Image Inpainting using Context Encoders and Diffusion Model

Mansi Nanavati
Electrical and Computer Engineering
A69036329

Ajjkumar Patel
Electrical and Computer Engineering
A69036699

Abstract

Image inpainting, the task of restoring missing or corrupted image regions, is indispensable for photoediting, cultural heritage preservation, and privacy-aware vision systems. This paper presents a comparative study of image inpainting techniques using a Context Encoder and Stable Diffusion models. We implemented a lightweight Context Encoder–GAN that learns to synthesize content inside large square masks by combining an encoder–decoder generator with a patch-level discriminator. Trained end-to-end on CelebA-HQ, Cityscapes, and a filtered subset of Places365, the model achieves PSNR 22.33 dB / L1 0.0279 on CelebA-HQ and maintains real-time inference (< 1 s for 128 × 128 images) on GPUs. Qualitative results demonstrate sharp, colour-consistent face completions and acceptable scene reconstructions despite the network's compact footprint. A comparison with a frozen Stable Diffusion v2 inpainting model reveals a clear trade-off: diffusion offers higher perceptual quality but has more latency and occasional hallucinations. These findings position context encoders as an attractive baseline for resource-constrained or interactive applications, and motivate future hybrid approaches that blend generative efficiency with diffusion-level fidelity. GitHub - https://github.com/ajjpatel/Image_Inpainting

1 Problem Statement

Image inpainting infers missing parts in an image based on available regions specified by a binary mask. To achieve this goal, inpainting approaches use generative models modified to condition on the available image regions to produce high-quality inferences, i.e., semantic inpainting. Let $I \in \mathbb{R}^{3*H*W}$ be an RGB image and $M \in \{0,1\}^{H*W}$ a binary mask that zeros out pixels to be filled. The goal is to learn a mapping, $f_\theta = (I, M) \longrightarrow \hat{I}$, such that the reconstructed image $\hat{I}$ is perceptually and semantically consistent with the ground-truth image I over the masked region, while leaving unmasked pixels intact. For example, it can be used in image editing to remove unwanted image content, while filling in the resulting space with plausible imagery. Previous deep learning approaches have focused on rectangular regions located around the center of the image using Convolutional Neural Networks.

Traditional encoder-decoder approaches like Context Encoder (CE)[1] effectively capture global semantic structure (see Figure 1) but often produce blurry reconstructions. On the other hand, diffusion models [2], particularly in latent space (e.g., RePaint[3], LatentPaint[4]), generate high-quality, detailed inpainting results but are computationally expensive and struggle with global coherence without sufficient conditioning.

2 Related Work

Context Encoders: Feature Learning by Inpainting (Pathak et al., 2016)[1] is a foundational work in deep learning-based image inpainting that frames the task as a self-supervised feature

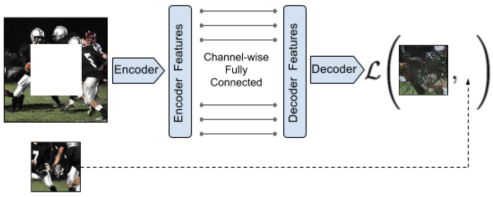


Figure 1: Context Encoder. The image is passed through the encoder to obtain features which are connected to the decoder using channel-wise fully-connected layer. The decoder produces the missing regions in the image.

learning problem. The model uses a convolutional encoder-decoder architecture trained to fill large missing regions conditioned on surrounding context, combining reconstruction and adversarial losses to produce semantically coherent outputs. Unlike traditional autoencoders that handle minor corruptions, Context Encoders address substantial image gaps, encouraging the network to learn high-level semantics. The learned features also generalize well to downstream tasks like classification and segmentation, outperforming earlier unsupervised approaches. This work has influenced many subsequent inpainting methods, particularly in leveraging generative models for both image restoration and representation learning.

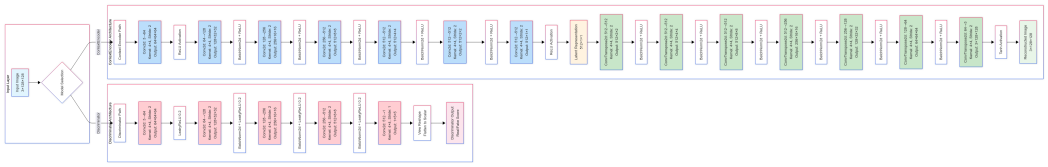


Figure 2: Context-Encoder Model

High-Resolution Image Synthesis with Latent Diffusion Models (Rombach et al.)[5] introduces latent diffusion models (LDMs), which perform the diffusion process not in pixel space but within the compressed latent space of a pretrained autoencoder. By doing so, they drastically reduce computational overhead while preserving high-frequency details, enabling high-quality synthesis at megapixel resolutions. Critically, the model employs cross-attention mechanisms that allow seamless conditioning on various inputs—such as text, segmentation maps, and inpainting masks. This conditioning capability yields state-of-the-art results in image inpainting, super-resolution, and semantic synthesis, outperforming pixel-based diffusion models both in fidelity and efficiency. By setting the foundation for models like Stable Diffusion, LDMs represent a major shift toward practical, versatile, and high-resolution generative approaches in inpainting research.

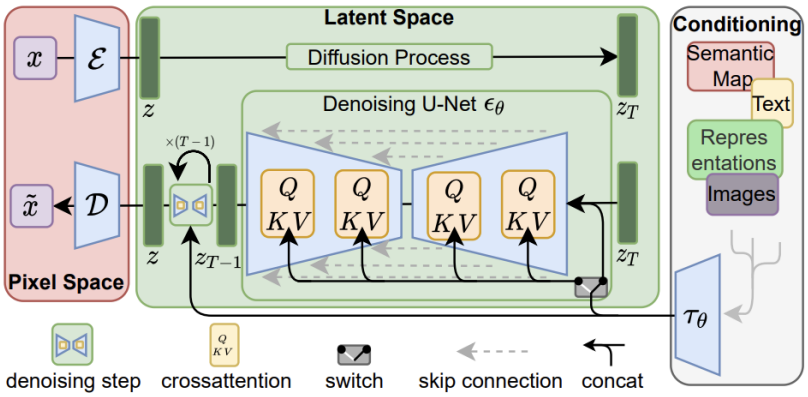


Figure 3: Latent Diffusion Model

3 Methodology

We have implemented two separate pipelines for the task of image inpainting. The first pipeline is a replication of the Context Encoder (CE) [1]. The second pipeline is image inpainting using a pre-trained stable diffusion model. The masking strategy is inspired by Pathak et al [1] and Liu et al.[6].

3.1 Context Encoder Pipeline

The Context Encoder proposed by Pathak et al.[1] employs a convolutional neural network (CNN) with an encoder-decoder architecture designed to reconstruct missing image regions based on surrounding context is implemented from scratch. The pipeline begins with an input image I of size 128x128, paired with a binary mask M identifying the region to inpaint. The encoder, composed of several convolutional layers with ReLU activations, downsamples the masked image into a compact latent representation. This representation captures spatial and semantic features from the visible pixels. The decoder then upsamples this latent code using transposed convolutions, reconstructing the full image $\hat{I}$. Figure 2 captures model architecture in detail. Formally, our model uses a combination of two loss functions for the generator (G) and one loss function for the discriminator (D). For the Generator (G), the total loss is:

$$L_G \;=\; \lambda_{\text{rec}}\, L_{\text{rec}} \;+\; \lambda_{\text{adv}}\, L_{\text{adv}} \quad \text{where} \tag{1}$$

$$L_{\text{rec}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \Big\| \big(G(x_{\text{masked}}) \odot M\big) \;-\; \big(x_{\text{original}} \odot M\big) \Big\|_2^2 \tag{2}$$

$$L_{\text{adv}} \;=\; -\mathbb{E}_{x\sim p_{\text{data}}}\big[\log D\big(G(x_{\text{masked}})\big)\big] \tag{3}$$

where, x_{masked} – input image containing the mask, M – binary mask, x_{original} – ground-truth image, $\odot$ – element-wise multiplication. And for the Discriminator (D), the loss is:

$$L_D \;=\; -\mathbb{E}_{x\sim p_{\text{data}}}\big[\log D(x_{\text{original}})\big] \;-\; \mathbb{E}_{x\sim p_{\text{data}}}\big[\log\big(1 - D(G(x_{\text{masked}}))\big)\big] \tag{4}$$

In our implementation, the Context Encoder was trained on three datasets: CelebA-HQ (resized to 256x256 then cropped/padded to 128x128), Cityscapes, and Places365. Training utilized the Adam optimizer with a learning rate of $2 * 10^{-4}$, targeting minimization of the loss over the masked region. We will compare the results of this experiment qualitatively and quantitatively with the evaluation metrics mentioned below.

3.2 Stable Diffusion Pipeline

Here, we have used the pre-trained model "stable-diffusion-2-inpainting" by Stability AI. This model is resumed from stable-diffusion-2-base and trained for another 200k steps. It followed the mask-generation strategy presented in LAMA [7], which, in combination with the latent VAE representations of the masked image, is used as an additional conditioning. Initially, the idea was to integrate the diffusion techniques in the context encoder, but that proved to be difficult because of Datahub constraints (10 GB storage, 1 GPU, 16GB RAM). In this pipeline, the inputs are the original image, the mask, and the text prompt. This information passes to the Variational Autoencoder (VAE), which compresses it to a "latent space" via an Encoder while conditioning with masked image latents. The UNet model denoises the latent representation, guided by text and inpainting masks. We are passing an empty text prompt because the pipeline performs the inpainting without extra instructions. The VAE decoder reconstructs the image from the latent space output of the UNet model, which is our final product.

3.3 Datasets

We use three datasets in this project: CelebA-HQ[8], Cityscapes [9] and the Places365 dataset[10]. CelebA-HQ is a high-quality dataset of human faces with 30,000 images at 1024×1024 resolution, perfect for a baseline comparison. The Places2 dataset contains more than 2.5 million images covering more than 205 scene categories, with more than 5,000 images per category. Since the Places2 dataset is quite huge, we intend to use Places365, a filtered version of the dataset. Places365 has 260k images with a similar level of diversity in images that Places2 offers. While preparing each image for training

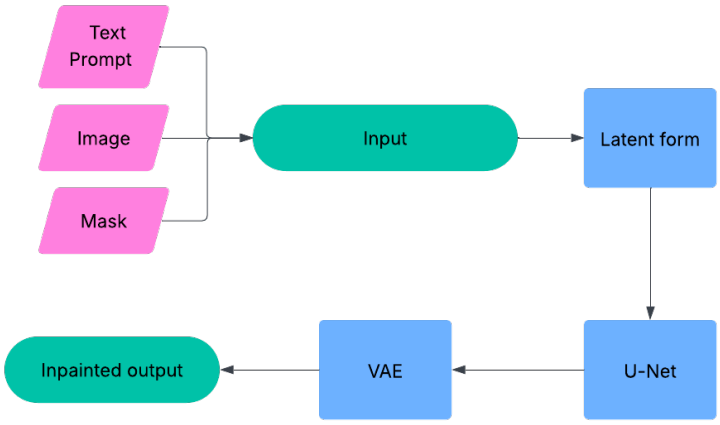


Figure 4: Pipeline for image inpainting using pre-trained diffusion model

in Context Encoder, we centered-cropped and resized it to (128 x 128), then normalized it. For the stable diffusion pipeline, we resized each image to (128 x 128) and converted it to a tensor.

3.4 Masking Strategy

The input to our architecture is an image with one or more of its regions "dropped out"; i.e., set to zero. Masks are generated that render regions of an image empty. We follow the masking strategies suggested by Pathak et al.[1] and Liu et al.[6]: central square patch, random blocks, and irregular masks (see 5). These masks are controlled by varying the hole-to-image area ratios. We generated many versions of the mask that vary in hole-to-image area ratio and the shape as part of our experiments. We have 6 kinds of masks: square, circle, triangle, ellipse, irregular, and random patches.

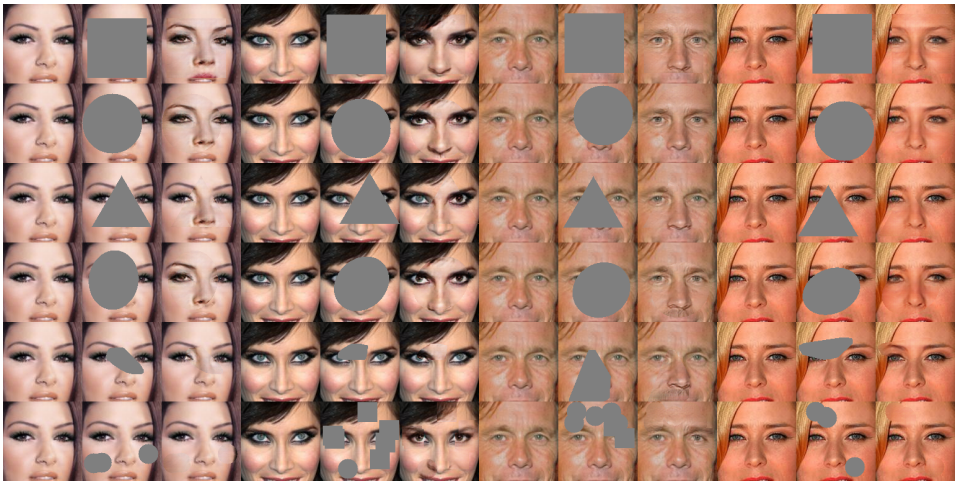


Figure 5: Various masking strategies applied on the images

3.5 Evaluation Metrics

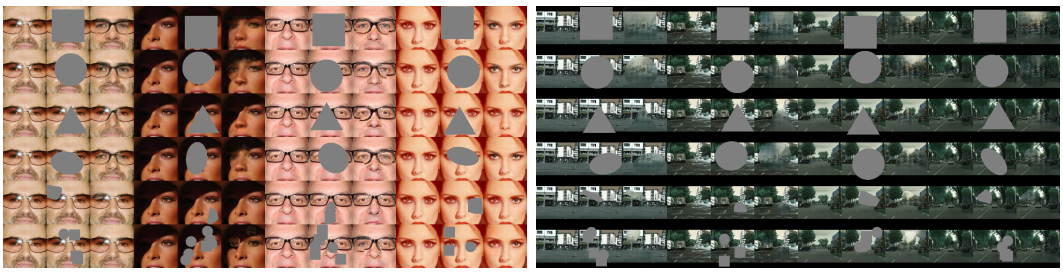
As part of qualitative comparison, we present visualizations that demonstrate the ability of the context encoder and the stable diffusion model. Common evaluation metrics include L1 and L2 loss functions as well as Peak Signal-to-Noise Ratio (PSNR). L1 loss (mean absolute error) and L2 loss (mean squared error) measure pixel-wise differences between the inpainted output and the ground truth, with L1 being more robust to outliers and often leading to sharper reconstructions. L2 loss, while more sensitive to large errors, emphasizes smoothness and is useful for penalizing significant deviations. In addition to these, PSNR is widely used to assess the perceptual quality of reconstructed images by quantifying the ratio between the maximum possible pixel value and the mean squared error. Higher PSNR values indicate better fidelity to the ground truth, making it a valuable metric for comparing inpainting models. Together, these metrics offer a balanced assessment of both the pixel accuracy and perceptual quality of inpainted results.

4 Results and Discussions

Model	Dataset	L1 Loss	L2 Loss	PSNR
NN-inpainting (Baseline)	Paris StreetView	9.37%	1.96 %	18.58 dB
Context Encoder	CelebA-HQ	3.22%	0.88%	21.28dB
Context Encoder	Cityscapes	6%	2.98%	16.17dB
Context Encoder	Places365	6.99%	3.93%	15.01dB
Stable Diffusion	**CelebA-HQ**	**2.79%**	**0.81%**	**22.33 dB**
Stable Diffusion	Cityscapes	3.52%	1.19%	19.90 dB

Table 1: L1, L2 losses and PSNR values for each model and dataset

In assessing the Context Encoder's inpainting capabilities across CelebA-HQ, Cityscapes, and Places365, we utilized Mean L1 Loss, Mean L2 Loss, and PSNR as key metrics. To test model training capabilites we evaluated CelebA-HQ results on train dataset. The model shines on CelebA-HQ (Figure 5, 6a), achieving a PSNR of 21.28 dB with Mean L1 and L2 losses of 0.0322 and 0.0088, respectively, owing to the structured, low-entropy nature of face images where symmetry and uniform textures are effectively captured. In contrast, its performance wanes on Cityscapes (Figure 6b) and Places365 (Figure 7), which present high-entropy scenes with diverse objects and textures, yielding PSNR values of 16.17 dB and 15.01 dB, and elevated loss values. Detailed results on evaluation are captured in Table 1. This drop-off reveals the model's difficulty in generalizing to complex visual content, often resulting in blurry or less detailed inpainted regions. When benchmarked against a baseline NN-inpainting model, which recorded a PSNR of 18.58 dB on Paris StreetView, the Context Encoder demonstrates a clear advantage on structured datasets like CelebA-HQ, yet it struggles to maintain this edge across the more intricate scenes of Cityscapes and Places365. These findings highlight the model's strengths in real-time, on-device applications for predictable, low-semantic images.



(a) CE results on CelebA-HQ dataset (b) CE results on Cityscapes dataset

Figure 6: Example output from context encoder model

Generally, the inpainted images from the stable diffusion model are of high quality. The pipeline not only performs inpainting but also scales the image and increases the size of the image from 128x128 to 512x512. It achieves the lowest L1 and MSE loss scores and has the highest PSNR score. This

Figure 7: CE results on blurry predictions for Places365 dataset

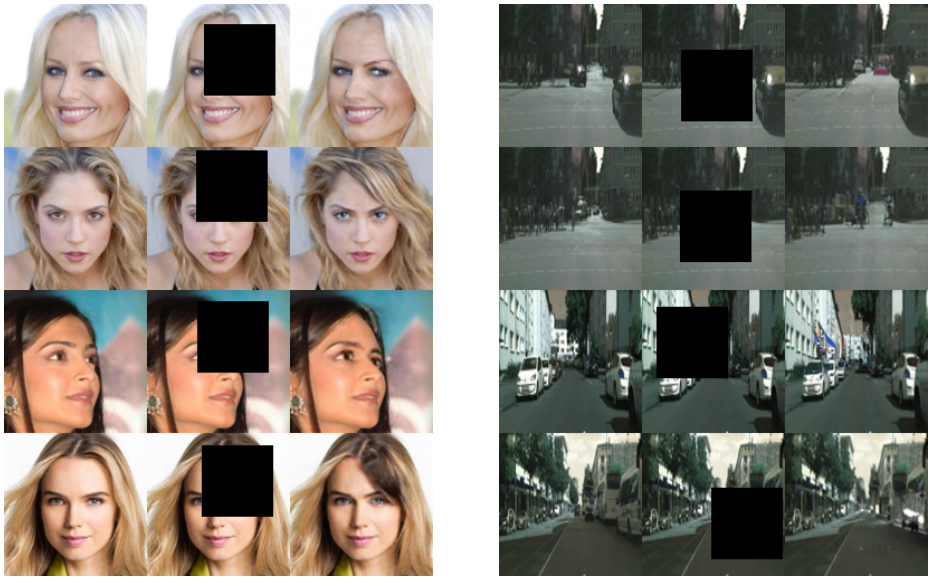
shows that it is able to successfully reconstruct a high-quality image. The model tends to hallucinate sometimes and generates garbage output in the masked area, a pattern very common with circular masks. The inference time for each image is quite high (around 5 seconds). We were unable to run the model on the Places365 dataset because of tight memory constraints on Datahub, so we generated results for CelebA-HQ and Cityscapes datasets.

(a) Original image (128x128)

(b) Masked image (128x128)

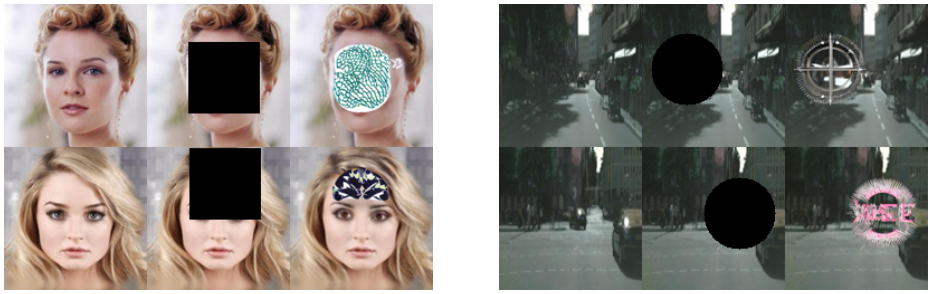
(c) Inpainted image (512x512)

Figure 8: Example output from stable diffusion pipeline



(a) CelebA dataset (b) Cityscapes dataset

Figure 9: Visualization of output from stable diffusion pipeline over datasets



(a) CelebA dataset (b) Cityscapes dataset

Figure 10: Limitations: Diffusion model hallucinates and generates garbage, especially with circular masks.

Acknowledgments

We acknowledge the use of LLMs in preparing this report. It was used to assist with language refinement, organization, formatting, code debugging and code formatting. All technical ideas, structure and methodology were developed and reviewed by the authors.

References

[1] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *CoRR*, vol. abs/1604.07379, 2016. arXiv: 1604.07379. [Online]. Available: http://arxiv.org/abs/1604.07379.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *CoRR*, vol. abs/2006.11239, 2020. arXiv: 2006.11239. [Online]. Available: https://arxiv.org/abs/2006.11239.

[3] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool, *Repaint: Inpainting using denoising diffusion probabilistic models*, 2022. arXiv: 2201.09865 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2201.09865.

[4] C. Corneanu, R. Gadde, and A. M. Martinez, "Latentpaint: Image inpainting in latent space with diffusion models," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 4322–4331. DOI: 10.1109/WACV57701.2024.00428.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, vol. abs/2112.10752, 2021. arXiv: 2112.10752. [Online]. Available: https://arxiv.org/abs/2112.10752.

[6] G. Liu, F. A. Reda, K. J. Shih, T. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," *CoRR*, vol. abs/1804.07723, 2018. arXiv: 1804.07723. [Online]. Available: http://arxiv.org/abs/1804.07723.

[7] R. Suvorov, E. Logacheva, A. Mashikhin, *et al.*, "Resolution-robust large mask inpainting with fourier convolutions," *arXiv preprint arXiv:2109.07161*, 2021.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *CoRR*, vol. abs/1710.10196, 2017. arXiv: 1710.10196. [Online]. Available: http://arxiv.org/abs/1710.10196.

[9] M. Cordts, M. Omran, S. Ramos, *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, *Places: An image database for deep scene understanding*, 2016. arXiv: 1610.02055 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1610.02055.