Degradation Methods for Real-World Video Super Resolution

Patrick Timons ptimons@mit.edu
Jacob McCarran mccarran@mit.edu
Elliot Chen echen02@mit.edu

Abstract

Video Super Resolution (VSR) aims to generate high-resolution video from degraded low-resolution input video. Generative models, particularly diffusion models, have demonstrated strong performance on VSR tasks when trained and benchmarked on synthetic data. However, there exists a significant distributional gap between datasets ubiquitously used in the VSR literature, and low-quality video likely to occur in real-world settings. Minimizing this gap is critical to ensuring the effectiveness of VSR models in real-world applications, such as in self-driving cars or enhancing security footage data. To address this issue, we propose a novel video data augmentation method which makes use of random rotation and intensity scaling to emulate camera shake and artifacts induced by glare. We employ this data augmentation strategy to create the Augmented Video Low Quality Dataset (AVLQ), which we use to benchmark the performance of the Motion Guided Latent Diffusion model (MGLD) against the Recurrent Video Restoration Transformer (RVRT) in more realistic settings. We find that both models have reduced performance on videos with our degradations. However, the MGLD diffusion model outperforms the RVRT transformer model with respect to all quantitative metrics and subjective inspection. Our results suggest that the diffusion model is more robust to real world noise compared to the transformer model and that future VSR models may benefit from training on data with our realistic degradations.

1. Introduction

Video super-resolution [1] aims to generate high-resolution (HR) video from its degraded low-resolution (LR) counterpart. Real world VSR has gained attention from researchers for its high potential value in practical applications such as camera-phone video enhancement and online video streaming. Compared with image super-resolution, VSR incurs increased computational complexity as it must aggregate information temporally across frame sequences rather than over a single frame as videos need to stay temporally consistent.

VSR can be broken down into synthetic and real-world subdomains. In synthetic VSR, LR-HR training pairs are formed using a known degradation approach. Typically, synthetic methods assume overly idealized degradations and use fixed blur and bicubic downsampling. Unfortunately, using these degradations in training leads to poor model generalizability since these convenient abstractions do not fully model real-world degradation processes. For example, viral internet videos often incur unknown amounts of lossy compression, camera shake, motion blur, and intensity artifacts. Unlike synthetic VSR, real-world VSR [4,23] aims to increase the resolution of videos whose degradation process is completely unknown.

Previous attempts to approximate realistic degradation processes have been successful. Wang *et al*. approximate these realistic degradations by applying two rounds of blur, downsampling, noise, and JPEG compression. They train Real-ESRGAN with their data augmentations, achieving qualitative improvements in noise removal [18]. We build off of Wang *et al*. contributions by investigating other realistic degradation processes likely to occur in real-world applications of VSR.

Diffusion models have shown promising results for real-world image and VSR tasks [16]. In particular, latent diffusion models (LDMs) have shown incredible promise with their strong generative priors which ensure that outputs remains consistent with the underlying data distribution [19]. Most VSR research neglects realistic data considerations to focus on novel model architectures. Consequently, benchmarks for real world video are scarce and the performance of real world VSR for LDMs remains underexplored. We believe training on more realistic data is necessary to improve model generalizability. To demonstrate this, we aim to evaluate whether the MGLD model maintains high performance under real world data augmentations like intensity scaling (glare) and random rotation (shake).

The contributions of our report are summarized as follows. First, we introduce a stochastic method for synthesizing a realistic Augmented Video Low-Quality (AVLQ) benchmark dataset with desirable qualities. Second, using our newly created benchmark dataset, we investigate the performance of a pre-trained LDM on our realistic data and

Figure 1. Example MGLD super-resolved results (lower-left triangles) versus low-resolution input (upper-right triangles) on video frames from the Augmented Video Low Quality dataset (AVLQ). The top row is non-augmented LR video. The bottom row has random rotation parameter δ and intensity scaling applied. To create this figure, we upsampled the LR video frames with bicubic interpolation.

compare it with Recurrent Video Restoration Transformer (RVRT), a transformer-based VSR model.

2. Related Work

Real-World Super-Resolution. Most existing VSR techniques are trained on synthetic data with simplistic degradation, such as bicubic resizing or downsampling after Gaussian blur [14, 17]. Additionally, these models are benchmarked on highly-curated synthetic datasets, such as REDS4 [10], UDM10 [21], and SPMCS [15]. These datasets cannot assess the generalizability of VSR methods as they are too small and lack meaningful real-world degradation.

There have been previous efforts to create more realistic benchmark datasets. RealVSR [20] filmed short videos on an iPhone to better capture real-world degradation. Chan *et al*. created the VideoLQ dataset [4], a compilation of 50 videos scraped from YouTube and Flicker, each 100 frames long. These datasets are effective in some cases, but their degradation methods can be labor-intensive and lack generalizability by containing camera-specific degradation.

Model Selection VSR has seen significant advancements powered by novel deep learning architectures [2, 3, 6]. Generative model classes, such as diffusion and transformer-based models, are particularly well-motivated for real-world VSR where degradation processes are impossible to explicitly model. Although these models have enabled the generation of high fidelity synthetic data, they are subject to poor performance at inference time if there are distributional changes between the training and test data. Because generative models have achieved unparalleled performance in other fields, we propose novel data augmentation strategies with the goal of more fully unlocking generative power for real-world VSR.

We benchmark a latent diffusion model, MGLD [19], on our AVLQ dataset and compare performance to an older transformer model. We aim to benchmark the diffusion model's performance on our AVLQ dataset because they have shown greater competence at reconstructing the fine details of HR content compared with other generative modeling methods such as general adversarial networks (GANs) [19]. Additionally, due to the relative novelty of diffusion models, less work has been published with respect to real-world VSR, leaving room for exploration.

3. Methodology

Given a real-world video sequence of N frames, our goal is to synthesize realistic LR-HR testing pairs and characterize the strength of MGLD when applied to real-world VSR problems. Inspired by the recent success of diffusion models to curated VSR datasets [4, 19], we investigate applying these generative priors to real-world testing data.

3.1. Latent Diffusion Model

Diffusion models [5] are a class of generative models which transform Gaussian noise into data samples through

a forward and backward Markovian process. The forward process, which transforms the data distribution into a prior distribution, can be described by:

$$q(z_t \mid z_{t-1}) := \mathcal{N}(z_t, (1-\beta_t)z_{t-1}, \beta_t I) \quad (1)$$

where Gaussian noise is iteratively applied to the sample z_t at time-step $t-1$ to produce z_{t-1}, guided by a variance scheduler defined by $\{\beta_t\}_{t=1}^{T}$ along the diffusion process. The backwards process seeks to reverse the forward process and generate a distribution that resembles the prior z_0. The formulation of the learnable transition kernel is:

$$p_\theta(z_{t-1} \mid z_t) := \mathcal{N}(z_{t-1} \mid \mu_\theta(z_t, \gamma_t), \Sigma_\theta(z_t, \gamma_t)) \quad (2)$$

where μ_θ and Σ_θ are learnable parameters. To guide the backward diffusion in learning the forward process, we minimize the Kullback-Leibler (KL) divergence of the joint distribution of the forward and reverse sequences.

Latent Diffusion Models (LDMs) [12] have significantly increased the efficiency of diffusion models by employing Variational Autoencoders (VAEs) [7] to map input data to a latent space. Because the diffusion model no longer optimizes in the pixel space, LDMs take less computational resources and may be trained on massive datasets. Pre-trained LDMs are equipped with powerful prior knowledge about natural images, which may be utilized by various domains to produce high-quality content. For this reason, we believe LDMs are best suited for tackling the domain of real-world VSR.

We utilize a pre-trained Motion Guided Latent Diffusion Model (MGLD) [19] for experimentation. The MGLD model has been specifically fine-tuned with a temporally-aware sequence decoder built on top of a pre-trained VAE decoder. Its design facilitates interleaved spatial-temporal interactions and restores details with high continuity while incurring minimal computational cost. This model has shown improved qualitative performance over alternative models. Despite this, the model was only tested on a few real-world videos, none of which contained significant amounts of shake or intensity degradation.

3.2. Capture of Real-World Video

A quality real-world VSR dataset must cover a wide range of degradations, content, and resolutions. We have compiled the ALQ dataset consisting of 50 short single-scene videos either filmed on an iPhone 13 or from YouTube. To ensure the diversity in potential degradation, we select videos with different scenes and resolutions. For each video, we extract a sequence that is 100-150 frames long. For this project we only investigate videos with no scene transitions in order to accurately benchmark methods that rely on longer-term propagation. Additionally, to incorporate existing research into our work, we add some of the VideoLQ dataset to our dataset. These videos have downsampling, compression, and noise pre-applied. Thus, when applying our own degradation to these videos, we only apply shake and glare methods.

3.3. Augmented Video Low Quality Dataset

We propose a stochastic degradation approach. Our main focus is adding two novel degradation types: shake and intensity scaling. We call video that contains these traits *augmented*. We believe these to be common traits of real-world video that must be taken into account when synthesizing LR-HR testing pairs. Prior real-world datasets [4, 11, 22] do not address these categories of degradation.We anticipate that the introduction of shake, which increases temporal inconsistencies, will significantly challenge the performance of VSR models.

Video Shake. Given a video with N frames and a specified maximum rotational angle of δ, our process begins by zooming into each frame to mitigate the appearance of black borders that appear when rotating pixels past the border of the frame. We use a $\delta = 5^\circ$ and then rotate each frame sequentially within the range of $[\delta, \delta]$, completing four full rotations across the entire video. To simulate camera shake, we introduce a random rotation to each frame, varying between $[-2^\circ, 2^\circ]$. The result is a video that appears both dynamically moving and shaky.

Intensity Scaling. To simulate glare, we choose a short interval of frames randomly and scale the pixel values by 2, clipping pixel values at 255.

Along with our two novel augmentations, we apply more common degradations, such as random blur, resize, and JPEG compression, to every frame of the video. Additionally, we compress the video to reduce its size. The randomized blur is applied using a Gaussian blur kernel with mean zero and uniformly sampled variance. The video is then resized (respecting the original aspect ratio) by a factor of $\frac{1}{\alpha}$ where $\alpha \in [2, 4]$. Video compression is applied by randomly selecting a codex and bitrate.

Fig. 1 displays video frames before and after applying our augmentation methods. The bottom row corresponds to augmented data. In particular, frames 0 and 120 show video shake, while frames 60 and 90 display intensity scaling.

4. Experiments

4.1. Experimental Settings

We apply our degradation methods to the AVLQ dataset and then perform the following three experiments. First, we quantitatively and qualitatively evaluate MGLD's HR reconstructed video to assess output quality and benchmark the model's performance. Second, we compare the MGLD's performance on data with shake and intensity scaling and data without shake and intensity scaling but with

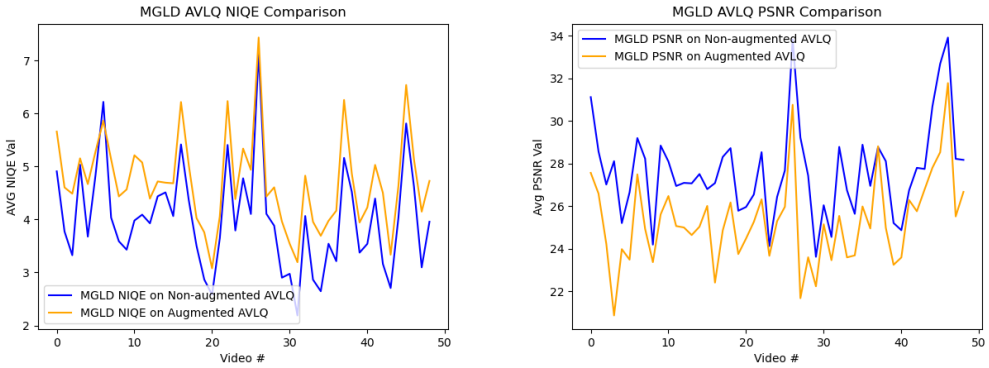

Figure 2. Comparison of augmented and non-augmented data. Augmented data refers to degraded video with shake and intensity scaling applied. Non-augmented data has standard degradations applied (blur, downsampling etc.) but does not have shake and intensity scaling. Note that higher PSNR represents a higher quality reconstruction and that a lower NIQE score reflects higher quality frames.

other common degradations applied. Lastly, we compare the performance of MGLD with that of the Recurrent Video Restoration Transformer (RVRT) [8], another VSR model that has achieved good results when trained on synthetic data.

Evaluation Metrics. To evaluate performance, we use two common metrics: Peak Signal to Noise Ratio (PSNR) and Natural Image Quality Evaluator (NIQE) [9]. Because PSNR and NIQE are image evaluation metrics, to evaluate video we take the average evaluation value between all frames of a HR reconstructed output.

PSNR is one of the most widely used techniques to evaluate image reconstruction quality. It represents the ratio between the maximum pixel value L and the Mean Squared Error (MSE) between the reconstructed HR image y' and the actual image y. The PSNR equation P is given by:

$$P(y, y') = 10 \cdot \log_{10} \left(\frac{L^2}{\frac{1}{N} \sum_{i=1}^{N} [y - y']^2} \right) \quad (3)$$

Despite being popular, PSNR does not accurately match human perception [13]. To remedy this and to make sure the output video has good perceptual quality, we also evaluate model performance using NIQE. NIQE is designed to approximate human perception and is a no-reference metric that only makes use of measurable deviations from statistical regularities observed in natural images. NIQE is commonly used as a blind evaluation metric in the literature. It should be noted that lower NIQE scores are desirable.

We did not focus on temporal evaluation when running our models. However, to the naked eye, MGLD performed very well preserving temporal consistency. Given more time, further research into this area would be desirable.

4.2. Experimental Results

Output Evaluation. We first show the quantitative evaluation results for MGLD's performance on the AVLQ dataset in Fig. 3. As shown from Fig. 3, MGLD's super-resolved output is a significant improvement over the LR input with respect to PSNR and NIQE evaluation scores and is consistent across the entire dataset. Across all 50 AVLQ videos, MGLD produced a mean PSNR score of 25.282 and an average NIQE score of 4.346. Qualitatively, we compare input LR video frames with super-resolved output frames and see that MGLD is capable of removing complex spatial-variant degradation and enhancing visual detail, but noticeably underperforms when the video frame has significant shake. For example, the structure of the edges of the plane are unclear in frame 0, but are quite clear in frame 30 when the video has no shake. Interestingly, MGLD handles intensity scaling quite well. We find that super-resolved output for intensity-scaled frames remain just as clear as their non-scaled counterparts. This is explicitly shown in frame 60 of Fig. 2 .

Shake and Intensity Scaling Comparison. Next, we compare the evaluation metrics between data augmented with shake and intensity scaling and data that only has more common degradations (blur, downsampling, etc.). Fig. 3 shows the PSNR and NIQE scores for augmented versus non-augmented data. We see that MGLD performs significantly worse on augmented data, indicating that the model struggles to generalize when the initial distribution is injected with heavy noise. The non-augmented data had an average PSNR score of 4.005 and an average NIQE score of 27.638. These results indicate that training on larger datasets with distributions more similar to AVLQ are necessary to improve model performance on general real-world data.

RVRT Comparison. Finally, we compare MGLD to

Figure 3. Top Row: Original video with degradation pipeline applied. Second Row: Super-resolved output with RVRT (transformer model). Bottom Row: Super-resolved output with MGLD (diffusion model).

RVRT and show that MGLD outperforms the transformer model by a significant margin. We chose RVRT as it integrates a recurrent neural network structure which helps maintain temporal consistency across video frames. Additionally, its transformer component enables the model to capture complex spatial and temporal correlations within the video data. This dual capability has made RVRT a foundational model within VSR literature. However, because RVRT was trained on simple degradations, it does not generalize well to the real-world AVLQ dataset. Our comparison results are shown in Tabs. 1 and 2. We see that MGLD has more desirable PSNR and NIQE scores across both augmented and non-augmented data. Qualitatively, RVRT fails to enhance finer details within video frames as shown in Fig. 2. Even with augmentations, MGLD clarifies the edges of the plane and smooths the body.

Table 1. Mean and median NIQE values for MGLD and RVRT on augmented (contains shake and intensity scaling) and non-augmented videos.

AVLQ Dataset	**MGLD**		**RVRT**	
	Mean	**Median**	**Mean**	**Median**
Non-augmented	4.005	3.958	5.458	5.257
Augmented	4.346	4.303	6.227	6.134

Table 2. Mean and median PSNR values. Compares same datasets and models as Tab. 1.

AVLQ Dataset	**MGLD**		**RVRT**	
	Mean	**Median**	**Mean**	**Median**
Non-augmented	27.638	27.503	26.336	26.533
Augmented	25.282	25.061	23.265	24.214

5. Conclusion

As our results show that both the MGLD and RVRT perform worse on our AVLQ dataset, which indicates the need for more realistic training data for VSR generative models. Our proposed method for synthesizing real-world data adds novel degradations, namely shake and intensity scaling, that should be an integral part of any real-world VSR dataset. Through our experimentation using our AVLQ dataset, we demonstrate that RVRT, a more foundational VSR model, does not generalize when faced with unknown real-world data. Instead, diffusion models show much more promise to successfully handle video data with unknown degradation. Thus, we believe that diffusion models have more potential with respect to real-world VSR and that future work should focus on assessing diffusion models by training and bench marking them on real world datasets like AVLQ. Improving the performance of diffusion models on such datasets could significantly advance not only VSR but also the broader field of computer vision by enhancing models' ability to generalize to real-world scenarios.

6. Division of Work

The work for this project was equally divided amongst the three members. Each member (Elliot, Patrick, Jacob) was responsible for a single experiment and the creation of a figure. Jacob coded a degradation pipeline that incorporated blur, downsampling, shake, and intensity scaling and was responsible for benchmarking the MGLD model. Pat was responsible for comparing augmented data and non-augmented data and helped source original data for the ALQ dataset. Elliot was responsible for comparing the RVRT model against the MGLD model, capturing real-world data, and running evaluation metric scripts. All members contributed equally by writing this report and creating presentation slides.

References

[1] Christopher M. Bishop, Andrew Blake, and Bhaskara Marthi. Super-resolution enhancement of video. In Christopher M. Bishop and Brendan J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pages 25–32. PMLR, 03–06 Jan 2003. Reissued by PMLR on 01 April 2021. 1

[2] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation, 2017. 2

[3] Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond, 2021. 2

[4] Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution, 2021. 1, 2, 3

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. 2

[6] Younghyun Jo, Seoung Oh, Jaeyeon Kang, and Seon Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. pages 3224–3232, 06 2018. 2

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 3

[8] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention, 2022. 4

[9] A. Mittal, Soundarajan, and Alan Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing letters*, 20:209–212, 01 2012. 4

[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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1996–2005, 2019. 2

[11] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1996–2005, 2019. 3

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 3

[13] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4713–4726, apr 2023. 4

[14] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 2

[15] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution, 2017. 2

[16] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution, 2023. 1

[17] Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks, 2019. 2

[18] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, 2021. 1

[19] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution, 2023. 1, 2, 3

[20] X. Yang, W. Xiang, H. Zeng, and L. Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4761–4770, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. 2

[21] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3106–3115, 2019. 2

[22] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3106–3115, 2019. 3

[23] Huanjing Yue, Zhiming Zhang, and Jingyu Yang. Real-rawvsr: Real-world raw video super-resolution with a benchmark dataset, 2022. 1