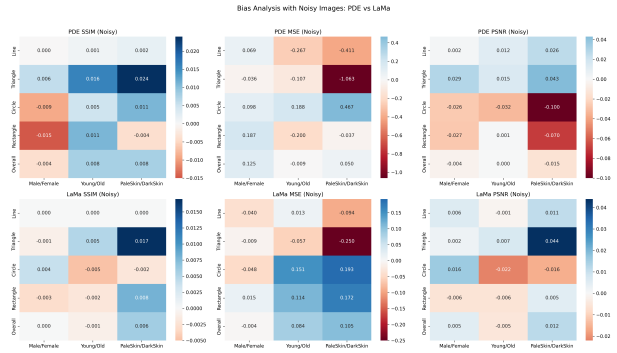

Figure 7: Comparison of the two models and their metrics on any potential bias on noisy images.