Benchmark on Video Super Resolution

Daksh K. Shah
Computer Science and Engineering
ID: 1953946
dakshah@ucsc.edu

Dylan J. Louie
Computer Science and Engineering
ID: 1909298
djlouie@ucsc.edu

Abstract

Video Super-Resolution (VSR) aims to enhance the spatial resolution of video frames while maintaining temporal consistency. These are major challenges, often leading to visual artifacts such as flickering and inconsistent motion in naive frame-by-frame approaches. Our benchmark will be based the REDS dataset, a widely used dataset focused on realistic degradations and real-world videos. This project will explore Generative Adversarial Networks (GANs) and Diffusion model-based approaches to VSR. We implement ESRGAN [17], Real-ESRGAN [16], SR3 [12], and StableVSR [11] architectures to investigate different approaches to VSR. We will evaluate the results using a variety of metrics including pixel-wise fidelity (PSNR, SSIM [18]), perceptual quality (LPIPS [20], DISTS [2]), and temporal consistency (Temporal LPIPS). We will be performing an in-depth analysis and discussion on different generative approaches and the introduction of temporal consistency techniques in tackling the challenges in VSR.

1 Introduction

1.1 Video Super Resolution

Single-image super resolution (SISR) is the process of generating high resolution (HR) images given low resolution (LR) image priors. Similarly, Video Super-Resolution (VSR) is the process of generating HR video sequences given LR video sequence priors. While both SISR and VSR seek to enhance visual quality, the crucial distinction lies in VSR's ability to exploit temporal information present across multiple sequential frames. This temporal context is vital for state-of-the-art VSR approaches, allowing them to synthesize details more effectively and, critically, to maintain coherence and smoothness throughout the video.

There are two overarching goals of VSR: Firstly, enhancing spatial details, where individual frames are sharp and perceptually pleasing. Secondly, achieving temporal consistency, where details are stable across frames with more natural motion, and visual artifacts are mitigated.

The demand for effective VSR is pervasive, as a significant portion of our digital video content originates or is transmitted in low resolution.This includes scenarios ranging from low-bandwidth video streaming, where higher resolutions are downscaled for efficient transmission, to footage acquired from surveillance systems or older camera technologies that inherently capture at lower resolutions. In these and numerous other applications, VSR offers a powerful means to significantly improve visual quality, thereby enhancing user experience, facilitating clearer analysis, and unlocking new possibilities for legacy or constrained video content.

1.2 Challenges in Video Super Resolution

Video super resolution faces many of the same challenges as single image super resolution which is why they use similar metrics like pixel-wise fidelity metrics such as PSNR and SSIM. Additionally

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

they must overcome the human perceptual tests and not just the pixel wise metrics which is where LPIPS and DISTS are good metrics which compare perceptual feature embeddings using image encoders, this is similar to how iBOT [21] learns by comparing perceptual embeddings during self distillation.

The main challenges that set video super resolution apart from single image super resolution is the temporal domain which is measured using temporal consistency metrics such as temporal optical flow (tOF) and temporal LPIPS (tLP), we demonstrate how models that specifically train with temporal consistency techniques such as StableVSR outperform single image super resolution models like ESRGAN. A common visual problem people will notice is flickering between generated frames which is caused by inconsistent lighting in consecutive generated frames. Specifically where computation is a limited resource, there can be a tradeoff in models focusing on high resolution output per frame or temporal consistency. In the related works, you can see different models focus on improving on different parts of challenges in VSR. Specifically VSR has much more limited compute than SISR as it much generate much more, for example one sample might have 400 more frames which leads to 400 times more generated image for one sample.

2 Related work

2.1 GAN-based approaches

2.1.1 SRGAN

In the context of SISR, Ledig et al. (2017) introduced SRGAN [8], a pioneering generative adversarial network (GAN) framework for 4x upscaling on individual images. When this paper was published, it was the first model that was capable of photo-realistic upscaling at 4x. Their approach uses both adversarial and content loss. The former utilizes the discrimator network to converge to photo-realistic images, while the latter utilizes a perceptual loss function.

2.1.2 ESRGAN

ESRGAN (Enhanced Super-Resolution Generative Adversarial Network) [17] is a deep learning model widely employed for single image super-resolution tasks. It builds upon the SRGAN architecture by introducing several major enhancements, including the removal of batch normalization layers, the adoption of a Residual-in-Residual Dense Block (RRDB) for richer feature extraction, and an improved perceptual loss function. These enable ESRGAN to generate more visually pleasing and realistic super-resolved outputs with finer textures and details, helping mitigate common artifacts found in traditional super-resolution methods when applied frame-by-frame.

2.1.3 Real-ESRGAN

Real-ESRGAN [16] extends the capabilities of ESRGAN by specifically addressing real-world image degradation, which is often more complex than the synthetic degradation used to train ESRGAN. Real-ESRGAN introduces a more sophisticated degradation model during training, which simulates realistic corruptions like various types of blur, noise, and compression artifacts. This enhanced training, coupled with an improved U-Net in the discriminator, allows Real-ESRGAN to produce more robust and visually superior results on real-world low-resolution videos, making it highly effective for practical video restoration by reducing common artifacts and improving overall visual clarity.

2.2 Diffusion-based approaches

2.2.1 Image Super-Resolution via Iterative Refinement (SR3)

SR3 is a diffusion-based model architecture that relies on conditional image generation to upscale images. Instead of directly predicting the high-resolution image, SR3 starts with pure Gaussian noise and iteratively refines it through a stochastic denoising process, conditioned on the low-resolution input. The iterative refinement process relies on a U-Net architecture which generates photo-realistic images, outperforming known GAN-based methods in human evaluation with a 50% fool rate.

While its cascaded nature presents a potential benefit for video super-resolution, our practical experimentation revealed significant challenges. However, when attempting inference using an unofficial implementation, we found that this would take over 24 hours to upscale a single sequence from REDS4, with mediocre results. We believe this may require more steps to upscale each image. Consequently, due to limited computational resources, we decided to exclude this approach from our primary experiments.

2.2.2 Cascaded Diffusion Models for High Fidelity Image Generation (CDM)

Cascading diffusion models innovate by having multiple diffusion models for increasing resolution instead of just one diffusion model taking in the input lower resolution directly to the high resolution [5]. They find that augmentation in the pipeline between diffusion models, conditioning augmentation, is crucial for preventing compounding errors. Using this technique they are able to get performance on ImageNet without auxillery classifiers.

2.2.3 Diffusion Posterior Sampling for General Noisy Inverse Problems

Diffusion posterior sampling is a technique for tackling denoising general non-linear transformations such as phase retrieval and non-uniform blur [1]. This improves on techniques trained and tested on specifically linear blur and other linear transforms.

2.3 Temporal Consistency-based approaches

2.3.1 EDVR: Video Restoration with Enhanced Deformable Convolutional Networks

From the same group that created ESRGAN and Real-ESRGAN, we have EDVR [15]. This approach uses both convolutions and spatial-attention. to handle temporal consistency. Its key innovation lies in effectively handling large and complex motions between video frames. This is achieved through a "Pyramid, Cascading and Deformable (PCD) alignment module" that aligns features at multiple scales using deformable convolutions, and a "Temporal and Spatial Attention (TSA) fusion module" that selectively combines information from aligned frames. EDVR has demonstrated superior performance and won multiple challenges in video restoration, showcasing its ability to produce high-quality restored videos.

2.3.2 StableVSR

On top of single image super resolution models, StableVSR introduces a temporal conditioning module on a pretrained diffusion module backbone [11]. Instead of focusing on pixel-wise fidelity metrics like PSNR and SSIM, this module aims to improve accuracy on temporal optical flow and temporal learned perceptual image patch similarity. The temporal conditioning module specifically uses temporal texture guidance, which brings in information from adjacent frames in generating process.

2.3.3 DiffVSR

VSR research is still happening at a rapid pace, here is an example of a recent model which does a progressive learning strategy in stages to achieve temporal consistency with minimal overhead. Its progressive learning strategy which focuses first on temporal consistency, then introducing complex degradations, and finally fine tunes the whole on high-quality video datasets combined with interweaved latent transitions supposedly does well on temporal consistency metrics despite severely degraded initial videos.[9]

3 Experiments

Our goal was to compare different diffusion and GAN models on video super resolution using pixel-wise fidelity metrics, perceptual quality metrics, and and temporal consistency metrics. The pixel-wise fidelity metrics we used were peak-signal to noise ratio (PSNR) and structural simularity index (SSIM). The perceptual quality metrics we used were deep image structure and texture simularity (DISTS) with VGGNet [13] and learned perceptual image patch simularity (LPIPS) with squeeze net [6]. The temporal consistency metric we used was temporal LPIPS (tLP).

We evaluated four models with these five metrics on the REDS dataset, specifically videos 000, 011, 015. and 020 which are often with held videos used for evaluation of REDS and not in training. The four models that we evaluated on were bicubic linear interpolation as a baseline, ESRGAN a GAN model [17], Real-ESRGAN an improved GAN model [16], and StableVSR [11], a diffusion model with temporal conditioning. As we used NRP Nautilus to help our limited compute, this work was supported in part by National Science Foundation (NSF) awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, OAC-2112167, CNS-2100237, CNS-2120019.

Our baseline is computed using bicubic interpolation on a $4\times$ scale from the LR.

We used the weights for ESRGAN, trained with a batch size of 16. There are two stages for training. Firstly, the PSNR-model using L1 loss, learning rate as 2×10^{-4} that anneals by a factor of 2 every 2×10^5 mini-batch updates. Secondly, the generator model is initialized using this, which is then trained using a specialized loss function comprised of perceptual loss, L1 loss, and custom adversarial loss as follows: [17]

$$L_G^{RA} = -\mathbb{E}_{x_r}[1 - \log(D_{Ra}(X_r, x_f))] - \mathbb{E}_{x_r}[\log(D_{Ra}(X_f, x_r))]$$

parameterized by $\lambda = 5 \times 10^{-3}$ and $\eta = 1 \times 10^{-2}$. The learning rate is 1×10^{-4} and halved at different iterations. Adam [7] optimizer is used with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. In the standard approach to training a GAN, the discriminator and generator networks alternate updates until convergence.

Real-ESRGAN is nearly identical in training parameters to ESRGAN, however it incorporates real-world degradations in the training process and upgrades the discriminator model from a VGG [13]-style approach to a U-Net-based one.

StableVSR is build using the Stable Diffusion x4 upscaler model [10]. This is built using a VAE decoder [3] for super-resolution. The Temporal Conditioning Module uses ControlNet [19] trained at 20k steps and RAFT for optical flow [14]. The Adam optimizer was used with batch size 32 and learning rate at 1×10^{-5}. Data augmentations include 256×256 random crops and horizontal flips. For training and inference, DDPM sampling is used [4] at 1000 iterations and 50 iterations, respectively.

Table 1: Quantitative Comparison of Super-Resolution Models: PSNR, SSIM, DISTS, LPIPS, and tLP metrics on REDS4 (Averaged) comparing Bicubic Interpolation, ESRGAN, Real-ESRGAN, and StableVSR on $180 \times 320 \rightarrow 720 \times 1280$ super-resolution.

Model	**PSNR (↑)**	**SSIM (↑)**	**DISTS (↓)**	**LPIPS (↓)**	**tLP (↓)**
Baseline	0.7330	25.9946	0.2447	0.1657	0.0973
ESRGAN	0.7383	45.6387	0.0714	0.0545	0.1318
Real-ESRGAN	0.6896	23.6045	0.1390	0.1032	0.1454
StableVSR	0.7952	27.3590	0.0644	0.0439	0.1316

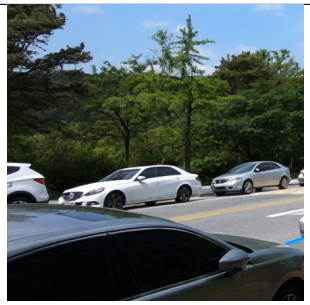
Note: ↑ indicates higher values are better, ↓ indicates lower values are better.

4 Discussion

Our experiments showed that StableVSR outperformed the other models across most metrics. We see the benefit of using a diffusion-based framework with temporal conditioning, particularly in the temporal LPIPS (tLP) metric, which measures frame-wise perceptual consistency. The use of temporal attention and noise modeling appears to provide this model with enhanced frame-to-frame coherence, resulting in fewer flickering artifacts and stable outputs.

Its to be expected that StableVSR which came out in 2023 and specifically trained to perform well using temporal conditioning would perform better than Real-ESRGAN (2021) and ESRGAN (2018). One of the challenges we encountered was specifically finding models with public code and models that were computationally efficient enough to run on our limited compute. One thing we noticed about diffusion models such as SR3 is that they often were very computationally expensive and took a long time to do inference to the point where we had to stop after it took 2 hours to generate a single frame which was not viable when we had 400 frames to generate for our test set.

Table 2: Comparison of 3 frames in sequence 000 from REDS4, zoomed in and cropped.

Model	Frame 0	Frame 1	Frame 2
Bicubic			
ESRGAN			
Real-ESRGAN			
StableVSR			

One interesting result was ESRGAN outperforming Real-ESRGAN in terms of PSNR, SSIM, and LPIPS, despite Real-ESRGAN producing visually cleaner and more realistic frames without typical GAN artifacts such as checkerboarding or over-sharpening. We believe this to be the result of ESRGAN's model being trained on synthetic degradations, which matches the (bicubic downsampling) degradations from REDS. ESRGAN exhibited lower tLP compared to Real-ESRGAN, but this may be from the consistent presence of artifacts, such as checkerboard patterns, across frames. These consistent patterns may reduce temporal variation in perceptual features, thus yielding better tLP despite lower visual quality. Real-ESRGAN often hallucinates plausible textures or introduces slight deviations from ground-truth geometry — improving perceived realism but worsening LPIPS scores due to these structural mismatches.

5 Conclusion

Overall, this experiment was an interesting forray into video super resolution and learning to use public published models, evaluate them using popular metrics, and comparing these models using metrics. We learned about different types of metrics for VSR that evaluate different parts of models and used them to gain insight into the different ways VSR models are good and bad. The model that performed best was Stable VSR whose more modern backbone and training methods with temporal consistency meant that on PSNR, SSIM, DISTS, LPIPS, and tLP it performed the best.

6 Project Resources

Project Github: https://github.com/dakshshah03/VSR-CSE244c

Google Drive with Dataset:
https://drive.google.com/drive/folders/1TbiJqGn_Ar4hyZRvMnHraqB2NW6_LMk5?usp=drive_link

References

[1] Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems, 2024.

[2] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020.

[3] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[5] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *CoRR*, abs/2106.15282, 2021.

[6] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, 2016.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[8] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2017.

[9] Xiaohui Li, Yihao Liu, Shuo Cao, Ziyan Chen, Shaobin Zhuang, Xiangyu Chen, Yinan He, Yi Wang, and Yu Qiao. Diffvsr: Revealing an effective recipe for taming robust video super-resolution against complex degradations, 2025.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[11] Claudio Rota, Marco Buzzelli, and Joost van de Weijer. Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models, 2024.

[12] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021.

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[14] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020.

[15] Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks, 2019.

[16] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, 2021.

[17] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks, 2018.

[18] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[20] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[21] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022.