Video Super-Resolution Benchmark: Evaluating Spatial Fidelity and Temporal Coherence Tradeoffs

Daksh K. Shah*
Computer Science and Engineering
University of California, Santa Cruz
dakshah@ucsc.edu

Dylan J. Louie*
Computer Science and Engineering
University of California, Santa Cruz
djlouie@ucsc.edu

Abstract

Video Super-Resolution (VSR) aims to enhance the spatial resolution of video frames while maintaining temporal consistency. Maintaining temporal consistency while enhancing resolution is a major challenge, often leading to visual artifacts such as flickering and inconsistent motion in naive frame-by-frame approaches. Our benchmark is based on the REDS [10] dataset, a widely used dataset focused on realistic degradations and real-world videos. This work explores Generative Adversarial Networks (GANs) and diffusion model-based approaches to VSR. We benchmark ESRGAN [17], Real-ESRGAN [16], and StableVSR [12] against a bicubic interpolation baseline to investigate different approaches to VSR. We evaluate 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 present an in-depth analysis of different generative approaches and the role 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.

*Equal contribution.

1.2 Challenges in Video Super Resolution

Video super-resolution faces many of the same challenges as single image super-resolution, which is why both tasks share pixel-wise fidelity metrics such as PSNR and SSIM. However, these metrics alone do not capture human perceptual quality. LPIPS [20] and DISTS [2] address this by measuring perceptual similarity through deep feature embeddings, better correlating with human judgment than pixel-wise comparisons alone.

The central challenge that sets VSR apart from SISR is the temporal domain, measured using metrics such as temporal optical flow (tOF) and temporal LPIPS (tLP). We demonstrate how models that incorporate explicit temporal consistency techniques, such as StableVSR, outperform frame-by-frame SISR models like ESRGAN on these metrics. A common visual artifact is flickering between consecutive frames, caused by inconsistent generation across frames. There is also an inherent compute tradeoff: models must balance per-frame spatial quality against temporal coherence, and VSR is substantially more expensive than SISR — a single test sequence with 400 frames requires 400 times the inference compute of a single image.

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 discriminator 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% fooling rate [**?**].

Although its cascaded nature presents potential benefits for video super-resolution, practical experimentation revealed significant computational bottlenecks. Inference using a widely available implementation proved prohibitively slow, requiring over 24 hours to upscale a single REDS4 sequence. Consequently, this approach was excluded from the primary quantitative benchmarking.

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 auxiliary 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 [12]. 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 progressing rapidly; for example, DiffVSR [9] employs 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-tuning the whole on high-quality video datasets combined with interweaved latent transitions, performs well on temporal consistency metrics despite severely degraded initial videos.

2.4 Model Selection for Benchmarking

While recent advancements like DiffVSR and EDVR offer sophisticated, multi-stage approaches to temporal consistency, and cascaded diffusion models (CDM) push the boundaries of spatial fidelity, our experiments focus on isolating the impact of explicit temporal conditioning versus traditional frame-by-frame generative priors. Therefore, we selected ESRGAN and Real-ESRGAN as representative state-of-the-art GAN baselines to evaluate performance on synthetic versus real-world degradations. We selected StableVSR as our primary diffusion-based candidate to evaluate whether its temporal conditioning module can overcome the inherent flickering artifacts of these established, frame-independent GAN paradigms.

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 temporal consistency metrics. The pixel-wise fidelity metrics we used were peak-signal to noise ratio (PSNR) and structural similarity index (SSIM). The perceptual quality metrics we used were deep image structure and texture similarity (DISTS) with VGGNet [13] and learned perceptual image patch similarity (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 held-out videos used for evaluation of REDS and not seen during training. The four approaches evaluated were bicubic linear interpolation as a baseline, ESRGAN [17], Real-ESRGAN [16], and StableVSR [12]. Quantitative results across all metrics are summarized in Table 1, and a visual comparison of cropped sequences is provided in Table 2. Computational resources for this study were provided by the National Research Platform (NRP) Nautilus cluster, 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. Training proceeds in two stages. First, a PSNR-oriented model is trained using L1 loss with an initial learning rate of 2×10^{-4}, which is annealed by a factor of 2 every 2×10^5 mini-batch updates. Second, the generator model is initialized using this PSNR-oriented model and is subsequently trained using a specialized total loss function comprised of perceptual loss (L_{percep}), pixel-wise L1 loss (L_1), and a custom relativistic adversarial loss (L_G^{RA}). The adversarial loss evaluates the probability that a real image (x_r) is relatively more realistic than a fake image (x_f), defined as:

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

The total generator loss is then calculated using the weighting parameters $\lambda = 5 \times 10^{-3}$ and $\eta = 1 \times 10^{-2}$ as follows: [17]

$$L_G = L_{percep} + \lambda L_G^{RA} + \eta L_1$$

The learning rate is 1×10^{-4} and halved at different iterations. The 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.

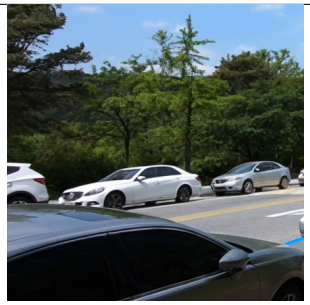
StableVSR is built using the Stable Diffusion x4 upscaler model [11]. This is built using a VAE decoder [3] for super-resolution. The Temporal Conditioning Module uses ControlNet [19] trained at 20k steps and RAFT for optical flow [14]. The Adam optimizer was used with batch size 32 and learning rate 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: SSIM, PSNR, 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	**SSIM (↑)**	**PSNR (↑)**	**DISTS (↓)**	**LPIPS (↓)**	**tLP (↓)**
Baseline	0.7330	25.9946	0.2447	0.1657	0.0973
ESRGAN	0.7383	25.6387	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.

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			

4 Discussion

Our experiments reveal a significant trade-off between spatial detail and temporal stability. While the **Baseline (Bicubic)** achieves the best temporal consistency metric ($tLP = 0.0973$), this is primarily due to its lack of high-frequency detail; because it does not attempt to "hallucinate" textures, there is no variance in generation to cause flickering. Among the generative models, **StableVSR** demonstrated the most effective balance, outperforming the GAN-based models in perceptual metrics ($DISTS, LPIPS$) while maintaining a tLP (0.1316) comparable to ESRGAN, despite producing much more complex textures.

The performance gap between older GAN architectures and recent diffusion-based methods is evident. StableVSR, leveraging a pretrained diffusion backbone and temporal conditioning, effectively

mitigates the flickering artifacts common in frame-by-frame approaches. A major hurdle in this study was the high computational cost of diffusion inference. For instance, **SR3** proved non-viable for our test set, as it required nearly two hours to generate a single frame, highlighting a significant barrier to the real-time adoption of diffusion for VSR tasks.

An unexpected result was the performance of the GAN-based models on pixel-wise fidelity metrics. Both ESRGAN and Real-ESRGAN achieved lower PSNR scores than the naive Bicubic baseline. This is a well-documented phenomenon in generative super-resolution: GANs hallucinate high-frequency textures that appear perceptually realistic but do not perfectly align with the ground truth at the pixel level, thereby penalizing their PSNR scores. Real-ESRGAN scored the lowest on PSNR (23.6045) because it is optimized for blind, real-world degradations; its outputs are visually cleaner but introduce structural deviations from the synthetic bicubic downsampling used in REDS. StableVSR, however, managed to achieve the highest PSNR (27.3590) while simultaneously dominating the perceptual metrics, indicating a superior ability to generate details that are both perceptually convincing and structurally accurate.

5 Conclusion

This study provided a comparative analysis of GAN and diffusion-based approaches to Video Super-Resolution. We found that standard pixel-wise fidelity metrics like PSNR often penalize generative models for hallucinating realistic textures, as evidenced by ESRGAN and Real-ESRGAN performing below the interpolation baseline on pixel-wise fidelity. However, **StableVSR** demonstrated that a diffusion backbone coupled with temporal conditioning can overcome this tradeoff, achieving the strongest overall performance across all evaluated spatial ($SSIM$, $PSNR$, $DISTS$, $LPIPS$) and temporal (tLP) metrics. Future work should focus on reducing the sampling iterations of diffusion-based VSR to make these high-fidelity outputs computationally accessible for longer sequences.

6 Project Resources

Project Github: `https://github.com/dakshshah03/VSR-CSE244c`
REDS Dataset: `https://seungjunnah.github.io/Datasets/reds.html`

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] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, June 2019.

[11] 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.

[12] 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.

[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.