

Motion-Guided Latent Diffusion for Temporally Consistent Real-world Video Super-resolution

Xi Yang^{1,2}, Chenhang He¹, Jianqi Ma^{1,2}, and Lei Zhang^{1,2*}

¹ The Hong Kong Polytechnic University, Hong Kong SAR, China

² OPPO Research Institute, Shenzhen, China

{xxxxi.yang, jianqi.ma}@connect.polyu.hk, chenhang.he@polyu.edu.hk,
cslzhang@comp.polyu.edu.hk

Abstract. Real-world low-resolution (LR) videos have diverse and complex degradations, imposing great challenges on video super-resolution (VSR) algorithms to reproduce their high-resolution (HR) counterparts with high quality. Recently, the diffusion models have shown compelling performance in generating realistic details for image restoration tasks. However, the diffusion process has randomness, making it hard to control the contents of restored images. This issue becomes more serious when applying diffusion models to VSR tasks because temporal consistency is crucial to the perceptual quality of videos. In this paper, we propose an effective real-world VSR algorithm by leveraging the strength of pre-trained latent diffusion models. To ensure the content consistency among adjacent frames, we exploit the temporal dynamics in LR videos to guide the diffusion process by optimizing the latent sampling path with a motion-guided loss, ensuring that the generated HR video maintains a coherent and continuous visual flow. To further mitigate the discontinuity of generated details, we insert temporal module to the decoder and fine-tune it with an innovative sequence-oriented loss. The proposed motion-guided latent diffusion (MGLD) based VSR algorithm achieves significantly better perceptual quality than state-of-the-arts on real-world VSR benchmark datasets, validating the effectiveness of the proposed model design and training strategies. Codes and models are available at <https://github.com/IanYeung/MGLD-VSR>.

Keywords: Video super-resolution · Motion-guided latent diffusion

1 Introduction

Video super-resolution (VSR) [3] aims at reconstructing a high-resolution (HR) video from a given low-resolution (LR) video sequence. Thanks to the rapid development of deep learning techniques, significant progress has been made for VSR in the past decade. Representative VSR methods include sliding window based methods like EDVR [49], recurrent based methods like BasicVSR [9] and

* Corresponding author. This work is supported by the Hong Kong RGC RIF grant (R5001-18) and the PolyU-OPPO Joint Innovation Lab.



Fig. 1: The VSR results of 4 consecutive frames in the Sequence 026 from VideoLQ. The top row shows the VSR results by RealBasicVSR [11], the middle row shows the naive VSR results by running StableSR [48] on each frame of the sequence, and the bottom row shows the VSR results produced by our proposed MGLD-VSR method. Our method generates realistic details while achieving good temporal consistency.

more recent transformer based methods [30]. However, most of the above works assume simple degradations (e.g., bicubic resizing, downsampling after Gaussian blur) between the LR and HR videos. As a result, the trained VSR models are hard to be generalized to real-world LR videos, whose degradations are unknown and much more complex.

Recently, real-world VSR [11, 53, 55] has gained attention from researchers for its high potential value in practical applications such as camera-phone video enhancement and online video streaming. Different from traditional VSR tasks, which focus on super-resolving clean LR videos with simple synthetic degradation (e.g., bicubic downsampling), real-world VSR aims to enhance videos with complex and unknown degradations that often appear in real-world low-quality videos. The main challenge lies in how to effectively reproduce video details while suppressing the visual artifacts caused by unknown degradations. Existing real-world VSR algorithms mostly attempt to trade-off between details and artifacts using techniques such as Laplacian pyramid generative adversarial loss [55], input pre-cleaning [11] and hidden state attention [53], etc. Despite these efforts and progresses, their performance on real-world testing sets remains limited due to the insufficient training data and the limited model capacity.

Since the seminal work of denoising diffusion probabilistic model (DDPM) [20], diffusion-based generative models have achieved great success in image gen-

eration. In particular, the latent diffusion models (LDMs) [42] have demonstrated compelling results not only in text-to-image generation, but also in downstream tasks such as image editing [24], inpainting [32], colorization [7], etc. Recently, researchers have also attempted to employ the powerful generative priors of LDMs for real-world image restoration tasks [48] and shown promising results. It is thus interesting to investigate whether the latent diffusion priors can be used to improve real-world VSR results. However, this is a non-trivial task because of the intrinsic randomness in the diffusion process. This randomness in turn gives rise to the temporal inconsistencies among video frames in the latent space. Such inconsistencies can be further amplified by the decoder of the LDM. As shown in the middle row of Figure 1, directly employing the LDM based image super-resolution algorithm [48] for VSR will lead to inconsistent details across consecutive frames, disturbing the visual perception of the reproduced video.

To address the above-mentioned challenges, we propose a Motion-Guided Latent Diffusion (MGLD) model for real-world VSR in this paper, aiming to produce high-quality HR video sequence with good temporal consistency. We propose to incorporate the motion dynamics of input LR video into the the generation of SR outputs. This can be done through a guided diffusion process where we can generate conditional samples from a guidance model based on its score function [14, 44]. In particular, we calculate the optical flows among adjacent LR frames, and use the calculated optical flows to warp the latent features of each frame to align the features of neighboring frames. The l_1 norm of the warping error is employed as the motion-guided loss, and its gradients can be incorporated in the sampling process to update the latent features.

To overcome the inconsistent details generated by the original decoder of LDM, we insert temporal modules to the decoder and fine-tune it with the ground-truth HR video sequences. We also introduce an innovative sequence-oriented loss to guide the decoder to improve the continuity of the details. As depicted in the bottom row of Figure 1, our proposed MGLD-VSR approach could mitigate the discontinuity of generated details while maintaining natural and visually appealing texture recovery.

The contributions of this paper are summarized as follows. First, we propose a diffusion sampling process based on motion-guided loss, allowing the temporal dynamics of the input frame to be used in generating temporally consistent latent features. Second, we propose a temporal-aware sequence decoder, along with two sequence-oriented losses, to further enhance the continuity of generated videos. Our proposed MGLD-VSR model achieves highly competitive real-world VSR results, exhibiting perceptually much more realistic details with fewer flickering artifacts than existing state-of-the-arts.

2 Related Work

Video Super-Resolution. The recent advances in VSR research largely attribute to the rapid development of deep-learning technologies [6, 9, 22, 30, 49]. Based on the paradigms, existing deep-learning based VSR algorithms could

be roughly classified into two categories. The first category performs VSR in a multiple-input single-output (MISO) fashion. Typical algorithms include DUF [22], EDVR [49], and MuCAN [28], etc. The MISO algorithms are usually sliding-window based, which aggregate information from neighboring frames to generate the super-resolved result of the center frame. The other category of methods adopt a multiple-input multiple-output (MIMO) design, such as BasicVSR [9], BasicVSR++ [10], VRT [30] and RVRT [31]. These methods take multiple LR frames as input and generate their corresponding SR results in one go. Compared with MISO, the amortized running cost of MIMO is usually much lower as it avoids a series of redundant operations. In addition, benefiting from the lower cost, the MIMO methods are able to aggregate information from more frames to improve the performance. We adopt the MIMO design in this work.

Real-world VSR. Most existing VSR methods are trained on synthetic data with simple degradations (e.g., bicubic resizing, downsampling after Gaussian blur). Due to the synthetic-to-real gap, models trained with such data cannot perform well on real-world videos. To tackle this problem, Yang *et al.* [55] collected LR-HR video pairs with iPhone cameras to better model real-world degradations. While being effective in some cases, it is relatively labor-intensive and may not generalize well to videos collected by other devices. Later works usually adopt an easier solution: synthesizing more realistic training data by mixing more types of degradations, such as blur, downsampling, noise and video compression [11, 53]. With the improved synthetic training data, the trained models could achieve better super-resolution performance on real-world videos. However, there is still much distribution gap between the synthetic LR data and real-world LR videos. Instead of working on synthesizing more realistic training data, in this work we investigate how to leverage the strong generative priors of pre-trained latent diffusion models to improve the real-world VSR performance.

Generative Prior for Image Restoration. The highly ill-posed nature of real-world restoration problem brings forward a great challenge on real-world image restoration algorithms. To boost the performance, a series of methods have introduced generative priors into the image restoration process. Early methods usually adopt pre-trained generative adversarial networks (GANs) as the generative prior [8, 19, 34, 46, 50]. Although effective, the priors adopted in these methods are usually tailored for specific tasks (e.g., face image restoration), limiting their capability in handling more general natural image restoration problems. Considering the higher generality of diffusion models than GANs in natural image generation, recently researchers have started to investigate how to leverage generative diffusion priors for image restoration, and demonstrated interesting results [17, 48]. In this work, we investigate the possibility of applying latent diffusion priors to real-world VSR, and develop effective techniques to tackle the flickering issue and improve the temporal consistency of generated videos.

Diffusion for Video Synthesis. Recently, diffusion based models have demonstrated its superiority in the video synthesis [4, 5, 18, 43]. As one of the pioneer works, Make-a-Video [43] leverages a pretrained text-to-image (T2I) model and build a text-to-video (T2V) model by inserting temporal modules into the

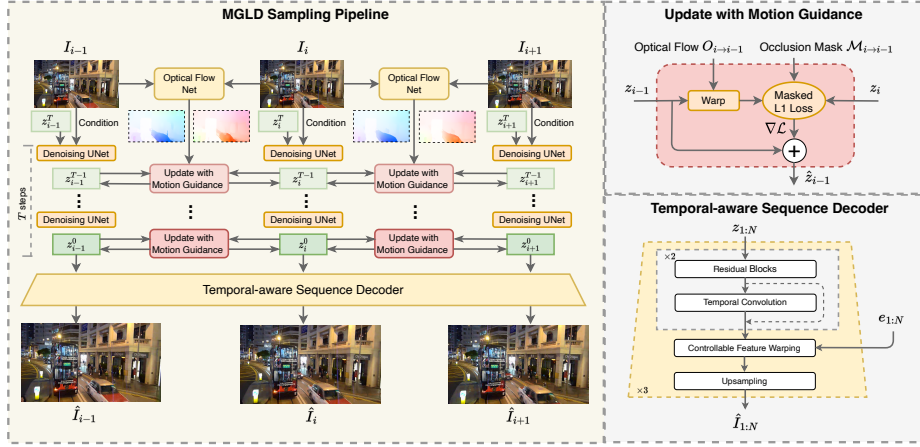


Fig. 2: Overview of the proposed MGLD-VSR framework for real-world VSR. We first estimate the forward and backward optical flow with a pre-trained optical flow estimation network and then employ a motion-guided loss to update the latent diffusion sampling process. The motion-guided loss is computed by summing the forward and backward masked warping error of the latent sequence at each time-step. After a number of T sampling steps, we obtain the generated latent sequence and feed it into the temporal-aware sequence decoder to reconstruct the VSR sequence.

T2I model and designing a cascaded architecture. VideoLDM [5] adopts a latent diffusion paradigm and train the T2V model on large-scale paired text-video datasets, greatly improving the generation results. Stable Video Diffusion [4] further improve the results by conducting training on more carefully constructed datasets and introducing more advanced training paradigm. In this work, we leverage the diffusion prior and build a powerful real-world VSR model, which aims to tackle the ill-posedness of real-world VSR problem with diffusion prior and obtain temporally consistent VSR results with rich details.

3 Methodology

Given an LR video sequence of N frames $I = \{I_1, I_2, \dots, I_N\}$, our goal is to reproduce an HR video sequence $\hat{I} = \{\hat{I}_1, \hat{I}_2, \dots, \hat{I}_N\}$ with enhanced visual quality while maintaining the content consistency across adjacent frames. Inspired by the recently developed diffusion-based image restoration methods [17, 48], we propose to harness the generative priors of the pre-trained SD model [42] to improve the real-world VSR performance. First, we propose a motion-guided diffusion sampling process, where the temporal dynamics from the LR sequence are introduced into the diffusion sampling process to enhance the consistency of latent features across frames. Then, we further improve the quality of generated details by designing a temporal-aware sequence decoder and fine-tuning

it with a video-oriented loss. The overall framework of our algorithm, namely motion-guided latent diffusion (MGLD), is illustrated in Figure 2.

3.1 Latent Diffusion Model

Diffusion models [21] are a class of generative models which transform Gaussian noise into data samples through an iterative reverse Markovian process. This process can be described by: $q(x_t|x_{t-1}) := \mathcal{N}(x_t, (1 - \beta_t)x_{t-1}, \beta_t I)$, where the Gaussian noise is iteratively applied to the sample x_{t-1} at time-step $t - 1$ to produce x_t , guided by a variance scheduler $\{\beta_t\}_{t=1}^T$ along the diffusion process. The key objective of diffusion models is to train a denoising U-Net ϵ_θ parameterized by θ and conditioned on the time-step t and an optional signal y . Given the noise target ϵ_t , the diffusion loss can be defined as: $L(\theta) = \mathbb{E}_{t \sim U(1, T), \epsilon_t \sim \mathcal{N}(0, I)} \|\epsilon_t - \epsilon_\theta(\mathbf{x}_t; t, y)\|_2^2$.

Latent Diffusion Models (LDMs) [42] have made a remarkable impact in the field of image generation. These models employ Variational Autoencoders (VAE) to map the images into a latent space, enabling them to be trained on large-scale text-image datasets. This equips LDMs with powerful prior knowledge about natural images, which can be leveraged by downstream tasks such as image restoration and super-resolution [48] to produce images with clear details and high-quality content. However, in the case of VSR, the stochastic nature of diffusion process often leads to the temporal inconsistencies in the restored video sequence. To solve this issue, we propose a MGLD model that achieves high-quality VSR by utilizing guidance during the sampling process and incorporating temporal awareness in the decoder.

3.2 Motion-guided Diffusion Sampling

From a score-based perspective [44], the sampling process of a diffusion model involves the application of Langevin Dynamics to both gradient of the data distribution $\log p(\mathbf{z})$ and a guidance model $g = G(z)$, which can be a network [15, 38] or an energy model [12]:

$$\nabla_{\mathbf{z}} \log p(\mathbf{z}, g) = \nabla_{\mathbf{z}} \log p(\mathbf{z}) + \nabla_{\mathbf{z}} \log p(g|\mathbf{z}). \quad (1)$$

In this work, we introduce a technique that utilizes the temporal dynamics in low-resolution input videos to create latent features that maintain temporal consistency, thereby enhancing VSR results. Specifically, we incorporate an innovative motion-guided module within the sampling process to measure the warping errors of latent features across frames. Given the LR frames, we first compute their optical flow with $\mathcal{F}$ [47] and downsample the flow map to the dimension of latent features. Given the forward flow O_f and backward flow O_b , we warp the latent features to their neighbouring frames and compute the accumulated warping error along both directions:

$$E(z)^t = \sum_{i=1}^{N-1} \|(\text{Warp}(z_i^t, O_{b,i}) - z_{i+1}^t)\|_1 + \sum_{i=2}^N \|(\text{Warp}(z_i^t, O_{f,i-1}) - z_{i-1}^t)\|_1. \quad (2)$$

We empirically find that the presence of occlusion will pose adverse impact on flow estimation, which in turn causes distraction in the sampling process and leads to artifacts in the final VSR results. To solve this issue, we further estimate the occlusion mask $\mathcal{M}$ in each frame [33] and ignore the contributions in the occluded region. Given the variance σ_t^2 of noise scheduler at step t , the motion-guided sampling process at sampling step t can be written as:

$$\hat{z}^t \leftarrow \text{DDPM}(z^{t+1}, t) - \eta \sigma_t^2 \nabla_z (\mathcal{M} \circ E(z)^t), \quad (3)$$

where the DDPM [21] scheduler is applied for diffusion reverse (denoising) process and $\eta = 10$ is a weighting factor to control the guidance strength. After T iterations, the final denoised latent sequence z^0 is fed into the VAE decoder to obtain the final video sequence.

3.3 Temporal-aware Decoder Fine-tuning

With the proposed motion-guided diffusion sampling, the VSR model could better exploit the temporal dynamics to generate temporally more consistent HR videos. However, since the guidance is performed in the low-resolution latent space, which is 8 times smaller in both height and width than the output image space, the video reconstruction through a VAE decoder may still lead to temporal discontinuities in the generated details of VSR output. Thus, we introduce a temporal-aware sequence decoder and fine-tune it with ground-truth video sequences to improve the smoothness and accuracy of output videos.

As shown in Figure 2, we build our sequence decoder on top of the pre-trained VAE decoder by incorporating 1D convolutions along the temporal dimension. This design facilitates interleaved spatial-temporal interactions and enables our module to restore details with high continuity while incurring minimal computational cost. We further incorporate the information e from the VAE encoder through a Controllable Feature Warping (CFW) module [48, 60] for better restoration and generation results. To leverage the restorative power of the pretrained VAE, we freeze their original spatial blocks and only update the parameters in the temporal convolution and CFW module.

We first employ the ℓ_1 loss and perceptual loss [23, 58] as the reconstruction loss L_{recon} for each frame, then utilize the GAN-loss and frame difference loss that tailored for video sequences. The frame difference loss L_{diff} calculates the difference between consecutive predicted frames $\hat{I}$ and ground truth frames I^{gt} , which is defined as:

$$L_{\text{diff}} = \sum \|(\hat{I}_{i+1} - \hat{I}_i) - (I_{i+1}^{gt} - I_i^{gt})\|_1. \quad (4)$$

It is noteworthy that diffusion models typically generate strong structures. To make the super-resolved video look more natural, we need to apply consistency constraints to those structured regions. We hence introduce a structure weighted consistency loss L_{swc} :

$$\begin{aligned} L_{\text{swc}} = & \sum \|\mathcal{M} \circ (\mathcal{W}_{i+1}(\text{Warp}(\hat{I}_i, O_{b,i}^{gt}) - \hat{I}_{i+1}))\|_1 \\ & + \sum \|\mathcal{M} \circ (\mathcal{W}_{i-1}(\text{Warp}(\hat{I}_i, O_{f,i}^{gt}) - \hat{I}_{i-1}))\|_1. \end{aligned} \quad (5)$$

where O_f^{gt}, O_b^{gt} denote the forward and backward flow computed on the ground-truth sequence. This loss helps the network to generate consistent details by modulating the loss through a weighted map $\mathcal{W} = (1 + wS)$ computed on the ground-truth frames, where $w = 3$ is a weighting factor and S is a structure (edge) map using the Sobel operator. Occlusion masking is also adopted. Given the GAN-based loss L_{GAN} and reconstruction loss L_{recon} , the total loss for decoder fine-tuning is defined as:

$$L_{\text{video}} = L_{\text{recon}} + \alpha L_{\text{diff}} + \beta L_{\text{swc}} + \gamma L_{\text{GAN}}, \quad (6)$$

where α, β and γ are empirically set to 0.5, 0.5 and 0.025.

4 Experiments

4.1 Experimental Settings

Implementation Details. The training of MGLD-VSR consists of two main stages. In the first stage, we fine-tune the denoising U-Net for latent space diffusion. The weight of the denoising U-Net is initialized from Stable Diffusion V2.1. We insert 1D temporal convolution into the U-Net to help temporal dynamic modeling. We then fix the weight in the denoising U-Net of SD model and train the conditioning and temporal modeling modules. For conditioning, we adopt a small time-aware encoder to encode LR condition and inject it into the denoising U-Net with spatial feature transform operation as in [48]. The training batch size, sequence length and latent image size are set to 24, 6 and 64×64 , respectively. In the second stage, we first generate the clean latent sequence with the proposed motion-guided diffusion sampling process, then fine-tune the temporal-aware sequence decoder with the LR sequence, the generated latent sequence and the HR sequence. We fix the VAE decoder and insert temporal modeling layers and CFW modules for training. The training batch size, sequence length and image size are set to 4, 5 and 512×512 , respectively.

Our models are trained on 4 NVIDIA A100 GPUs with the PyTorch [40] framework. Adam [26] is chosen as the optimizer. During inference, due to memory constrains, we divide the LR video into multiple sequences to run. To handle output with arbitrary resolution, we apply a progressive patch aggregation sampling algorithm [48]. For each sequence, we run the sampling for 50 steps.

Training and Testing Datasets. As in previous works [11, 49, 53], we merge the training set and validation set of REDS [37] in training and leave 4 sequences for testing (REDS4). We follow the degradation pipeline of RealBasicVSR [11] to synthesize the training sequence pairs, which involves blur, noise, downsampling and compression degradations. We evaluate our method on both synthetic and real-world datasets. To synthesize LR-HR testing sequence pairs, we apply the RealBasicVSR degradation pipeline on several widely used VSR datasets, including REDS4 [37], UDM10 [56] and SPMCS [45]. REDS4 contains 4 video sequences, each having 100 frames. UDM10 consists of 10 sequences, each

having 32 frames. SPMCS has 30 sequences, each having 31 frames. For real-world dataset, we adopt VideoLQ [11] for testing, which contains 50 real-world sequences with complex degradations.

Evaluation Metrics. On synthetic datasets, we could adopt full-reference metrics to evaluate different real-world VSR methods. The perceptual quality metrics LPIPS [58] and DISTS [16] are selected. Fidelity-oriented metrics PSNR and SSIM are also calculated for reference. Apart from image-based metrics, the video quality assessment metric VMAF [29] is also employed.

On real-world datasets, as ground truths are not available, we compute the no-reference image quality metrics NIQE [36] and BRISQUE [35] for evaluation. In addition, we also include a deep learning based image-based metric MUSIQ [25] and a state-of-the-art video quality assessment metric DOVER [52].

4.2 Experimental Results

To demonstrate the effectiveness of the our MGLD-VSR algorithm, we compare it with seven state-of-the-art methods, including three real-world image super-resolution models (BSRGAN [57], RealESRGAN [51], StableSR [48]), one classic VSR model (RVRT [31]), and three blind or real-world VSR models (DBVSR [39], RealVSR [55], RealBasicVSR [11]).

Quantitative Comparison. We first show the quantitative comparison on the sythetic datasets and real-world video benchmarks in Table 1. As shown from Table 1, our method achieves the best results in term of full-reference perceptual based metrics LPIPS and DISTS on all synthetic test datasets, indicating that the proposed method is able to reconstruct high-quality plausible details from sequences degraded with complex degradations. Although methods like DBVSR may achieve better performance in terms of PSNR or SSIM on some datasets, they tend to generate blurry results, as demonstrated in the LPIPS and DISTS metrics. In terms of VMAF, our model achieves the best or second best results. As for real-world VSR dataset VideoLQ, our methods achieves the best performance in terms of NIQE, BRISQUE, and DOVER and the second best results in terms of MUSIQ, which reflects its strong capability to enhance the real-world videos and generate realistic details and textures.

Qualitative Comparison. To demonstrate the effectiveness of MGLD-VSR, we visualize the VSR results on synthetic datasets in Figure 3 and the results on real-world VideoLQ in Figure 4. One can see that MGLD-VSR can remove complex spatial-variant degradations and generate realistic details, outperforming other state-of-the-art real-world VSR algorithms. On synthetic datasets, as shown in Figure 3, MGLD-VSR can well reconstruct the structure while generating plausible details under complex degradations. On real-world videos, as shown by the example of sequence 008 in Figure 4, MGLD-VSR reconstructs clearer characters and patterns than competing methods. For the example of sequence 013, MGLD-VSR not only reconstructs the structure of the windows, and at the same time generates plausible exterior structure of the brick wall. More visual comparisons can be found in the **supplementary materials**.

Table 1: Quantitative comparison of different VSR models evaluated on popular VSR benchmarking datasets. **Bold** and underline represent the best and second best score, respectively. For synthetic datasets (REDS4, UDM10, SPMCS), we adopt full-reference metrics to evaluate different methods. For real-world video datasets (VideoLQ), we adopt no-reference metrics to compare different methods.

Datasets	Metrics	Bicubic	RVRT	BSRGAN	RealESRGAN	StableSR	DBVSR	RealVSR	RealBasicVSR	Ours
REDS4	PSNR $\uparrow$	22.13	22.29	22.02	21.34	22.04	22.29	21.78	22.35	22.46
	SSIM $\uparrow$	0.5730	0.5759	0.5702	0.5647	0.5683	<u>0.5760</u>	0.5663	0.5846	0.5723
	LPIPS $\downarrow$	0.6277	0.6221	0.4802	0.4509	0.4553	0.6221	0.6267	<u>0.4190</u>	0.3776
	DISTS $\downarrow$	0.3667	0.3554	0.2118	0.1750	0.1845	0.3556	0.3620	<u>0.1581</u>	0.1151
	VMAF $\uparrow$	2.1255	4.8005	26.2413	33.9955	26.6953	2.6075	2.5903	34.9884	<u>34.3214</u>
UDM10	PSNR $\uparrow$	25.78	26.18	25.29	24.94	25.58	<u>26.14</u>	25.09	25.56	25.99
	SSIM $\uparrow$	0.7559	<u>0.7629</u>	0.7323	0.7439	0.7592	0.7635	0.7524	0.7448	0.7548
	LPIPS $\downarrow$	0.4441	0.4413	0.4217	0.3884	<u>0.3684</u>	0.4398	0.4330	0.3897	0.3491
	DISTS $\downarrow$	0.2730	0.2575	0.1926	0.1651	<u>0.1497</u>	0.2585	0.2608	0.1637	0.1369
	VMAF $\uparrow$	5.4647	10.1724	31.0874	36.9158	33.8264	7.2048	6.0079	<u>38.2638</u>	39.3892
SPMCS	PSNR $\uparrow$	22.28	<u>22.51</u>	22.12	21.66	22.26	22.50	22.03	22.41	22.66
	SSIM $\uparrow$	0.5691	0.5753	0.5730	0.5694	<u>0.5844</u>	0.5764	0.5738	0.5736	0.5960
	LPIPS $\downarrow$	0.5572	0.5471	0.4509	0.4435	<u>0.4321</u>	0.5458	0.5487	0.4420	0.3990
	DISTS $\downarrow$	0.3593	0.3415	0.2307	0.2150	0.2351	0.3435	0.3479	<u>0.2130</u>	0.1934
	VMAF $\uparrow$	0.9656	5.3533	22.4828	<u>27.6935</u>	21.8060	3.7576	2.7024	27.4519	28.7032
VideoLQ	NIQE $\downarrow$	9.1319	6.5888	4.2104	4.2036	4.4442	6.7504	6.3331	<u>3.6915</u>	3.5346
	BRISQUE $\downarrow$	72.7405	62.1768	25.2400	29.8428	26.3092	61.1897	57.2765	<u>24.5484</u>	21.9839
	MUSIQ $\uparrow$	21.6789	28.4204	52.2213	49.8380	47.8045	29.0039	24.2189	55.9691	<u>52.7812</u>
	DOVER $\uparrow$	0.3677	0.4907	0.7184	0.7309	0.7108	0.4998	0.5202	<u>0.7401</u>	0.7481

Temporal Consistency. Compared with single image super-resolution, one major aspect of VSR is the temporal consistency of the generated videos. Following most previous works, in this paper, we adopt the average warping error (WE) [13,27] of the sequence to quantitatively measure the temporal consistency: $WE = \frac{1}{N-1} \sum_{i=1}^{N-1} \|\hat{I}_{i+1} - \text{Warp}(\hat{I}_i, O_{b,i})\|_1$, where $\hat{I}_i$ is the predicted frame and $O_{b,i}$ is the estimated backward optical flow [54] from ground-truth video. On real-world video datasets without ground-truth, the optical flow is computed on the predicted frames.

It is worth mentioning that, despite the wide use of WE for temporal consistency evaluation, it may not be able to faithfully reflect the real human perception on the video. For example, WE could be easily attacked by blurring the sequence, resulting much higher WE scores but lower video quality. The WE scores of competing methods are given in Table 2. One can see that our diffusion prior based method MGLD-VSR is not good at this metric because it tends to generate more details and textures, which are not favourable by WE. Nonetheless, the WE of MGLD-VSR is better than StableSR, which is also driven by the latent diffusion priors. To better showcase the effectiveness of our method, we visualize the temporal profiles of the VSR results produced by different algorithms in Figure 5. One can see that MGLD-VSR reconstructs more accurate details while maintaining good temporal consistency.

User Study. We also conduct a user study to further examine the effectiveness of the proposed algorithm in comparison with RealBasicVSR [11] and StableSR [48]. The study is conducted in pairs: given the LR video, the subject

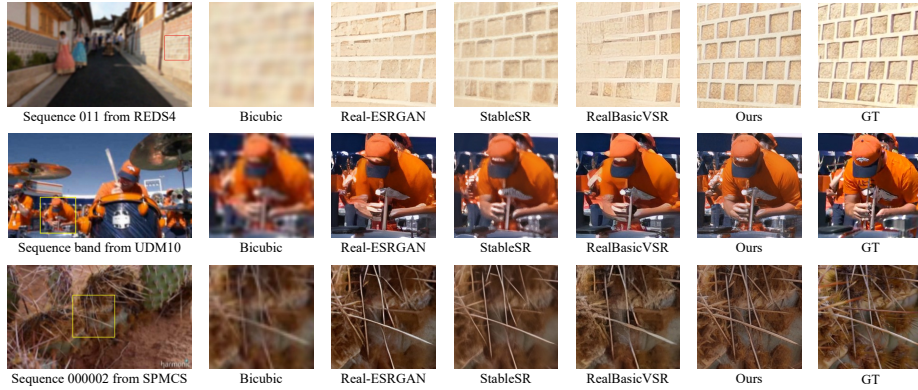


Fig. 3: Qualitative comparison on synthetic datasets (REDS4, UDM10, SPMCS) for $\times 4$ VSR. (Zoom-in for best view.)

Table 2: Temporal consistency comparison of competing VSR methods by using WE $\downarrow$. **Bold** and underline represent the best and second best scores, respectively.

Datasets	Bicubic	RVRT	BSRGAN	RealESRGAN	StableSR	DBVSR	RealVSR	RealBasicVSR	Ours
REDS4	8.57	6.50	10.25	11.91	11.81	<u>8.51</u>	8.61	10.09	10.08
UDM10	3.10	3.06	4.82	4.99	4.80	<u>3.04</u>	2.75	4.33	4.55
SPMCS	<u>3.01</u>	3.58	5.18	7.12	4.62	3.58	4.18	2.72	3.87
VideoLQ	5.48	6.38	7.72	7.64	9.33	5.98	<u>5.74</u>	6.76	7.56

is asked to choose the better one from two HR videos obtained by different methods, considering both the reconstruction quality and the temporal consistency. As shown in Figure 6, our method gains more votes in the competitions.

4.3 Ablation Study

In this section, we conduct a series of analyses on the proposed framework, including its Motion-guided Diffusion Sampling (MDS) module and Temporal-aware Sequence Decoder (TSD) module, and employ the VideoLQ dataset for experiments. We adopt WE to measure temporal consistency and no-reference metrics NIQE, BRISQUE, MUSIQ and DOVER to measure perceptual quality.

Different Components of MGLD-VSR. As can be seen from Table 3, the introduction of MDS improves all the metrics over the baseline (i.e., without using MDS and TSD), suggesting that MDS helps not only in reducing the temporal differences (i.e., WE), but also enhancing the perceptual quality of the reconstructed video sequence.

The incorporation of MDS aids in maintaining perceptual continuity by ensuring that the restored frames could transition smoothly from one to the next. This contributes to more natural and visually pleasing VSR results. The incorporation of TSD also yields better results across all the metrics compared to

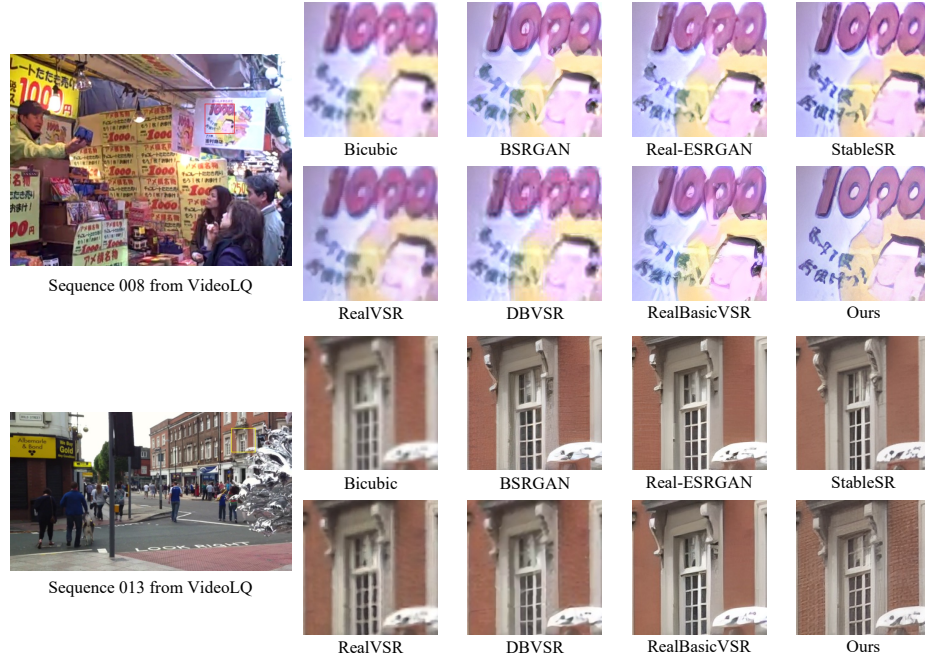


Fig. 4: Qualitative comparison on the real-world video dataset (VideoLQ) for $\times 4$ video SR. (Zoom-in for best view.)

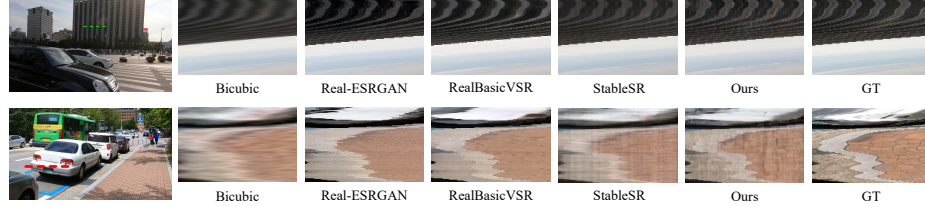


Fig. 5: Temporal profiles of competing real-world VSR methods.

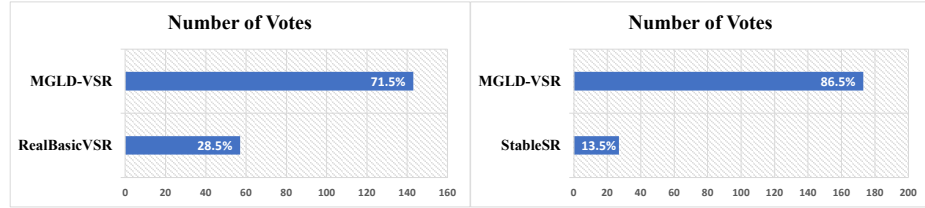


Fig. 6: User study on 20 videos from VideoLQ evaluated by 10 subjects. MGLD-VSR gains more votes than competing methods.

Table 3: Ablation study on the different components of the proposed algorithm. "MDS" refers to Motion-guided Diffusion Sampling and "TSD" refers to Temporal-aware Sequence Decoder.

MDS	TSD	WE ↓	NIQE ↓	BRISQUE ↓	MUSIQ ↑	DOVER ↑
		9.3069	4.1098	25.3249	51.5677	0.7412
✓		8.9848	3.9876	24.0840	52.4143	0.7435
	✓	8.3679	3.6281	22.0806	52.6874	0.7468
✓	✓	7.5478	3.5346	21.9839	52.7812	0.7481

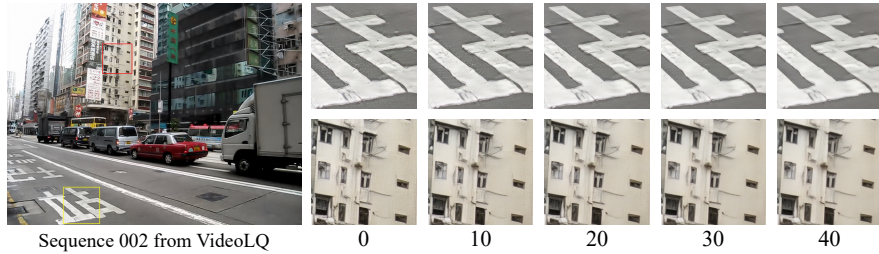


Fig. 7: Impact of guidance strength during sampling. The numbers below the cropped patches indicate the motion guidance weighting factor η .

the baseline. This occurs because TSD is capable of extracting richer contextual information from neighboring frames. When both MDS and TSD are used, the results show further improvements. The full MGLD-VSR approach achieves the best scores on all the metrics, indicating the synergy of the two proposed components for real-world VSR.

Impact of Flow Network. As shown in Table 4, the flow estimation network in motion guided sampling does not show big influence on the results. This might be because the motion guided sampling is applied in low-resolution space.

Impact of Guidance Strength. During guided sampling process, we introduce a weighting factor to control the guidance strength. As shown in Figure 7, a moderate guidance strength could reduce the artifacts and improve the temporal consistency of the generated results. A large guidance strength will blur the results and degrade the reconstructed details. To balance the generation power and temporal consistency, we set the weighting factor $\eta = 10$.

The Design of VAE Decoder. To validate the effectiveness of the proposed architecture and loss design for the VAE decoder, we conduct a series of study and the results are shown in Table 5. As can be seen from the table, the introduction of the adaptive temporal modeling module improves the temporal consistency as well as perceptual quality (measured by no-reference metrics), and incorporating the feature from encoder to decoder during decoder fine-tuning further improves the results. By introducing two sequence based losses during fine-tuning, we improve the temporal consistency of the results while maintaining or improving the perceptual quality.

Table 4: Ablation study on the flow network in motion guided diffusion sampling.

Flow Network	WE ↓	NIQE ↓	MUSIQ ↑	DOVER ↑
SpyNet [41]	7.6734	3.5545	52.6944	0.7478
RAFT [47]	7.5478	3.5346	52.7812	0.7481

Table 5: Ablation study on the different designs of VAE decoder. "TM" refers to the temporal modeling module. "CFW" refers to the controllable feature warping module that injects VAE encoder feature into VAE decoder.

Models	TM	CFW	WE ↓	NIQE ↓	MUSIQ ↑	DOVER ↑	Losses	L_{diff}	L_{swc}	WE ↓	NIQE ↓	MUSIQ ↑	DOVER ↑
(a)			9.8670	4.0671	50.2707	0.7360	(a)			7.9480	3.5359	52.7894	0.7438
(b)		✓	8.6846	3.8451	52.5408	0.7431	(b)		✓	7.7579	3.5409	52.5648	0.7449
(c)	✓		8.9842	3.9236	52.4386	0.7426	(c)	✓		7.7688	3.5376	52.7456	0.7458
(d)	✓	✓	7.5478	3.5346	52.7812	0.7481	(d)	✓	✓	7.5478	3.5346	52.7812	0.7481

4.4 Limitations

Our proposed MGLD-VSR method leverages the diffusion prior from pre-trained SD model, aiming to achieve perceptually realistic and temporally consistent VSR results. Due to the scale of pre-trained SD model, it does not have advantages in terms of number of parameters and inference time compared with those non-diffusion VSR methods (comparisons can be found in the supplementary materials). In addition, since MGLD-VSR will generate more visual details than those non-diffusion VSR methods, it does not have advantages in terms of temporal consistency metrics either. In the future, we can resort to model distillation [2] and more advanced diffusion sampling [1, 59] techniques to improve the efficiency of MGLD-VSR, and resort to the recently developed video synthesis models [4] to improve the temporal consistency.

5 Conclusion

We presented a motion-guided latent diffusion algorithm, namely MGLD-VSR, for high-quality and temporally consistent real-world VSR. Our method exploited the strong generative prior of pre-trained latent diffusion model to restore video sequence with realistic details. To ensure the content consistency among frames, we developed a motion-guided sampling mechanism, where the motion dynamics of input video were incorporated in the diffusion sampling process. To further improve the continuity of generated details, we introduced a temporal module into the decoder, and fine-tuned it by high quality video sequences with delicately designed sequence-based losses. The fine-tuned sequence decoder mitigated the flickering artifacts and further improved the perceptual quality. Our proposed MGLD-VSR achieved state-of-the-art visual results on the widely adopted real-world VSR benchmarks, validating the great potentials of latent diffusion priors in enhancing real-world video quality.

References

1. Lu et al., C.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *NeurIPS* **35**, 5775–5787 (2022)
2. Luo et al., S.: Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378* (2023)
3. Bishop, C.M., Blake, A., Marthi, B.: Super-resolution enhancement of video. In: *AISTATS* (2003)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023)
5. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22563–22575 (2023)
6. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4778–4787 (2017)
7. Carrillo, H., Clément, M., Bugeau, A., Simo-Serra, E.: Diffusart: Enhancing line art colorization with conditional diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3485–3489 (2023)
8. Chan, K.C., Wang, X., Xu, X., Gu, J., Loy, C.C.: Glean: Generative latent bank for large-factor image super-resolution. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14245–14254 (2021)
9. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4947–4956 (2021)
10. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5972–5981 (2022)
11. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5962–5971 (2022)
12. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373* (2023)
13. Chu, M., Xie, Y., Mayer, J., Leal-Taixé, L., Thurey, N.: Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)* **39**(4), 75–1 (2020)
14. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*. vol. 34, pp. 8780–8794. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) *Advances*

- in Neural Information Processing Systems. vol. 34, pp. 8780–8794. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf
16. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence* **44**(5), 2567–2581 (2020)
 17. Fei, B., Lyu, Z., Pan, L., Zhang, J., Yang, W., Luo, T., Zhang, B., Dai, B.: Generative diffusion prior for unified image restoration and enhancement. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9935–9946 (2023)
 18. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 22930–22941 (2023)
 19. Gu, J., Shen, Y., Zhou, B.: Image processing using multi-code gan prior. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3012–3021 (2020)
 20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
 21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) *Advances in Neural Information Processing Systems*. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
 22. Jo, Y., Oh, S.W., Kang, J., Kim, S.J.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3224–3232 (2018)
 23. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. pp. 694–711. Springer (2016)
 24. Kavar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6007–6017 (2023)
 25. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5148–5157 (2021)
 26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
 27. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 170–185 (2018)
 28. Li, W., Tao, X., Guo, T., Qi, L., Lu, J., Jia, J.: Mucan: Multi-correspondence aggregation network for video super-resolution. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. pp. 335–351. Springer (2020)
 29. Li, Z., Bampis, C., Novak, J., Aaron, A., Swanson, K., Moorthy, A., Cock, J.: Vmaf: The journey continues. *Netflix Technology Blog* **25**(1) (2018)

30. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. arXiv preprint arXiv:2201.12288 (2022)
31. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems* **35**, 378–393 (2022)
32. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11461–11471 (2022)
33. Meister, S., Hur, J., Roth, S.: Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 32 (2018)
34. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 2437–2445 (2020)
35. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* **21**(12), 4695–4708 (2012)
36. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012)
37. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 0–0 (2019)
38. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) *Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 162, pp. 16784–16804. PMLR (17–23 Jul 2022), <https://proceedings.mlr.press/v162/nichol122a.html>
39. Pan, J., Bai, H., Dong, J., Zhang, J., Tang, J.: Deep blind video super-resolution. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4811–4820 (2021)
40. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019)
41. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4161–4170 (2017)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
43. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
44. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: *Internat-*

- tional Conference on Learning Representations (2021), <https://openreview.net/forum?id=PXTIG12RRHS>
45. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE international conference on computer vision. pp. 4472–4480 (2017)
 46. Tao Yang, Peiran Ren, X.X., Zhang, L.: Gan prior embedded network for blind face restoration in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
 47. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)
 48. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015 (2023)
 49. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
 50. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9168–9178 (2021)
 51. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: International Conference on Computer Vision Workshops (ICCVW) (2021)
 52. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20144–20154 (2023)
 53. Xie, L., Wang, X., Shi, S., Gu, J., Dong, C., Shan, Y.: Mitigating artifacts in real-world video super-resolution models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2956–2964 (2023)
 54. Xu, H., Yang, J., Cai, J., Zhang, J., Tong, X.: High-resolution optical flow from 1d attention and correlation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10498–10507 (2021)
 55. Yang, X., Xiang, W., Zeng, H., Zhang, L.: Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4781–4790 (2021)
 56. Yi, P., Wang, Z., Jiang, K., Jiang, J., Ma, J.: Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3106–3115 (2019)
 57. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: IEEE International Conference on Computer Vision. pp. 4791–4800 (2021)
 58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
 59. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems* **36** (2024)

60. Zhou, S., Chan, K.C., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. In: NeurIPS (2022)