Image Inpainting using Context Encoders and Diffusion Model

Mansi Nanavati, Ajjkumar Patel

UC San Diego

Problem Statement

- Given a masked image with missing content, learn a function $f(I,M) \rightarrow \hat{I}$ that reconstructs the masked region.
- Various datasets
- Model selection

Motivation

- Real World Need
- Limitations of Classical Methods
- Trade-off Exploration

Approach

Two different techniques

- Implemented context encoders from scratch
- Built a pipeline with pretrained stable diffusion model

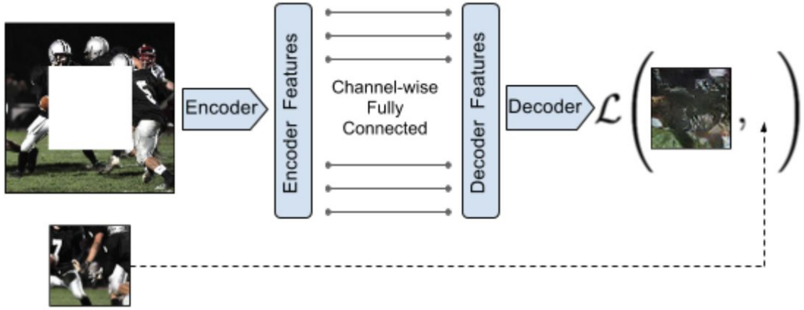
Dataset

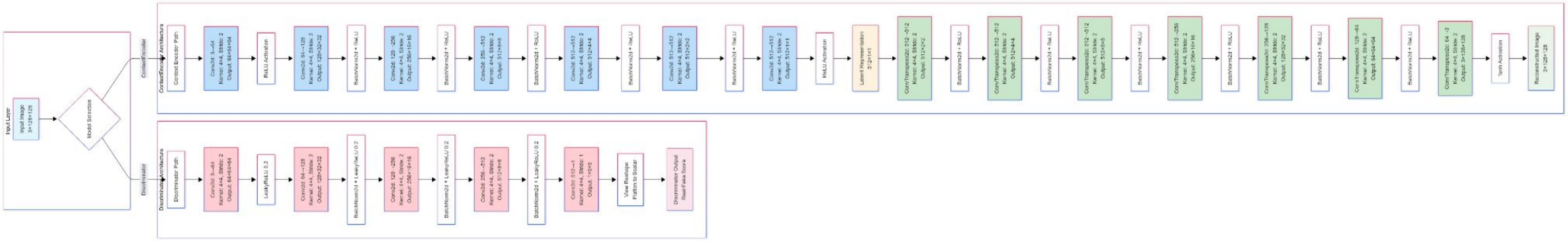
- Images of good quality
- Smaller dataset size and image size (128 x 128)
- Easy to use

We decided to go with 3 datasets:

❖ CelebA-HQ resized to 256x256
❖ Cityscapes
❖ Places365 (filtered version of Places2 dataset)

Context Encoder – Overview





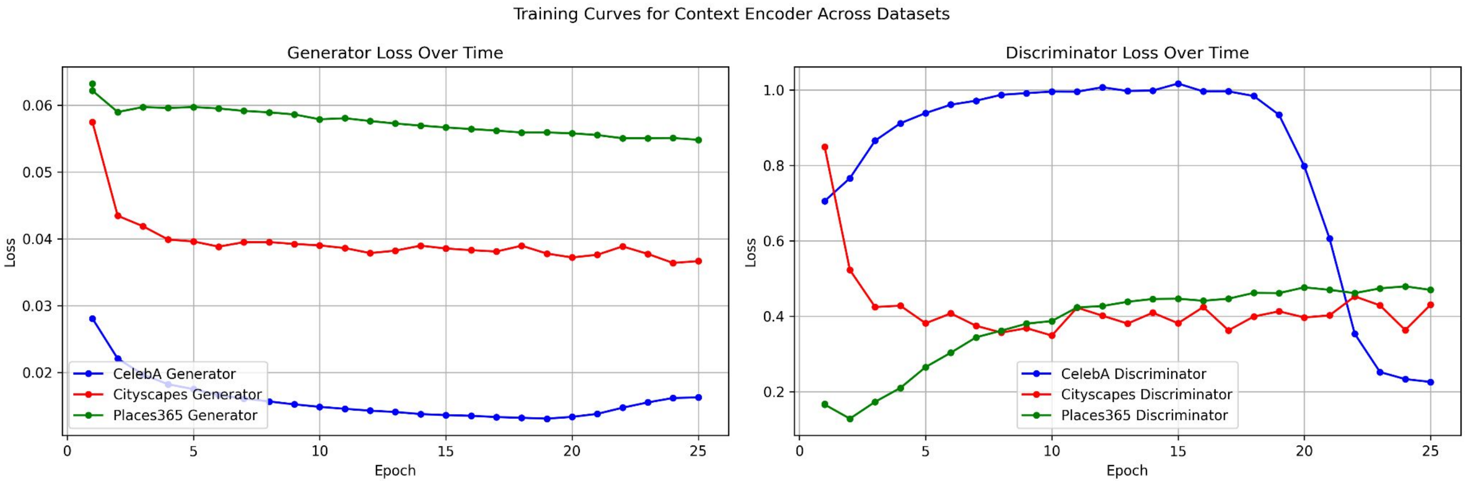
Context Encoder - Model Architecture

Encoder	Decoder	Discriminator
Conv2d(3, 64, 4, 2, 1), ReLU	ConvTranspose2d(512, 512, 4, 2, 1), BatchNorm2d(512), ReLU	Conv2d(3, 64, 4, 2, 1), LeakyReLU(0.2)
Conv2d(64, 128, 4, 2, 1), BatchNorm2d(128), ReLU	ConvTranspose2d(512, 512, 4, 2, 1), BatchNorm2d(512), ReLU	Conv2d(64, 128, 4, 2, 1), BatchNorm2d(128), LeakyReLU(0.2)
Conv2d(128, 256, 4, 2, 1), BatchNorm2d(256), ReLU	ConvTranspose2d(512, 512, 4, 2, 1), BatchNorm2d(512), ReLU	Conv2d(128, 256, 4, 2, 1), BatchNorm2d(256), LeakyReLU(0.2)
Conv2d(256, 512, 4, 2, 1), BatchNorm2d(512), ReLU	ConvTranspose2d(512, 256, 4, 2, 1), BatchNorm2d(256), ReLU	Conv2d(256, 512, 4, 2, 1), BatchNorm2d(512), LeakyReLU(0.2)
Conv2d(512, 512, 4, 2, 1), BatchNorm2d(512), ReLU	ConvTranspose2d(256, 128, 4, 2, 1), BatchNorm2d(128), ReLU	Conv2d(512, 1, 4, 1, 0)
Conv2d(512, 512, 4, 2, 1), BatchNorm2d(512), ReLU	ConvTranspose2d(128, 64, 4, 2, 1), BatchNorm2d(64), ReLU	
Conv2d(512, 512, 4, 2, 1), ReLU	ConvTranspose2d(64, 3, 4, 2, 1), Tanh	

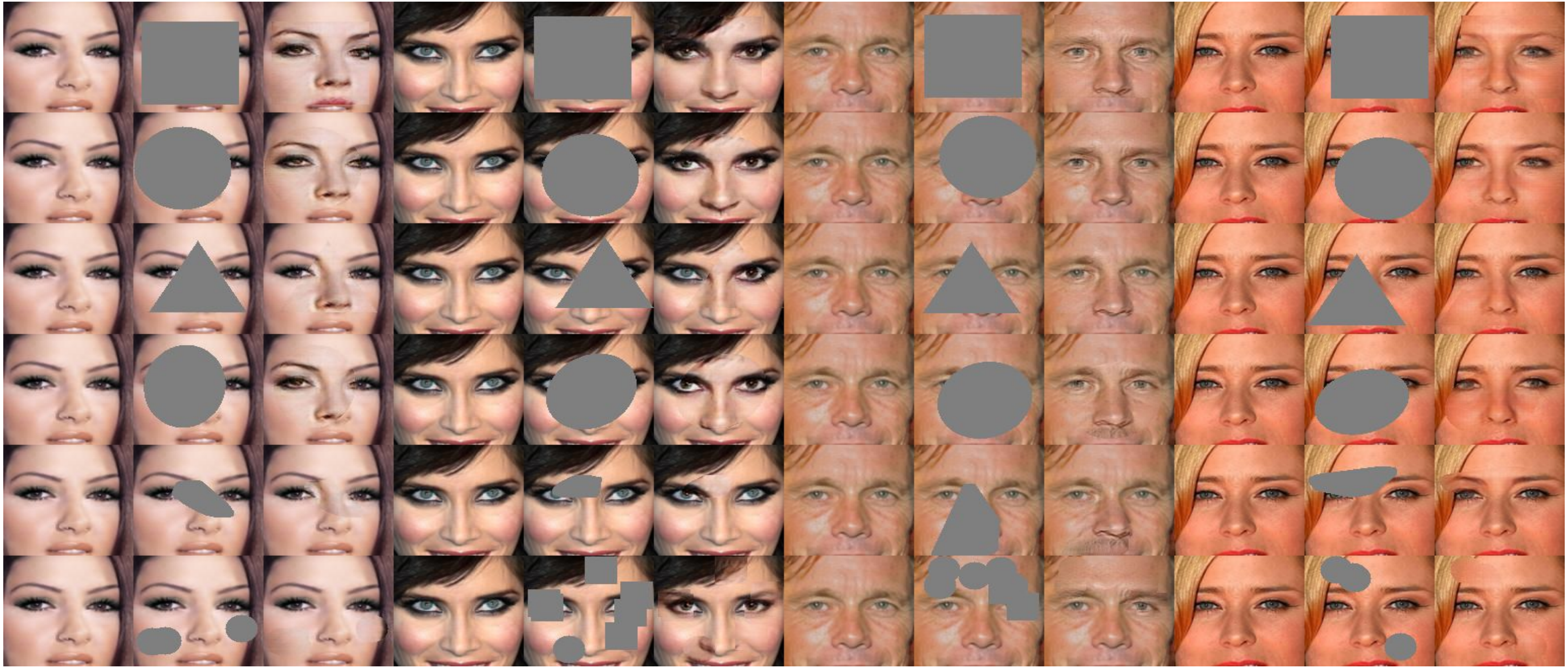
Context Encoder - Training Details

- Faces are a comparatively low-entropy manifold, so a context encoder can memorise structural priors quickly.
- We tested Celeba dataset on training data to verify training model. Cityscape and Places365 results are tested on validation dataset.

Dataset	Mean L1 Loss	Mean L2 Loss	PSNR (higher better)
Celeba - Train	0.0322	0.0088	21.28
Cityscape - Val	0.0600	0.0298	16.17
Places365(Filtered) - Val	0.0699	0.0393	15.01



Results: CelebA - HQ

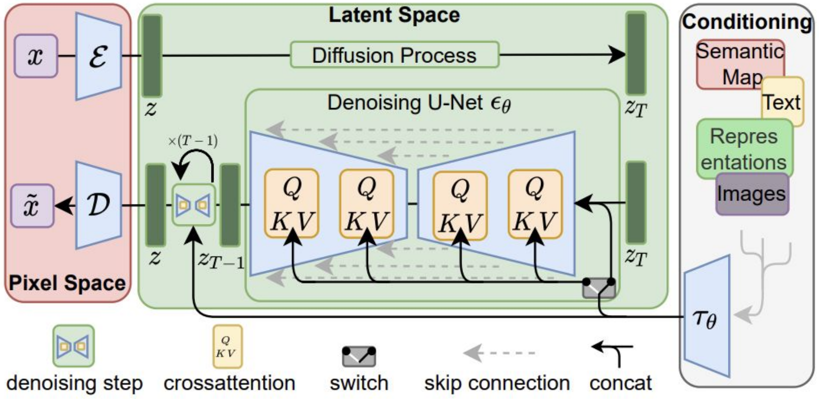


Results - Places365

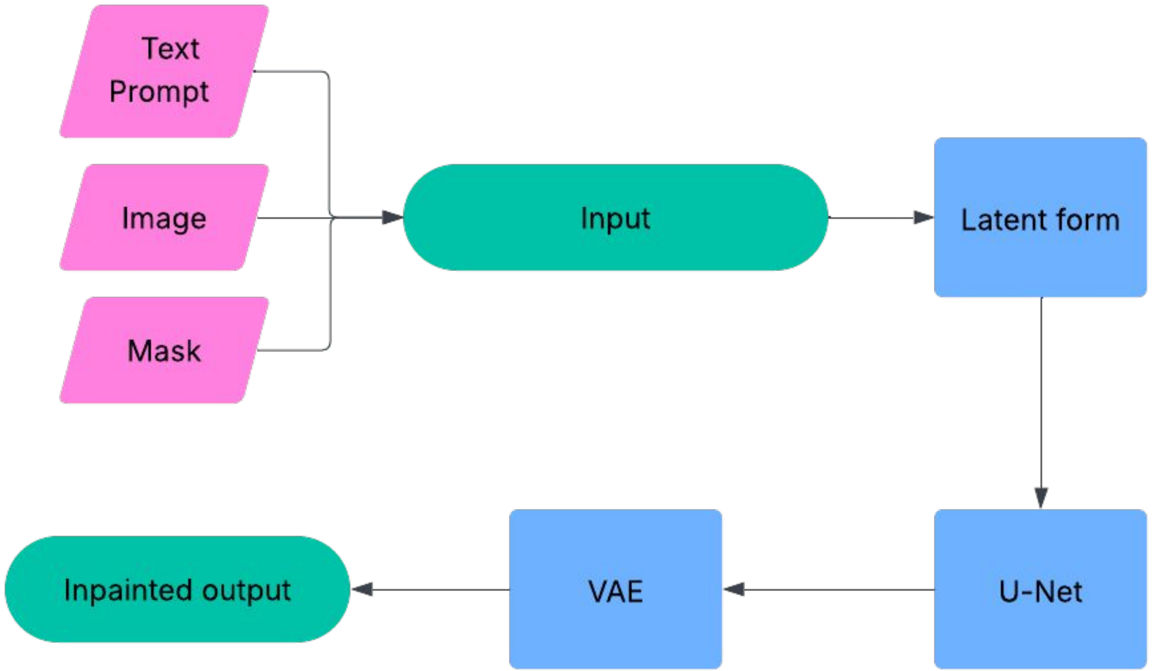
Stable Diffusion

UC San Diego

- Training a diffusion model needs a lot of compute which is a constraint
- Stabilityai's Stable Diffusion v2 pretrained model
- Latent Diffusion Model
- The inpainting model was trained on stable diffusion v2 base for another 200k steps with masking strategy from LaMa (Large Mask Inpainting with Fourier Convolutions)
- Output size: 512 x 512
- The output does more than inpainting



Stable Diffusion



Results

Original
(128 x 128)

Masked
(128 x 128)

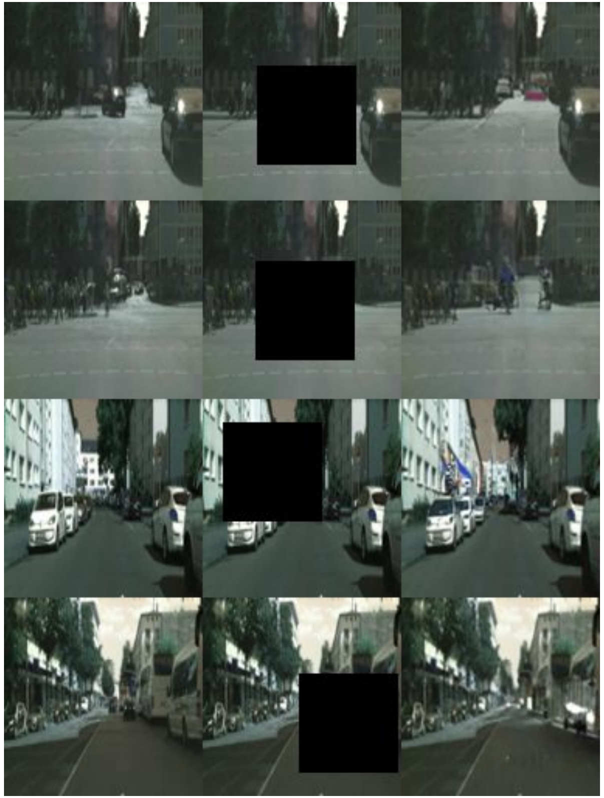
Inpainted (512 x 512)

Results

Mask Type: square
Mask Size: 64
Image Size: 128

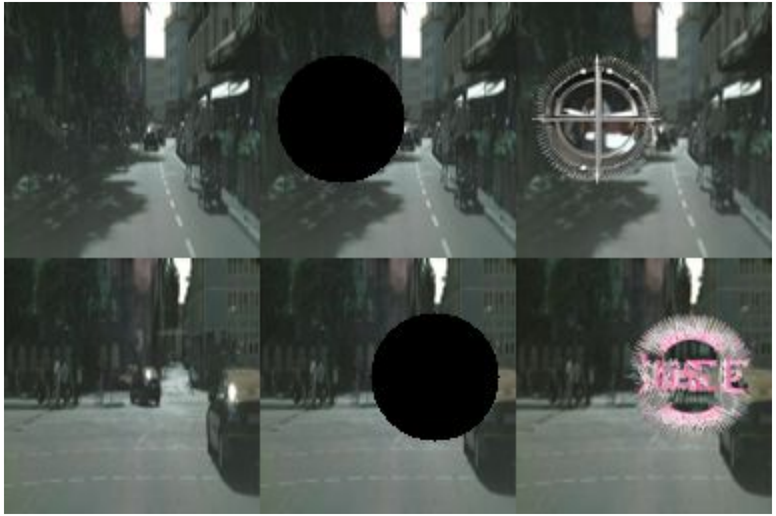
CelebA-HQ:
Mean L1 Loss: 0.0279
Mean L2 Loss: 0.0081
PSNR: 22.33 dB

Cityscapes:
Mean L1 Loss: 0.0352
Mean L2 Loss: 0.0119
PSNR: 19.90 dB



UC San Diego

Limitations



- Sometimes it hallucinates and generates garbage while inpainting, especially with circular masks.
- Moderate inference time (<5 seconds)

Final Results + Analysis

UC San Diego

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

Final Results + Analysis

UC San Diego

- Vanilla Context Encoder:
  - Excel on local, low-semantic images - Texture continuity and colour harmony are consistently high for Celeba dataset.
  - Blurry output on complex images from Cityscapes and Places365 dataset.
  - <1 sec for inference - Ideal for on-device, real-time use cases.

- Stable Diffusion inpainting:
  - Outperforms the baseline and the context encoder
  - Diffusion models that are pre-trained on vast datasets hallucinate occasionally
  - Moderate inference time