Learning Spatial Adaptation and Temporal Coherence in Diffusion Models for Video Super-Resolution*

Zhikai Chen[†], Fuchen Long[§], Zhaofan Qiu[§], Ting Yao[§], Wengang Zhou[†], Jiebo Luo[‡], and Tao Mei[§]
[†]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
[‡]University of Rochester, Rochester, NY USA [§]HiDream.ai Inc.
czk654@mail.ustc.edu.cn, {longfuchen, qiuzhaofan, tiyao}@hidream.ai
zhwg@ustc.edu.cn, jluo@cs.rochester.edu, tmei@hidream.ai

Abstract

Diffusion models are just at a tipping point for image super-resolution task. Nevertheless, it is not trivial to capitalize on diffusion models for video super-resolution which necessitates not only the preservation of visual appearance from low-resolution to high-resolution videos, but also the temporal consistency across video frames. In this paper, we propose a novel approach, pursuing Spatial Adaptation and Temporal Coherence (SATeCo), for video super-resolution. SATeCo pivots on learning spatial-temporal guidance from low-resolution videos to calibrate both latent-space high-resolution video denoising and pixel-space video reconstruction. Technically, SATeCo freezes all the parameters of the pre-trained UNet and VAE, and only optimizes two deliberately-designed spatial feature adaptation (SFA) and temporal feature alignment (TFA) modules, in the decoder of UNet and VAE. SFA modulates frame features via adaptively estimating affine parameters for each pixel, guaranteeing pixel-wise guidance for high-resolution frame synthesis. TFA delves into feature interaction within a 3D local window (tubelet) through self-attention, and executes cross-attention between tubelet and its low-resolution counterpart to guide temporal feature alignment. Extensive experiments conducted on the REDS4 and Vid4 datasets demonstrate the effectiveness of our approach.

1. Introduction

In recent years, diffusion models [11, 36, 37, 55] have shown great progress in revolutionizing image generation. In between, a series of image super-resolution works [36, 46, 52] benefit from leveraging knowledge prior embedded in diffusion models to upscale low-resolution (LR) images into high-resolution (HR) ones. Compared to 2D images, videos have one more time dimension, bringing more challenges when capitalizing on diffusion models for video super-resolution (VSR). One natural way is to utilize the pre-trained diffusion models for image super-resolution (ISR), e.g., StableSR [46], to magnify each video frame. The representative advances [46, 52] manifest that diffusion models for ISR could synthesize more details than traditional regression models, e.g., VRT [23]. As depicted in Figure 1, the edges of the windows in the building produced by StableSR are much clearer than those generated by VRT. Nevertheless, the inherent stochasticity of diffusion models might jeopardize the spatial fidelity and hallucinate some extra visual content. Moreover, the independent frame-wise super-resolution overlooks the relation across consecutive frames, resulting in the issue of frame inconsistency in the high-resolution videos. For instance, the traffic signs in Figure 1 are totally different between the two adjacent frames generated by StableSR.

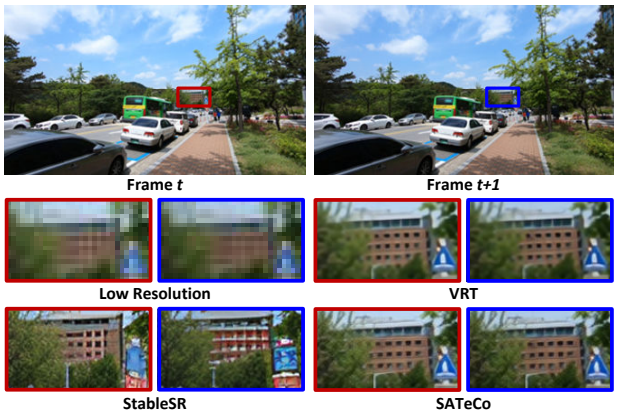

Figure 1. An illustration of video super-resolution by using different approaches of StableSR [46], VRT [23] and our SATeCo to generate two adjacent frames. The region in the same local position is presented in the zoom-in view.

In general, the difficulty of exploring diffusion models for video super-resolution originates from two aspects: 1) how to alleviate the stochasticity in diffusion process to pre-

*This work was performed at HiDream.ai.

serve visual appearance? 2) how to guarantee the temporal consistency across frames in the HR videos? We propose to address the two issues through learning spatial-temporal guidance from low-resolution videos to manage diffusion procedure for video super-resolution. To regulate spatial adaptation, we estimate affine parameters on the LR frame features to modulate each pixel in HR frames. As such, the pixel-wise guidance is employed to nicely learn the feature of every pixel in HR frames and better improve spatial fidelity. In an effort to temporally cohere video frames, we strengthen feature interaction across HR frames, and feature calibration between HR frames and LR counterpart via the attention mechanism. Moreover, large receptive field is attained by conducting the self-attention and cross-attention on the features within a 3D local window (tubelet), thereby facilitating temporal feature alignment.

To materialize our idea, we present a new SATeCo method to carry out Spatial Adaptation and Temporal Coherence for video super-resolution. Technically, SATeCo uses a transformer-based video upscaler to up-sample the input LR video. The VAE encoder then extracts the video features and latent code of LR video, which are further exploited for diffusion calibration. SATeCo deliberately devises spatial feature adaptation (SFA) and temporal feature alignment (TFA) modules, and inserts the two modules into each decoder block of UNet and VAE, for latent-space video denoising and pixel-space video reconstruction. In the regularization of latent-space video denoising, SFA exploits two convolutional layers on the latent code of each up-sampled LR frame, to predict a scale and bias to modulate the pixel-wise feature of HR frame. TFA first executes self-attention on HR video latent code within a tubelet to enhance feature interaction, and further performs cross-attention between the tubelet and its LR counterpart for feature calibration in HR video. The LR video features are exploited in the same way to regulate the HR video feature learning in pixel-space video reconstruction. SATeCo finally refines the decoded HR video by referring to the up-sampled LR video via a neural network to balance synthesized quality and fidelity.

The main contribution of this paper is the proposal of SATeCo to explore spatial adaptation and temporal coherence in diffusion models for video super-resolution. The solution also leads to the elegant views of how to leverage pixel-wise information from LR videos for visual appearance preservation, and how to achieve frame consistency in HR video generation. Extensive experiments on REDS4 and Vid4 verify the superiority of SATeCo in terms of both spatial quality and temporal consistency.

2. Related Work

Video super-resolution. Modern VSR approaches are mainly based on deep neural networks and can be grouped into two categories, i.e., sliding window-based methods and recurrent methods. Early sliding window-based VSR techniques [1, 22, 50, 51, 53] rely on 2D or 3D CNNs [19, 20] which incorporate a sequence of LR frames to predict center HR frame. To fully utilize the complementary information across adjacent frames, the deformable convolutions [43, 48] are employed for feature alignment. Inspired by the success of transformer architecture in various computer vision tasks [6, 27–29], self-attention emerges to be integrated into the VSR frameworks [14, 23, 26, 47]. One representative is VRT [23] which plugs the temporal mutual attention block into transformer backbone to facilitate motion estimation, feature alignment and fusion. Nevertheless, the sliding window-based approaches are difficult to capture long-range dependencies which could limit the performance of video super-resolution. In contrast to aggregate information from adjacent frames in a short term, recurrent approaches [2, 3, 15, 17, 18, 24, 38, 39, 54] utilize a hidden state to sequentially propagate information from all previous frames to the current frame, benefiting the frame restoration. For instance, Chan *et al.* [2] adopt a bidirectional propagation scheme with flow-based feature alignment to maximize information gathering in super resolution. Despite having the great capacity of the recurrent models for temporal information gathering, the local details are still hard to be restored when the LR video encounters significant degradation in a long temporal range.

Diffusion models for super-resolution. Impressive performances of image synthesis achieved by diffusion models [7, 11, 16, 30, 34, 55] encourage the deployment on image super-resolution. These explorations [9, 10, 13, 21, 31, 42, 49, 57] leverage the knowledge prior embedded in the pre-trained diffusion models to magnify images. For example, StableSR [46] integrates a time-aware encoder into Stable-Diffusion [36] model without altering the pre-trained weights, and achieves promising super-resolution results. To further enhance the reconstruction of image texture details, Yang *et al.* [52] introduce an attention-based control module to maintain pixel consistency between LR and HR images. Different from the advances which optimize a small part of inserted parameters, several approaches [13, 21, 49] fix all weights in the pre-trained synthesis model and attempt to incorporate constraints into the reverse diffusion process to guide image restoration. Although the effectiveness of knowledge prior has been manifested in various diffusion-based ISR methods, it is still a grand challenge to employ diffusion models for video super-resolution and preserve spatial fidelity and temporal consistency.

In summary, our work mainly focuses on diffusion models for video super-resolution. The proposal of SATeCo contributes by exploring not only how to preserve spatial fidelity through modulating HR frame features, but also how to calibrate HR video features with LR counterpart for better temporal feature alignment.

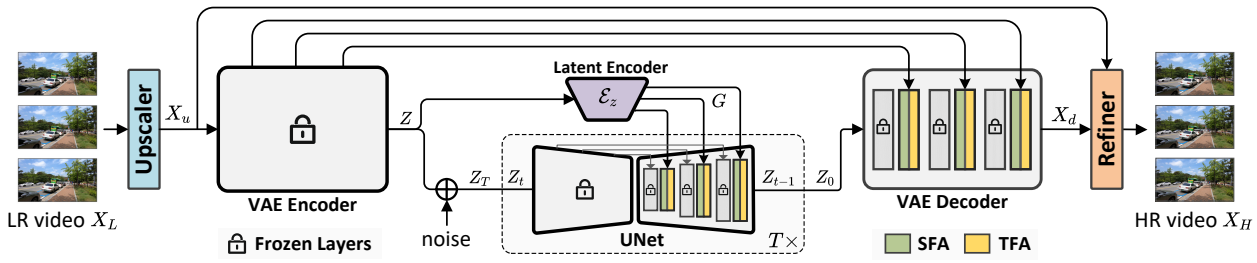

Figure 2. An overview of our SATeCo architecture. The input LR video X_L is first up-sampled to the target resolution via a transformer-based video upscaler. Then, the up-sampled video X_u is fed into the VAE encoder to extract the video features and latent code Z. Next, the Gaussian noise is added into Z according to the diffusion scheduler, and the noisy video latent code is then restored by UNet for quality enhancement. In latent space, a latent encoder extracts the LR latent feature maps G on the LR latent code Z, followed by spatial feature adaptation (SFA) and temporal feature alignment (TFA) modules in each decoder block of UNet for spatial-temporal guidance learning. Given the denoised video latent code Z_0, the VAE decoder decodes the video X_d based on the guidance learnt by SFA and TFA on LR video features. Finally, the decoded video X_d is adjusted by a video refiner via referring to X_u for final HR video X_H synthesis.

3. Our Approach

In this section, we present our newly-minted SATeCo, pursuing Spatial Adaptation and Temporal Coherence in diffusion models for video super-resolution. Figure 2 depicts an overview of the architecture. SATeCo begins with a video upscaler to increase the resolution of the input LR video. Then, the up-sampled video is fed into VAE encoder for video feature extraction and latent code prediction. After that, a spatial feature adaptation (SFA) and a temporal feature alignment (TFA) module are leveraged to learn spatial-temporal guidance on latent code and features of LR video, to calibrate latent-space video denoising and pixel-space video reconstruction. As such, the two modules are plugged into each block of the decoder in UNet and VAE. In procedure of video latent code denoising, SFA estimates the affine parameters on LR video latent code to modulate each pixel of the HR video latent code. TFA first performs self-attention on the HR video latent code within a tubelet, and further enhances latent code by executing cross-attention between the tubelet and its LR counterpart. Similarly, SFA and TFA are conducted in the VAE decoder to guide HR video reconstruction with the LR video features. Finally, SATeCo designs a video refiner to adjust the decoded HR video by referring to the up-sampled video for a good trade-off between synthesized quality and fidelity.

3.1. Video Upscaler

Most existing VSR approaches [39, 51] first upscale the input LR videos through a resampling operation, and then improve their visual quality. Nevertheless, the widely adopted resampling operations, e.g., Bilinear and Bicubic sampling, might damage the original visual patterns [39] in LR frames, having a negative impact on the subsequent video enhancement. Therefore, we exploit the recipe of reducing frame degradation ahead of the feature learning [4] in neural networks and propose a video upscaler, which generates more accurate up-sampled videos for the following quality enhancement by diffusion models.

Given the input LR video X_L, we utilize a transformer-based video upscaler for video up-scaling as illustrated in Figure 3(a). It consists of two cascaded temporal mutual self-attention (TMSA) blocks [23] to temporally aggregate video features, and a pixel-shuffle layer [40] to increase video spatial resolution via feature reshaping. The up-sampled video $X_u = \{x_u^i\}_{i=1}^L$ with L frames is then fed into the diffusion model for video quality enhancement.

3.2. Spatial Feature Adaptation Module

The inherent stochasticity [52] of diffusion models might result in the distortion of texture details in image super-resolution. A natural way of employing diffusion models for super-resolution is to learn the spatial-level condition via convolution-based [46] or transformer-based [52] structure to guide latent code denoising in UNet. Such kind of mechanism only manages feature regularization in latent space, posting difficulty to learn sufficient inductive bias and provide precise guidance for high-resolution image restoration. The similar issue also exists in video super-resolution. To alleviate this, we introduce a spatial feature adaptation (SFA) module which dynamically learns pixel-wise guidance from the input LR videos for diffusion calibration. In the meanwhile, the SFA module emphasizes the inductive bias learning in both of the latent-space video denoising (i.e., training of UNet) and pixel-space video reconstruction (i.e., training of VAE).

Figure 3(c) illustrates our SFA module. Given the up-sampled LR video X_u, the VAE encoder first encodes X_u into the video latent code $Z = \{z^i\}_{i=1}^L$. Next, we exploit a convolution-based latent encoder $\mathcal{E}_z$ to extract the LR latent feature maps $G = \mathcal{E}_z(Z)$, which are further utilized to guide the HR feature learning in UNet decoder. Formally, we denote the HR intermediate feature maps in UNet and the LR latent feature maps in latent encoder as

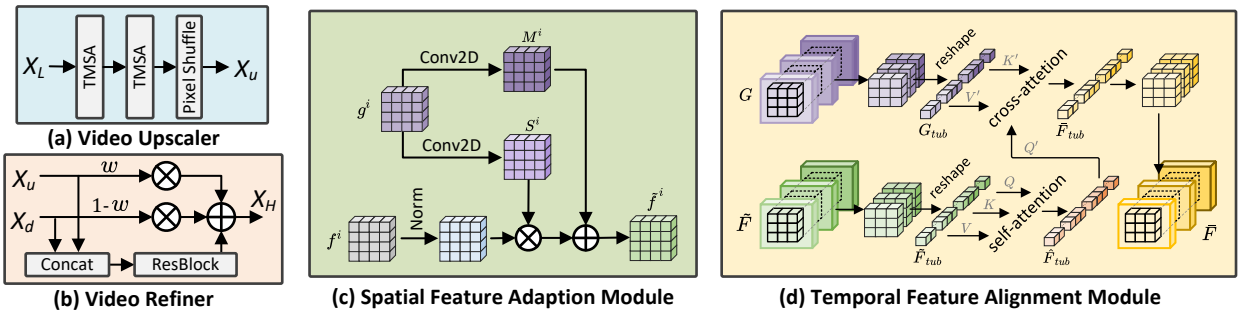

Figure 3. An illustration of (a) video upscaler, (b) video refiner, (c) spatial feature adaptation and (d) temporal feature alignment module.

$F = \{f^i\}_{i=1}^L$ and $G = \{g^i\}_{i=1}^L$, respectively. For the i-th frame, we measure a scale ratio $S^i \in \mathbb{R}^{H\times W\times C}$ and a bias $M^i \in \mathbb{R}^{H\times W\times C}$ for each pixel of HR intermediate feature map $f^i \in \mathbb{R}^{H\times W\times C}$ based on the LR latent feature map $g^i \in \mathbb{R}^{H\times W\times C}$ via two 2D convolution layers:

$$M^i = \text{Conv2D}(g^i), \quad S^i = \text{Conv2D}(g^i). \tag{1}$$

Then, the output HR feature map $\tilde{f}^i$ in UNet is generated by modulating the normalized HR intermediate feature map f^i with S^i and M^i as:

$$\tilde{f}^i = S^i \odot \frac{f^i - \mu^i}{\sigma^i} + M^i, \tag{2}$$

where $\odot$ denotes point-wise multiplication. μ^i and σ^i are the mean and standard deviation values of the feature map f^i. Hence, the affine parameters estimated on the latent feature maps of LR videos calibrate the intermediate feature maps of HR videos in latent code denoising, which adaptively injects the pixel-wise information into the video latent code to preserve the visual appearance. For video feature learning in pixel space, SFA module is inserted into each block of VAE decoder. Similarly, the extracted video features of LR videos are taken as the guidance to estimate the affine parameters in SFA module to adjust HR video feature learning for video reconstruction. We take all the modulated intermediate feature maps $\tilde{f}^i$ from SFA module as $\tilde{F} = \{\tilde{f}^i\}_{i=1}^L$, which is employed for the following temporal feature alignment in the UNet and VAE decoders.

3.3. Temporal Feature Alignment Module

Frame-wisely conducting ISR models for video super-resolution could amplify the differences of blurry patterns [39] across frames, leading to content inconsistency such as the object shape deformation. The issue originated from solely relying on spatial level super-resolution and lacking temporal coherence modeling across frames. To facilitate visual content alignment in video super-resolution, a temporal feature alignment (TFA) module is devised after each SFA module in UNet and VAE decoder, for the temporal feature interaction and calibration.

Figure 3(d) depicts the learning procedure of TFA module. Given the input HR intermediate feature maps $\tilde{F} = \{\tilde{f}^i\}_{i=1}^L$ from the SFA module in UNet decoder, we first partition the feature map $\tilde{f}^i$ of each frame into N non-overlapping windows with the spatial resolution of $h \times w$. $N = \frac{HW}{hw}$ is the total window number. Then, we link all features within a local window across L frames to form a HR feature tubelet $\tilde{F}_{tub} \in \mathbb{R}^{L\times h\times w\times C}$. We reshape the dimension of each HR feature tubelet into $hwL \times C$ and execute the standard self-attention on it:

$$\begin{aligned} Q, K, V &= \text{Conv3D}(\tilde{F}_{tub}), \\ \hat{F}_{tub} &= \text{Attention}(Q, K, V), \end{aligned} \tag{3}$$

where $Q, K, V \in \mathbb{R}^{hwL\times C}$ are the *query*, *key* and *value* matrices, respectively. Each of them is predicted by a 3D convolution layer. The self-attention conducted on the HR feature tubelet enables the feature interaction across different frames, mitigating the temporal feature misalignment in local regions. To further conduct the temporal feature calibration, we leverage the counterpart of HR feature tubelet, i.e., the feature tubelet G_{tub} of the LR latent feature maps, as a reference for feature adjustment. We perform the cross-attention between $\hat{F}_{tub}$ and G_{tub} to obtain the output feature tubelet $\bar{F}_{tub}$:

$$\begin{aligned} Q' &= \text{Conv3D}(\hat{F}_{tub}), \quad K', V' = \text{Conv3D}(G_{tub}), \\ \bar{F}_{tub} &= \text{Attention}(Q', K', V'), \end{aligned} \tag{4}$$

where the query Q' is learnt on the HR feature tubelet $\hat{F}_{tub}$ and the key/value K'/V' is estimated on the LR counterpart via 3D convolution layers, respectively. We collect all the output feature tubelets from the TFA module and reshape them into the original size as $\bar{F} \in \mathbb{R}^{L\times H\times W\times C}$. The output feature $\bar{F}$ is then fed into the next block of the decoder in UNet or VAE for video latent denoising or reconstruction.

In this way, the coupled SFA and TFA modules in UNet and VAE decoder not only emphasize the pixel-wise feature adaptation for visual appearance preservation but also strengthen the temporally feature interaction and calibration for temporal coherence modeling.

Table 1. Performance comparisons in terms of pixel-based (PSNR and SSIM) and perception-based (LPIPS, DISTS, NIQE and CLIP-IQA) evaluation metrics on the REDS4 and Vid4 datasets. The width and height of the LR videos are rescaled by 4 times through different VSR approaches. We follow VRT [23] to set the frame number as 6 in each clip for HR video inference.

Datasets	Metrics	Bicubic	StableSR [46]	TOFlow [51]	EDVR-M [48]	BasicVSR [2]	VRT [23]	IconVSR [2]	SATeCo
REDS4	PSNR↑	26.14	24.79	27.98	30.53	31.42	31.60	**31.67**	31.62
	SSIM↑	0.7292	0.6897	0.7990	0.8699	0.8909	0.8888	**0.8948**	0.8932
	LPIPS↓	0.3519	0.2412	0.3104	0.2312	0.2023	0.2077	0.1939	**0.1735**
	DISTS↓	0.1876	0.0755	0.1468	0.0943	0.0808	0.0823	0.0762	**0.0607**
	NIQE↓	7.257	4.116	6.260	4.544	4.197	4.252	4.117	**4.104**
	CLIP-IQA↑	0.6045	0.6579	0.6176	0.6382	0.6353	0.6379	0.6162	**0.6622**
Vid4	PSNR↑	23.78	22.18	25.89	27.10	27.24	**27.93**	27.39	27.44
	SSIM↑	0.6347	0.5904	0.7651	0.8186	0.8251	**0.8425**	0.8279	0.8420
	LPIPS↓	0.3947	0.3670	0.3386	0.2898	0.2811	0.2723	0.2739	**0.2291**
	DISTS↓	0.2201	0.1385	0.1776	0.1468	0.1442	0.1372	0.1406	**0.1015**
	NIQE↓	7.536	5.237	7.229	5.528	5.340	5.242	5.392	**5.212**
	CLIP-IQA↑	0.6817	**0.7644**	0.7365	0.7380	0.7410	0.7434	0.7411	0.7451

3.4. Video Refiner

Recent advance [8] reveals that images synthesized by diffusion model conditioning on visual contents might loss some original color information in local regions. To address this problem, StableSR [46] performs a non-parametric post-processor to refine the generation with reference to original input for achieving color preservation. Instead, we propose a trainable video refiner to emphasize the adjustment of decoded HR video from VAE decoder, by leveraging the information from up-sampled LR video.

Figure 3(b) details the structure of our video refiner. We first concatenate the decoded video X_d and the up-sampled LR video X_u along channel dimension, and then feed it into a residual block. The refined HR video X_H is generated by fusing X_u, X_d and the output feature mapping of residual block as:

$$X_H = wX_u + (1-w)X_d + ResBlock([X_u, X_d]), \quad (5)$$

where w is a trade-off parameter. The devised video refiner balances the original visual contents of the up-sampled LR video and the synthesized contents of decoded HR video via feature fusion learning. Accordingly, our design is more powerful in terms of color preservation, and achieves a good trade-off between the synthesized quality and fidelity.

3.5. Training Strategy

We construct our SATeCo for video super-resolution based on the Stable Diffusion [36] model. There are four training stages to optimize the whole architecture. In the first stage, we train the video upscaler using the Charbonnier loss [5] to optimize the video reconstruction of HR videos. After that, we follow the standard setting in [36] to train UNet for the optimization of the inserted SFA and TFA modules. We fix all parameters of UNet except for the two kinds of modules during training. For the optimization of SFA and TFA modules in VAE decoder, we take the video latent codes of the HR videos as the input, and optimize the similarity between the decoded videos and ground-truth HR videos. Finally, we freeze all the parameters in video upscaler, UNet and VAE, and train the video refiner using the pairs of decoded and ground-truth HR videos.

4. Experiments

4.1. Experimental Settings

Datasets. We empirically evaluate the effectiveness of our SATeCo on two widely-used datasets: REDS [33] and Vid4 [25]. The **REDS** dataset consists of 240, 30 and 30 video clips for training, validation and testing. Each video clip contains 100 frames with the resolution of $1,280 \times 720$. We employ the standard protocols in [2, 3, 48] and select four video clips from the validation set as the testing data, namely **REDS4**. The **Vid4** dataset also includes four video clips, and there are about 40 frames in each clip with the resolution of 720×480. Following the standard settings [3, 23], we employ all the videos in Vid4 for evaluation and choose the video data in the training set of Vimeo-90K [51] for model optimization. There are $64,612$ training clips and each clip has 7 frames with the resolution of 448×256.

Implementation Details. We implement our SATeCo on the PyTorch platform by using Diffusers [44] library. The noise scheduler is set as linear scheduler ($\beta_1 = 0.00085$, $\beta_T = 0.0120$, and $T = 1,000$). The trade-off parameters w in video refiner is determined as 0.5 by cross validation. We empirically set the window size in TFA as $h = 8$, $w = 8$. The frame number L of input clip is 6. The model is trained with AdamW optimizer and the learning rate is 5.0×10^{-5}.

Evaluation Metrics. We evaluate the VSR models via two kinds of metrics, i.e., pixel-based and perception-based metrics. The pixel-based metrics include PSNR and SSIM which calculate the similarity of every pixel between the generated and ground-truth HR videos. There are also some

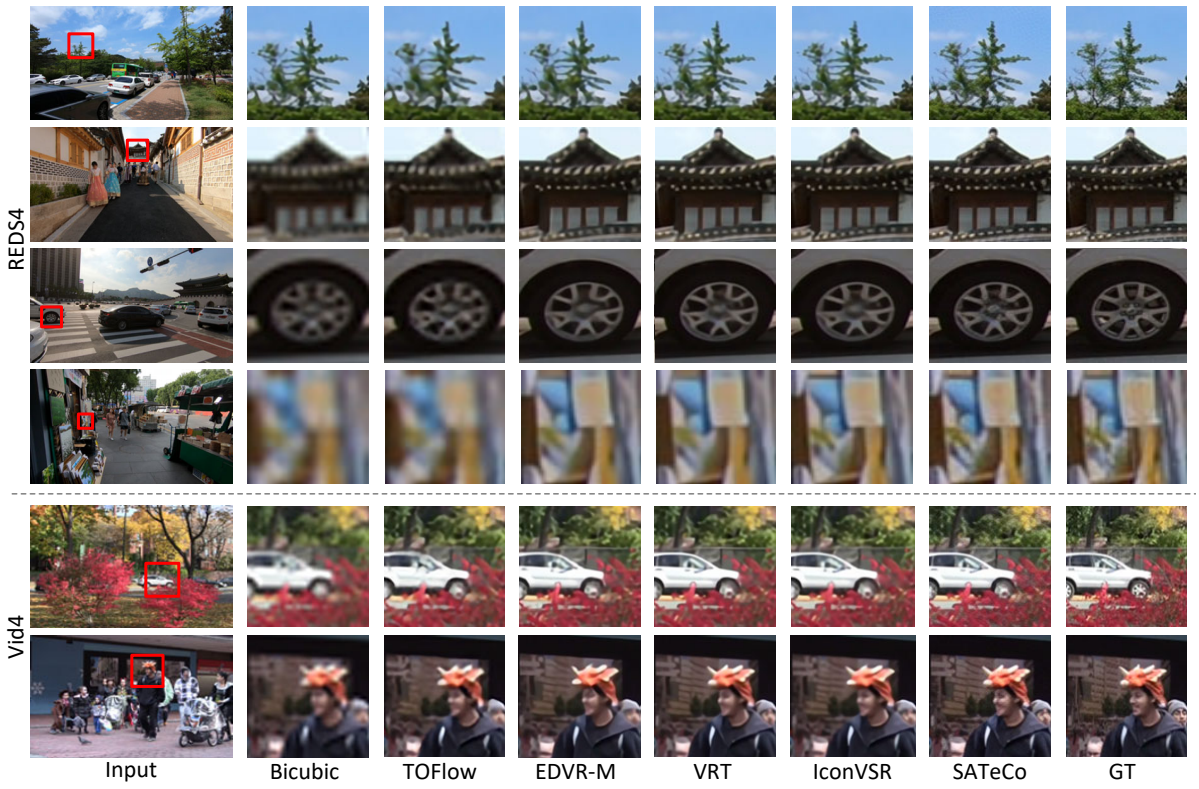

Figure 4. Six visual examples of video super-resolution results by different approaches on the REDS4 and Vid4 datasets. The region in the red box is presented in the zoom-in view for comparison.

perception-based evaluation metrics for super-resolution. These metrics mainly measure video quality from the viewpoint of human perceptual preference, and we adopt LPIPS [56], DISTS [12], NIQE [32] and CLIP-IQA [45] in this paper. Specifically, LPIPS utilizes VGG [41] model to extract frame features and measures the feature similarity between the synthesized and ground-truth videos. DISTS also computes the feature similarity between video pairs via a variant of VGG model, but the emphasis is on image texture. For NIQE and CLIP-IQA, the scores are directly predicted by the learnt models without using the ground-truth HR videos. NIQE measures the similarity of feature distribution between synthesized frames and a realistic image set [32], while CLIP-IQA computes cosine similarity between generated frames and text prompts (e.g., "High Resolution") via CLIP model [35]. In addition, we conduct a user study to verify human preference on different models.

4.2. Comparisons with State-of-the-Art Methods

We compare our SATeCo with several state-of-the-art techniques, including Bicubic Interpolation, StableSR [46], TOFlow [51], EDVR-M [48], BasicVSR [2], VRT [23] and IconVSR [2], on REDS4 and Vid4 datasets.

Quantitative Evaluation. Table 1 summarizes the performances of different VSR approaches in terms of the six metrics over the two datasets. Overall, SATeCo achieves the best performances across all perception-based metrics (i.e., LPIPS, DISTS, NIQE and CLIP-IQA) on REDS4. These metrics emphasize the quality judgment from human perceptual aspect and the results demonstrate the advantage of exploiting abundant knowledge prior in the pre-trained diffusion models to generate high-quality HR videos with better visual perception. In terms of pixel-based metrics, recent advances [46, 52] manifest that the stochasticity in diffusion models could hurt the preservation of visual appearance in HR videos, resulting in inferior performances to traditional regression models. Our SATeCo, by capitalizing on pixel-wise guidance from LR videos to modulate HR frame feature synthesis, alleviates the downsides and obtains the PSNR of 31.62dB. Notably, such performance is very comparable to that of IconVSR [2], which is the SOTA baseline of regression VSR models. The performance trends on Vid4 are similar with those on REDS4. In particular, SATeCo attains the DISTS of 0.1015, which relatively reduces that of the best competitor VRT [23] by 26.0%. The results indicate that SATeCo benefits from learning pixel-wise spatial adaption in diffusion to preserve frame-wise image texture for achieving better video fidelity.

Qualitative Evaluation. Figure 4 visualizes the video super-resolution with six examples from REDS4 and Vid4.

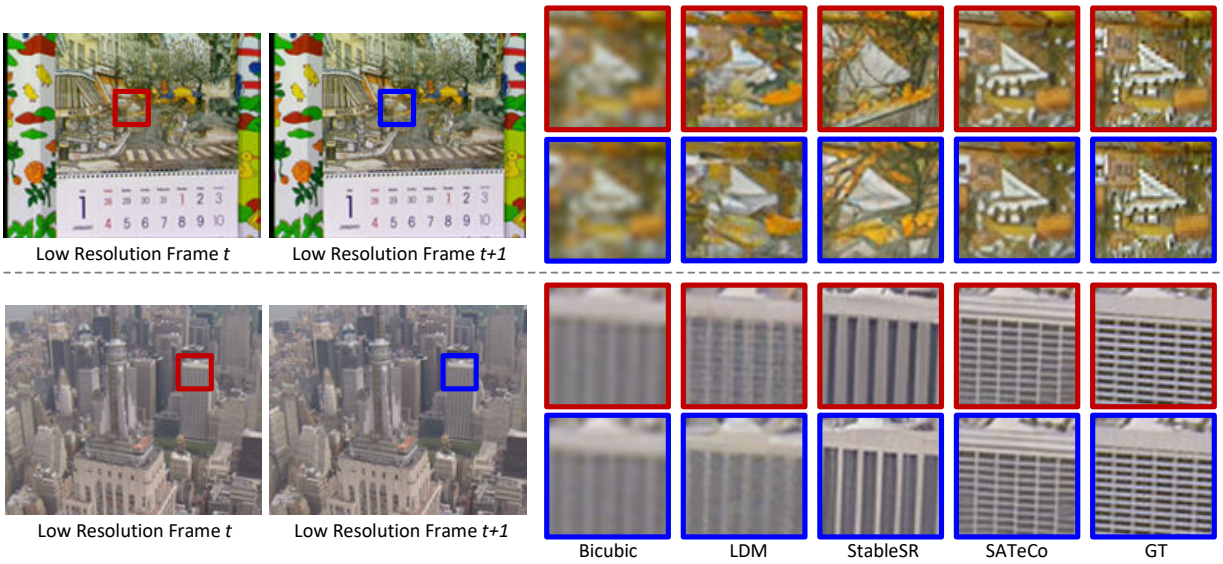

Figure 5. Video super-resolution results of two videos in the Vid4 dataset. The region in the same local position across two adjacent frames (i.e., regions highlighted by red and blue boxes) is scaled up to show more details.

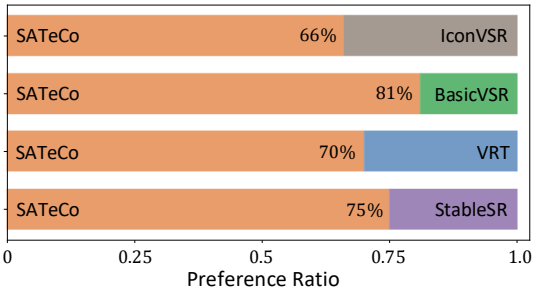

Figure 6. Human evaluation of user preference ratios between SATeCo and other baselines on REDS4 and Vid4.

Compared to other baselines, SATeCo can successfully restore more local details (e.g., the sharp edges in the eave and spoke of the 2nd and 3rd case) in frames with high fidelity. Even with large blurriness (e.g., the 4th case), SATeCo still exhibits the strong restoration ability for video super-resolution, which again confirms the effectiveness of leveraging rich knowledge prior of diffusion models and learning spatial adaptation. To further validate the temporal coherence learnt by SATeCo, we visualize two adjacent frames of two synthesized HR videos by using different diffusion-based super-resolution approaches in Figure 5. As observed in the figure, LDM and StableSR synthesize different visual contents across the two frames, e.g., the small windows in the building. In contrast, our SATeCo predicts the HR videos with higher frame consistency and preserves the visual fidelity. That basically validates the merit of performing tubelet-based self-attention within HR videos and cross-attention between HR videos and LR counterparts to achieve better temporal feature interaction and calibration.

Human Evaluation. Next, we further conduct human study to verify the HR video generation quality by using different VSR approaches with respect to the user preference. We invite 100 evaluators on the Amazon MTurk platform, and ask each evaluator to choose the better one from two synthetic HR videos generated by two different methods given the same LR video. Figure 6 depicts the user preference ratios on all eight videos in the REDS4 and Vid4 datasets. SATeCo clearly wins the traditional regression models of IconVSR, BasicVSR and VRT, and the diffusion model of StableSR. The results indicate SATeCo nicely magnifies LR videos with better visual quality and temporal coherence through the spatial feature adaptation and temporal feature alignment design in video diffusion procedure.

Table 2. Performance comparisons on REDS4 among variants with different integration of SFA and TFA modules.

Model	UNet		VAE		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓
	SFA	TFA	SFA	TFA					
A					28.56	0.7925	0.2159	0.0758	4.404
B	✓				28.93	0.8087	0.2042	0.0693	4.349
C	✓	✓			29.45	0.8398	0.1892	0.0620	4.324
D	✓	✓	✓	✓	**31.62**	**0.8932**	**0.1735**	**0.0607**	**4.104**

4.3. Model Analysis

Analysis on SFA and TFA modules. We first investigate how the SFA and TFA modules influence the overall performances of video super-resolution. Table 2 lists the performance comparisons among variants with different integration ways of SFA and TFA modules. We start from the basic diffusion model **A**, which leverages the zero-initialized convolution [55] in UNet/VAE to learn the spatial guidance from LR videos for super-resolution. The model **B** and **C**

Table 3. Ablation studies on the design of video upscaler and video refiner in SATeCo. The performances are reported on REDS4.

Model		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓
Upscaler	PixelShuffle	29.77	0.8426	0.1979	0.0720	4.298
	Ours	**31.62**	**0.8932**	**0.1735**	**0.0607**	**4.104**
Refiner	$w = 0$	30.36	0.8572	**0.1581**	**0.0339**	**3.457**
	$w = 0.5$	**31.62**	**0.8932**	0.1735	0.0607	4.104
	$w = 1.0$	28.99	0.8001	0.1815	0.0652	4.488

gradually upgrade the basic model A through plugging SFA and TFA modules into the UNet, which improves the PSNR from 28.56dB to 29.45dB. Compared to zero-initialized convolution which simply conducts weighted summation of LR frame features and HR ones to guide spatial-level diffusion learning, the combination of SFA and TFA not only enhances the spatial adaptation via feature modulation but also strengthens temporal feature alignment by the tubelet-based attention. As such, the higher PSNR and SSIM which measure the spatial fidelity is attained by the model C. Finally, the model **D**, i.e., our SATeCo, by further exploiting SFA and TFA in VAE to regulate pixel-space video reconstruction, shows the best performances in PSNR and SSIM. In view of perception-based evaluation metrics, SATeCo also constantly obtains improvements over other variants, indicating the potential benefit from spatial-temporal guidance learning to enhance visual perception in HR videos. Furthermore, Figure 7 showcases video super-resolution in a local region of one example in two adjacent frames. SATeCo reconstructs the HR videos with high-quality visual appearance and promising temporal consistency among adjacent frames, proving the impact of exploring feature adaptation and alignment in diffusion for super-resolution.

Analysis on Video Upscaler. Then, we study the effectiveness of the video upscaler in the SATeCo. One alternative is to employ the pre-trained Pixel Shuffle layer [40] as the video upscaler. The upper part of Table 3 details the performances of the two approaches on REDS4. Our approach exhibits better performances against PixelShuffle across all evaluation metrics, especially in PSNR and SSIM. Technically, PixelShuffle resamples videos via directly conducting a 2D convolution layer on the input frames. Instead, ours delves into the formulation of frame-wise correlation through temporal mutual self-attention, which is more effective in pixel feature enhancement for video resampling. As such, ours effectively preserves the visual contents in LR videos and facilitates the subsequent video diffusion.

Analysis on Video Refiner. The video refiner in SATeCo aims for adjusting the decoded HR videos from diffusion model by referring up-sampled original LR videos to alleviate color degradation. The trade-off parameter w in video refiner balances the impact of the visual contents between the decoded videos and the LR videos. To evaluate the influence of the parameter w, we list the VSR performances by varying w in the lower part of Table 3. When w is 0, the performances on perception-based metrics are the best, but there is a slight performance drop on PSNR and SSIM. The performances indicate that the synthesized visual contents by diffusion models are more acceptable for human visual system. In contrast, employing a large value of w (e.g., 1.0) for video refinement considers the information of LR videos more and weakens the contribution of diffusion models, affecting the quality of visual content generation. Therefore, we empirically set the w as 0.5 to seek a good trade-off between the synthesized contents and original visual appearance.

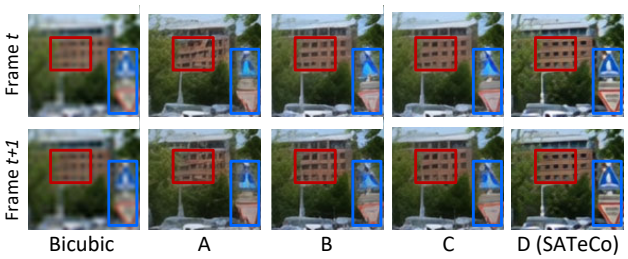

Figure 7. Zoom-in view of two adjacent frames in one video super-resolution result synthesized by variants of SATeCo.

5. Conclusions

We have presented SATeCo that explores spatial adaptation and temporal coherence in diffusion models for video super-resolution. In particular, we study the problem of learning spatial-temporal guidance from low-resolution videos to calibrate high-resolution video diffusion procedure. To materialize our idea, SATeCo freezes all the parameters in the pre-trained UNet/VAE, and plugs the spatial feature adaptation (SFA) and temporal feature alignment (TFA) modules in each decoder block to regulate latent-space video denoising and pixel-space video reconstruction. Through learning affine parameters on the guidance of low-resolution videos, SFA modulates the high-resolution features of each pixel to achieve spatial adaptation. TFA performs self-attention within a tubelet to enhance feature interaction and further conducts cross-attention between the tubelet and its low-resolution counterpart to guide temporal feature alignment learning. Experiments conducted on two video datasets, i.e., REDS4 and Vid4, validate the effectiveness of the proposed SATeCo for video super-resolution in terms of both spatial fidelity and temporal consistency.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Contract 62021001, and in part by the Fundamental Research Funds for the Central Universities under contract WK3490000007. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution of USTC.

References

[1] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation. In *CVPR*, 2017. 2

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In *CVPR*, 2021. 2, 5, 6

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. In *CVPR*, 2022. 2, 5

[4] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating Tradeoffs in Real-World Video Super-Resolution. In *CVPR*, 2022. 3

[5] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two Deterministic Half-quadratic Regularization Algorithms for Computed Imaging. In *ICIP*, 1994. 5

[6] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. AnchorFormer: Point Cloud Completion from Discriminative Nodes. In *CVPR*, 2023. 2

[7] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *ICCV*, 2021. 2

[8] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception Prioritized Training of Diffusion Models. In *CVPR*, 2022. 5

[9] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving Diffusion Dodels for Inverse Problems Using Manifold Constraints. In *NeurIPS*, 2022. 2

[10] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *ICLR*, 2023. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021. 1, 2

[12] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE TPAMI*, 2020. 6

[13] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *CVPR*, 2023. 2

[14] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. RSTT: Real-time Spatial Temporal Transformer for Space-Time Video Super-Resolution. In *CVPR*, 2022. 2

[15] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent Back-Projection Network for Video Super-Resolution. In *CVPR*, 2019. 2

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *ICLR*, 2023. 2

[17] Yan Huang, Wei Wang, and Liang Wang. Video Super-Resolution via Bidirectional Recurrent Convolutional Networks. *IEEE TPAMI*, 2017. 2

[18] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video Super-Resolution with Recurrent Structure-Detail Network. In *ECCV*, 2020. 2

[19] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video Super-resolution with Temporal Group Attention. In *CVPR*, 2020. 2

[20] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation. In *CVPR*, 2018. 2

[21] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising Diffusion Restoration Models. In *NeurIPS*, 2022. 2

[22] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. MuCAN: Multi-Correspondence Aggregation Network for Video Super-Resolution. In *ECCV*, 2020. 2

[23] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A Video Restoration Transformer. *arXiv:2201.12288*, 2022. 1, 2, 3, 5, 6

[24] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent Video Restoration Transformer with Guided Deformable Attention. In *NeurIPS*, 2022. 2

[25] Ce Liu and Deqing Sun. On Bayesian Adaptive Video Super Resolution. *IEEE TPAMI*, 2013. 5

[26] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning Trajectory-Aware Transformer for Video Super-Resolution. In *CVPR*, 2022. 2

[27] Fuchen Long, Zhaofan Qiu, Yingwei Pan, Ting Yao, Jiebo Luo, and Tao Mei. Stand-Alone Inter-Frame Attention in Video Models. In *CVPR*, 2022. 2

[28] Fuchen Long, Zhaofan Qiu, Yingwei Pan, Ting Yao, Chong-Wah Ngo, and Tao Mei. Dynamic Temporal Filtering in Video Models. In *ECCV*, 2022.

[29] Fuchen Long, Ting Yao, Zhaofan Qiu, Lusong Li, and Tao Mei. PointClustering: Unsupervised Point Cloud Pre-training using Transformation Invariance in Clustering. In *CVPR*, 2023. 2

[30] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. VideoDrafter: Content-Consistent Multi-Scene Video Generation with LLM. *arXiv:2401.01256*, 2024. 2

[31] Xiangming Meng and Yoshiyuki Kabashima. Diffusion Model Based Posterior Sampling for Noisy Linear Inverse Problems. *arXiv:2211.12343*, 2022. 2

[32] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "Completely Blind" Image Quality Analyzer. *IEEE SPL*, 2012. 6

[33] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study. In *CVPRW*, 2019. 5

[34] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards Photorealistic

Image Generation and Editing with Text-guided Diffusion Models. In *ICML*, 2022. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 6

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1, 2, 5

[37] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-Image Diffusion Models. In *ACM SIGGRAPH*, 2022. 1

[38] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-Recurrent Video Super-Resolution. In *CVPR*, 2018. 2

[39] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking Alignment in Video Super-Resolution Transformers. In *NeurIPS*, 2022. 2, 3, 4

[40] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *CVPR*, 2016. 3, 8

[41] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015. 6

[42] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided Diffusion Models for Inverse Problems. In *ICLR*, 2022. 2

[43] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution. In *CVPR*, 2020. 2

[44] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art Diffusion Models, 2022. 5

[45] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI*, 2023. 6

[46] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv:2305.07015*, 2023. 1, 2, 3, 5, 6

[47] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep Video Super-Resolution using HR Optical Flow Estimation. *IEEE TIP*, 2020. 2

[48] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks. In *CVPRW*, 2019. 2, 5, 6

[49] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. In *ICLR*, 2023. 2

[50] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal Modulation Network for Controllable Space-Time Video Super-Resolution. In *CVPR*, 2021. 2

[51] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. Video Enhancement with Task-Oriented Flow. *IJCV*, 2019. 2, 3, 5, 6

[52] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-Aware Stable Diffusion for Realistic Image Super-Resolution and Personalized Stylization. *arXiv:2308.14469*, 2023. 1, 2, 3, 6

[53] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive Fusion Video Super-Resolution Network via Exploiting Non-Local Spatio-Temporal Correlations. In *ICCV*, 2019. 2

[54] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Tao Lu, Xin Tian, and Jiayi Ma. Omniscient Video Super-Resolution. In *ICCV*, 2021. 2

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023. 1, 2, 7

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018. 6

[57] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising Diffusion Models for Plug-and-Play Image Restoration. In *CVPRW*, 2023. 2