Enhancing Perceptual Quality in Video Super-Resolution Through Temporally-Consistent Detail Synthesis Using Diffusion Models

Claudio Rota[1](✉), Marco Buzzelli[1], and Joost van de Weijer[2]

[1] University of Milano-Bicocca, Milan, Italy
{claudio.rota,marco.buzzelli}@unimib.it
[2] Universitat Autònoma de Barcelona, Barcelona, Spain
joost@cvc.uab.es

Abstract. In this paper, we address the problem of enhancing perceptual quality in video super-resolution (VSR) using Diffusion Models (DMs) while ensuring temporal consistency among frames. We present StableVSR, a VSR method based on DMs that can significantly enhance the perceptual quality of upscaled videos by synthesizing realistic and temporally-consistent details. We introduce the Temporal Conditioning Module (TCM) into a pre-trained DM for single image super-resolution to turn it into a VSR method. TCM uses the novel Temporal Texture Guidance, which provides it with spatially-aligned and detail-rich texture information synthesized in adjacent frames. This guides the generative process of the current frame toward high-quality and temporally-consistent results. In addition, we introduce the novel Frame-wise Bidirectional Sampling strategy to encourage the use of information from past to future and vice-versa. This strategy improves the perceptual quality of the results and the temporal consistency across frames. We demonstrate the effectiveness of StableVSR in enhancing the perceptual quality of upscaled videos while achieving better temporal consistency compared to existing state-of-the-art methods for VSR. The project page is available at https://github.com/claudiom4sir/StableVSR.

Keywords: Video super-resolution · Perceptual quality · Temporal consistency · Diffusion models

1 Introduction

Video super-resolution (VSR) aims to increase the spatial resolution of a video enhancing its level of detail and clarity. Recently, many VSR methods based on

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73254-6_3.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15070, pp. 36–53, 2025.
https://doi.org/10.1007/978-3-031-73254-6_3

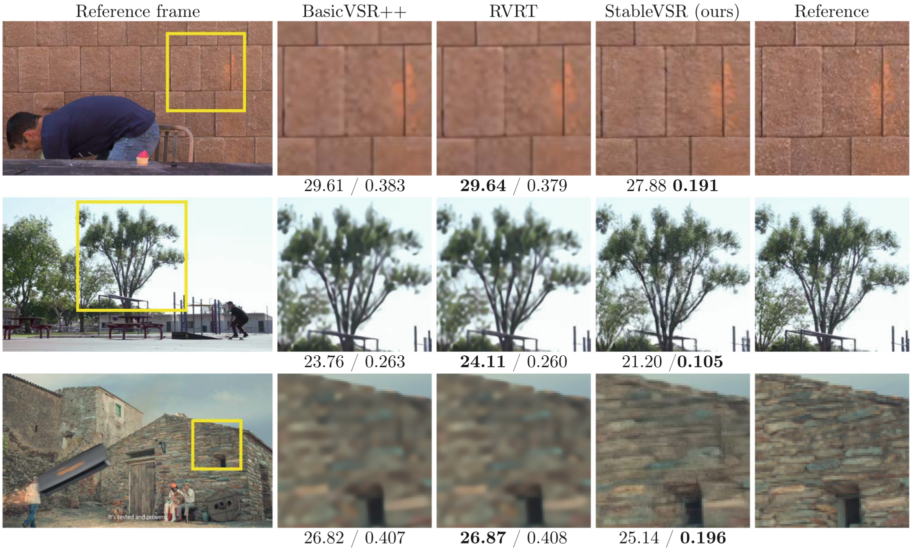


Fig. 1. Reconstruction metrics, such as PSNR, evaluate the pixel-wise difference and do not correlate well with human perception. Perceptual metrics, such as LPIPS, better capture the perceptual quality. Existing methods lack generative capability and focus on reconstruction quality, often producing perceptually unsatisfying results. The proposed StableVSR enhances the perceptual quality by synthesizing realistic details, leading to better visual results. Results reported as PSNR/LPIPS using ×4 upscaling. Best results in bold text. PSNR: the higher, the better. LPIPS: the lower, the better.

deep learning techniques have been proposed [24]. Ideally, a VSR method should generate plausible new contents that are not present in the low-resolution frames. However, existing VSR methods lack generative capability and cannot synthesize realistic details. According to the perception-distortion trade-off, under limited model capacity, improving reconstruction quality inevitably leads to a decrease in perceptual quality [2]. Existing VSR methods mainly focus on reconstruction quality. As a consequence, they often produce perceptually unsatisfying results [19]. As shown in Fig. 1, frames upscaled with recent state-of-the-art VSR methods [4,22] have high reconstruction quality but low perceptual quality, exhibiting blurriness and lack of details [42].

Diffusion Models (DMs) [15] are a class of generative models that transform random noise into images through an iterative refinement process. Inspired by the success of DMs in generating high-quality images [8,15,30,33], several works have been recently proposed to address the problem of single image super-resolution (SISR) using DMs [13,16,20,32,34,39]. They show the effectiveness of DMs in synthesizing realistic textures and details, contributing to enhancing the perceptual quality of the upscaled images [19]. Compared to SISR, VSR requires the integration of information from multiple closely related but misaligned frames to obtain temporal consistency over time. Unfortunately, applying a SISR method to individual video frames may lead to suboptimal results and may introduce temporal inconsistency [31]. Different approaches to encourage temporal consistency in video generation using DMs have been recently studied [1,10,44,46]. However, these methods do not specifically address VSR and do not use fine-texture temporal guidance. As a consequence, they may fail to achieve temporal consistency at fine-detail level, essential in the context of VSR.

In this paper, we address these problems and present *Stable Video Super-Resolution* (StableVSR), a novel method for VSR based on DMs. StableVSR enhances the perceptual quality of upscaled videos by synthesizing realistic and temporally-consistent details. StableVSR exploits a pre-trained DM for SISR [30] to perform VSR by introducing the *Temporal Conditioning Module* (TCM). TCM guides the generative process of the current frame toward the generation of high-quality and temporally-consistent results over time. This is achieved by using the novel *Temporal Texture Guidance*, which provides TCM with spatially-aligned and detail-rich texture information from adjacent frames: at every sampling step t, the predictions of the adjacent frames are projected to their initial state, *i.e.* $t = 0$, and spatially aligned to the current frame. At inference time, StableVSR uses the novel *Frame-wise Bidirectional Sampling strategy* to avoid error accumulation problems and balance information propagation: a sampling step is first taken on all frames before advancing in sampling time, and information is alternately propagated forward and backward in video time.

In summary, our main contributions are the following:

- We present *Stable Video Super-Resolution* (StableVSR): the first work that approaches VSR under a generative paradigm using DMs. It significantly enhances the perceptual quality of upscaled videos while ensuring temporal consistency among frames;
- We design the *Temporal Texture Guidance* containing detail-rich and spatially-aligned texture information synthesized in adjacent frames. It guides the generative process of the current frame toward the generation of detailed and temporally consistent frames;
- We introduce the *Frame-wise Bidirectional Sampling strategy* with forward and backward information propagation. It balances information propagation across frames and alleviates the problem of error accumulation;
- We quantitatively and qualitatively demonstrate that the proposed StableVSR can achieve superior perceptual quality and better temporal consistency compared to existing methods for VSR.

2 Related Work

Video Super-Resolution. Video super-resolution based on deep learning has witnessed considerable advances in the past few years [24]. ToFlow [45] fine-tuned a pre-trained optical flow estimation network with the rest of the framework to achieve more accurate frame alignment. TDAN [36] proposed the use of deformable convolutions [50] for spatial alignment as an alternative to optical flow computation. EDVR [40] extended the alignment module proposed in

TDAN [36] to better handle large motion and used temporal attention [37] to balance the contribution of each frame. BasicVSR [3] revised the essential components for a VSR method, *i.e.* bidirectional information propagation and spatial feature alignment, and proposed a simple yet effective solution. BasicVSR++ [4] improved BasicVSR [3] by adding second-order grid propagation and flow-guided deformable alignment. RVRT [22] combined recurrent networks with the attention mechanism [37] to better capture long-range frame dependencies and enable parallel frame predictions. RealBasicVSR [5] proposed to use a pre-cleaning module before applying a variant of BasicVSR [3], and the use of a discriminator model [41] to improve the perceptual quality of the results.

Diffusion Models for Single Image Super-Resolution. The success of Diffusion Models in image generation [8,15,30,33] inspired the development of single image super-resolution methods based on DMs [13,16,20,32,34,39]. SRDiff [20] and SR3 [34] demonstrated DMs can achieve impressive results in SISR. SR3+ [32] extended SR3 [34] to images in the wild by proposing a higher-order degradation scheme and noise conditioning augmentation. LDM [30] proposed to work in a VAE latent space [11] to reduce complexity requirements and training time. CDM [16] proposed to cascade multiple DMs to achieve SISR at arbitrary scales. IDM [13] proposed to introduce the implicit image function in the decoding part of a DM to achieve continuous super-resolution. StableSR [39] leveraged prior knowledge encapsulated in a pre-trained text-to-image DM to perform SISR avoiding intensive training from scratch.

3 Background on Diffusion Models

Diffusion Models [15] convert a complex data distribution $x_0 \sim p_{data}$ into a simple Gaussian distribution $x_T \sim \mathcal{N}(0, I)$, and then recover data from it. A DM is composed of two processes: the diffusion process and the reverse process.

Diffusion Process. The diffusion process is a Markov chain that corrupts data $x_0 \sim p_{data}$ until they approach Gaussian noise $x_T \sim \mathcal{N}(0, I)$ after T diffusion steps. It is defined as:

$$q(x_1, ..., x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{1}$$

where t represents a diffusion step and $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}(x_{t-1}), \beta_t I)$, with β_t being a fixed or learnable variance schedule. At any step t, x_t can be directly sampled from x_0 as:

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \tag{2}$$

where $\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\epsilon \sim \mathcal{N}(0, I)$.

Reverse Process. The reverse process is a Markov chain that removes noise from $x_T \sim \mathcal{N}(0, I)$ until data $x_0 \sim p_{data}$ are obtained. It is defined as:

$$p_\theta(x_0, ..., x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{3}$$

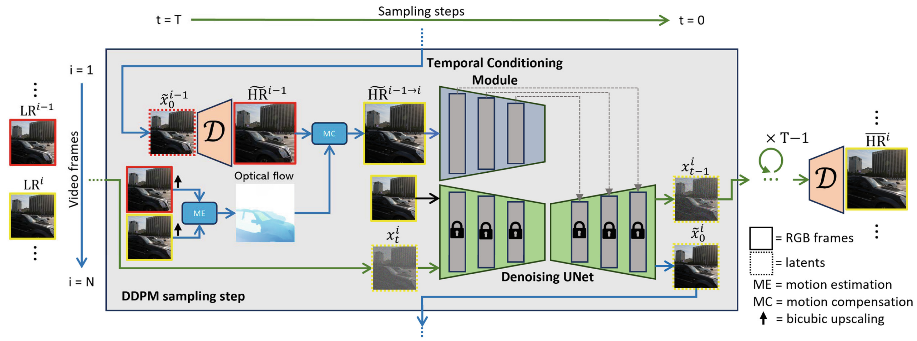


Fig. 2. Overview of the proposed StableVSR. We use the Temporal Conditioning Module (Sect. 4.1) to turn a single image super-resolution LDM (denoising UNet) into a video super-resolution method. TCM exploits the novel Temporal Texture Guidance (Sect. 4.2), which provides TCM with spatially-aligned and detail-rich texture information synthesized in adjacent frames. The sampling step is taken using the novel Frame-wise Bidirectional Sampling strategy (Sect. 4.3). $\mathcal{D}$ represents the VAE decoder. Green lines refer to progression in sampling time, while blue lines refer to progression in video time. (Color figure online)

where $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta I)$. The variance Σ_θ can be a learnable parameter [29] or a time-dependent constant [15]. A neural network ϵ_θ is trained to predict ϵ from x_t, and it can be used to estimate $\mu_\theta(x_t, t)$ as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\overline{\alpha}_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(x_t, t) \right). \tag{4}$$

As a consequence, we can sample $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z, \tag{5}$$

where $z \sim \mathcal{N}(0, I)$ and σ_t is the variance schedule. In practice, according to Eq. 2, we can directly predict $\tilde{x}_0$ from x_t via projection to the initial state $t = 0$ as:

$$\tilde{x}_0 = \frac{1}{\sqrt{\overline{\alpha}_t}} \left(x_t - \sqrt{1-\overline{\alpha}_t} \epsilon_\theta(x_t, t) \right). \tag{6}$$

4 Methodology

We present Stable Video Super-Resolution (StableVSR), a method for video super-resolution based on Latent Diffusion Models (LDMs) [30]. StableVSR enhances the perceptual quality in VSR through temporally-consistent detail synthesis. The overview of the method is shown in Fig. 2. Given a sequence of N low-resolution frames $\{\mathrm{LR}\}_{i=1}^{N}$, the goal is to obtain the upscaled sequence

$\{\overline{\mathrm{HR}}\}_{i=1}^{N}$. StableVSR is built upon a pre-trained LDM for single image super-resolution [30], which is turned into a VSR method through the design and the addition of the Temporal Conditioning Module. TCM uses detail and structure information synthesized in adjacent frames to guide the generative process of the current frame. It allows obtaining high-quality and temporally-consistent frames over time. We design the Temporal Texture Guidance to provide TCM with rich texture information about the adjacent frames: at every sampling step, their predictions are projected to their initial state via Eq. 6, converted into RGB frames, and aligned with the current frame via optical flow estimation and motion compensation. We introduce in StableVSR the Frame-wise Bidirectional Sampling strategy, where a sampling step is taken on all frames before advancing in sampling time, and information is alternately propagated forward and backward in video time. This alleviates the problem of error accumulation and balances the information propagation over time. A brief description of the pre-trained LDM for SISR [30] is provided in the supplementary material.

4.1 Temporal Conditioning Module

Applying the SISR LDM [30] to individual video frames introduces temporal inconsistency, as each frame is generated only based on the content of a single low-resolution frame. In addition, this approach does not exploit the content shared among multiple video frames, leading to suboptimal results [31]. We address these problems by introducing the Temporal Conditioning Module into the SISR LDM [30]. The goal is twofold: (1) enabling the use of spatio-temporal information from multiple frames, improving the overall frame quality; (2) enforcing temporal consistency across frames. We use the information generated by the SISR LDM [30] in the adjacent frames to guide the generative process of the current frame. In addition to obtaining temporal consistency, this solution provides additional sources of information to handle very small or occluded objects. TCM injects temporal conditioning into the decoder of the denoising UNet, as proposed in ControlNet [47].

4.2 Temporal Texture Guidance

The Temporal Texture Guidance provides TCM with the texture information synthesized in adjacent frames. The goal is to guide the generative process of the current frame toward the generation of high-quality and temporally-consistent results.

Guidance on $\tilde{\boldsymbol{x}}_{\mathbf{0}}$. Using results of the previous sampling step $\{x_t\}_{i=1}^{N}$ as guidance to predict $\{x_{t-1}\}_{i=1}^{N}$, as proposed in [1,26], may not provide adequate texture information along the whole reverse process. This is because x_t is corrupted by noise until t approaches 0, as shown in Fig. 3. We address this problem by using a noise-free approximation of x_t, *i.e.* $\tilde{x}_0$, to be used as guidance when taking a given sampling step t [12]. This is achieved by projecting x_t to its initial state, *i.e.* $t = 0$, using Eq 6. Since $\tilde{x}_0 \approx x_0$, it contains very little noise. In addition, it provides detail-rich texture information that is gradually refined as t approaches 0, as shown in Fig. 3.

Temporal Conditioning. We need to use information synthesized in adjacent frames to ensure temporal consistency. We achieve this by using $\tilde{x}_0$ obtained from the previous frame, *i.e.* $\tilde{x}_0^{i-1}$, as guidance when generating the current frame. As $\tilde{x}_0^{i-1}$ is computed from x_t^{i-1} using $\epsilon_\theta(x_t^{i-1}, t, \mathrm{LR}^{i-1})$ via Eq. 6, it contains the texture information synthesized in the previous frame at sampling step t.

Spatial Alignment. Spatial alignment is essential to properly aggregate information from multiple frames [3]. The texture information contained in $\tilde{x}_0^{i-1}$ may not be spatially aligned with respect to the current frame due to video motion. We achieve spatial alignment via motion estimation and compensation, computing optical flow on the respective low-resolution frames LR^{i-1} and LR^{i}. Directly applying motion compensation to $\tilde{x}_0^{i-1}$ in the latent space may introduce artifacts, as shown in Fig. 4. We address this problem by converting $\tilde{x}_0^{i-1}$ from the latent space to the pixel domain through the VAE decoder $\mathcal{D}$ [11] and then applying motion compensation.

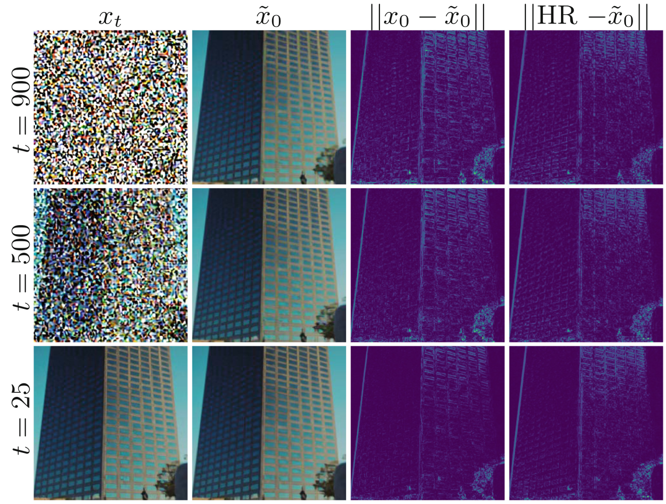


Fig. 3. Comparison between guidance on x_t and $\tilde{x}_0$. Compared to x_t (first column), $\tilde{x}_0$ computed via Eq. 6 contains very little noise regardless of the sampling step t (second column). We can observe $\tilde{x}_0$ is closer to x_0 as t decreases (third column). Here, x_0 corresponds to the last sampling step, *i.e.* when $t = 1$. In addition, $\tilde{x}_0$ increases its level of detail as t decreases (fourth column).

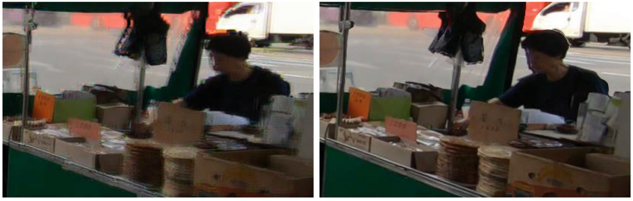
Formulation. Given the previous and the current low-resolution frames LR^{i-1} and LR^{i}, the current sampling step t and the latent of the previous frame x_t^{i-1}, the Temporal Texture Guidance $\widetilde{\mathrm{HR}}^{i-1\to i}$ is computed as:

$$\widetilde{\mathrm{HR}}^{i-1\to i} = \mathrm{MC}(\mathrm{ME}(\mathrm{LR}^{i-1}, \mathrm{LR}^{i}), \mathcal{D}(\tilde{x}_0^{i-1})), \tag{7}$$

where MC is the motion compensation function, ME is the motion estimation method, $\mathcal{D}$ is the VAE decoder [11] and $\tilde{x}_0^{i-1}$ is computed using $\epsilon_\theta(x_t^{i-1}, t, \mathrm{LR}^{i-1})$ via Eq. 6.

4.3 Frame-Wise Bidirectional Sampling Strategy

Progressing all the sampling steps on one frame and using the result as guidance for the next frame in an auto-regressive manner, as proposed in [46], may introduce the problem of error accumulation. In addition, unidirectional information propagation from past to future frames may lead to suboptimal results [3]. We address these problems by proposing the Frame-wise Bidirectional Sampling strategy: we take a given sampling step t on all the frames before taking the next sampling step $t-1$, alternately propagating information forward and backward in video time. The pseudocode is detailed in Algorithm 1. Given the latent x_t^i at a sampling step t, the Temporal Texture Guidance $\widetilde{\mathrm{HR}}^{i-1\to i}$ used by TCM is alternately computed via Eq. 7 using $\tilde{x}_0^{i-1}$ or $\tilde{x}_0^{i+1}$, respectively related to the previous or the next frame. Information is propagated forward and backward in



Motion compensation applied to $\tilde{x}_0$

Motion compensation applied to $\mathcal{D}(\tilde{x}_0)$

Fig. 4. Comparison between applying motion compensation to $\tilde{x}_0$ in the latent space and to $\mathcal{D}(\tilde{x}_0)$ in the pixel domain. $\mathcal{D}$ represents the VAE decoder. In the first scenario, visible artifacts are introduced.

Algorithm 1. Frame-wise Bidirectional Sampling strategy. ME and MC are "motion estimation" and "motion compensation", respectively.

```
Input: Sequence of low-resolution frames $\{\mathrm{LR}\}_{i=1}^N$; pre-trained $\epsilon_\theta$ for VSR, VAE decoder $\mathcal{D}$; method
    for ME.
for $i = 1$ to $N$ do
    $x_T^i = \mathcal{N}(0, I)$
end for
for $t = T$ to $1$ do
    for $i = 1$ to $N$ do                                                    ▷ Take sampling step $t$ on all the frames
        $\widetilde{\mathrm{HR}}^{i-1\to i} = \mathrm{MC}(\mathrm{ME}(\mathrm{LR}^{i-1}, \mathrm{LR}^i), \mathcal{D}(\tilde{x}_0^{i-1}))$ if $i > 1$             ▷ Eq. 7
        $\tilde{\epsilon} = \epsilon_\theta(x_t^i, t, \mathrm{LR}^i, \widetilde{\mathrm{HR}}^{i-1\to i})$ if $i > 1$ else $\epsilon_\theta(x_t^i, t, \mathrm{LR}^i)$
        $z = \mathcal{N}(0, I)$ if $t > 1$ else $0$
        $x_{t-1}^i = \frac{1}{\sqrt{\alpha_t}}\left(x_t^i - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{\epsilon}\right) + \sigma_t z$                          ▷ Eq. 5
        $\tilde{x}_0^i = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t^i - \sqrt{1-\bar{\alpha}_t}\tilde{\epsilon}\right)$                                  ▷ Eq. 6
    end for
    Reverse sequence order of $\{x_{t-1}\}_{i=1}^N$, $\{\tilde{x}_0\}_{i=1}^N$ and $\{\mathrm{LR}\}_{i=1}^N$
end for
return $\{\overline{\mathrm{HR}}\}_{i=1}^N = \{\mathcal{D}(x_0)\}_{i=1}^N$
```

video time: the current frame is conditioned by past frames during forward propagation, and by future frames during backward propagation. Additional details are provided in the supplementary material. The first and the last frames of the sequence do not use TCM during forward and backward propagation, respectively. This is in line with other methods [3,4].

4.4 Training Procedure

StableVSR is built upon a pre-trained LDM for SISR [30], hence we only need to train the Temporal Conditioning Module.

Algorithm 2. Training procedure. ME and MC are "motion estimation" and "motion compensation", respectively.

```
Input: Dataset $D$ with (LR, HR) pairs; pre-trained $\epsilon_\theta$ for SISR, method for ME.
repeat
    $(\mathrm{LR}^{i-1}, \mathrm{HR}^{i-1}), (\mathrm{LR}^{i}, \mathrm{HR}^{i}) \sim D$
    $x_0^{i-1}, x_0^{i} = \mathcal{E}(\mathrm{HR}^{i-1}), \mathcal{E}(\mathrm{HR}^{i})$
    $\epsilon^{i-1}, \epsilon^{i} \sim \mathcal{N}(0, I)$
    $t \sim \{0, ..., T\}$
    $\tilde{\epsilon}^{i-1} = \epsilon_\theta(\sqrt{\overline{\alpha}_t} x_0^{i-1} + \sqrt{1-\overline{\alpha}_t}\epsilon^{i-1}, t, \mathrm{LR}^{i-1})$
    $\tilde{x}_0^{i-1} = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t^{i} - \sqrt{1-\overline{\alpha}_t}\tilde{\epsilon}^{i-1}\right)$    ▷ Eq. 6
    $\widetilde{\mathrm{HR}}^{i-1\to i} = \mathrm{MC}(\mathrm{ME}(\mathrm{LR}^{i-1}, \mathrm{LR}^{i}), \mathcal{D}(\tilde{x}_0^{i-1}))$    ▷ Eq. 7
    Take gradient descent step on:
        $\nabla_\theta(||\epsilon^{i} - \epsilon_\theta(\sqrt{\overline{\alpha}_t} x_0^{i} + \sqrt{1-\overline{\alpha}_t}\epsilon^{i}, t, \mathrm{LR}^{i}, \widetilde{\mathrm{HR}}^{i-1\to i})||)$
until convergence
```

We extend the ControlNet [47] training procedure by adding a step to compute the Temporal Texture Guidance $\widetilde{\mathrm{HR}}^{i-1\to i}$ from the previous frame to be used for the current one. The pseudocode is detailed in Algorithm 2. Given two (LR, HR) pairs of consecutive frames (LR^{i-1}, HR^{i-1}) and (LR^{i}, HR^{i}), we first compute x_0^{i-1} and x_0^{i} by converting HR^{i-1} and HR^{i} into the latent space using the VAE encoder $\mathcal{E}$ [11]. We add $\epsilon \sim \mathcal{N}(0, I)$ to x_0^{i-1} via Eq. 2, obtaining x_t^{i-1}. We then compute $\tilde{x}_0^{i-1}$ using x_t^{i-1} and $\epsilon_\theta(x_t^{i-1}, t, \mathrm{LR}^{i-1})$ via Eq. 6, and we obtain $\widetilde{\mathrm{HR}}^{i-1\to i}$ to be used for the current frame via Eq. 7. The training objective is:

$$\mathbb{E}_{t, x_0^i, \epsilon, \mathrm{LR}^i, \widetilde{\mathrm{HR}}^{i-1\to i}}[||\epsilon - \epsilon_\theta(x_t^i, t, \mathrm{LR}^i, \widetilde{\mathrm{HR}}^{i-1\to i})||_2], \tag{8}$$

where $t \sim [1, T]$ and x_t^i is obtained by adding $\epsilon \sim \mathcal{N}(0, I)$ to x_0^i via Eq. 2.

5 Experiments

5.1 Implementation Details

StableVSR is built upon Stable Diffusion ×4 Upscaler[1] (SD×4Upscaler), which uses the low-resolution images as guidance via concatenation. SD×4Upscaler

[1] https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler.

uses a VAE decoder [11] with ×4 upscaling factor to perform super-resolution. We use the same decoder in our StableVSR. The architecture details are described in the supplementary material. In all our experiments, the results are referred to ×4 super-resolution. We add the Temporal Conditioning Module via ControlNet [47] and train it for 20000 steps. The training procedure is described in Algorithm 2. We use RAFT [35] for optical flow computation. We use 4 NVIDIA Quadro RTX 6000 for our experiments. We use the Adam optimizer [18] with a batch size set to 32 and the learning rate fixed to $1e-5$. Randomly cropped patches of size 256×256 with horizontal flip are used as data augmentation. We use DDPM [15] sampling with $T = 1000$ during training and $T = 50$ during inference.

Table 1. Quantitative comparison with state-of-art methods for VSR. Perceptual metrics are marked with ⋆, reconstruction metrics with ⋄, and temporal consistency metrics with •. Best results in bold text. All the perceptual metrics highlight the proposed StableVSR achieves better perceptual quality. Temporal consistency metrics show that StableVSR achieves better temporal consistency.

VSR method	Vimeo-90K-T						REDS4					
	tLP•↓	tOF•↓	LPIPS⋆↓	DISTS⋆↓	PSNR⋄↑	SSIM⋄↑	tLP•↓	tOF•↓	LPIPS⋆↓	DISTS⋆↓	PSNR⋄↑	SSIM⋄↑
Bicubic	12.47	2.23	0.289	0.209	29.75	0.848	22.72	4.04	0.453	0.186	26.13	0.729
ToFlow	4.96	1.53	0.152	0.150	32.28	0.898	–	–	–	–	–	–
EDVR	–	–	–	–	–	–	9.18	2.85	0.178	0.082	31.02	0.879
TDAN	4.89	1.50	0.120	0.122	34.10	0.919	–	–	–	–	–	–
MuCAN	4.85	1.50	0.097	0.108	35.38	0.934	9.15	2.85	0.185	0.085	30.88	0.875
BasicVSR	4.94	1.54	0.103	0.113	35.18	0.931	9.91	2.87	0.165	0.081	31.39	0.891
BasicVSR++	4.35	1.75	0.092	0.105	35.69	0.937	9.02	2.75	0.131	0.068	32.38	0.907
RVRT	4.28	1.42	0.088	0.101	**36.30**	**0.942**	8.97	2.72	0.128	0.067	**32.74**	**0.911**
RealBasicVSR	–	–	–	–	–	–	6.44	4.74	0.134	0.060	27.07	0.778
StableVSR (ours)	**3.89**	**1.37**	**0.070**	**0.087**	31.97	0.877	**5.57**	**2.68**	**0.097**	**0.045**	27.97	0.800

5.2 Datasets and Evaluation Metrics

We adopt two benchmark datasets for the evaluation of the proposed StableVSR: Vimeo-90K [45] and REDS [28]. Vimeo-90K [45] contains 91701 7-frame video sequences at 448 × 256 resolution. It covers a broad range of actions and scenes. Among these sequences, 64612 are used for training and 7824 (called Vimeo-90K-T) for evaluation. REDS [28] is a realistic and dynamic scene dataset containing 300 video sequences. Each sequence has 100 frames at 1280 × 720 resolution. Following previous works [3,4], we use the sequences 000, 011, 015, and 020 (called REDS4) for evaluation and the others for training.

We evaluate perceptual quality using LPIPS [48] and DISTS [9]. The results evaluated with additional perceptual metrics [17,27,38] are reported in the supplementary material. For temporal consistency evaluation, we adopt tLP [7] and tOF [7], using RAFT [35] for optical flow computation. We also report reconstruction metrics like PSNR and SSIM [43] for reference.

5.3 Comparison with State-of-the-Art Methods

We compare StableVSR with other state-of-the-art methods for VSR, including ToFlow [45], EDVR [40], TDAN [36], MuCAN [21], BasicVSR [3], BasicVSR++ [4], RVRT [22], and RealBasicVSR [5]. Note that RealBasicVSR [5] is a generative method based on GANs [14]. The quantitative comparison is reported in Table 1.

Frame Quality Results. As shown in Table 1, StableVSR outperforms the other methods in perceptual quality metrics. This is also confirmed by the qualitative results shown in Fig. 5: the frames upscaled by StableVSR look more natural and realistic. Additional results are reported in the supplementary material. StableVSR and RealBasicVSR [5], due to their generative nature, can synthesize details that cannot be found in the spatio-temporal frame neighborhood. This is because they capture the semantics of the scenes and synthesize miss-

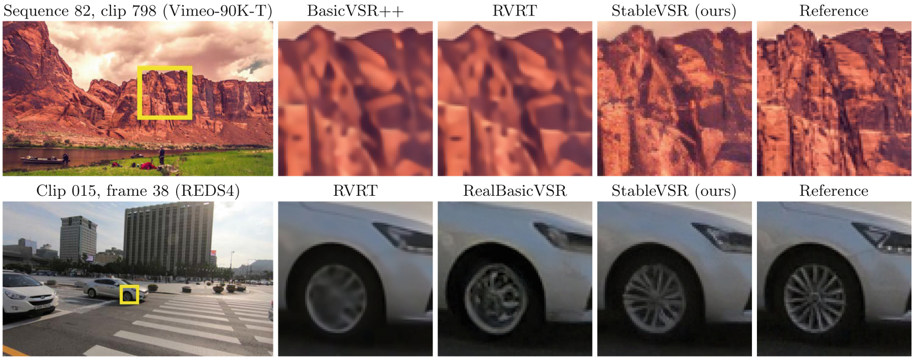


Fig. 5. Qualitative comparison with state-of-the-art methods for VSR. The proposed StableVSR better enhances the perceptual quality of the upscaled frames by synthesizing more realistic details.

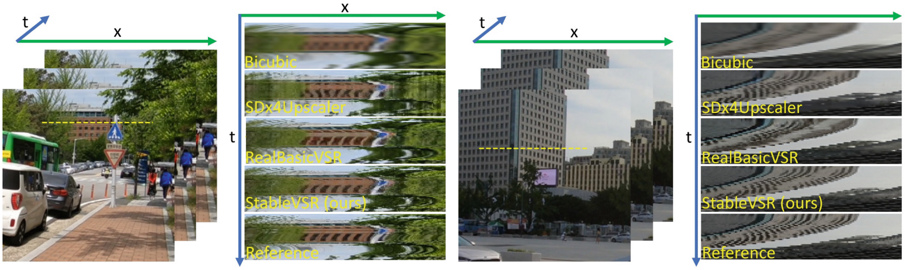


Fig. 6. Comparison of temporal profiles. We consider a frame row and track the changes over time. The temporal profile of StableVSR is more regular than SD×4Upscaler and more similar to the reference profiles than RealBasicVSR, reflecting a better consistency over time. Results on sequences 000 and 015 of REDS4, respectively.

ing information accordingly. Compared to RealBasicVSR [5], StableVSR generates more natural and realistic details, leading to higher perceptual quality. In Table 1, we can observe StableVSR has poorer performance in PSNR and SSIM [43]. This is in line with the perception-distortion trade-off [2]. Nevertheless, StableVSR achieves better reconstruction quality than bicubic upscaling and RealBasicVSR [5].

Temporal Consistency Results. Both temporal consistency metrics in Table 1 show StableVSR achieves more temporally-consistent results. We provide some demo videos as supplementary material to qualitatively assess this aspect. In Fig. 6, we show a comparison among temporal profiles of RealBasicVSR [5], which is the second-best method on REDS4 [28] according to tLP [7] in Table 1, and the proposed StableVSR. We also report the temporal profiles of SD × 4Upscaler, which represents the baseline model used by our method. The temporal profiles of StableVSR are more regular and consistent with the reference profiles compared to the other methods, reflecting better consistency. In Fig. 7, we compare the optical flow computed on consecutive frames obtained from RVRT [22], which represents the second-best method on REDS4 [28] according to tOF [7] in Table 1, and the proposed StableVSR. We also report SD × 4Upscaler and RealBasicVSR [5] results. We can observe the proposed StableVSR allows obtaining an optical flow more similar to the reference flow than the other methods. RealBasicVSR [5] obtains second-best and worst results on REDS4 [28] according to tLP [7] and tOF [7], respectively. Instead, the proposed StableVSR obtains best performance according to both the metrics.

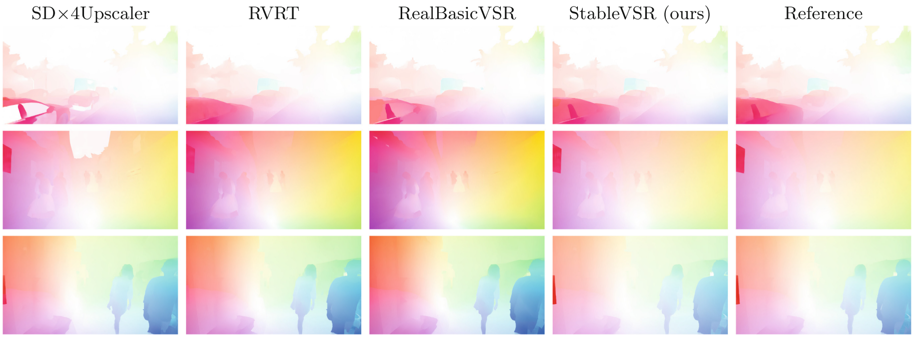


Fig. 7. Comparison of optical flow (visualized) computed using RAFT on different state-of-the-art methods. Note that the hue represents the flow direction, while the saturation represents the flow magnitude. The optical flow computed on StableVSR results is more similar to the reference flow than the other methods. Results on sequences 000, 011 and 020 of REDS4, respectively.

5.4 Ablation Study

Temporal Texture Guidance. We evaluate the effectiveness of the Temporal Texture Guidance design by removing one of the operations involved in its computation. Quantitative and qualitative results are shown in Table 2 (upper part) and Fig. 8a, respectively. Using guidance on x_t instead of $\tilde{x}_0$ leads to very noisy frames. These noisy frames cannot provide adequate information when t is far from 0. With no motion compensation, the spatial information is not aligned with respect to the current frame and cannot be properly used. Applying motion compensation in the latent space introduces distortions in the guidance, as also shown in Fig. 4. In all these cases, temporal consistency at fine-detail level cannot be achieved. The proposed approach provides detail-rich and spatially-aligned texture guidance at every sampling step t, leading to better temporal consistency. Additional results are reported in the supplementary material.

Frame-Wise Bidirectional Sampling Strategy. We compare the Frame-wise Bidirectional Sampling strategy with: single-frame sampling, *i.e.* no temporal conditioning; auto-regressive sampling, *i.e.* the previous upscaled frame is used as guidance for the current one; frame-wise unidirectional sampling, *i.e.* only forward information propagation. The results are quantitatively and qualitatively evaluated in Table 2 (bottom part) and Fig. 8b, respectively. Single-frame sampling leads to poor results and introduces temporal inconsistency due to the differences in the synthesized frame details. The auto-regressive approach has the problem of error accumulation, which is propagated to the next frames. Unidirectional sampling unbalances the information propagation, as only future frames receive information from the past ones, limiting the overall performance. The proposed Frame-wise Bidirectional Sampling solves these problems, leading to better and more consistent results.

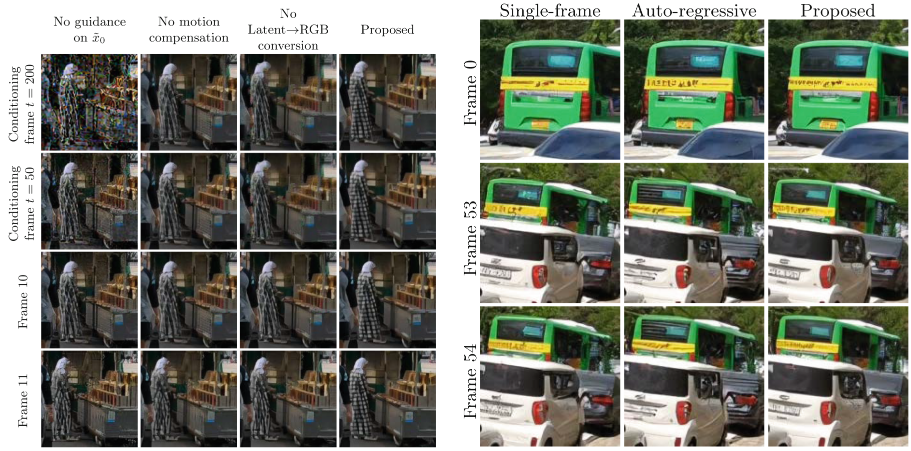
Table 2. Ablation experiments, quantitative results. Perceptual metrics are marked with $\star$, reconstruction metrics with $\diamond$, and temporal consistency metrics with $\bullet$. Best results in bold text. For "No guidance on $\tilde{x}_0$" experiment, we use guidance on x_t. In these experiments, the proposed solution achieves better results in terms of frame quality and temporal consistency. Results computed on center crops of 512 × 512 resolution of REDS4.

Ablated component	Experiment name	tLP$\bullet\downarrow$	tOF$\bullet\downarrow$	LPIPS$\star\downarrow$	DISTS$\star\downarrow$	PSNR$\diamond\uparrow$	SSIM$\diamond\uparrow$
Temporal Texture Guidance	No guidance on $\tilde{x}_0$	38.16	3.34	0.132	0.094	24.74	0.698
	No motion comp.	18.97	3.47	0.116	0.077	25.70	0.749
	No Latent → RGB conv.	21.17	3.32	0.113	0.076	25.78	0.752
	Proposed	**6.16**	**2.84**	**0.095**	**0.067**	**27.14**	**0.799**
Frame-wise Bidirectional Sampling	Single-frame	14.67	3.99	0.121	0.087	25.49	0.729
	Auto-regressive	8.61	3.39	0.120	0.082	25.78	0.745
	Unidirectional	6.36	2.94	0.097	0.069	27.08	0.769
	Proposed	**6.16**	**2.84**	**0.095**	**0.067**	**27.14**	**0.799**

6 Discussion and Limitations

Reconstruction Quality Results. We focus on using DMs to enhance the perceptual quality in VSR. Under limited model capacity, improving perceptual quality inevitably leads to a decrease in reconstruction quality [2]. Recent works on single image super-resolution using DMs [13,20,34] reported lower reconstruction quality when compared to regression-based methods [6,23]. This is related to the high generative capability of DMs, which may generate some patterns that help improve perceptual quality but negatively affect reconstruction quality. Although most VSR methods target reconstruction quality, various studies [24,31] highlight the urgent need to address perceptual quality. We take a step in this direction. We believe improving perceptual or reconstruction quality is a matter of choice: for some application areas like the military, reconstruction error is more important, but for many areas like the film industry, gaming, and online advertising, perceptual quality is key.

Model Complexity. The overall number of model parameters in StableVSR is about ×35 higher than the compared methods, with a consequent increase in inference time and memory requirements. The iterative refinement process of DMs inevitably increases inference time. StableVSR takes about 100 s to upscale a video frame to a 1280 × 720 target resolution on an NVIDIA Quadro RTX 6000



(a) Ablation experiments on Temporal Texture Guidance. Using x_t propagates noise. When motion compensation is not used, fine-detail information cannot be correctly used. Applying motion compensation in the latent space leads to undesired artifacts in the guidance. The proposed guidance solves these problems.

(b) Ablation experiments on Frame-wise Bidirectional Sampling strategy. Single-frame sampling introduces temporal inconsistency. Auto-regressive sampling shows the error accumulation problem. The proposed sampling solves both problems.

Fig. 8. Ablation experiments, qualitative results. For "No guidance on $\tilde{x}_0$" experiment, we use guidance on x_t. For "No Latent→RGB conversion" experiment, the aligned latent is converted to RGB just for visualization.

using 50 sampling steps. In future works, we plan to incorporate current research in speeding up DMs [25,49], which allows reducing the number of sampling steps and decreasing inference time.

7 Conclusion

We proposed StableVSR, a method for VSR based on DMs that enhances the perceptual quality while ensuring temporal consistency through the synthesis of realistic and temporally-consistent details. We introduced the Temporal Conditioning Module into a pre-trained DM for SISR to turn it into a VSR method. TCM uses the Temporal Texture Guidance with spatially-aligned and detail-rich texture information from adjacent frames to guide the generative process of the current frame toward the generation of high-quality results and ensure temporal consistency. At inference time, we introduced the Frame-wise Bidirectional Sampling strategy to better exploit temporal information, further improving perceptual quality and temporal consistency. We showed in a comparison with state-of-the-art methods for VSR that StableVSR better enhances the perceptual quality of upscaled frames while ensuring superior temporal consistency.

Acknowledgments. We acknowledge projects TED2021-132513B-I00 and PID2022-143257NB-I00, financed by MCIN/AEI/10.13039/501100011033 and FSE+ by the European Union NextGenerationEU/PRTR, and the Generalitat de Catalunya CERCA Program. This work was partially supported by the MUR under the grant "Dipartimenti di Eccellenza 2023–2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

References

1. Blattmann, A., et al.: 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)
2. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6228–6237 (2018)
3. 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)
4. 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)
5. 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)
6. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8628–8638 (2021)

7. Chu, M., Xie, Y., Mayer, J., Leal-Taixé, L., Thuerey, N.: Learning temporal coherence via self-supervision for GAN-based video generation. ACM Trans. Graph. **39**(4), 75-1 (2020)
8. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Adv. Neural. Inf. Process. Syst. **34**, 8780–8794 (2021)
9. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: unifying structure and texture similarity. IEEE Trans. Pattern Anal. Mach. Intell. **44**(5), 2567–2581 (2020)
10. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7346–7356 (2023)
11. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12873–12883 (2021)
12. Fei, B., et al.: 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)
13. Gao, S., et al.: Implicit diffusion models for continuous super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10021–10030 (2023)
14. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
16. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**(1), 2249–2281 (2022)
17. 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)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
20. Li, H., et al.: SRDiff: single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
21. Li, W., Tao, X., Guo, T., Qi, L., Lu, J., Jia, J.: MuCAN: multi-correspondence aggregation network for video super-resolution. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12355, pp. 335–351. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58607-2_20
22. Liang, J., et al.: Recurrent video restoration transformer with guided deformable attention. Adv. Neural. Inf. Process. Syst. **35**, 378–393 (2022)
23. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
24. Liu, H., et al.: Video super-resolution based on deep learning: a comprehensive survey. Artif. Intell. Rev. **55**(8), 5981–6035 (2022)
25. Liu, X., Zhang, X., Ma, J., Peng, J., et al.: InstaFlow: one step is enough for high-quality diffusion-based text-to-image generation. In: International Conference on Learning Representations (2023)

26. Luo, Z., et al.: VideoFusion: decomposed diffusion models for high-quality video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10209–10218 (2023)
27. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2012)
28. Nah, S., et al.: 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. 1996–2005 (2019)
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR (2021)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
31. Rota, C., Buzzelli, M., Bianco, S., Schettini, R.: Video restoration based on deep learning: a comprehensive survey. Artif. Intell. Rev. **56**(6), 5317–5364 (2023)
32. Sahak, H., Watson, D., Saharia, C., Fleet, D.: Denoising diffusion probabilistic models for robust image super-resolution in the wild. arXiv preprint arXiv:2302.07864 (2023)
33. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural. Inf. Process. Syst. **35**, 36479–36494 (2022)
34. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4713–4726 (2022)
35. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part II. LNCS, vol. 12347, pp. 402–419. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_24
36. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: TDAN: temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3360–3369 (2020)
37. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
38. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 2555–2563 (2023)
39. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. Int. J. Comput. Vision 1–21 (2024)
40. 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. 1954–1963 (2019)
41. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 1905–1914 (2021)
42. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 606–615 (2018)
43. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)

44. Wu, J.Z., et al.: Tune-a-video: one-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7623–7633 (2023)
45. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. Int. J. Comput. Vision **127**, 1106–1125 (2019)
46. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18456–18466 (2023)
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847 (2023)
48. 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)
49. Zheng, H., Nie, W., Vahdat, A., Azizzadenesheli, K., Anandkumar, A.: Fast sampling of diffusion models via operator learning. In: International Conference on Machine Learning, pp. 42390–42402. PMLR (2023)
50. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable ConvNets v2: more deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9308–9316 (2019)