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

Chad Weatherly
Department of Computer Science
University of Houston
Houston, TX, USA
cdweathe@central.uh.edu

Paramjit S. Kainth
Department of Mechanical Engineering
University of Houston
Houston, TX, USA
pskainth@central.uh.edu

Abhigna Sowgandhika Vadlamudi
Department of Computer Science
University of Houston
Houston, TX, USA
avadlam2@central.uh.edu

Abstract

With the ongoing need for high-fidelity videos in different areas of entertainment, Video Super-Resolution (VSR) has become a pivotal area of research, transcending its roots in gaming to broader applications such as medical imaging and surveillance. This paper 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 prioritize architectural innovations, our research emphasizes the training process to enhance perceptual quality, a metric that better aligns with human visual perception than pixel accuracy. 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 Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (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. Our study paves the way for future explorations into complex models that could benefit from a mixed approach to loss functions, potentially leading to superior VSR techniques. This project represents a collective effort to reconceptualize the training of VSR models, advocating for a paradigm shift towards perceptual quality-centric methods.
Note: Our code is on our GitHub.

1 Introduction

Within the last decade, deep learning (DL) has been used in an attempt to solve the problem of Single Image Super-Resolution (SISR), which takes a low resolution (LR) image and attempts to construct a corresponding high resolution (HR) image. This problem was introduced to the mainstream with NVIDIA's introduction of Deep Learning Super-Sampling (DLSS) in their GPU's for playing video games. DLSS is more of a Video Super-Resolution (VSR) task, which gives more complexity to the SISR issue by introducing inter-frame information. The VSR task can be described as the reconstruction of high-resolution (HR) videos when only given low-resolution (LR) counterpart video. In the video game landscape, there is currently a drive to have games played at a high frame rate and with high resolution and video fidelity. This is extremely computationally expensive due to the sheer number of physics calculations each second. Improvements have been added every year or so since its inception in 2019. AMD and Intel have since introduced their own rival version of this type of model software, Fidelity SuperResolution (FSR) and XeSS, respectively, demonstrating the growth in this field in the past 3-4 years. The DLSS problem can also be transposed to other fields besides video games, such as "medical imagery reconstruction... remote sensing... panorama video super-resolution... surveillance systems... high-definition television" [Liu+22].

1.1 Prior Work

Up until this point, there have been a few notable papers that have introduced new architectures to the field, based on other DL structures [Cha+21], [Cha+22], [Lia+], [RMW23], with most all papers using similar datasets like Vimeo-90K [Xue+19] and REDS [Nah+21]. Notably, significant improvements have been found by introducing deformable convolutions [Dai+16], but most improvements come by training models to reduce pixel-loss functions, which often does not result in images that appear better to the human eye. Perceptual Quality (PQ) was introduced as a metric for image/video restoration problems [JAF16], [Zha+18], which aims for a chosen model architecture to have similar feature maps to those of other state-of-the-art (SOTA) image/CNN models, thereby enhancing perceptual quality. To explain this concretely, imagine a new DL model that needs to be trained for a VSR task. Given a LR frame/image, 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 (VGG, ImageNet, ResNet, ...). The loss measures the difference between the activation maps (tensors) created by passing both frames through this model. The idea is that if two images are perceptually similar, they will "look" similar to a model that performs well on extracting features from images, i.e. the 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 [Zha+18].

Frustratingly, the training process is little emphasized in current SOTA VSR models, which is bizarre, as pixel-loss functions (like MSE) don't capture the problem in its entirety, making the problem ill-posed. Many of the most cited papers for VSR tasks emphasize updated architectures, rather than viewing the VSR task holistically [Liu+22]. In most cases, a model's acceptability is judged based on two criteria: Peak Signal-Noise Ratio (PSNR) and Structural Similarity Matrix (SSIM), as defined below for any two given images of the same size, $\widetilde{I}$ and $\hat{I}$:

$$SSIM(\widetilde{I},\hat{I}) = \frac{2\mu_{\hat{I}}\mu_{I} + k_1}{\mu_{\hat{I}}^2 + \mu_{\widetilde{I}}^2 + k_1} \cdot \frac{2\sigma_{\hat{I}I} + k_2}{\sigma_{\hat{I}}^2 + \sigma_{\widetilde{I}}^2 + k_2} \tag{1}$$

where $\mu_{\widetilde{I}}/\mu_{\hat{I}}$ and $\sigma_{\widetilde{I}}/\sigma_{\hat{I}}$ are the mean and standard deviation, respectively, of image $\widetilde{I}$ / $\hat{I}$, $\sigma_{\hat{I}\widetilde{I}}$ is the covariance of the two images, and k_1 and k_2 are constants, usually set to 0.01 and 0.03, respectively.

$$PSNR(\widetilde{I},\hat{I}) = 10 * log_{10}\left(\frac{L^2}{MSE(\widetilde{I},\hat{I})}\right) \tag{2}$$

where L represents the maximum range of color values, which is usually 255, and the Mean-Squared Error (MSE) is defined as:

$$MSE(\widetilde{I}, \hat{I}) = \frac{1}{N} \sum_{i=1}^{N} (\widetilde{I}_i - \hat{I}_i)^2 \tag{3}$$

for each pixel value i in all pixel values N.

2 Problem Statement

In this project, the aim is to assess how adjusting the training process to use perceptual loss might affect the final performance of a simple DL model trained on the VSR task. This projects strives to train a novel, but simple, VSR model, first on pixel-loss metrics, second using perceptual quality.

3 Technical Approach / Methodology

Originally, the intention was to create a diffusion model, similar to the one by Rota et al. [RMW23], that would perform VSR on the REDS dataset [Nah+21], but this was changed for two reasons: the data necessitated a large model that was infeasible to train on a local machine, and the diffusion model needed much compute in order to be trained effectively. We did not have the time to feasibly train the model on this data. For these reasons, we decided to go with a simpler model architecture and a simpler data set, where individual frames were smaller.

3.1 Data

We chose to use the Moving MNIST dataset, because its small frame size (64 x 64) was well suited to our task. Also, the fact that it only has one channel further motivated the decision. Each sample consists of 20 frames in a sequence, with 10,000 samples/videos in total. We used 8,000 samples for training and 2,000 for testing.

During training, 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. These 5 images were passed to the model, with the aim of predicting the middle frame's HR counterpart. For a video sequence consisting of 20 frames, inference can be done by upsampling the central 16 frames.

3.2 Model Architecture

Due to the technical and time constraints of a class project, we propose to create our own network that would be effective in evaluating the loss/objective metrics, but simple enough to be trained on local machines. The model takes in 5 frames (as it will be a video), and the middle frame of these 5 is the one we wish to reconstruct as a HR image (Fig. 1). First, to capture feature information about each of the images, the 5 frames are each passed through 3 convolutional layers, where each layer is Max Pooled and activated using the Exponential Linear Unit (ELU, Eq. 4). Each image tensor is flattened to a 1-Dimensional vector. Then, the vectors are each reshaped to be size (32, 256). These 2D vectors are stacked as "channels", creating one image, which is finally passed through two more convolutional layers and activated with ReLU. Finally, the entire vector is flattened to 1D, and passed through a linear layer, which reduces its size to 4096. This final vector is reshaped to be size (64, 64) and passed as the output of the model.

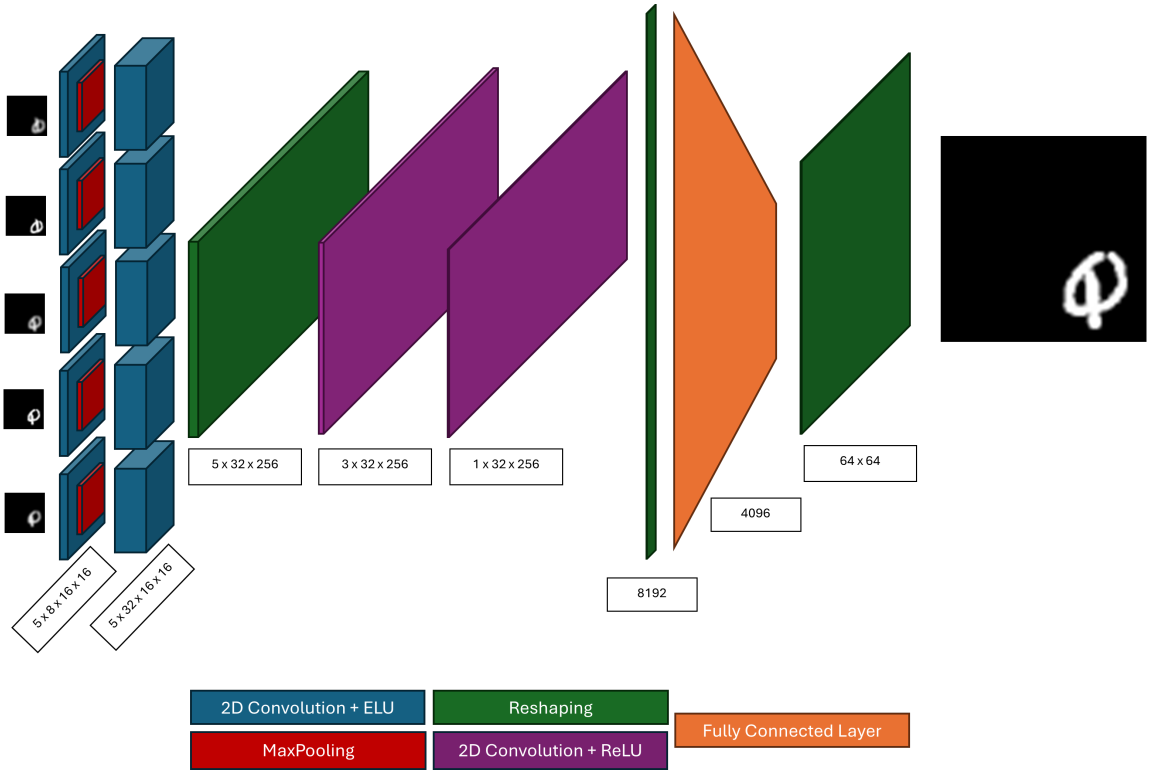

Figure 1: Model Architecture

$$ELU(x) = \begin{cases} x, & if \;\; x > 0 \\ e^x - 1, & if \;\; x \leq 0 \end{cases} \tag{4}$$

3.3 Training

The novelty of our approach is found in the training process. We trained two versions of the model architecture (Fig. 1) from scratch. Model 1 was trained on a perceptual loss metric (Eq. 5), as proposed by Johnson et al. [JAF16]. Our implementation of perceptual loss was done with the VGG-19 model [SZ15], 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. Model 2 was trained on the MSE (Eq.3) difference between the two image pixel values. Both models were trained on 50 epochs using the Adam optimizer. The learning rate was adjusted as epochs were passed: 0.0001 at start, 0.00001 after 5 epochs, and 0.000001 after 15 epochs.

$$Perceptual\ Loss(\hat{I}, \widetilde{I}) = \sum_{a=1}^{N} \left[\frac{1}{T} \sum_{t=1}^{T} \left| VGG_{a,t}(\hat{I}) - VGG_{a,t}(\widetilde{I}) \right| \right] \tag{5}$$

for T values in activation layers A, where $VGG_{a,t}(I)$ is the ReLU output value t from activation layer a when image I is passed through.

Table 1: Model Results

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

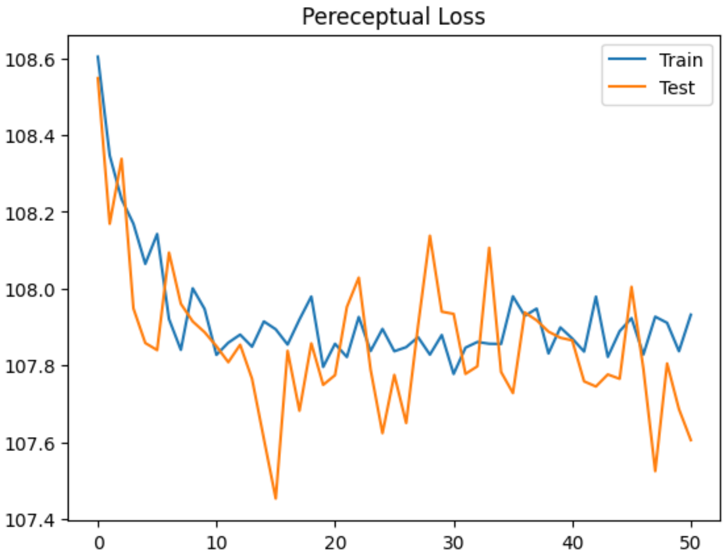

Figure 2: Avg. Perceptual Loss for a single image across training epochs

4 Results

Figures 2 and 3 show how the loss functions converged for each of the functions. As can be seen, the perceptual loss model (PLM) was able to somewhat train and converge, while the MSE Model (MSEM) didn't seem to find any sort of consensus, regardless of learning rate. Interestingly, both the PSNR and SSIM values were similar for each of the two model types (Table 1), with the MSEM performing slightly better in both metrics (although both models performed poorly compared to SOTA methods). Finally, Figure 4 shows a sample of how the models' predictions actually looked.

5 Discussion

Given the simplicity of the model and the difficulty in the VSR task, it's no shock that both models performed poorly. Our model didn't take temporal information into account much, besides concatenating all vectors together, but in their failures, some interesting distinctions can be made. It would appear that the PLM was able to actually learn a little, and the model was able to produce forms that somewhat resemble their ground-truth counterparts. This might signal the fact that the PLM learned lower-level features, with the low complexity of our model unable to fully capture high-level details. To a human observer, the PLM-predicted images at least capture 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 the initial assertion that PSNR and SSIM do not fully capture an image's quality,

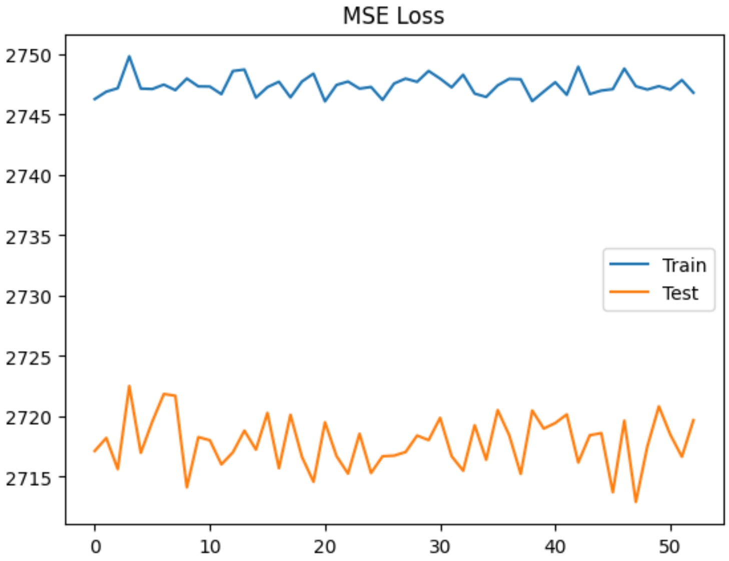

Figure 3: Avg. MSE Loss for a single image across training epochs

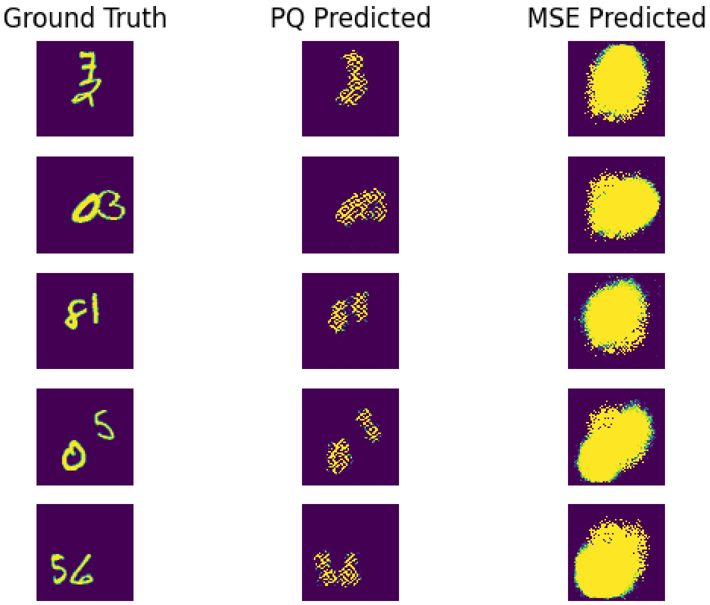

Figure 4: Model Final Predictions

as the MSEM performed slightly better (higher is better for both SSIM and PSNR) in both metrics (Table 1). We presume that given a more complex model, the differences in both loss functions might become even more apparent. In the future, given more time and computational resources, it would be fascinating to see how comparing perceptual loss to other loss functions might work, or even attempting to strike a balance, with a complex model that is trained on a combination of loss functions in order to generate high-fidelity super-resolved images and videos.

5.1 Contributions

This project was done with equal effort from all three team members.

References

[SZ15] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. en. arXiv:1409.1556 [cs]. Apr. 2015. URL: http://arxiv.org/abs/1409.1556 (visited on 04/17/2024).

[Dai+16] Jifeng Dai et al. “Deformable Convolutional Networks”. en. In: (2016).

[JAF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. “Perceptual Losses for Real-Time Style Transfer and Super-Resolution”. en. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 694–711. ISBN: 978-3-319-46475-6. DOI: 10.1007/978-3-319-46475-6_43.

[Zha+18] Richard Zhang et al. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric”. en. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, June 2018, pp. 586–595. ISBN: 978-1-5386-6420-9. DOI: 10.1109/CVPR.2018.00068. URL: https://ieeexplore.ieee.org/document/8578166/ (visited on 02/01/2024).

[Xue+19] Tianfan Xue et al. *Video Enhancement with Task-Oriented Flow*. 2019. URL: http://toflow.csail.mit.edu/ (visited on 02/17/2024).

[Cha+21] Kelvin C.K. Chan et al. “BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond”. en. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, June 2021, pp. 4945–4954. ISBN: 978-1-66544-509-2. DOI: 10.1109/CVPR46437.2021.00491. URL: https://ieeexplore.ieee.org/document/9577681/ (visited on 02/02/2024).

[Nah+21] Seungjun Nah et al. “NTIRE 2021 Challenge on Image Deblurring”. en. In: (2021).

[Cha+22] Kelvin C. K. Chan et al. “BasicVSR++: Improving Video Super-Resolution With Enhanced Propagation and Alignment”. en. In: 2022, pp. 5972–5981. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Chan_BasicVSR_Improving_Video_Super-Resolution_With_Enhanced_Propagation_and_Alignment_CVPR_2022_paper.html (visited on 01/29/2024).

[Liu+22] Hongying Liu et al. “Video super-resolution based on deep learning: a comprehensive survey”. en. In: *Artificial Intelligence Review* 55.8 (Dec. 2022), pp. 5981–6035. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10147-y. URL: https://doi.org/10.1007/s10462-022-10147-y (visited on 12/19/2023).

[RMW23] Claudio Rota, Marco Buzzelli, and Joost van de Weijer. *Enhancing Perceptual Quality in Video Super-Resolution through Temporally-Consistent Detail Synthesis using Diffusion Models*. en. 2023. URL: https://ar5iv.labs.arxiv.org/html/2311.15908 (visited on 02/01/2024).

[Lia+] Jingyun Liang et al. “Recurrent Video Restoration Transformer with Guided Deformable Attention”. en. In: ().