UNIVERSITY of HOUSTON

COLLEGE of NATURAL SCIENCES & MATHEMATICS

Comparing Perceptual Quality vs. Pixel Loss on a Simple Video Super-Resolution Model

Chad Weatherly
cdweathe@central.uh.edu

Paramjit Kainth
pskainth@central.uh.edu

Abhigna Vadlamudi
avadlam2@central.uh.edu

Single Image Super Resolution (SISR)

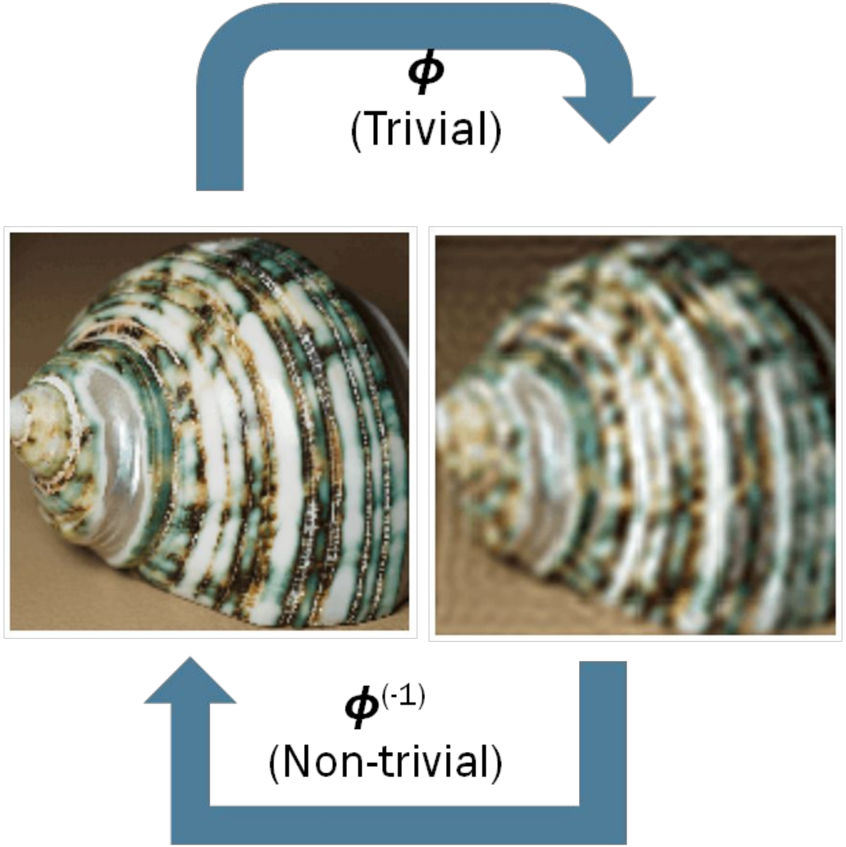
- Single Image Super-Resolution, (SISR) is the process of producing a high-resolution image/frame from a low-resolution one. It is a type of frame restoration.

- Assumes a degradation function, $\boldsymbol{\phi}$

  LR = $\boldsymbol{\phi}$(HR ; $\boldsymbol{\theta}$) ,

  LR = Low-Resolution Image
  HR = High-Resolution Image
  $\boldsymbol{\theta}$ = Parameters for degradation factors, such as noise, motion blur, and downsampling

- How can we find an approximation to $\boldsymbol{\phi}$(-1) that will be perceptually appealing to a human observer?



$\boldsymbol{\phi}$

(Trivial)

$\boldsymbol{\phi}^{(-1)}$

(Non-trivial)

Image Credit: https://en.wikipedia.org/wiki/Image_scaling

Video Super-resolution (VSR)

- The VSR task can be described as the ***reconstruction of high-resolution (HR) videos when only given low-resolution (LR) counterpart video.***

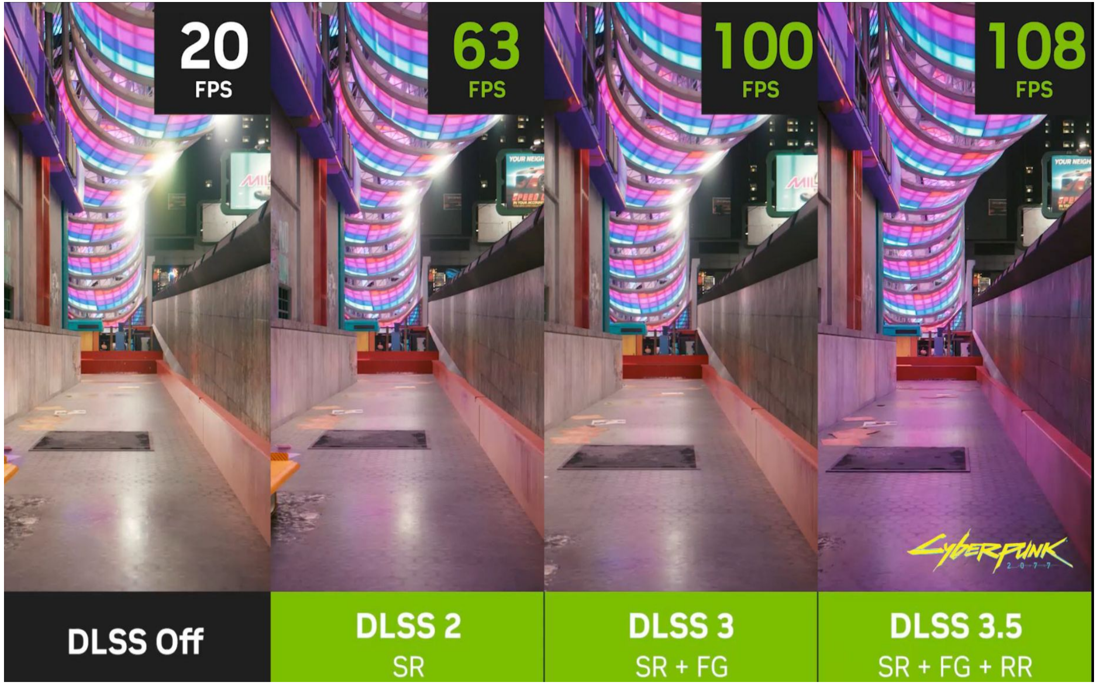
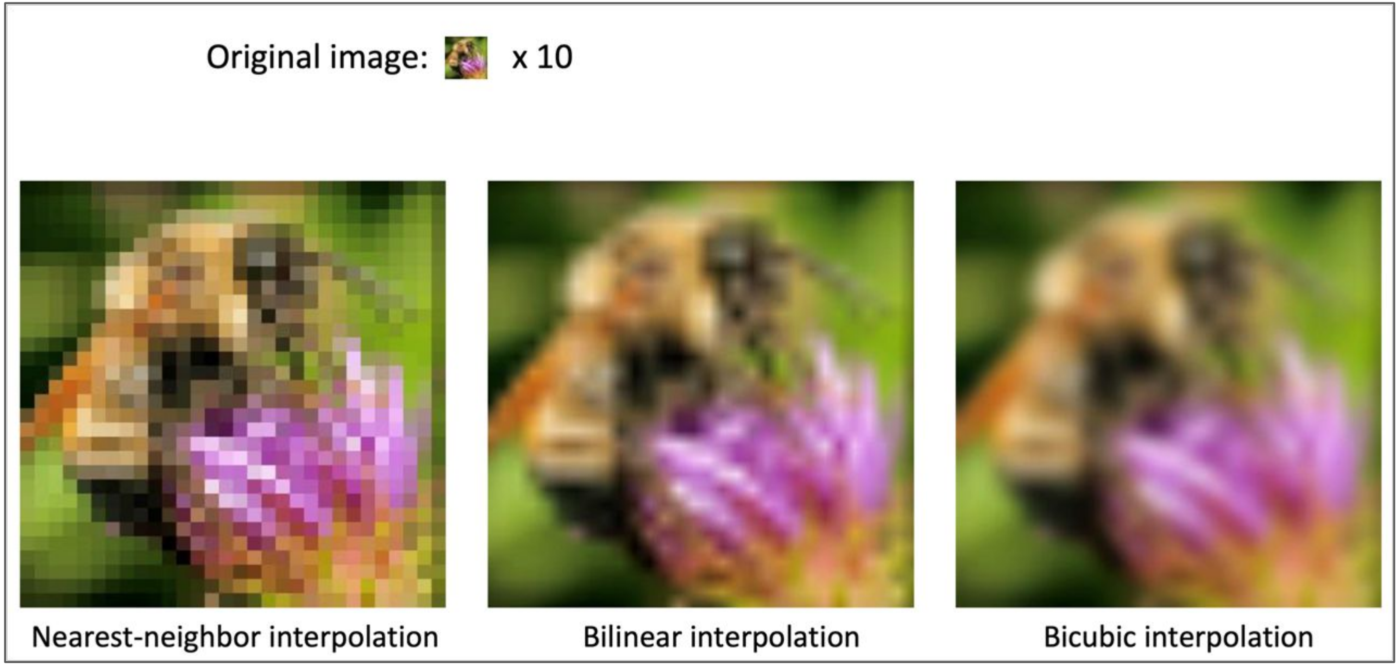


Image Credit: https://developer.nvidia.com/rtx/dlss

Prior Work

Frame Interpolation / upsampling: buggy, grainy, blurry



Original image: x 10

Nearest-neighbor interpolation

Bilinear interpolation

Bicubic interpolation

Image Credit: https://www.cs.toronto.edu/~guerzhoy/320/lec/upsampling.pdf

Prior Work

Deep Learning: better results but more difficult/ computationally heavy to train

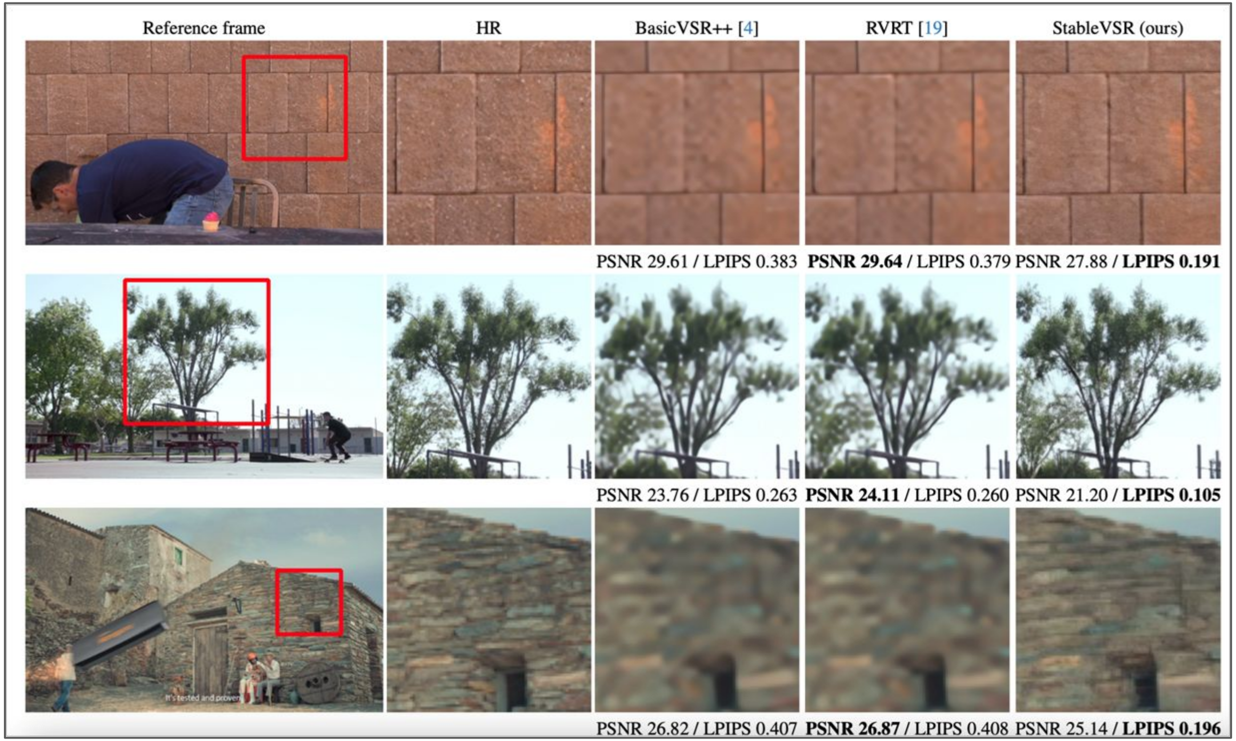
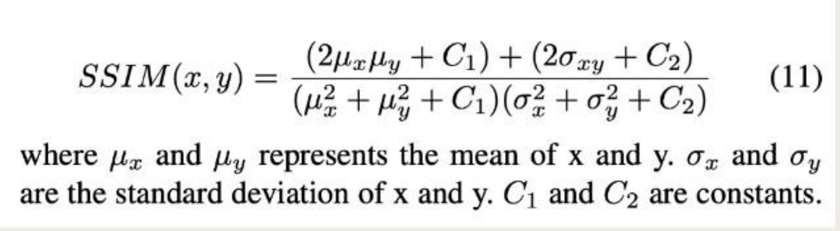

Reference frame
HR
BasicVSR++ [4]
RVRT [19]
StableVSR (ours)
PSNR 29.61 / LPIPS 0.383
PSNR 29.64 / LPIPS 0.379
PSNR 27.88 / **LPIPS 0.191**
PSNR 23.76 / LPIPS 0.263
PSNR 24.11 / LPIPS 0.260
PSNR 21.20 / **LPIPS 0.105**
PSNR 26.82 / LPIPS 0.407
PSNR 26.87 / LPIPS 0.408
PSNR 25.14 / **LPIPS 0.196**
It's tested and proven

Image Credit: https://arxiv.org/pdf/2311.15908.pdf

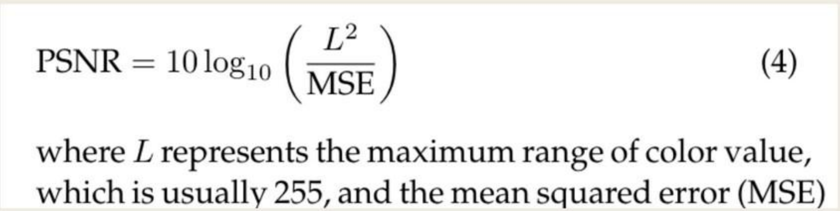
Prior Work

- ***Emphasised updated architectures***, rather than viewing the VSR task holistically
- **During training, <u>pixel-loss functions (like MSE)</u>, are used as the objective**, which often does not result in images that appear better to the human eye.
- In most cases, a model's ***acceptability is judged based on two criteria***: Peak Signal-Noise Ratio (PSNR) and Structural Similarity Matrix (SSIM)



$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{11}$$

where μ_x and μ_y represents the mean of x and y. σ_x and σ_y are the standard deviation of x and y. C_1 and C_2 are constants.



$$\mathrm{PSNR} = 10\log_{10}\left(\frac{L^2}{\mathrm{MSE}}\right) \tag{4}$$

where L represents the maximum range of color value, which is usually 255, and the mean squared error (MSE)

Image Credit: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10164823

Our Problem statement

- We presents a *comparative study of training a simple VSR model utilizing a traditional **pixel- loss function** against a model of the same architecture trained on a **perceptual loss function**.*

- While most state-of-the-art approaches prioritise ***architectural innovations***, *our research emphasises the training process to enhance perceptual quality*, a metric that better aligns with human visual perception than pixel accuracy.

Perceptual Loss

- Imagine a new DL model that needs to be trained for a VSR task. Given a LR frame sequence, it will try to create a corresponding HR frame.

- This is done by passing both the ground truth and predicted HR frame through a given model that performs well on image-related tasks (eg. VGG)
  - The loss measures the ***difference between the activation maps (tensors) created by passing both frames through this model.***
  - If two images are perceptually similar, they will "look" similar to a model that performs well on extracting features from images, i.e. th***e goal is for the hidden features of two images to be similar, instead of just a pixel-to-pixel recreation***.

- In testing, perceptual loss tends to create images that look "better" to human observers

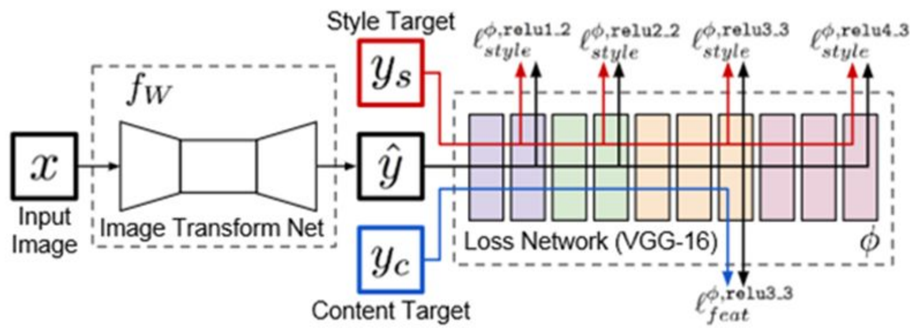

Style Target
y_s
$\ell_{style}^{\phi,relu1\_2}$
$\ell_{style}^{\phi,relu2\_2}$
$\ell_{style}^{\phi,relu3\_3}$
$\ell_{style}^{\phi,relu4\_3}$
f_W
x
Input Image
Image Transform Net
$\hat{y}$
y_c
Loss Network (VGG-16)
ϕ
Content Target
$\ell_{feat}^{\phi,relu3\_3}$

Fig. 2. System overview. We train an *image transformation network* to transform input images into output images. We use a *loss network* pretrained for image classification to define *perceptual loss functions* that measure perceptual differences in content and style between images. The loss network remains fixed during the training process.

Image Credit: https://ieeexplore.ieee.org/document/8578166/

Dataset → Moving MNIST

- The Moving MNIST dataset contains ***10,000 grayscale video sequences, each consisting of 20 frames***.
  - In each video sequence, two digits move independently around the frame, which has a spatial ***resolution of 64×64 pixels***.
  - The digits frequently intersect with each other and bounce off the edges of the frame

- **Data preprocessing →**
  - for a given video sample *i*, a random sequence of *5 consecutive frames* was chosen from the sample.
  - These frames were downsampled using *bicubic interpolation to be of size (32 x 32)* and then processed through a Gaussian blur.

Image Credit: https://www.cs.toronto.edu/~nitish/unsupervised_video/

Model Architecture

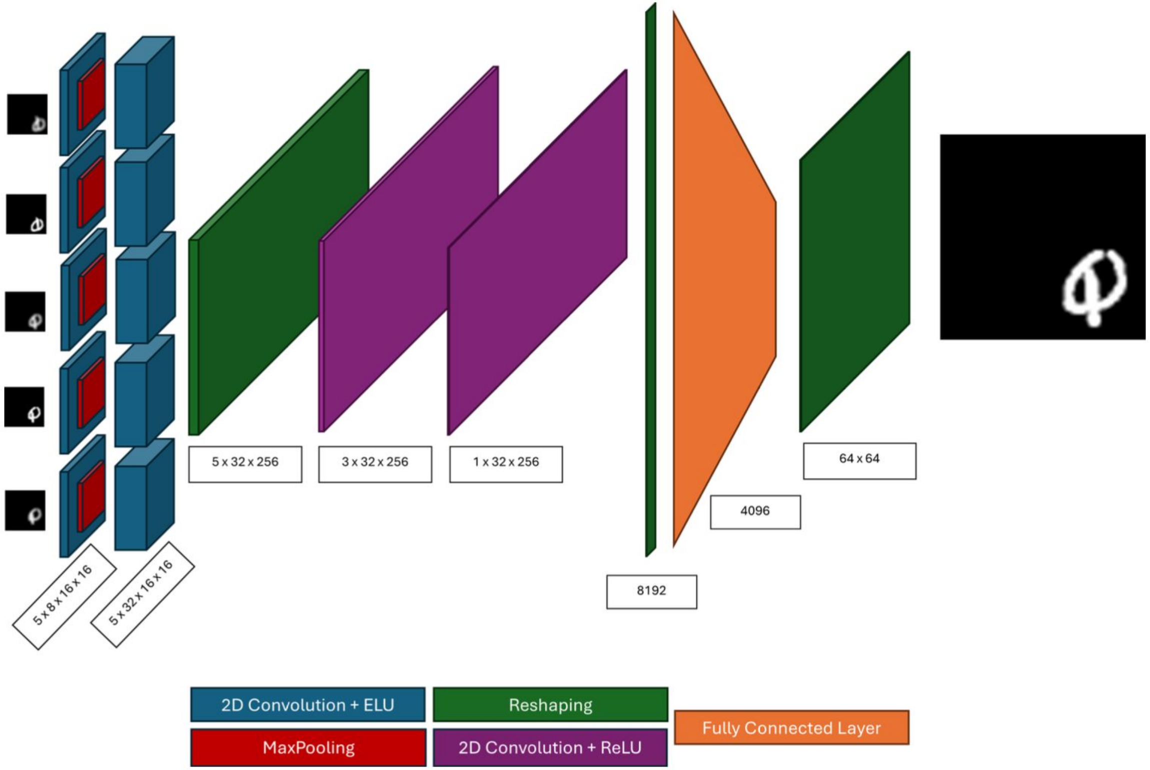


Figure 1: Model Architecture

Training (novelty of our approach)

Model 1→ was trained on a **perceptual loss metric**

- Our implementation of perceptual loss was done with the ***VGG-19 model***, where activations were computed by passing each image (the ground truth and predicted HR images) through the VGG-19 model and extracting the values from each ReLU activation.

$$Perceptual\ Loss(\hat{I},\tilde{I}) = \sum_{a=1}^{N}\left[\frac{1}{T}\sum_{t=1}^{T}|VGG_{a,t}(\hat{I}) - VGG_{a,t}(\tilde{I})|\right]$$

Model 2 →was trained on the MSE difference between the **image pixel values**.

$$MSE(\tilde{I},\hat{I}) = \frac{1}{N}\sum_{i=1}^{N}(\tilde{I}_i - \hat{I}_i)^2$$

Perceptual Loss Model(PLM) Results

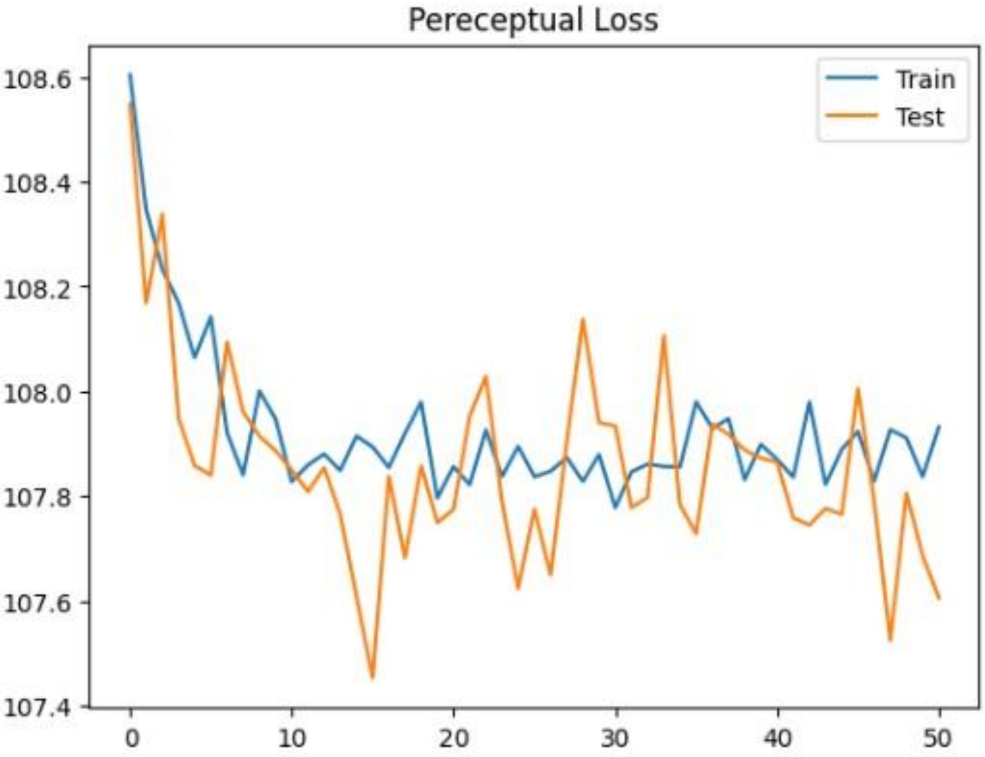


Figure 2: Avg. Perceptual Loss for a single image across training epochs

perceptual loss model (PLM) was able to somewhat train and converge

MSE Model (MSEM) Results

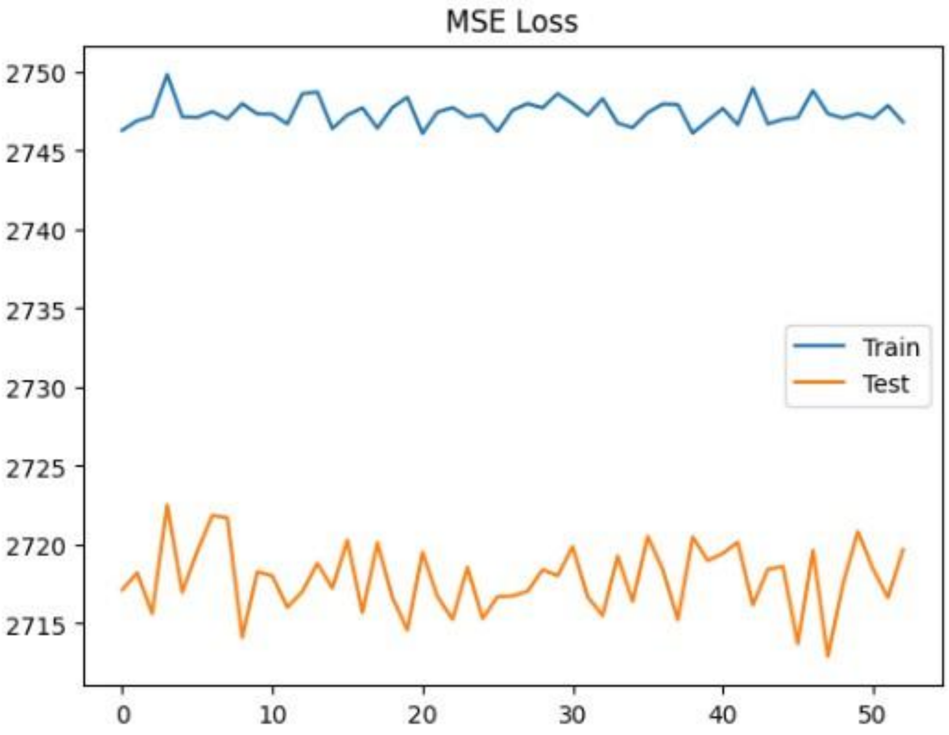


Figure 3: Avg. MSE Loss for a single image across training epochs

MSE Model (MSEM) didn't seem to find any sort of consensus, regardless of learning rate.

PSNR and SSIM (Metrics)

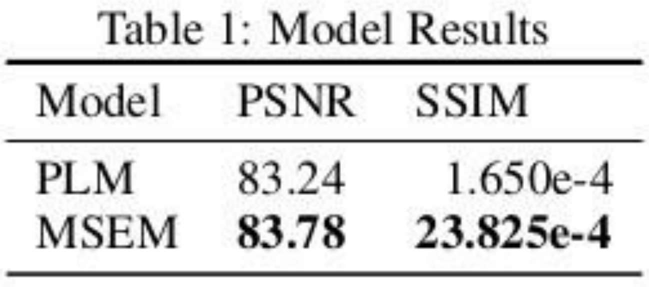


Table 1: Model Results

Model	PSNR	SSIM
PLM	83.24	1.650e-4
MSEM	**83.78**	**23.825e-4**

- both the **PSNR and SSIM values were similar** for each of the two model types

- ***MSEM performed slightly better*** in both metrics

Comparison

- **PLM was able to actually learn a little**
  - *PLM learned lower-level features*, with the low complexity of our model unable to fully capture high-level details.
  - PLM captured some information contained in the LR images, managing to form a general shape that resembles the ground truth image.

- Meanwhile, the ***MSEM-predicted images are no more than smatterings of pixels***.
  - This also proves that PSNR and SSIM do not fully capture an image's quality, as the MSEM performed slightly better (higher is better for both SSIM and PSNR) in both metrics (Table 1).

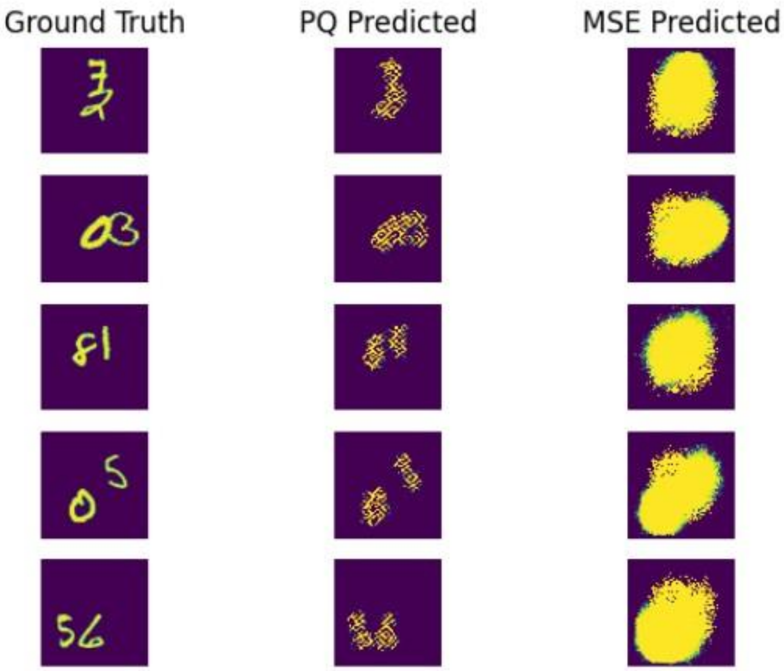


Figure 4: Model Final Predictions

Observation

Reasons for Poor Overall Performance of both models →

- Given the time and compute restraints of a class project, the model had to be ***simple***.

- We presume that given a more **complex model**, the differences in both loss functions might become even more apparent.

- With ***more time and computational resources***, it would be fascinating to see how comparing perceptual loss to other loss functions might work.

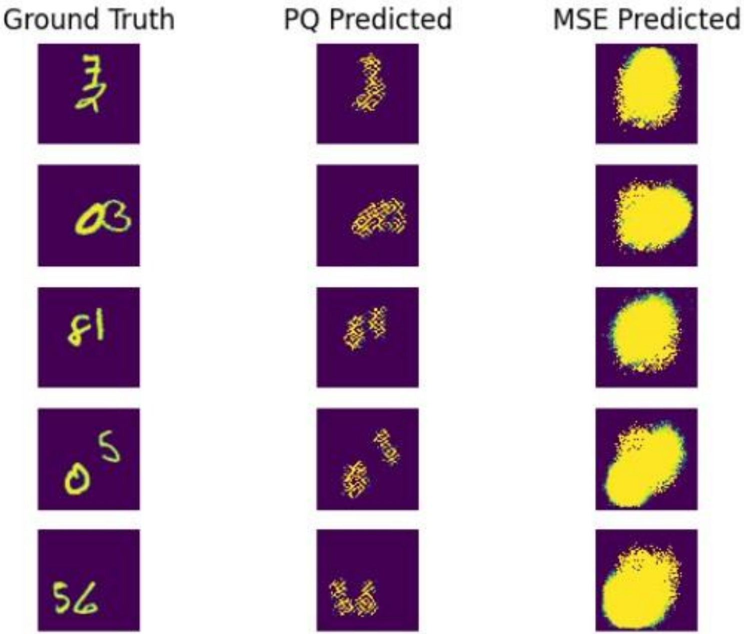


Figure 4: Model Final Predictions

Conclusion

- Despite the simplicity of the model and inherent limitations of the VSR task, our findings illustrate that the ***perceptual loss-trained model produces more visually coherent images*** than its MSE counterpart,

- even though traditional metrics like (PSNR) and (SSIM) showed similar or slightly better values for the MSE model. This discrepancy highlights the ***inadequacy of PSNR and SSIM as sole arbiters of image quality.***

Contributions of Members

Each member contributed equally to this project in terms of time and effort given

This project represents a collective effort to ***reconceptualize the training of Image/Video Restoration models***, advocating for a paradigm shift towards perceptual quality-centric methods.

Thank you!