Using Image Outpainting to Blend Images

By Wentao Ye, Mitchell Krieger, and Sebastian Jay

Outpainting: Can we extrapolate beyond the borders of the image?

Challenges:

- Extends into unknown regions with less context
- Unknown regions must attend to inputs near the edges and far from the edges
- Unknown regions should be similar but different from inputs

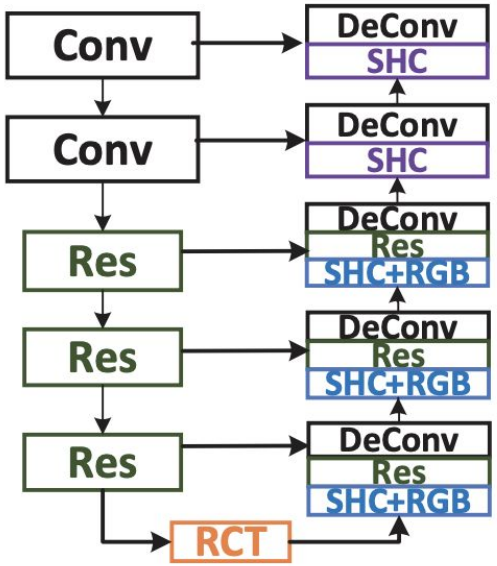
(a) Horizontal extrapolation. (b) Generalised image outpainting.

"Generalised Image Outpainting with U-Transformer" by Penglei Gao et al

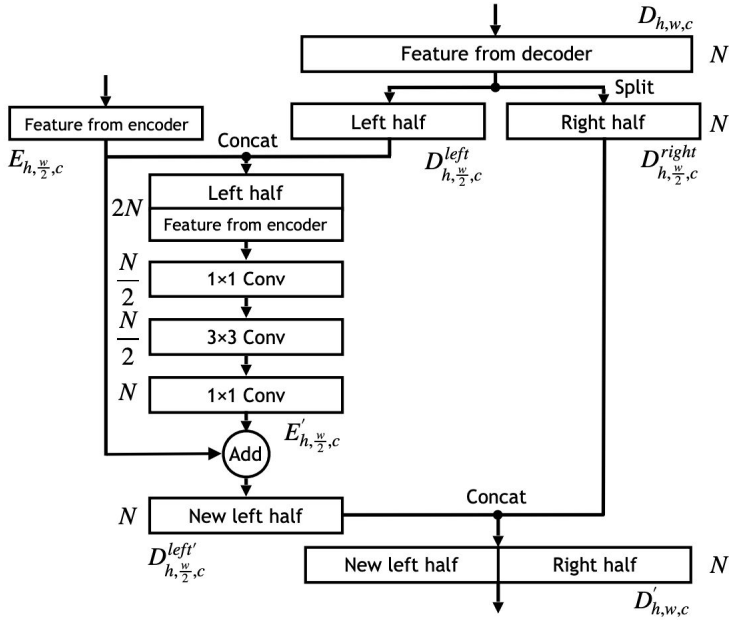
Our Challenge

If we provide two images, can we outpaint towards each other so that the images blend together? (Similar to Lu et al. 2021)

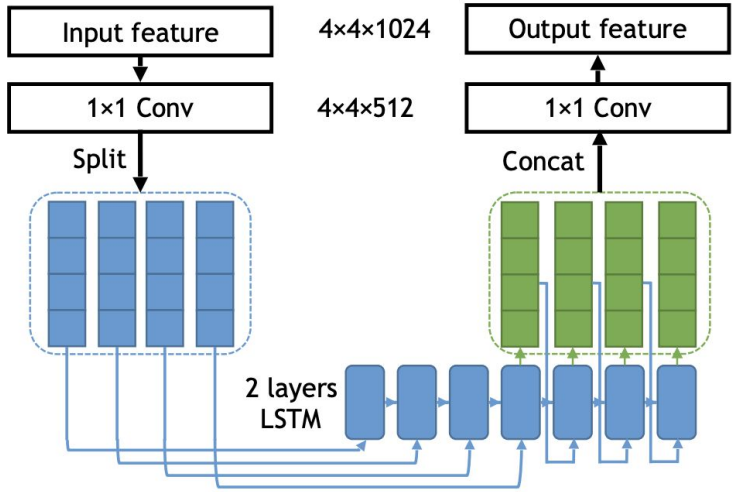
Prior Approaches: Unets



Adjusted Unet for outpainting



Skip horizontal connections (SHC): Allows for the spatial size of the encoder feature to be different than the decoder feature



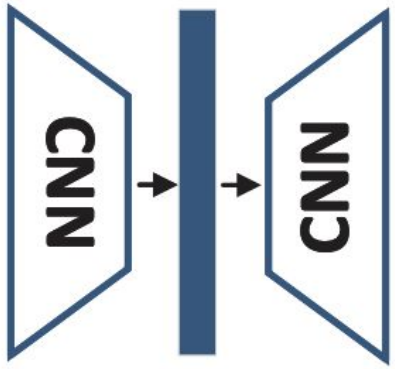
Recurrent Content Transfer (RCT): An LSTM based sequence prediction

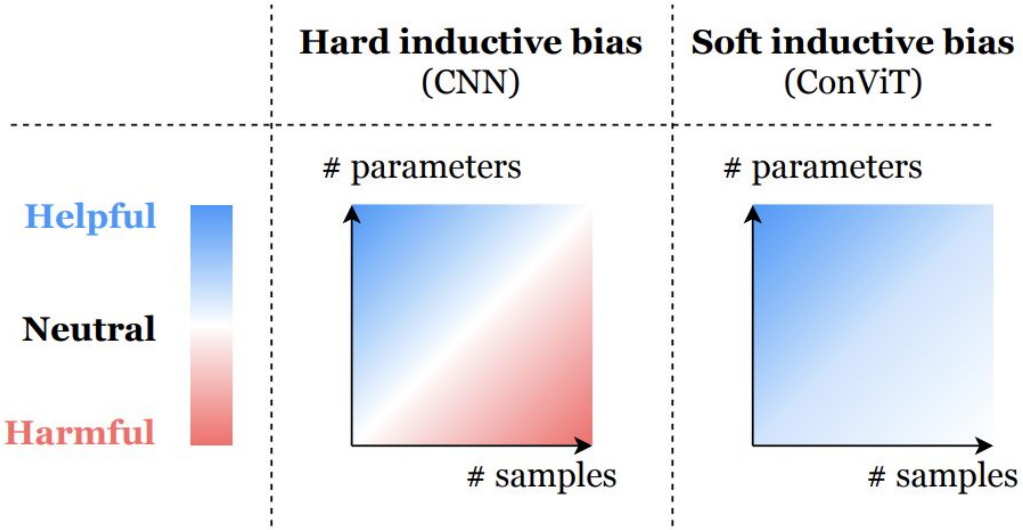
Very Long Natural Scenery Image Prediction by Outpainting Yang et al. (2019)

Prior Approaches: Convolutional & Transformer

CNN, Vision Transformers & ConVit:

- CNNs have hard inductive biases, meaning that when data is easily available at scale the structure of CNNs learning spatial relationships is overly restrictive
- Vision Transformers use self attention and don't have this same inductive bias but requires pre-training on large amounts of data
- ConVit introduces a soft inductive bias to bridge the between CNNs and Vision transformers
- Despite this, these methods are still limited by locality and have difficulty capturing global features and long-range semantic information





ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases
d'Ascoli et al. (2021)

Prior Approaches: GANs

UNets combined with GAN architecture and attention lead to increased performance!

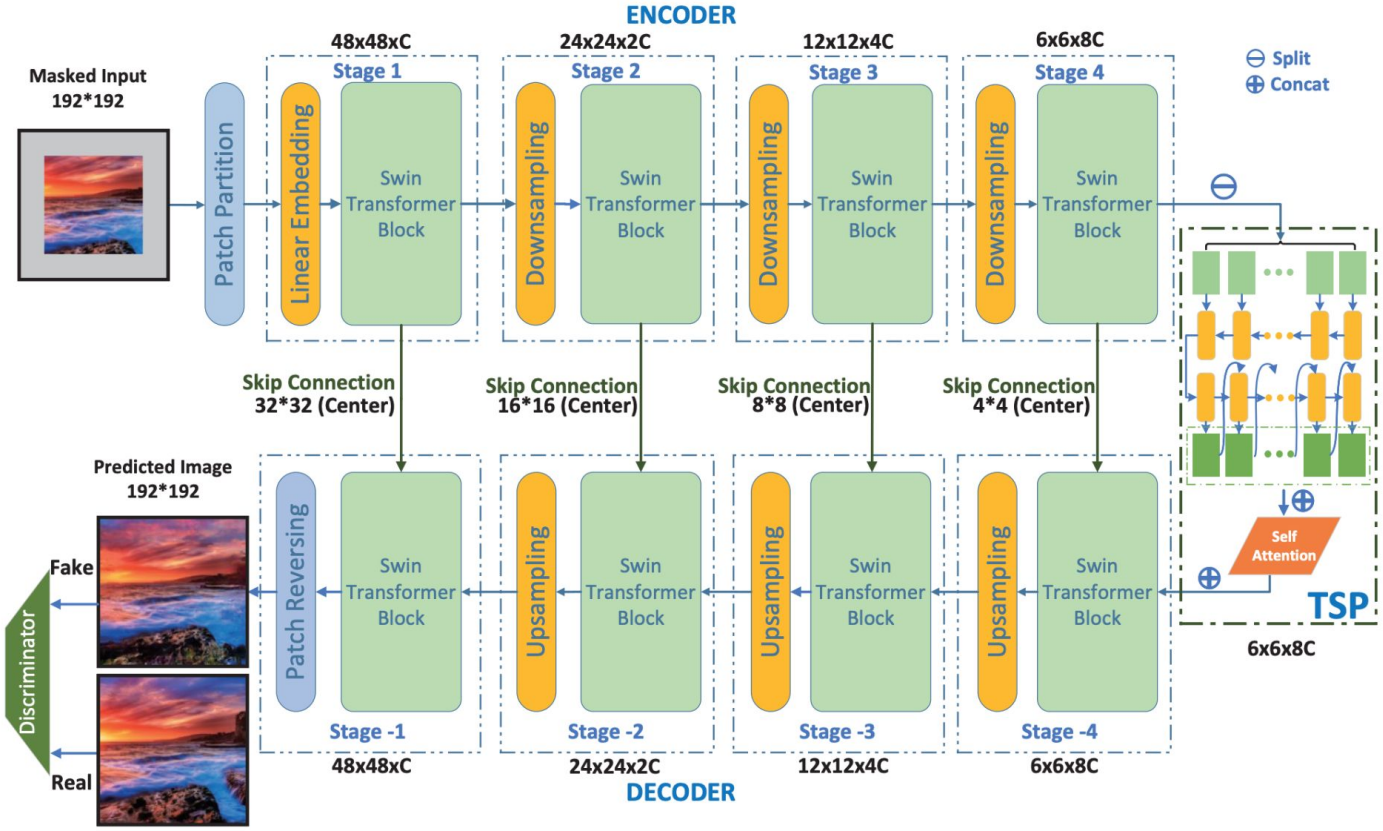
Lu et al (2021), attempted to outpaint images towards each other and that they blend together to create an entire panorama:

Bridging the Visual Gap: Wide-Range Image Blending, Lu et al. (2021)

Prior Approaches: UTransformer

- Uses Unet & GAN architecture with Swin Transformer.
- Swin (**S**hifted **Win**dows) an adaption of the transformer specifically for computer vision which leverages hierarchical feature extraction to handle spatial tasks in an efficient way by applying attention in windows

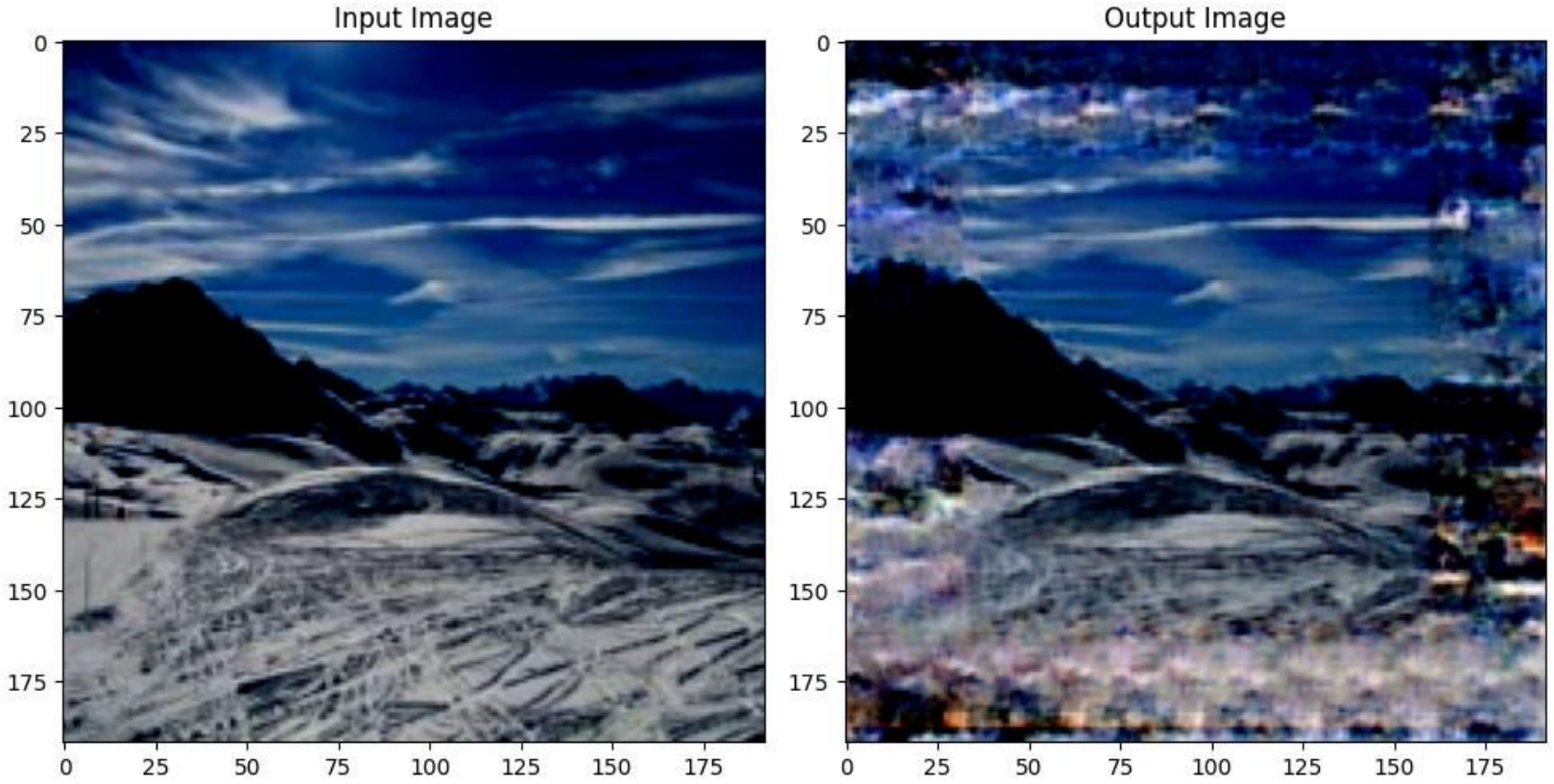
- Introduces a Temporal Spatial Predictor (TSP) as a bottleneck to ensure that spatial and temporal relationships are propagated
- TSP splits encoded features into bars and a 2-layer LSTM is used to process the sequence for each bar and then is passed to a self attention layer.



"Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", Liu et al. (2021)
"Generalised Image Outpainting with U-Transformer" by Penglei Gao et al (2022)

UTransformer

We were able to replicate the results of the UTransformer paper.



Results after 14 epochs

“Generalised Image Outpainting with U-Transformer” by Penglei Gao et al (2022)

Drawbacks to U-Transformer Model

1. Very slow and compute-intensive to train
   - 1 hour to run 14 epochs on a NVIDIA A100
   - 3 hours to train 1 epoch on RTX 2070.
2. Fixed size for center-expansion
   - 128 * 128 -> 192 * 192
   - Hard code in model

Our Models

1. GAN Model: Inspired by Image-to-Image Translation with Conditional Adversarial Networks

2. Diffusion Model: Inspired by “Repaint: Inpainting using Denoising Diffusion Probabilistic Models”, Lugmayr et al. (2022)

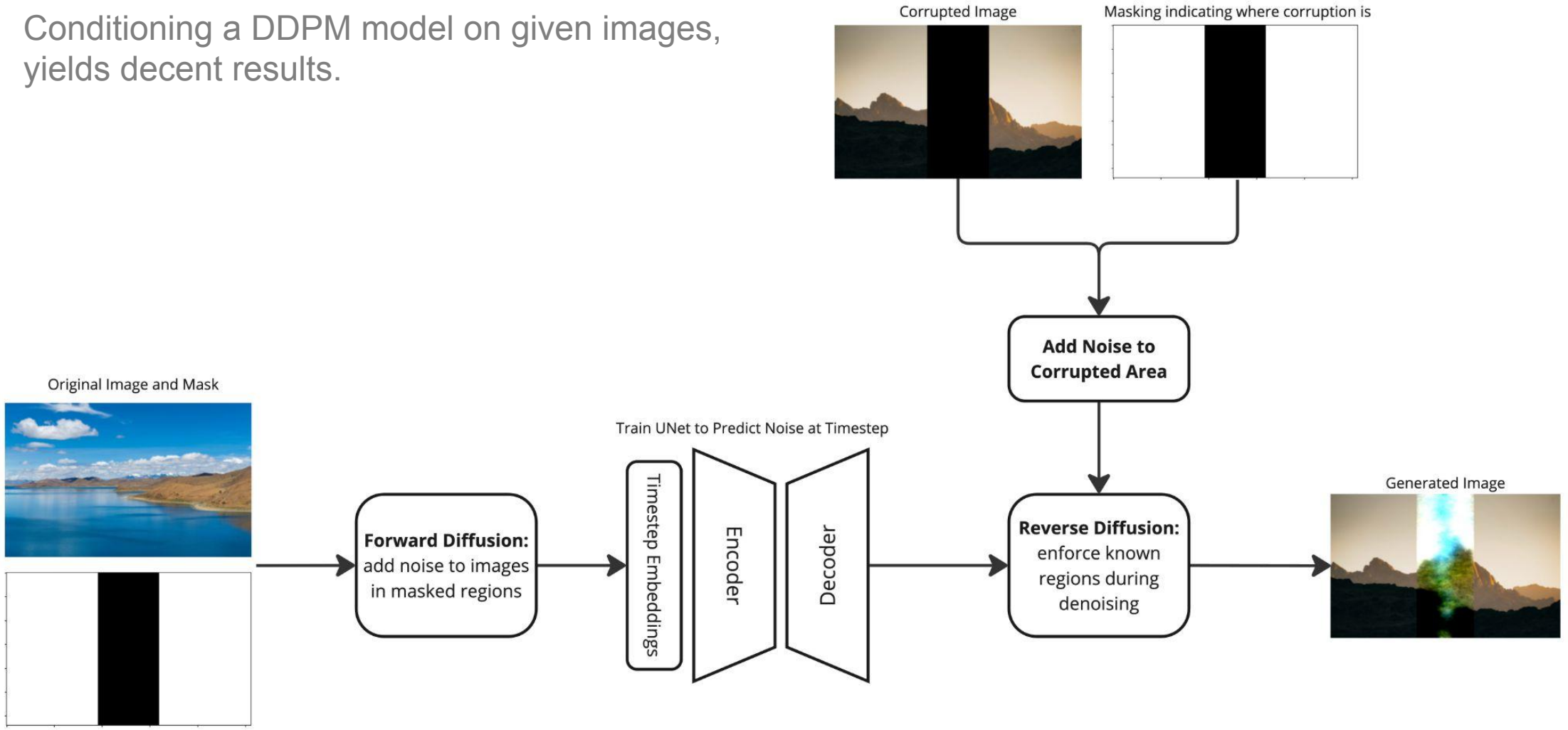
Our Dataset: NS-Outpainting

6,000 very wide images of scenery

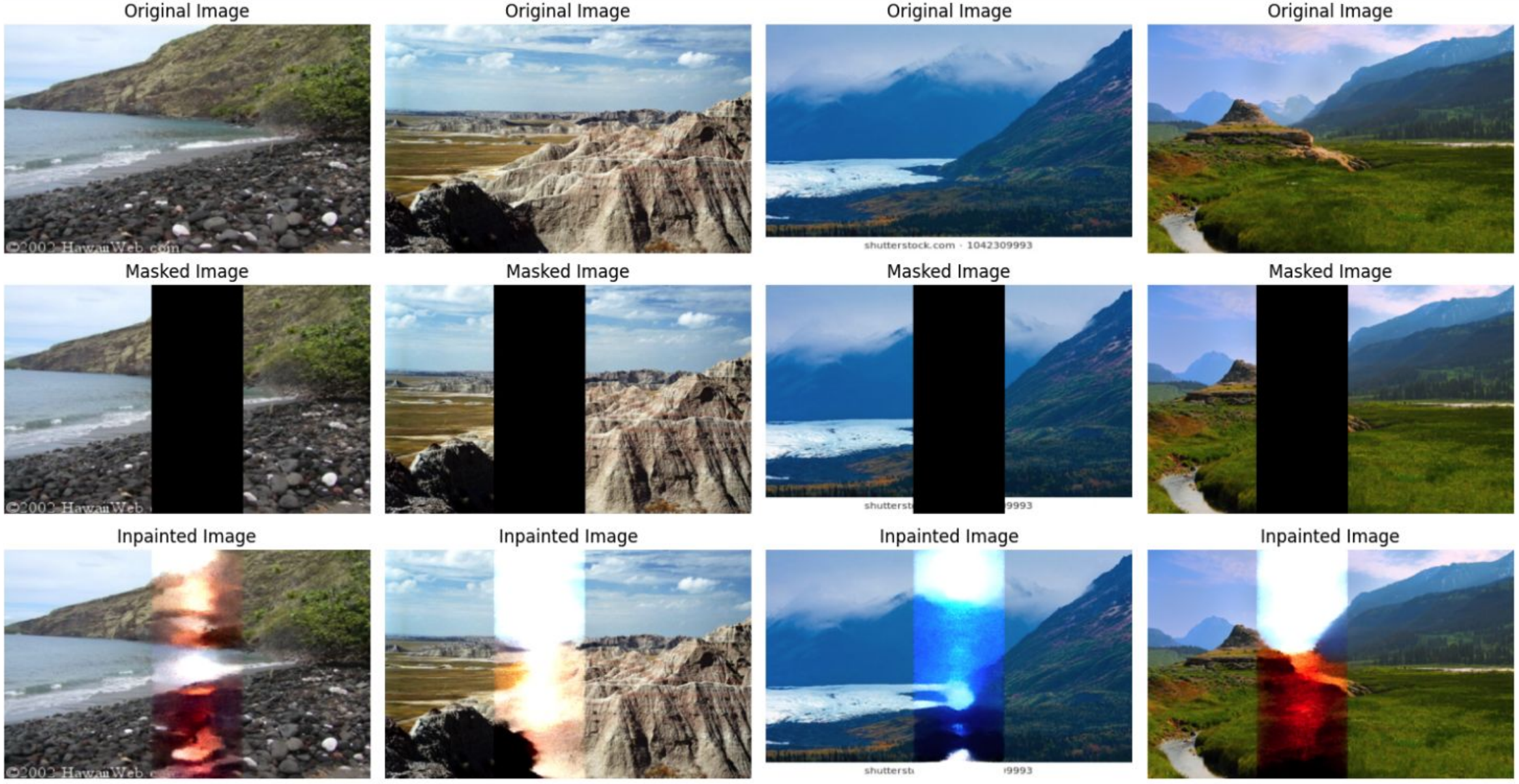
Very Long Natural Scenery Image Prediction by Outpainting, Yang et al (2019)

Diffusion Model Architecture

Conditioning a DDPM model on given images, yields decent results.



Diffusion Model Results



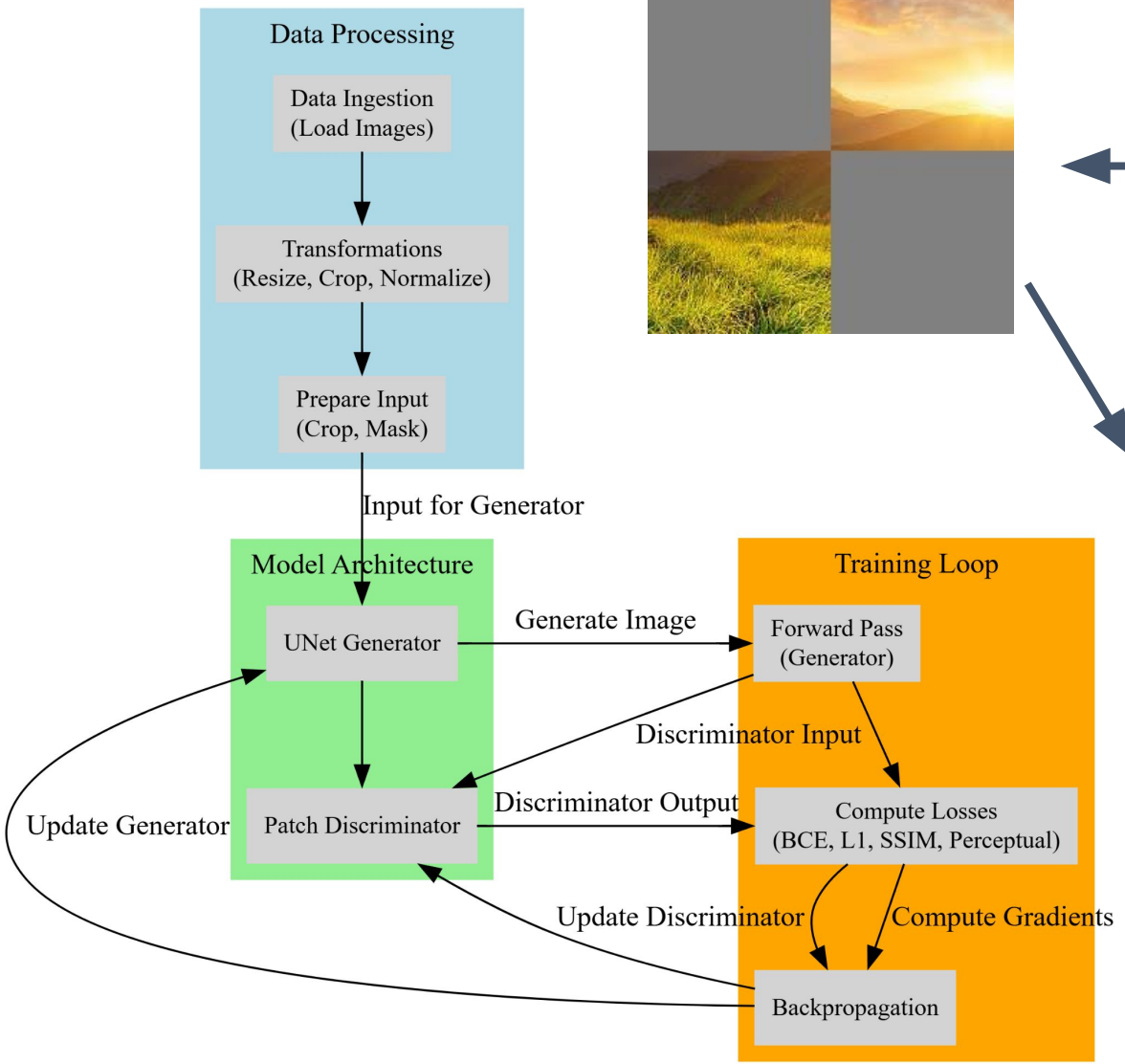
Diffusion Model Evaluation

Challenges & findings:

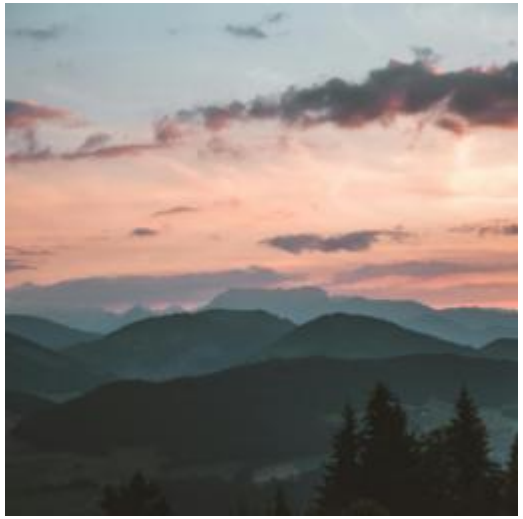
- Difficulty blurring the edges of given images and generated fill (attempted to fix with gaussian blur)
- Center of the fill image was usually quite blurry
- Initially only used one mask, which lead to poorer quality results. Specifically, desert and ocean scenes could take on a forest like texture
- Color tones in early epoch were good, but in later epoch could turn reddish

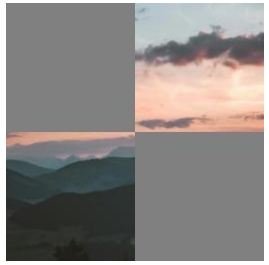
Metric	Value
Peak Signal-to-Noise Ratio (PSNR)	19.381
Structural Similarity Index Measure (SSIM)	0.762
Fréchet inception distance (FID)	505.608

GAN Model

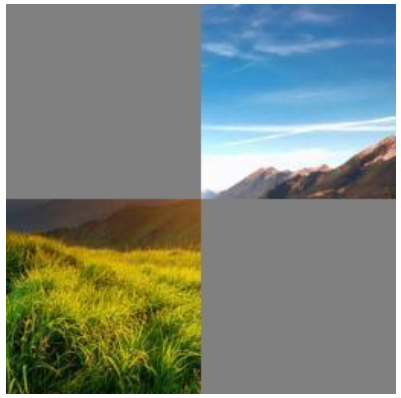

Data Processing
Data Ingestion
(Load Images)
Transformations
(Resize, Crop, Normalize)
Prepare Input
(Crop, Mask)
Input for Generator
Model Architecture
UNet Generator
Patch Discriminator
Update Generator
Generate Image
Discriminator Output
Training Loop
Forward Pass
(Generator)
Discriminator Input
Compute Losses
(BCE, L1, SSIM, Perceptual)
Update Discriminator
Compute Gradients
Backpropagation

GAN Results





GAN Results-Two images



GAN Results - Two images

GAN Model Metrics

Metric	Value
Peak Signal-to-Noise Ratio (PSNR)	18.8690
Structural Similarity Index Measure (SSIM)	0.7305
Fréchet inception distance (FID)	69.5485

GAN Results-Larger Mask

Our Model VS U-Transformer

1. Training Costs
   - U-Transformer
     i. 3 hours to train 1 epoch on RTX 2070.
     ii. 500 epochs needed/
   - Our model
     i. 1 hour to train 10 epochs on RTX 2070
     ii. 20 epochs needed for 0.5 mask
     iii. 50 epochs for 0.6 mask.

2. Image Size
   - U-Transformer
     i. 128 * 128 -> 192 * 192
     ii. Only for center-expansion, hard to change for other mask
   - Our model
     i. 256 * 256 base image
     ii. The mask can be set anywhere with configuration

Future Potential

Bigger image size

Reduce the blur

CORNELL TECH

JACOBS INSTITUTE

Thank You!

Appendix: GAN Model Next Steps

- Update Model Structure to optimize loss.
- Bigger image size
- Generate only the missing part, instead of using mask (hard based on GAN)
- Random crop region (hard based on GAN)

Appendix: Diffusion Model Next Steps

- Try switching to a DDIM architecture
- Use Repaint's resampling strategy that jumps back and forth between timesteps
- Try using pre-trained embeddings
- Add more crops to training
- Revisit gaussian blur implementation

Appendix: GAN Model Architecture