SUPER RESOLUTION

CV Spring 24 Course Project

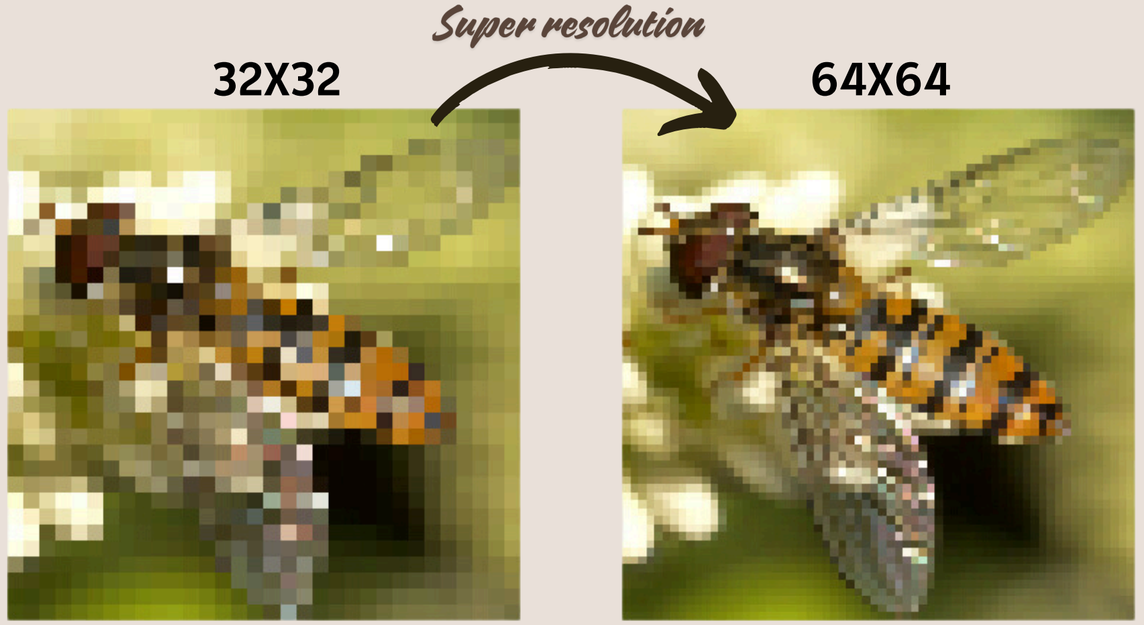
4th May 2024

Roja Sahoo (2021111014)
Sannidhya Gupta (2021112012)

CONTENTS

What is Super Resolution (SR)?

Image super-resolution (SR) is a process of enhancing the resolution and quality of an image, typically from a low-resolution (LR) version to a higher-resolution (HR) version. The goal of super-resolution is to generate an HR image that contains more details and clearer features than the original LR image.



FOCUSED SR TECHNIQUES

Transformer based

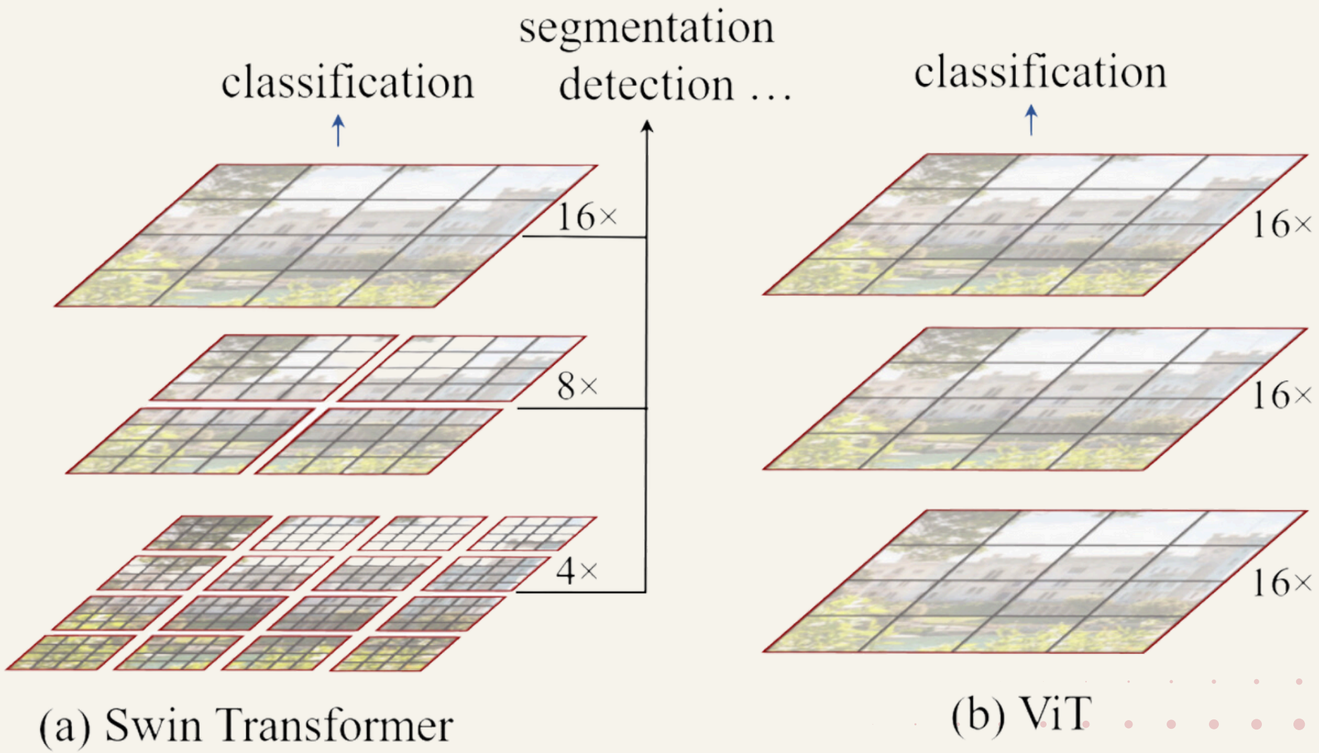
SWIN-IR:
Shifted Windows transformer

CNN based

RCAN:
Residual Channel Attention Network

SWIN IR – Transformer based

Shifted Windows (SWIN) Transformer is a hierarchical transformer architecture that processes images in a multi-scale fashion using a shifted window mechanism, enabling efficient modeling of long-range dependencies in large images while maintaining computational scalability and performance. It achieves state-of-the-art results on various vision tasks including image classification and object detection.



*https://arxiv.org/abs/2103.14030v2

SWIN IR - 3 Main Blocks

Shallow Feature Extraction

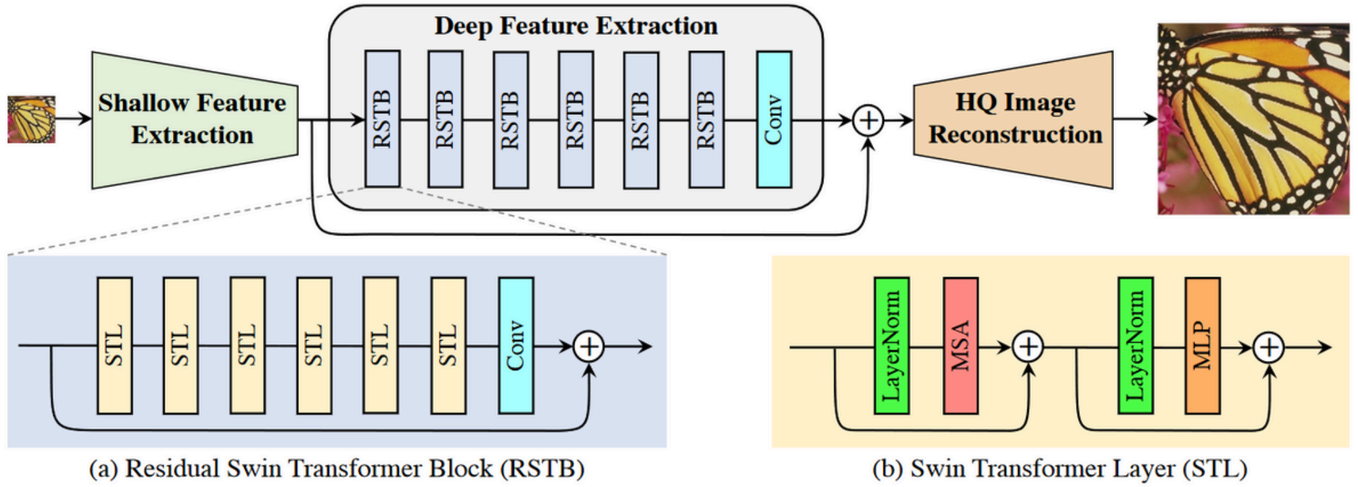
Uses a convolution layer to extract shallow features which is directly transmitted to the reconstruction module so as to preserve low-frequency information

Deep Feature Extraction

Consists several Residual Swin Transformer blocks (RSTB). Each RSTB has several Swin Transformer layers together with a residual connection. Uses local attention and cross-window interaction

High-quality image reconstruction

Both shallow and deep features are fused in the reconstruction module for high-quality image reconstruction.



(a) Residual Swin Transformer Block (RSTB)

(b) Swin Transformer Layer (STL)

*https://arxiv.org/abs/2108.10257

RCAN - CNN based

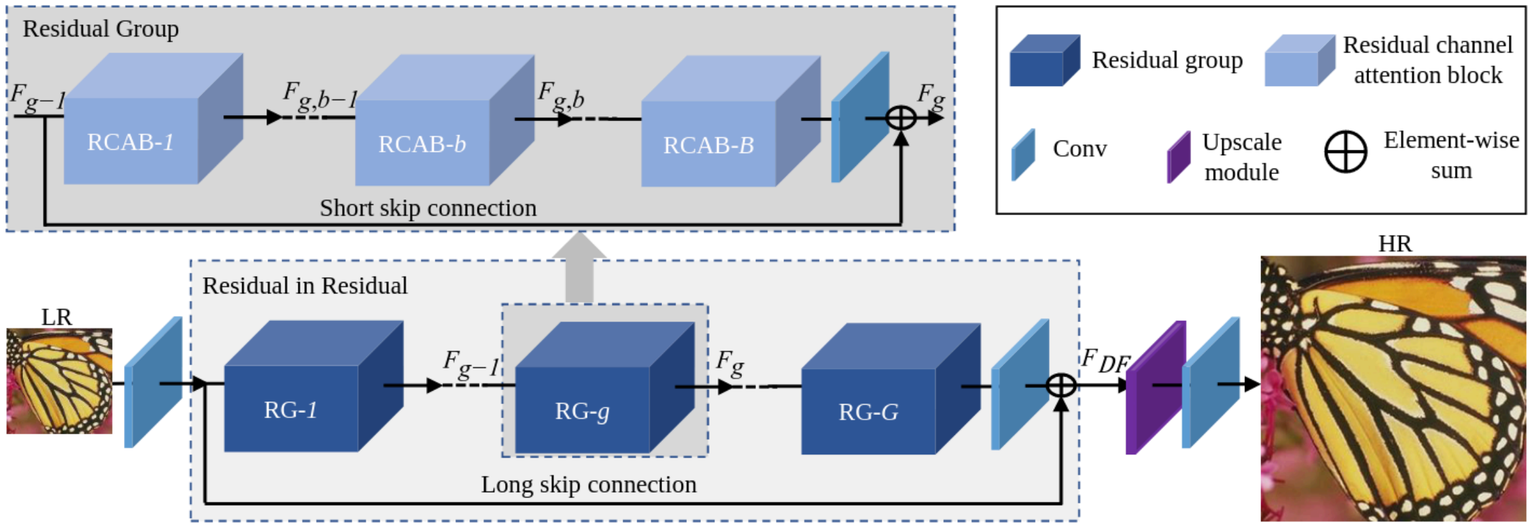
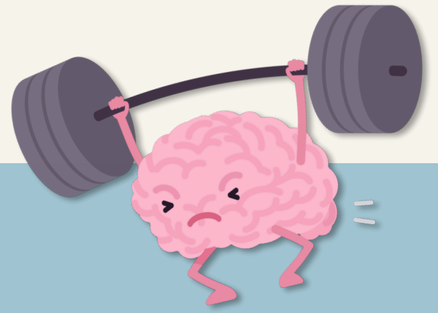


Fig. 2. Network architecture of our residual channel attention network (RCAN)

RCAN (Residual Channel Attention Networks) is a type of deep learning architecture specifically designed for image super-resolution tasks. The RCAN architecture builds upon the residual learning framework and incorporates channel attention mechanisms to effectively exploit the information in feature maps.

*https://arxiv.org/abs/1807.02758

Training



- **Dataset used**

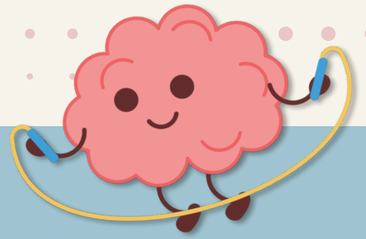
  We used a mixture of Urban100, General100, Set14 datasets with a train-val-test split with total dataset size of 160 images.

- **Input and output image size**

  This could be extended to higher resolutions, but we performed x2 upscaling of images from 32x32 to 64x64 and 64x64 to 128x128, with the computation resouces we had.

- **Parameters**

  We tested on a variety of parameters, and played around with different number of epochs and learning rates to get better results



Training Parameters

- **L1 v/s MSE**

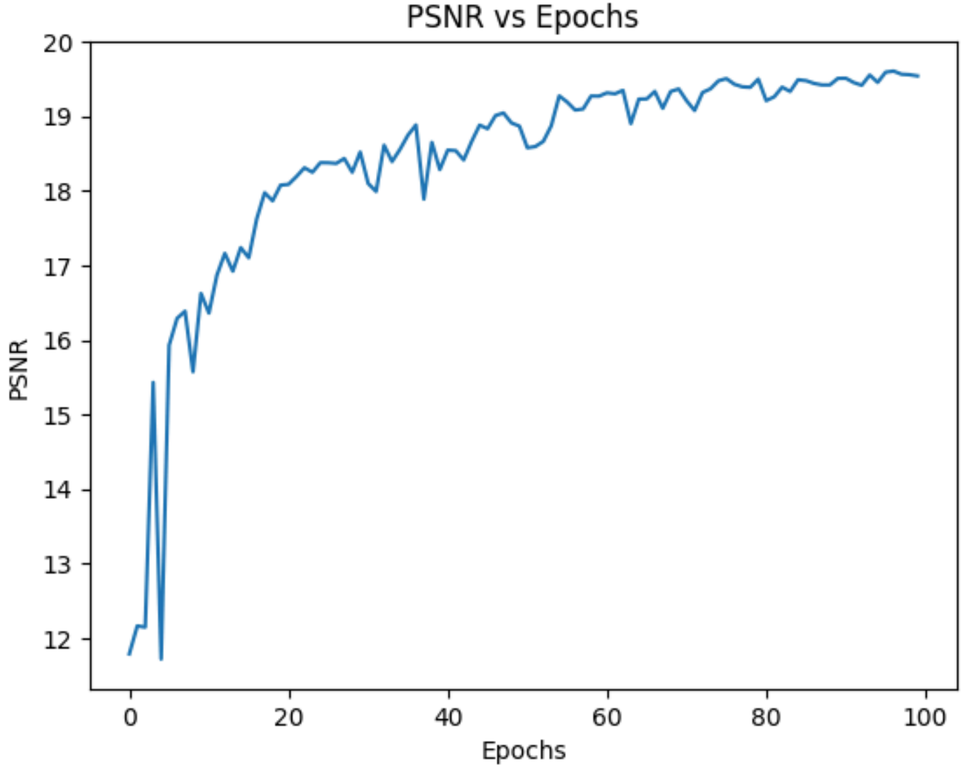
  MSE is able to converge quickly & reliably on train dataset, but performs much worse than L1 on test set. MSE tends to overfit train data and is bad at generalising for this task.

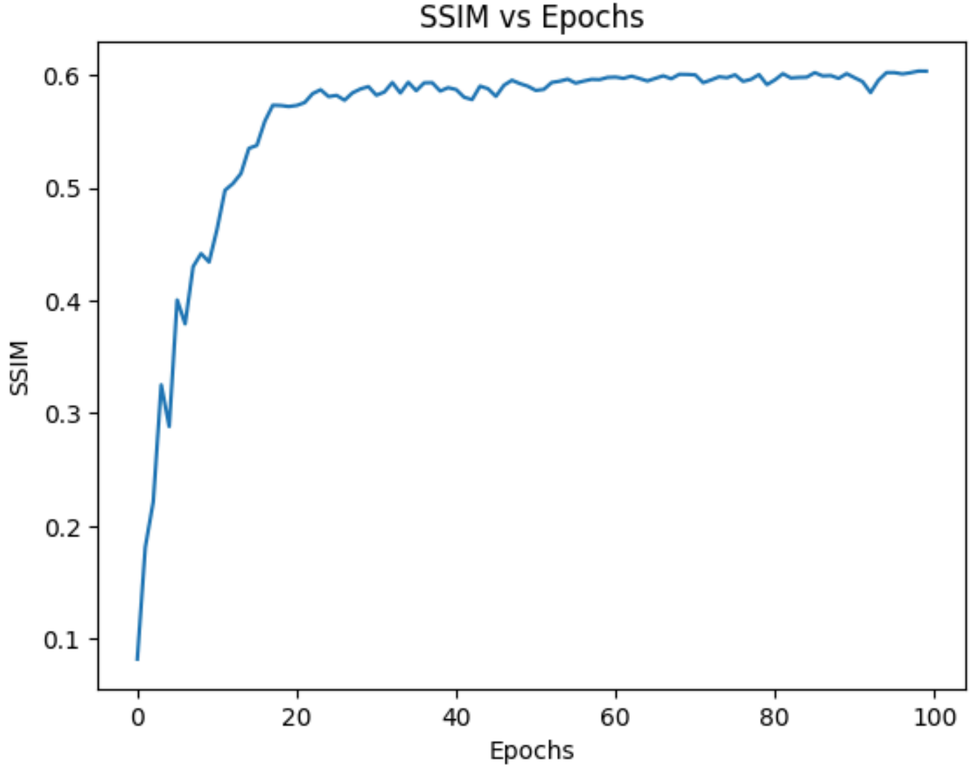
- **Tweaking LR: reducing on plateau v/s multi step**

  ReduceLROnPlateau produced better results for our training than MultiStepLR as it gives a condition(on plateau) on when to change / update the LR whereas MultiStepLR scheduling just updates in regardless.

• Training loop

SWIN (64 -> 128)



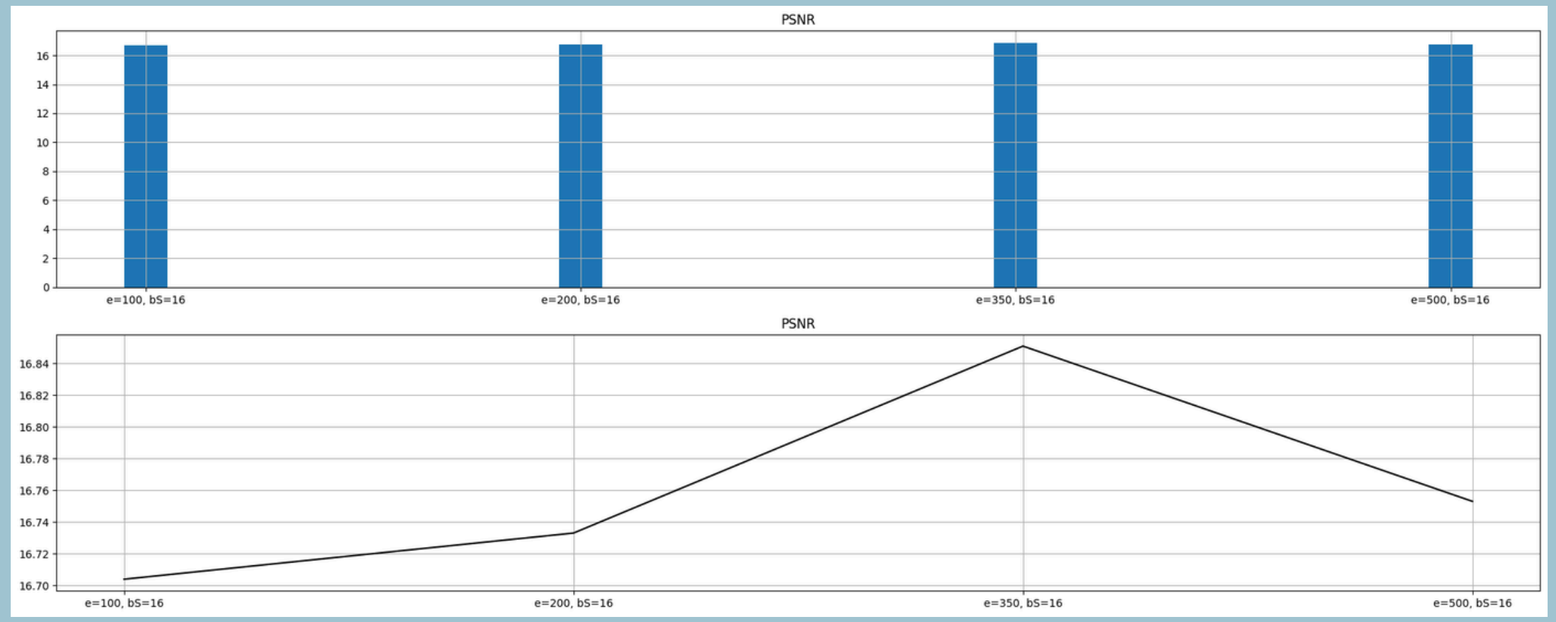


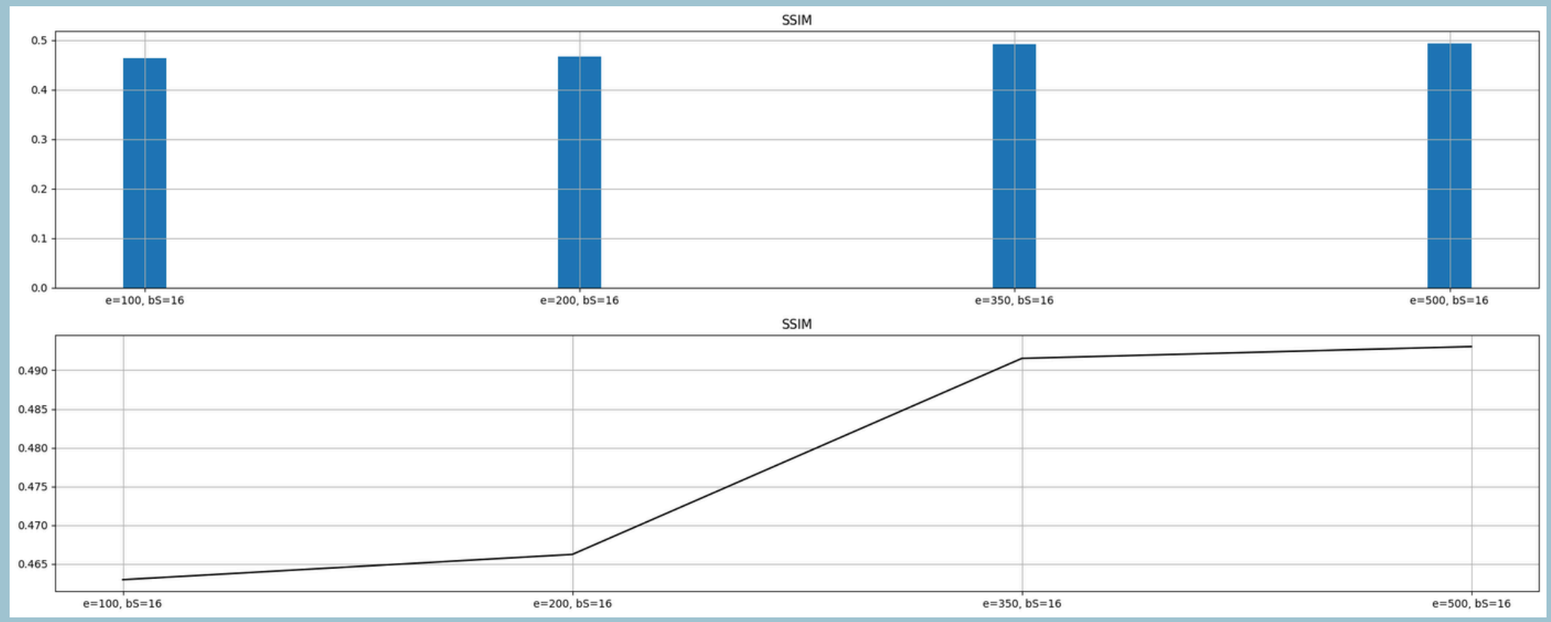
RCAN

Training - 32x32

Used metrics:

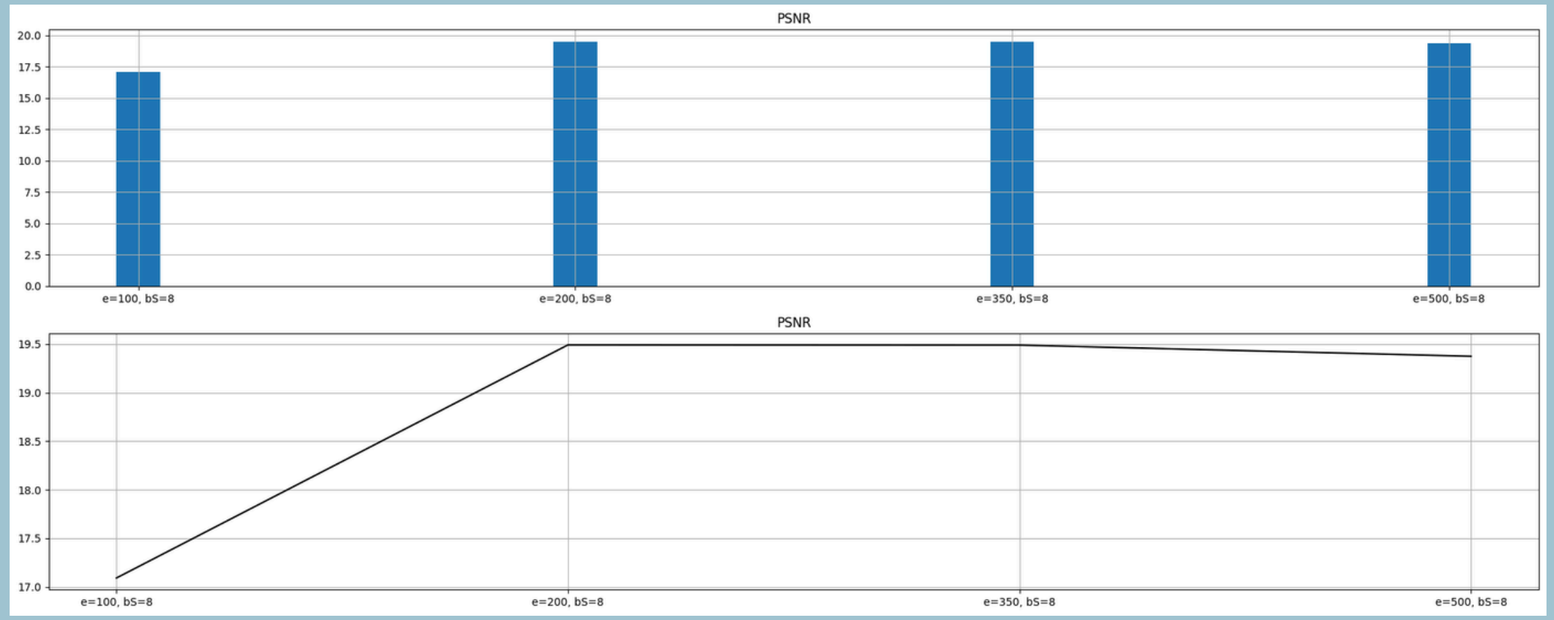
- **PSNR - peak signal to noise ratio**
- **SSIM - structural similarity index measure.**

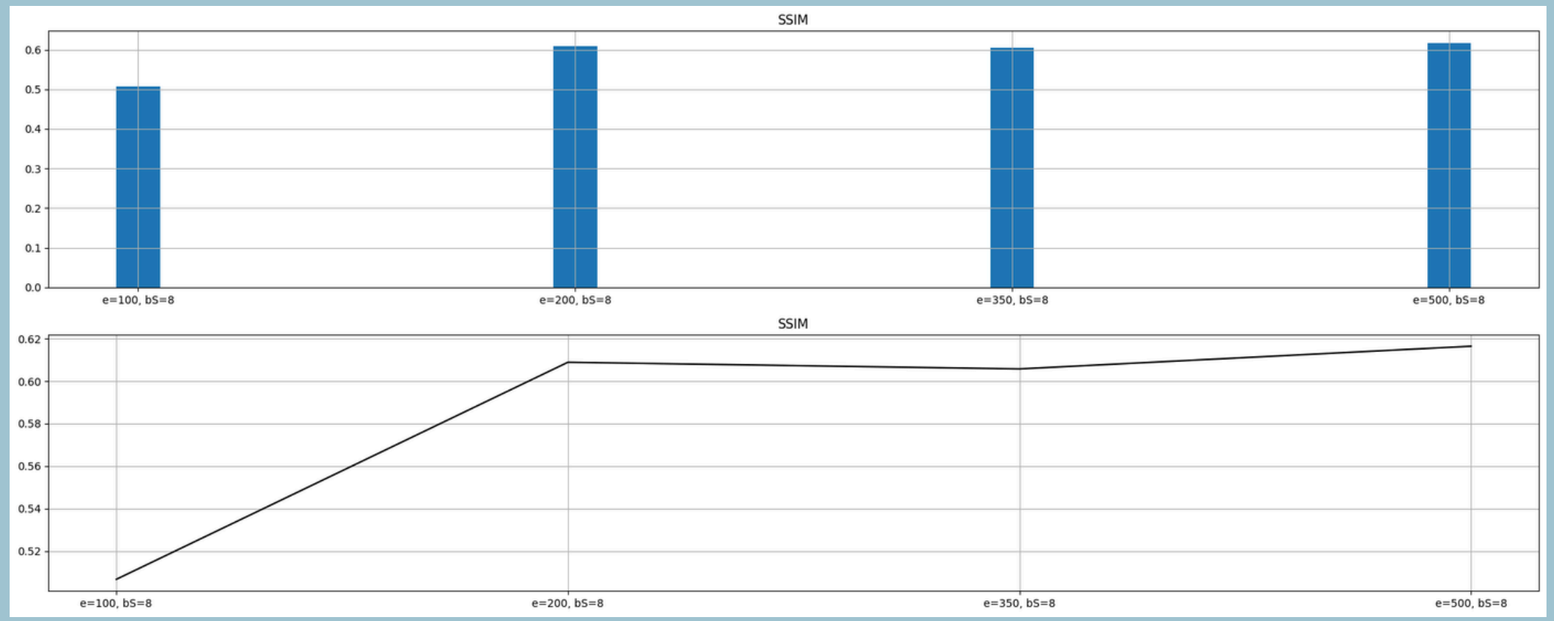




RCAN

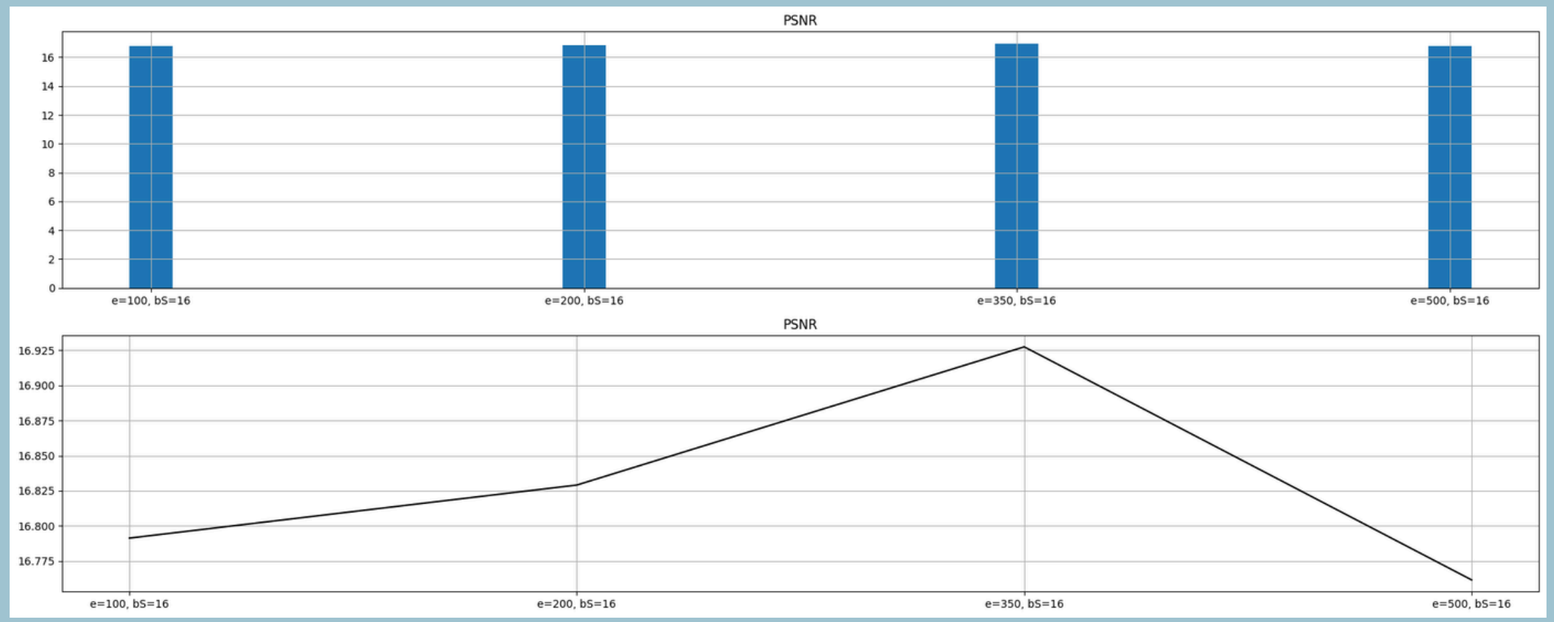
Training - 64x64

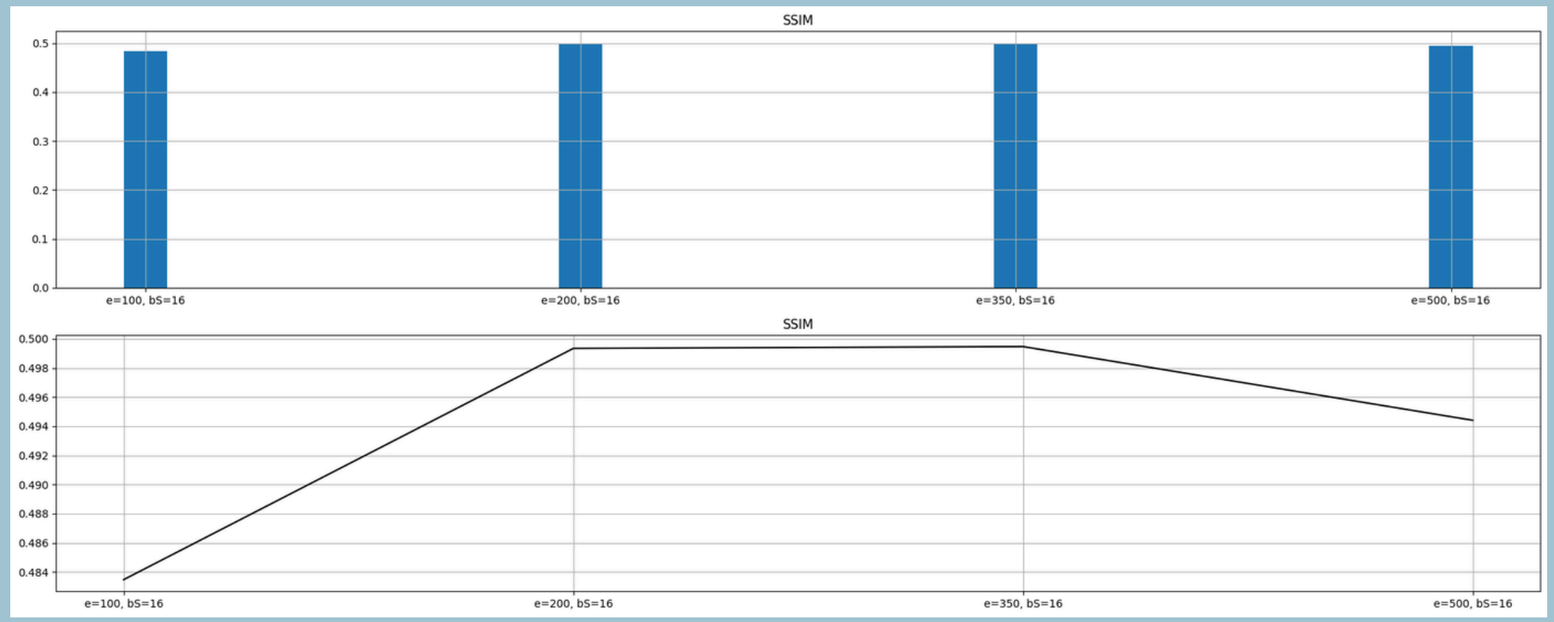




SWINIR

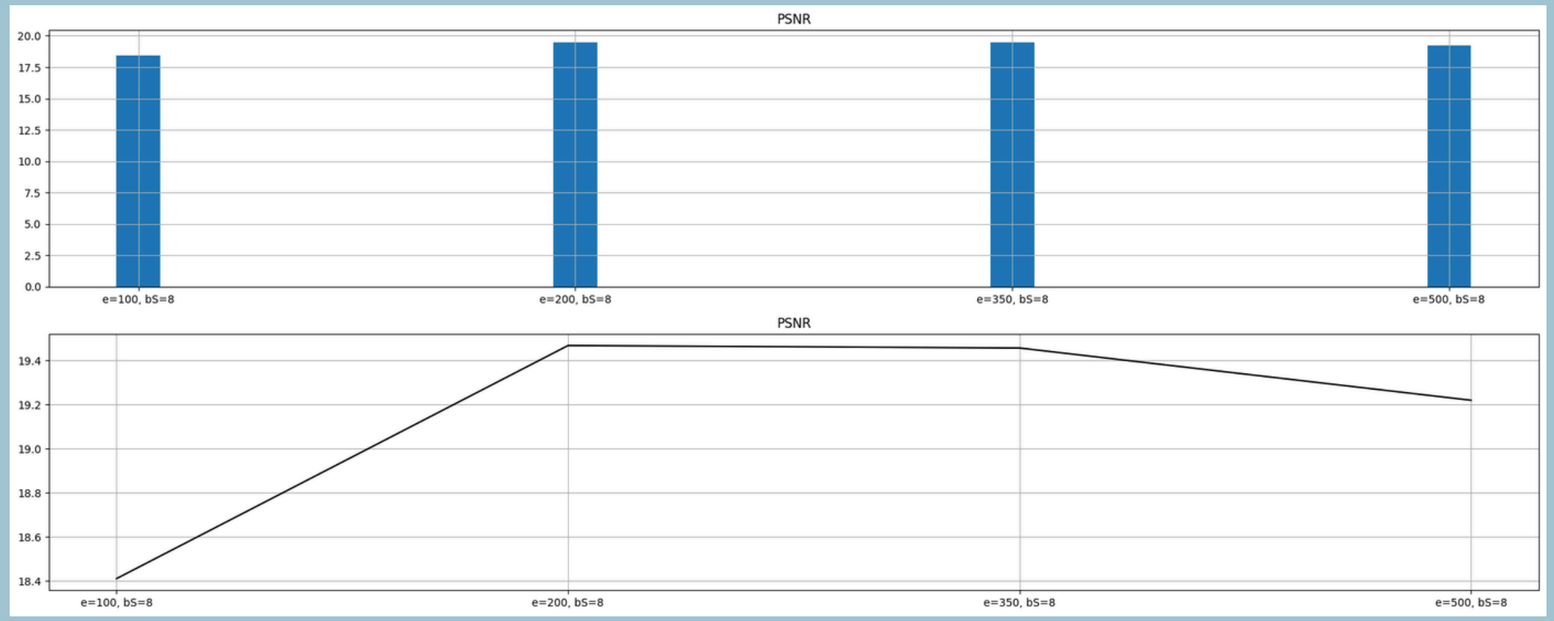
Training - 32x32

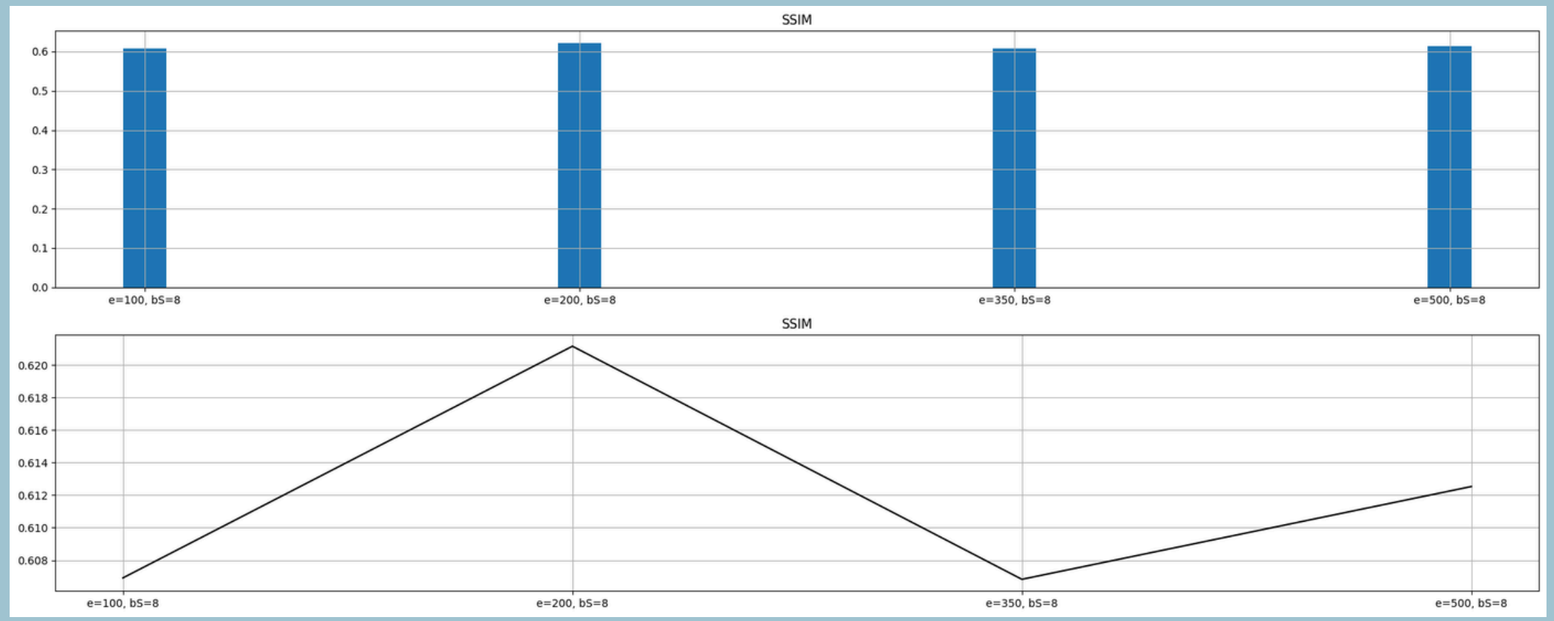




SWINIR

Training - 64x64



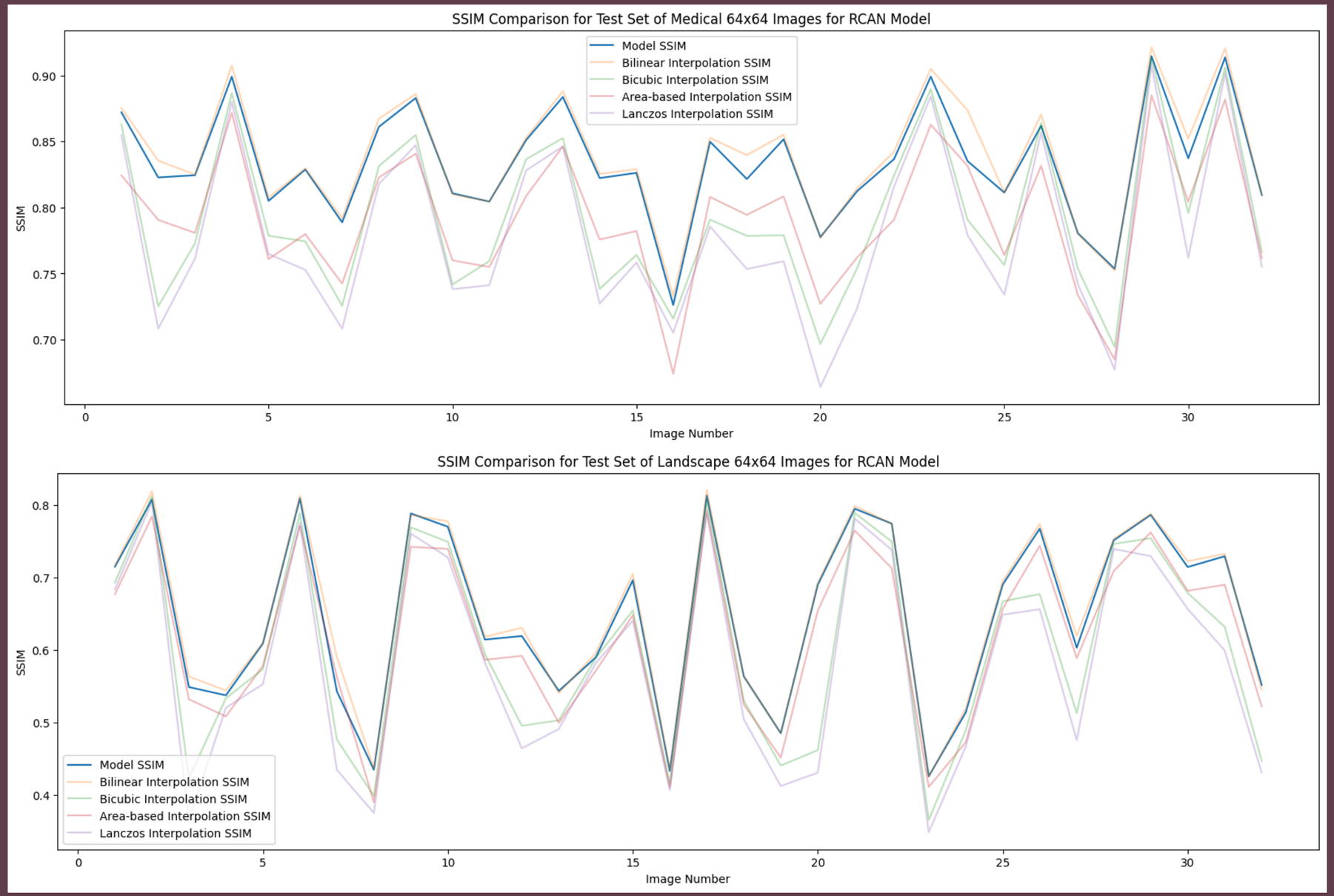


Testing & Results

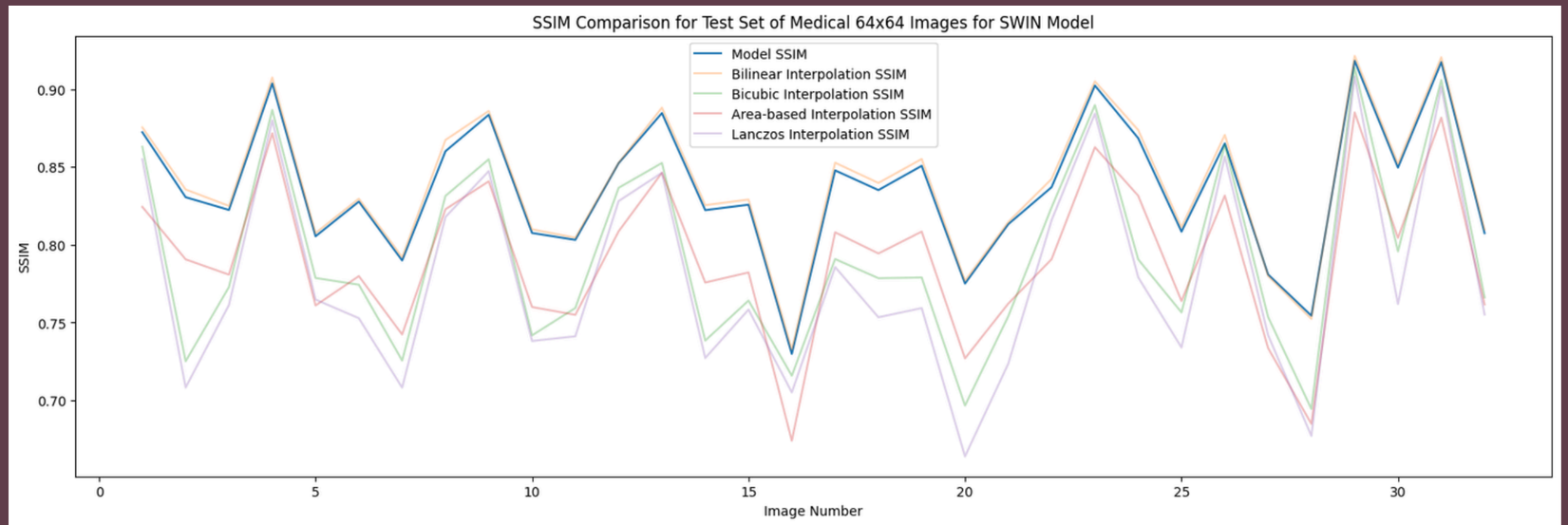
- **Testing on vastly different types of image data**
- **SR models v/s interpolation**

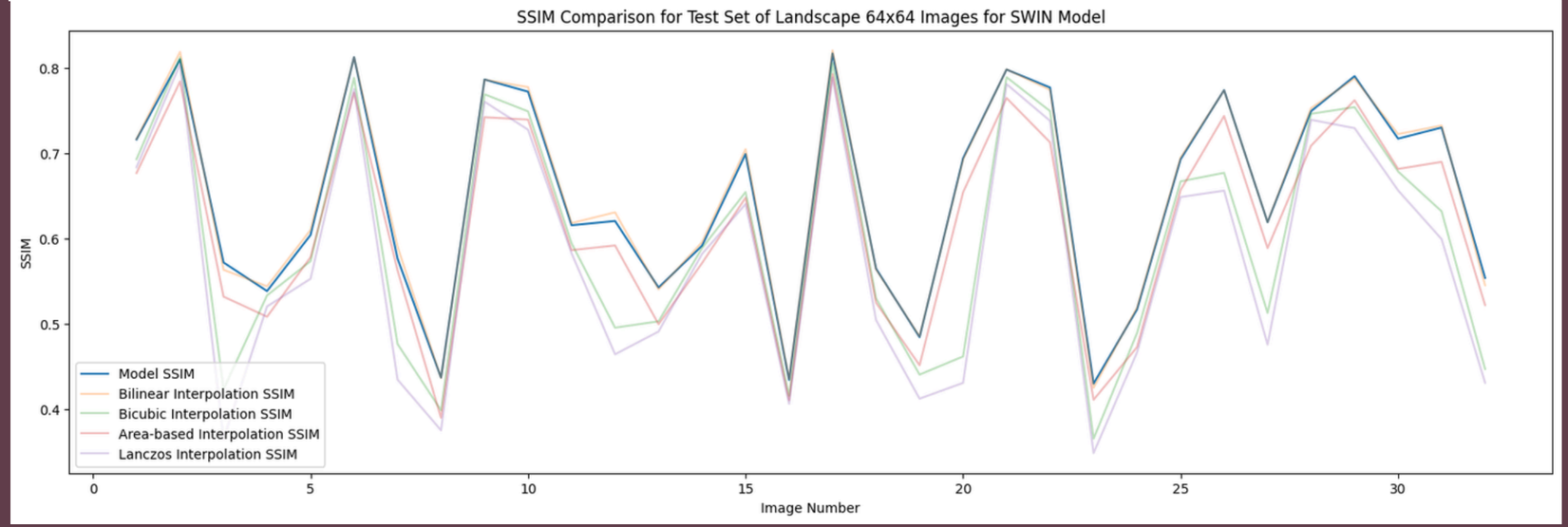
  We compared the models analyzed against the common interpolation techniques to compare the performance difference
- **Plotting Occlusion Ascription Maps (Proposed approach to visualise SR)**

RCAN (64->128) v/s Interpolation



SWINIR 64->128 vs Interpolation





Medical data:

Grayscale images, so easier to reconstruct.

Landscape data:

Model also performs well on different type of unseen data.

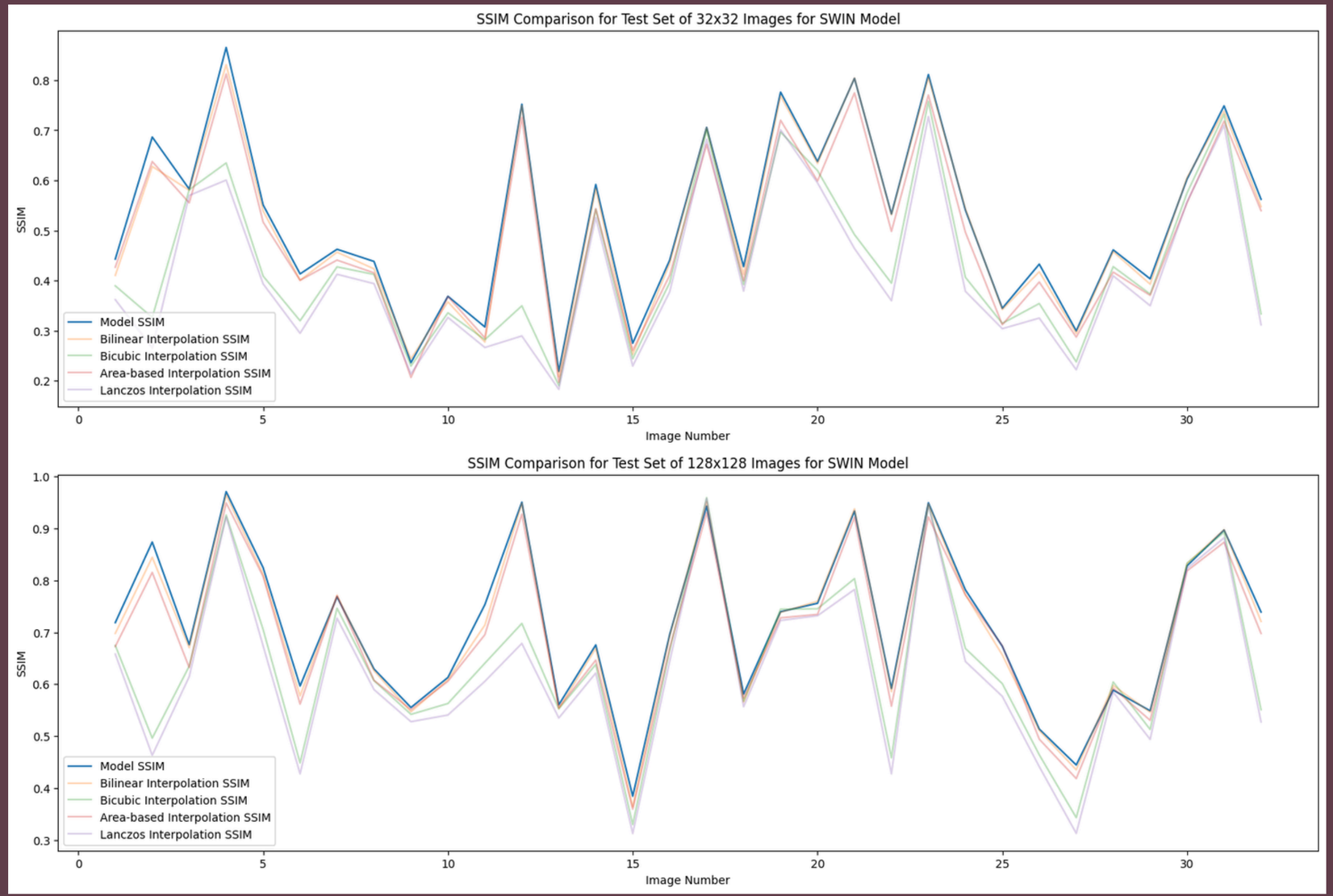
SWINIR 64->128 vs Interpolation

Doing 32->64:

Images have way too little information

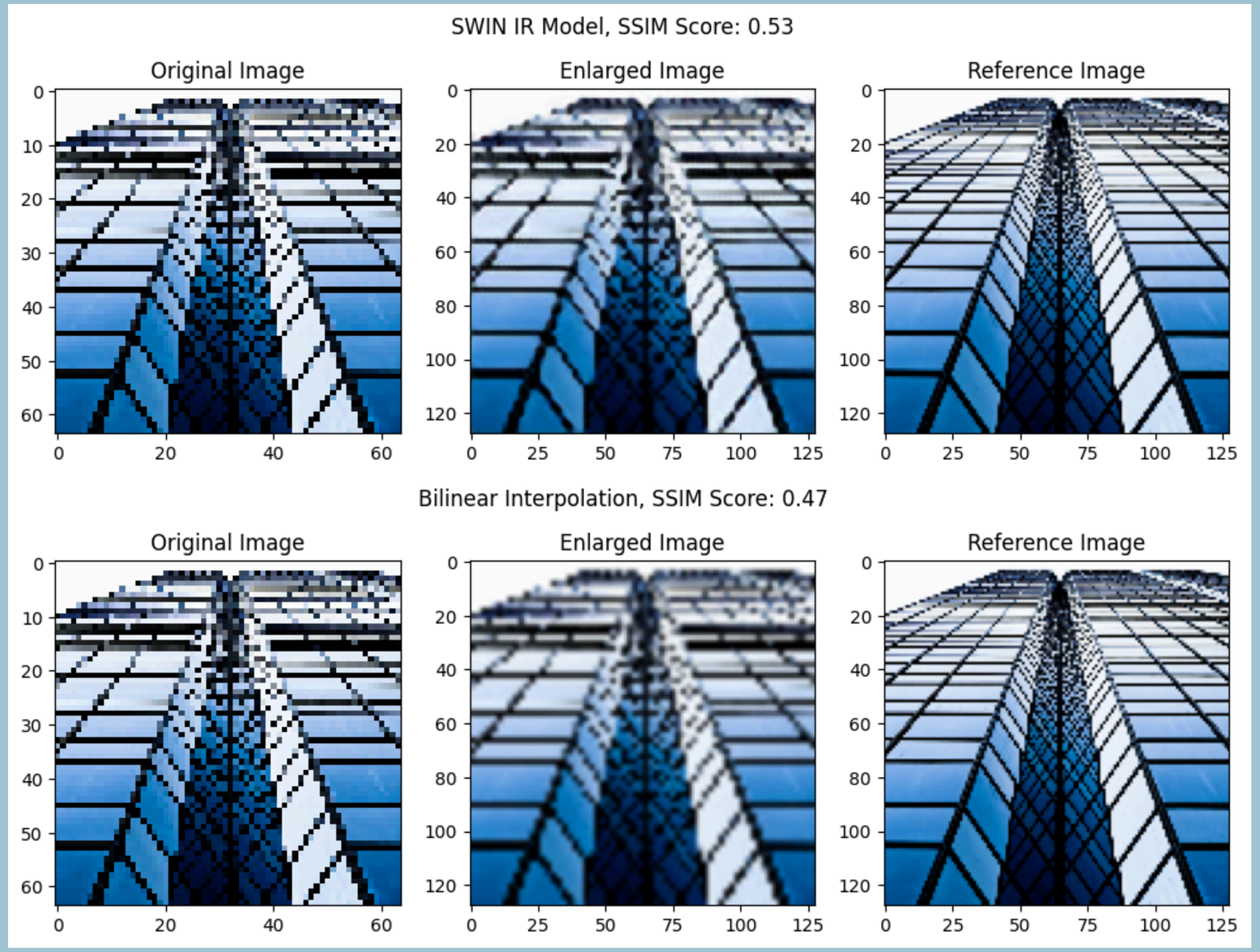
Doing 128->256:

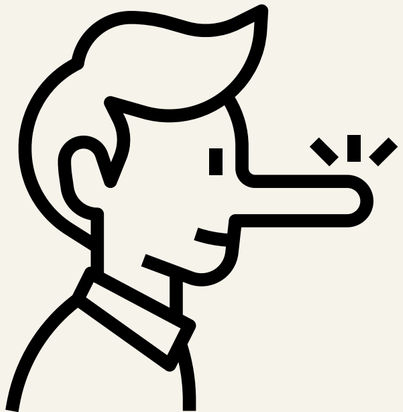
Not trained enough to extract enough info for 128x128 images.



SR (SWINIR) vs Bilinear Interpolation: Visual Results

SWINIR transformer



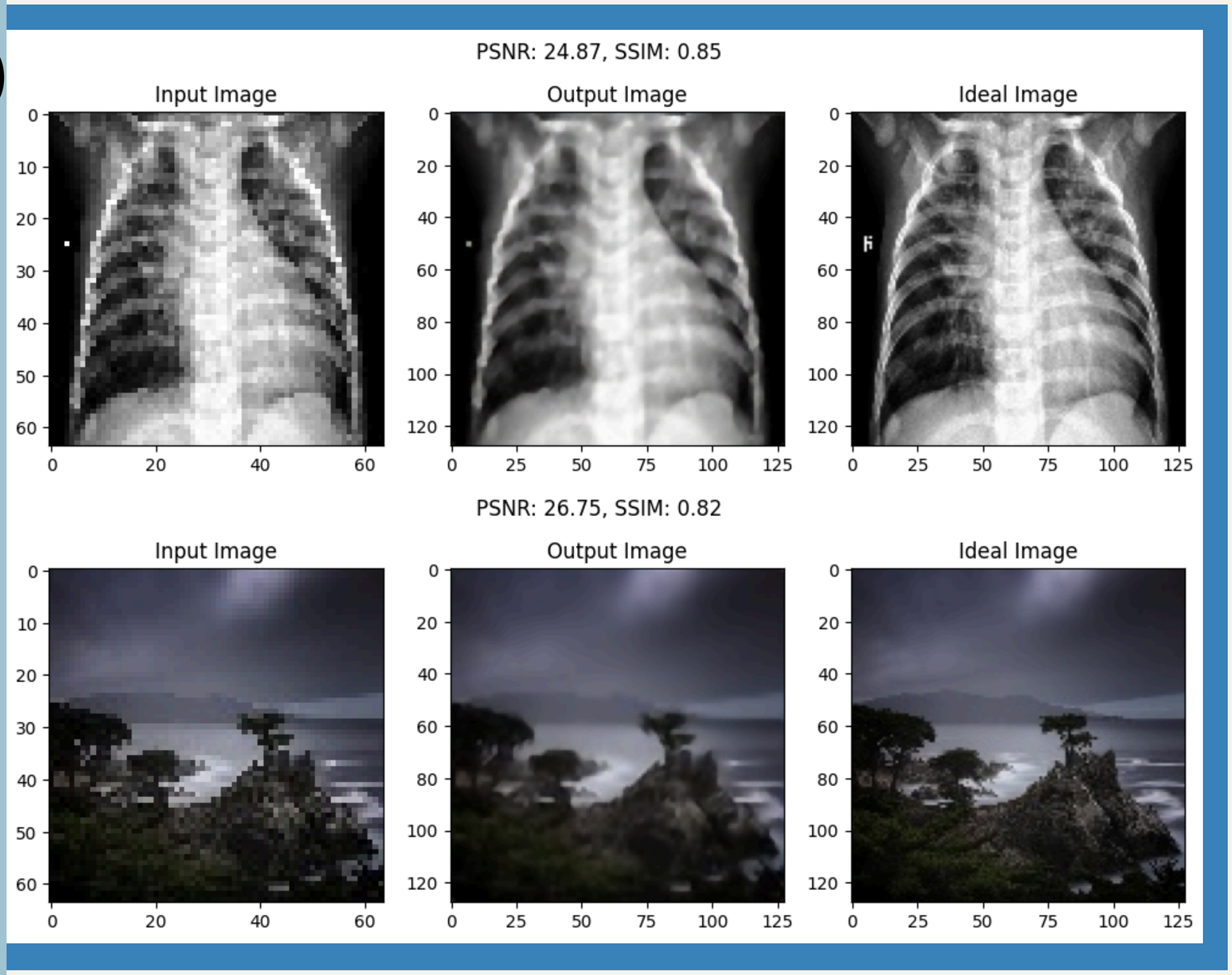


Interpolation

SWINIR (64->128)

Performance on vastly different Test Data: visual representation

Grayscale & “easy” images. Hence the model performs decently well



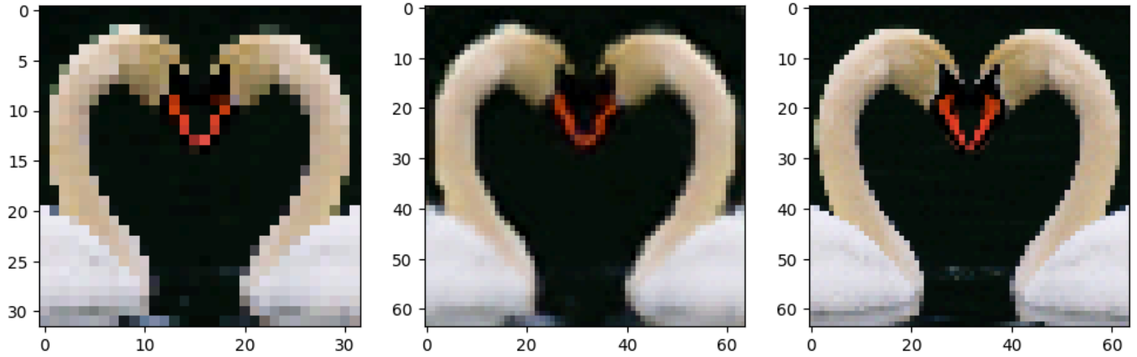
TRAINED TO EXTRACT ENOUGH INFORMATION HENCE PERFORMS BETTER!

SWINIR (64x64) to (128x128)

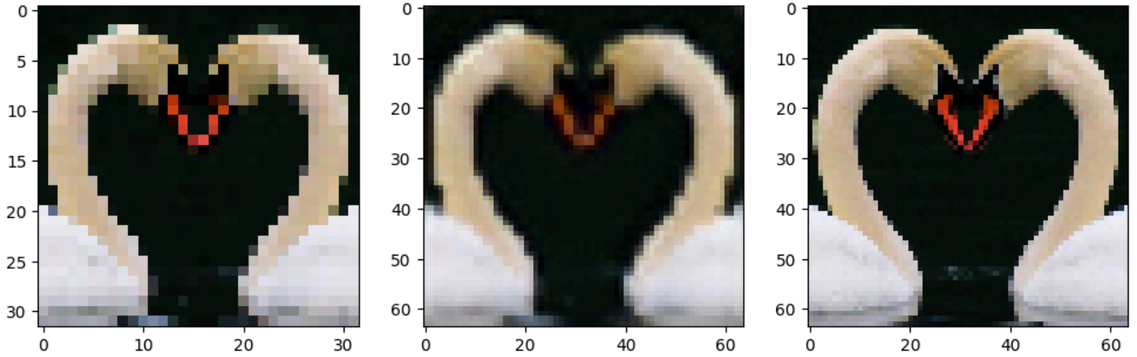
SWINIR (32x32) to (64x64)

32 -> 64

PSNR: 21.59, SSIM: 0.87

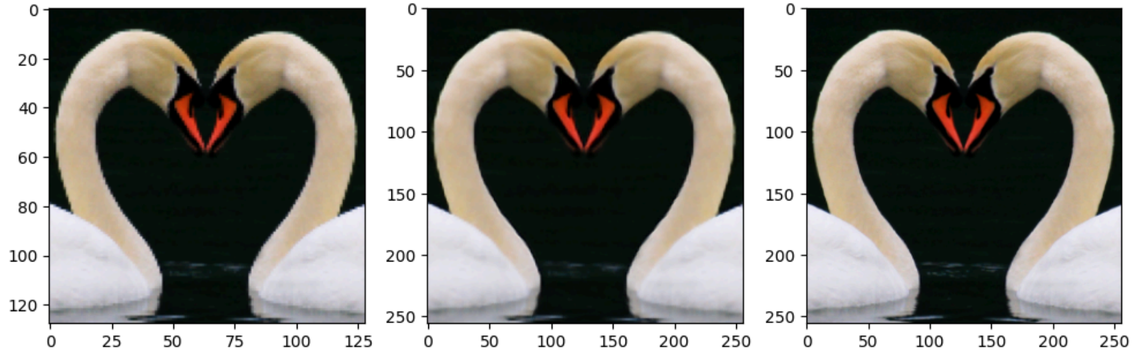


PSNR: 20.77, SSIM: 0.84

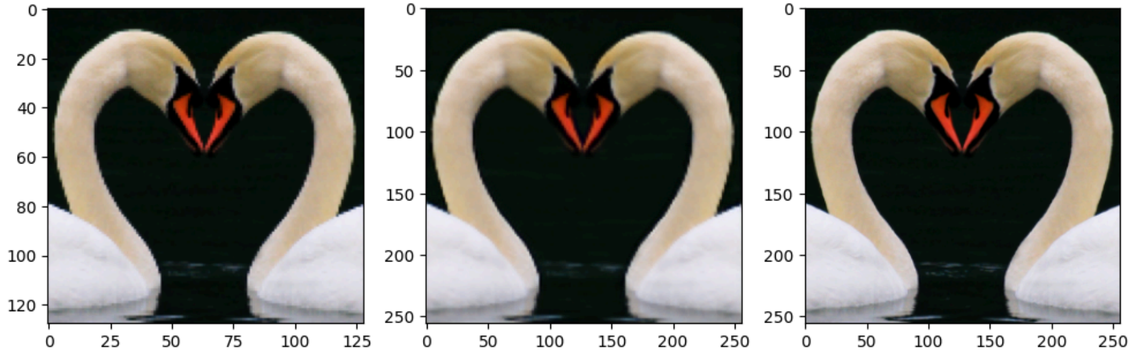


64->128

PSNR: 33.65, SSIM: 0.97



PSNR: 31.20, SSIM: 0.96



Occlusion based ascription map

Proposed algorithm

IMAGE

Select a patch size and get all the possible patch locations in the image

Make copies of the image, and selectively occlude one patch each at a different location in the image

Pass all the patch-occluded images through the model

Select a window of which ascription map is required

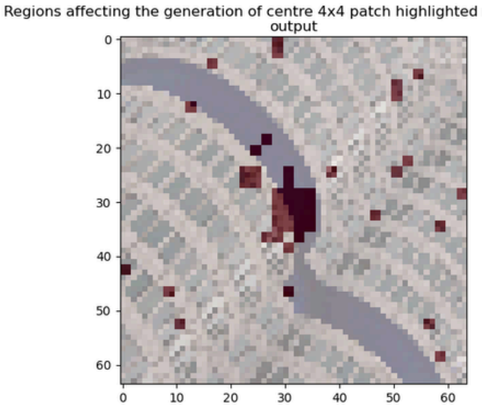
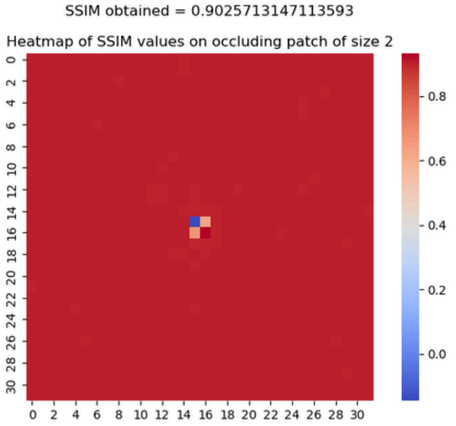
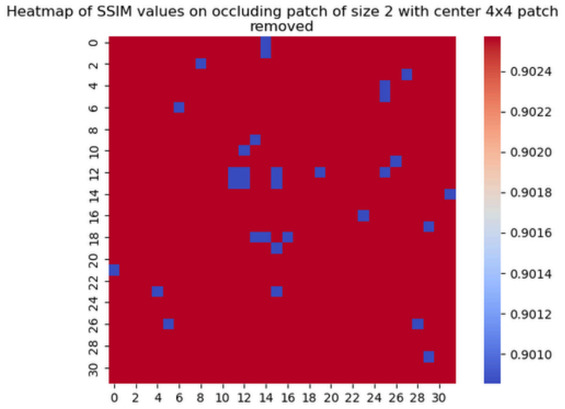
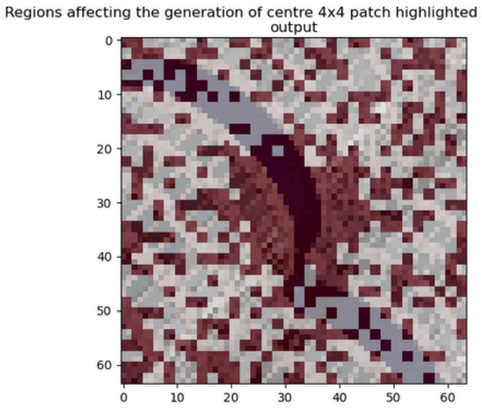
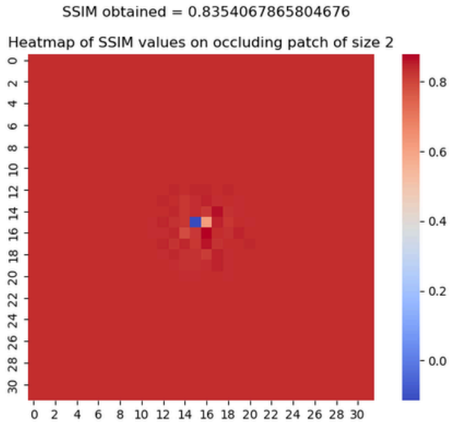
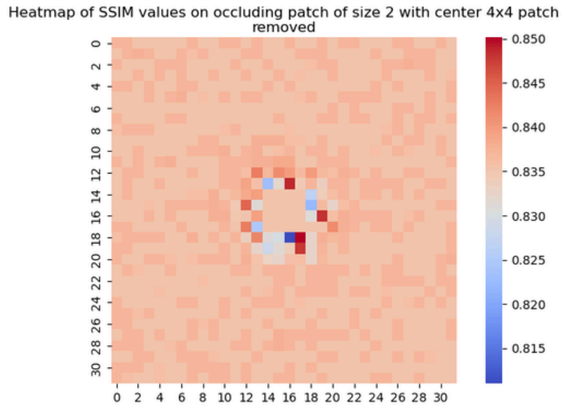
Crop all the output images and the label to the required window, and calculate ssim scores for each

These SSIM scores are mapped to occlusion patch locations and a heatmap is generated

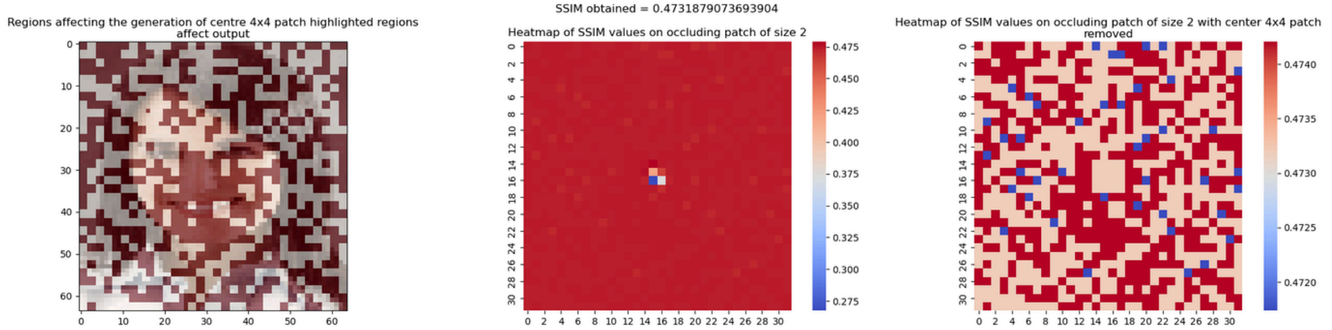
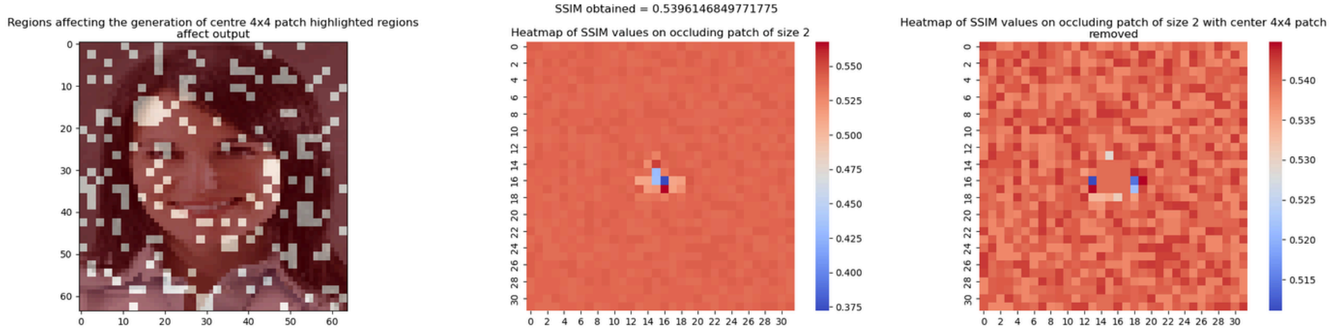
ASCRIPTION MAP

Occlusion based ascription map

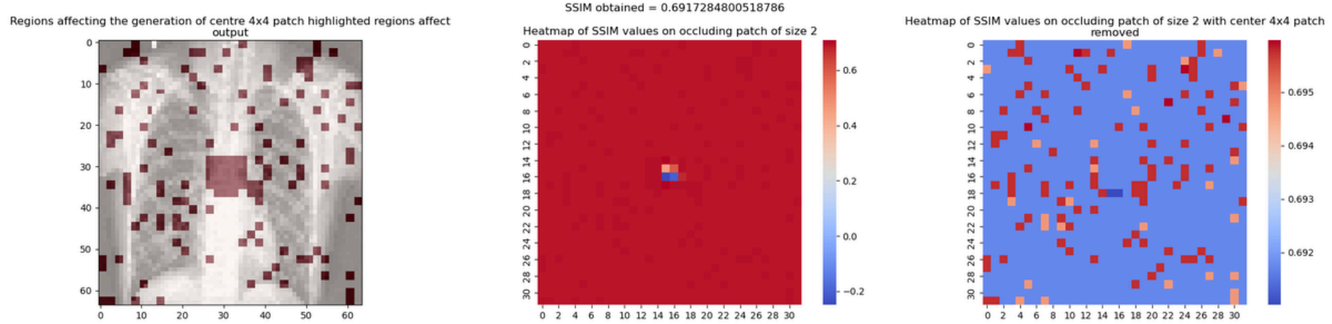
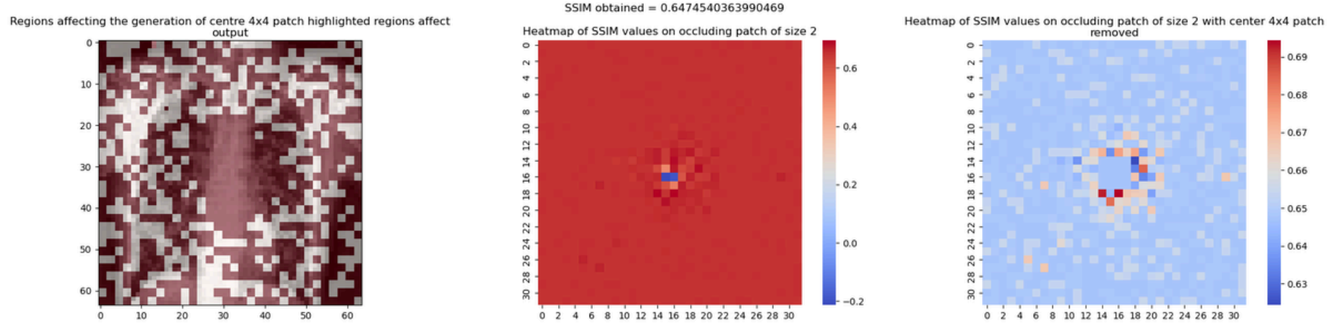
Obtained results

MODEL				SSIM
SWIN				0.90
RCAN				0.83

Facial SR

MODEL		SSIM
SWIN		0.47
RCAN		0.53

Medical SR

MODEL		SSIM
SWIN		0.69
RCAN		0.64

Occlusion based ascription map

Inference

SWINIR

Performs better (SSIM) but uses less global information to reconstruct the given patch

RCAN

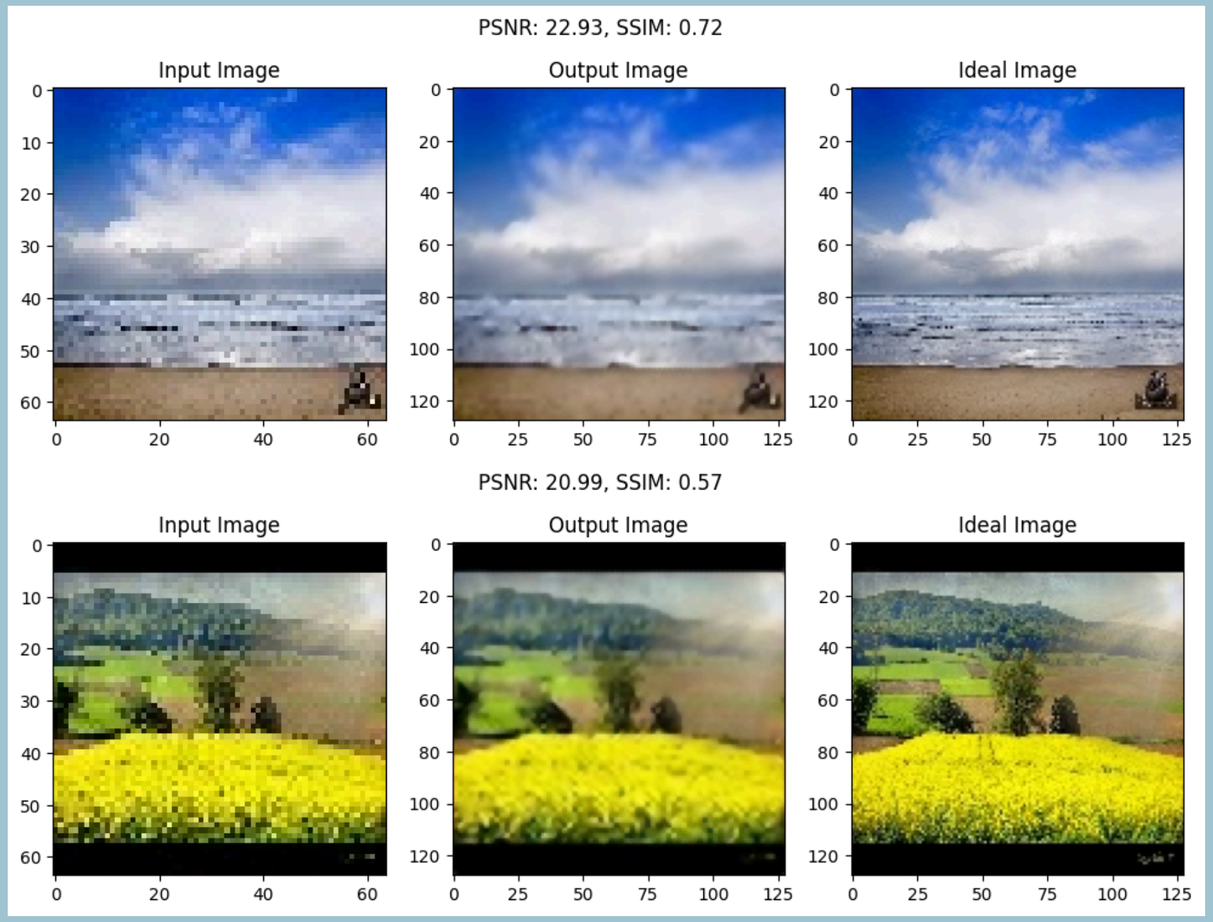
Performs worse (SSIM) but uses more global information to reconstruct the given patch

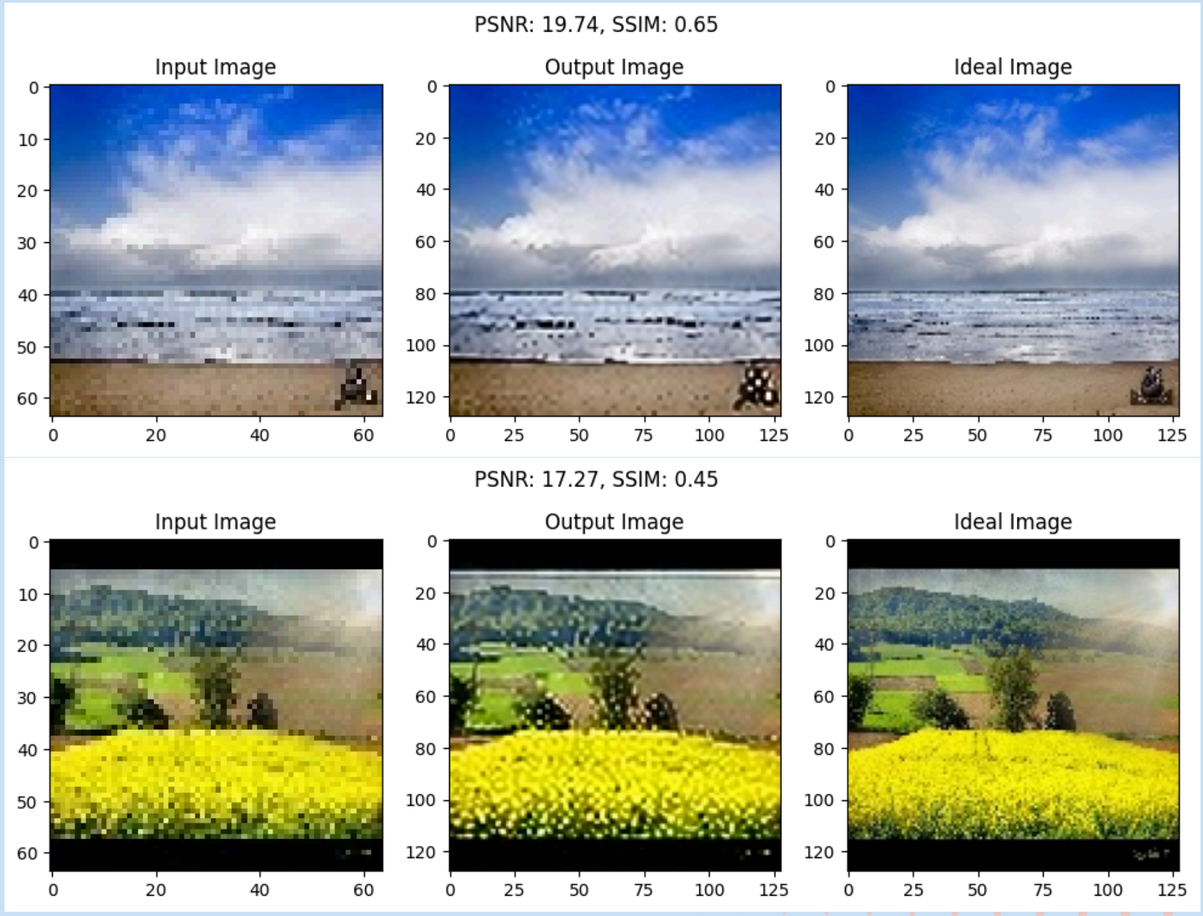
SWINIR vs SWINIR

trained from scratch

pre-trained

SURPRISING!





Avg SSIM: 0.65

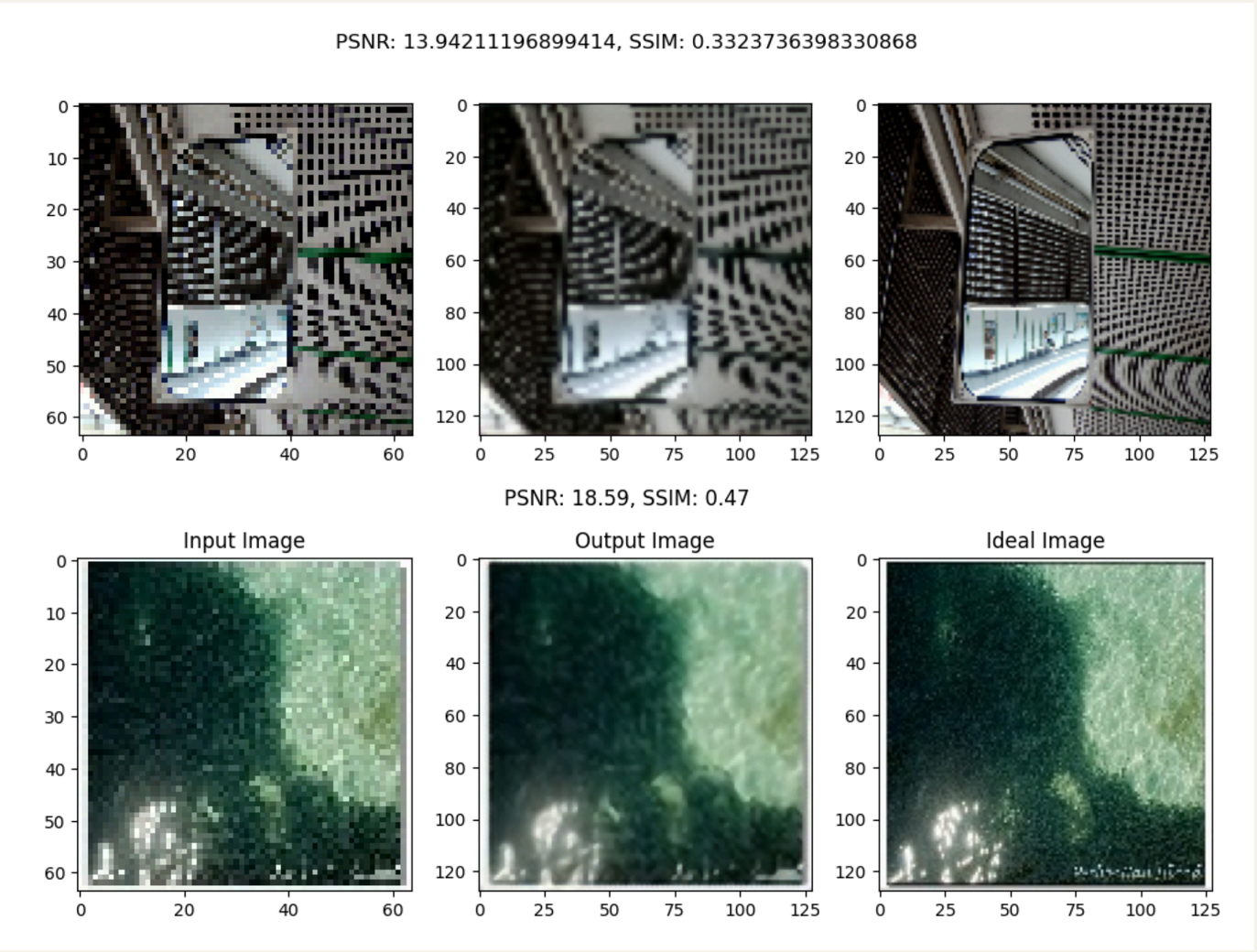
Avg SSIM: 0.56

Where SWINIR 64x64 fails

Highly complex

Cluttered

Where SWINIR 64x64 fails



Degradation of input

Limited context

SWINIR or RCAN

which is better?

SWINIR

Best Model:
Avg SSIM: 0.6211
Avg PSNR: 19.47

No. of parameters: ~11M

Inference from Ascription Map: Works with less global information

RCAN

Best Model:
Avg SSIM: 0.6164
Avg PSNR: 19.37

No. of parameters: ~15M

Inference from Ascription Map: Needs more global info to produce results

That's why vision transformers are the future of Computer Vision!

Future extension

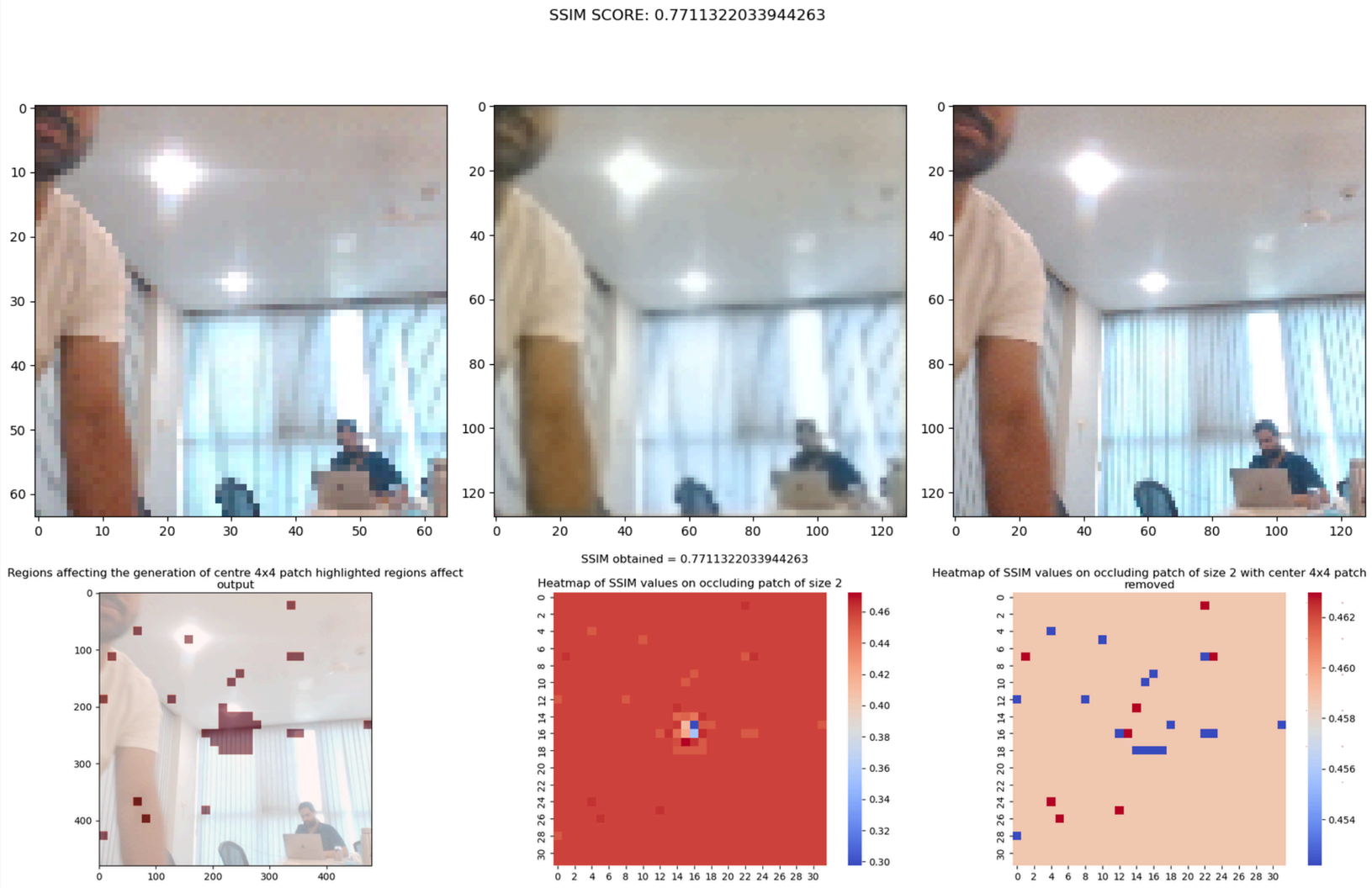
- Can apply pre-processing (Unsharp Masking, Edge Enhancement, histogram equalization) to improve quality on input.
  - We weren't able to achieve better results within our testing of pre-processing techniques.

- Image De-blurring

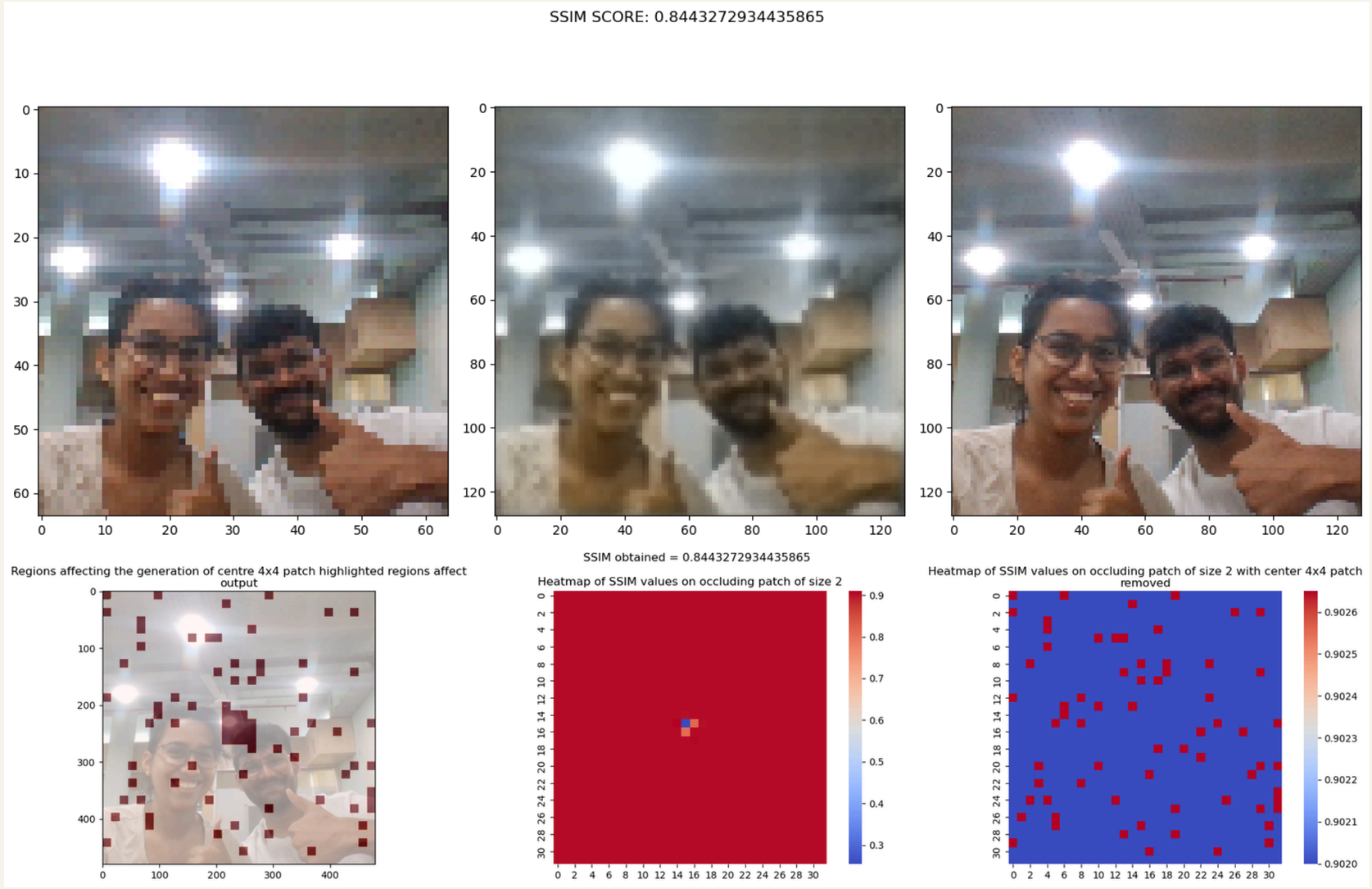
Thank you.

: Now we can move on to a short live demo

Live Demo Results I



Live Demo Results II



Our Codebase:

Github Repo link:
https://github.com/roja26/SuperRes

References

https://arxiv.org/abs/2108.10257:
SwinIR: Image Restoration Using Swin Transformer
https://arxiv.org/abs/1807.02758:
Image Super-Resolution Using Very Deep Residual Channel Attention Networks
https://arxiv.org/abs/2103.14030v2:
Swin Transformer: Hierarchical Vision Transformer using Shifted Windows
https://ieeexplore.ieee.org/document/5596999:
Image Quality Metrics: PSNR vs. SSIM
https://www.atlantis-press.com/proceedings/iccsee-13/4822:
Comparison of Commonly Used Image Interpolation Methods

Fin.