Learning Data-Driven Vector-Quantized Degradation Model for Animation Video Super-Resolution

Zixi Tuo[*1], Huan Yang[†2], Jianlong Fu[2], Yujie Dun[1], Xueming Qian[1,3]

[1]Xi'an Jiaotong University   [2]Microsoft Research Asia

[3]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd

zixit99@gmail.com, {huayan, jianf}@microsoft.com, {dunyj, qianxm}@mail.xjtu.edu.cn

Abstract

*Existing real-world video super-resolution (VSR) methods focus on designing a general degradation pipeline for open-domain videos while ignoring data intrinsic characteristics which strongly limit their performance when applying to some specific domains (e.g., animation videos). In this paper, we thoroughly explore the characteristics of animation videos and leverage the rich priors in real-world animation data for a more practical animation VSR model. In particular, we propose a multi-scale **V**ector-**Q**uantized **D**egradation model for animation video **S**uper-**R**esolution (VQD-SR) to decompose the local details from global structures and transfer the degradation priors in real-world animation videos to a learned vector-quantized codebook for degradation modeling. A rich-content **R**eal **A**nimation **L**ow-quality (RAL) video dataset is collected for extracting the priors. We further propose a data enhancement strategy for high-resolution (HR) training videos based on our observation that existing HR videos are mostly collected from the Web which contains conspicuous compression artifacts. The proposed strategy is valid to lift the upper bound of animation VSR performance, regardless of the specific VSR model. Experimental results demonstrate the superiority of the proposed VQD-SR over state-of-the-art methods, through extensive quantitative and qualitative evaluations of the latest animation video super-resolution benchmark. The code and pre-trained models can be downloaded at https://github.com/researchmm/VQD-SR.*

1. Introduction

Animation video super-resolution (VSR) [35], which is a subdiscipline of video super-resolution (VSR) [1, 2, 3, 19, 28, 37], aims to restore high-resolution (HR) videos from their low-resolution (LR) counterparts in animation domain. As a special type of visual art, animation video shows great
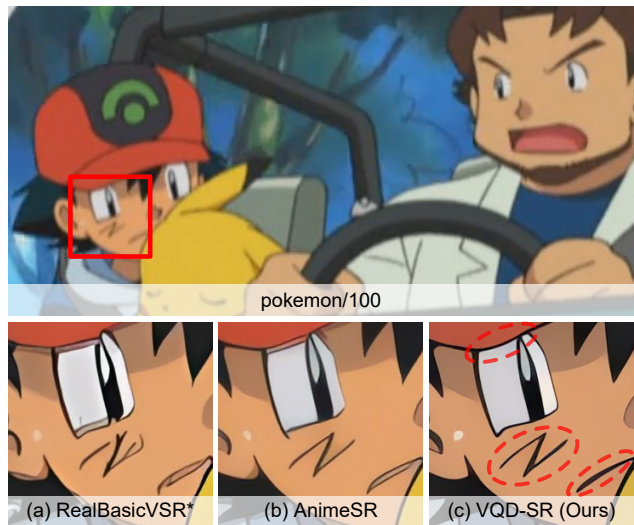

Figure 1. **A comparison between our proposed VQD-SR and other SOTA VSR methods in real scenario.** As indicated by the red dotted circles, our VQD-SR could restore more appealing details by considering the degradation priors in real animation videos. ('*' denotes fine-tune on animation videos).

entertainment, cultural, educational and commercial values. However, most of them available on the Web are in limited resolutions and have visually unappealing artifacts due to the passing ages, lossy compression and transmission. Dedicated to real applications, animation VSR is a blind SR [21, 32, 46, 40, 43] task which is more challenging than conventional VSR tasks [1, 2, 14, 19, 28, 36], because of the complicated and agnostic real-world degradations. Thus, practical VSR methods for improving real-world LR animation videos are in highly demanded.

Blind SR model for open-domain has long been the focus. As it's hard to collect LR-HR pairs in real scenarios, recent learning-based approaches try to model the real-world degradations from HR to LR, thus generating pseudo LR-HR pairs to train SR models. These approaches can

---

be roughly divided into two categories, explicit modeling and implicit modeling, with regard to the degradation procedures. Explicit modeling [9, 20, 32, 46] methods try to combine basic degradation operators like noise, blur and rescaling to imitate the real-world degradation pipeline. Instead, implicit modeling methods employ a neural network, which is usually in the manner of generative adversarial networks (GAN) [7], as an alternative to the combination of basic degradation operators. However, these general degradation methods designed for open-domain videos are not quite suitable for the animation domain as they ignore the intrinsic characteristics of the data. Directly adopting existing degradation models for open-domain to animation VSR would lead to unpleasant results (as shown in Fig. 1 a), even though they are fine-tuned on the animation videos.

Specific to animation VSR, AnimeSR [35] moves one step further to propose a learnable basic operator (LBO) with an "input-rescaling" strategy that considers the data characteristics and demonstrates their advantages over open-domain VSR methods. However, the proposed "input-rescaling" strategy of constructing pseudo HR-LR pairs for the LBO training requires human annotations, which prevents their method from scaling up to modeling various real-world degradations. In AnimeSR [35], only three LBOs are trained with three human-annotated videos respectively, leading to unappealing results (as shown in Fig. 1 b). Thus, how to effectively leverage large-scale real-world animation videos based on their data characteristics is still an open problem.

In this paper, we propose a multi-scale **V**ector-**Q**uantized **D**egradation model for animation video **S**uper-**R**esolution (VQD-SR), which leverages the data characteristics of animation videos. We observe that different from open-domain videos with complex textures and irregular illumination conditions, animation videos are roughly composed of basic visual patterns like *smooth color patches and clear lines* which can be easily captured with high fidelity by vector-quantization (VQ) [5, 25, 29]. Such data characteristics could also be used for encoding and transferring degradation priors from low-quality animation videos and results in a degradation codebook. Given any clean LR animation frame, the degraded LR counterpart can be easily obtained by looking up the pre-trained degradation codebook. Specifically, a novel multi-scale VQGAN for degradation modeling is proposed to learn and transfer real-world degradation priors from a large-scale **R**eal **A**nimation **L**ow-quality (RAL) video dataset collected from the Web. To improve the generalization ability of VQD-SR, we propose a two-stage training pipeline and a stochastic top-k VQ strategy to match up with the multi-scale VQGAN, yielding visually appealing results (as shown in Fig. 1 c).

To further lift the upper bound of existing animation VSR methods, we propose an HR-SR strategy to enhance the quality of HR training videos for more ideal ground truths. This is motivated by our observations that existing HR animation videos are mostly collected from the Web which contains conspicuous compression artifacts and such artifacts could be easily reduced by existing SR methods. Our contributions are summarized as follows:

- We collect a large-scale real LR animation dataset, and conduct a comprehensive study on real animation data to thoroughly explore the challenges and data characteristics in animation VSR task.
- We propose a novel multi-scale VQ degradation model to transfer real-world degradation priors for synthesizing pseudo LR-HR training pairs. We further propose an HR-SR data enhancement strategy to improve the performance of existing VSR methods for animation.
- Extensive experiments demonstrate the effectiveness of our proposed methods which significantly improve the recent SOTA animation VSR method in both quantitative and qualitative comparisons.

2. Related Work

2.1. Video Super-Resolution

Recent years have witnessed a prosperity in deep learning based VSR methods [2, 3, 14, 24, 28, 31, 38]. Different from single-image super-resolution (SISR), the auxiliary temporal information is available in VSR for the restoration of inner-frame features. Sliding-window structure [28, 31, 38, 42] and recurrent structure [1, 2, 6, 12, 14] are two common paradigms in VSR to utilize the inter-frame information. However, it also faces challenges while taking the advantages. The degradations could be more complicated in videos and may accumulate along the time sequence [3, 37], causing the performance drop in real-world scenarios.

There have been several attempts in blind-SR field to deal with real-world degradations, which could be categorized into explicit and implicit modeling. Explicit methods model the degradation as the effects of objective operators (or kernels). Methods like [9, 47] assume these degradations to be simple combinations of blur, noise and down-sampling kernels, therefore conducting conditional restoration based on the prediction of kernel parameters. Operators like JPEG/FFmpeg [20] for lossy compression, $sinc$ filter [32] for ringing artifacts and strategies like high-order modeling [32], random shuffling [46], are further investigated in latest studies to simulate the real degradation pipeline and expand the synthesis space. However, these methods merely depend on predefined operators, and ignore to utilize the information contained in real data, thus there's still a gap between the synthesis LR and real LR data. Implicit methods [21, 22, 34, 44] discard the objective degradation operators, but learn the distribution of real data,

therefore realize the degradation by converting the HR input to fit the real LR distribution. Yet, the learned distribution space is uncontrollable, thus the degradation network may easily project unseen out-of-distribution samples to unpredictable suboptimal space causing annoying artifacts.

2.2. Animation Video Processing

Animation is a type of visual art that is represented in a nonrealistic artistic style. A diversity of animation processing technologies based on deep learning have been developed these years, such as animation style transfer [4, 13, 30], animation video colorization [18, 27, 45], animation video generation [11] and animation video interpolation [26]. AnimeSR [35] is a recent study for real-world animation VSR tasks, which develops LBO to expand the basic degradation operators. In this paper, we take a further step to explore the characteristics of animation videos, and make special designs in consideration of these characteristics to improve the performance of real-world animation VSR.

2.3. Vector-Quantized Codebook

Vector-quantized(VQ) codebook was first proposed in VQ-VAE [29] to learn the discrete representation for images in latent space. Later, a hierarchical VQ-VAE-2 [25] is further proposed. To improve the perceptual quality, a discriminator and perceptual loss were added in VQGAN [5]. Recently in VQFR [10], a VQ-based face restoration method was also proposed. Although initially designed to avoid the"posterior collapse" issue of VAE or tokenize input images for a more effective transformer modeling, we discover that the idea of VQ is naturally suitable for representing animation video frames and the form of codebook is capable to internalize the real degradation priors.

3. Methodology

In this section, we first discuss the characteristics of real animation videos and the challenges for animation VSR task in Sec. 3.1. Based on the observations in Sec. 3.1, we introduce the proposed real-world animation degradation model in Sec. 3.2. Finally, we illustrate our HR training data enhancement strategy in Sec. 3.3.

3.1. Real-World Degradation in Animation Videos

To thoroughly understand the characteristics of animation videos and the challenges in animation VSR task, we collect a **R**eal **A**nimation **L**ow-quality (RAL) video dataset. RAL contains over 10K LR frames extracted from 441 real-world low-quality animation videos. To the best of our knowledge, RAL is the largest real-world low-quality video dataset in the animation domain. Example frames are shown in Fig. 2 (statistics of RAL can be found in the

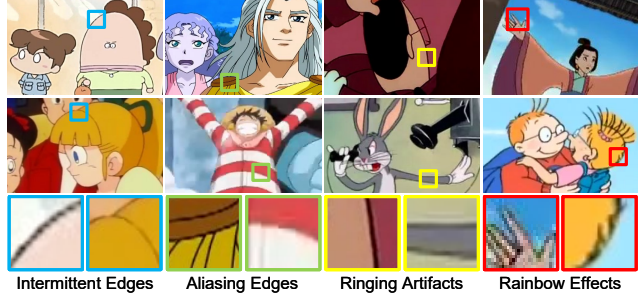| Intermittent Edges | Aliasing Edges | Ringing Artifacts | Rainbow Effects |

Figure 2. Samples of RAL and typical degradation phenomena in real-world LR animation videos. RAL is a real-world animation video dataset contains over 10K LR frames. Intermittent edges, aliasing edges, ringing artifacts and rainbow effects are four typical degradation phenomena in real-world animation videos.

supplementary). To construct such a dataset, we first download numerous earlier animation videos from multiple video sites such as YouTube and Bilibili. Then we manually select videos according to their resolutions, quality, contents and styles. To ensure the dataset contains rich animation structures and degradation priors, we adopt a simple scene filtering procedure to remove the frames which have subtle motions by evaluating their average absolute differences, thus reducing the redundancy.

Animation videos are roughly composed of *smooth color patches and clear lines*, while the latter ones usually play an important role in portraying objects and expressing visual semantics. In contrast to the fact that degradation losses for open domain videos are mainly about high-frequency textures, degradations for animation videos mostly happen around the edges which are more sensitive according to the human perceptual. Such degradations usually act as intermittent edges, aliasing edges, ringing artifacts, and rainbow effects, as shown in Fig. 2. These observations of the proposed RAL dataset inspire us that the key challenge in animation VSR is to restore visually natural and clear edges.

Degradation modeling has been the focus of many previous blind-SR works. Because of the obvious differences between animation and natural style videos, degradation methods designed for natural videos cannot properly handle degradations in animation videos. As a result, these methods suffer from amplified artifacts causing phenomena like perturbation edges, bleeding edges and inaccurate colorization along edges, even though they are fine-tuned on the animation videos (as shown in Fig. 1 a).

3.2. Vector-Quantized Degradation Modeling

Instead of imitating the real-world degradation pipeline by simply combining and adjusting some predefined basic operators [32, 46], we propose to explore the data intrinsic characteristics further and utilize degradation priors of real-world animation videos. As illustrated in Sec. 3.1,
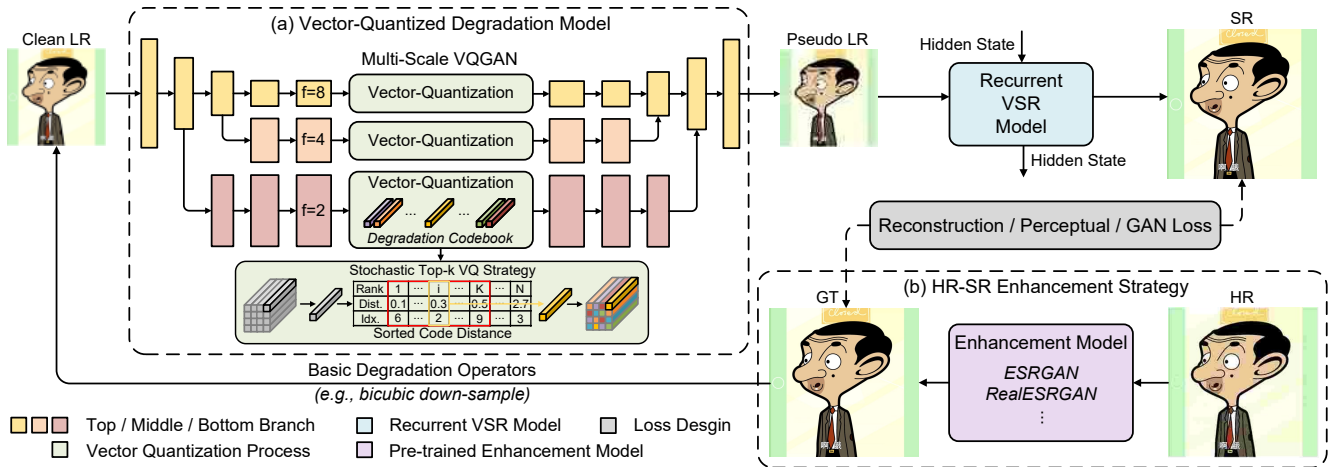
Figure 3. The overview of VQD-SR for training the animation VSR model. (a) Vector-quantized degradation model is a multi-scale VQGAN which transfers the real-world degradations from the learned VQ codebook to clean LR video frames to synthesize pseudo LR videos for VSR training, with the stochastic top-k VQ strategy. (b) HR-SR enhancement strategy eliminates the artifacts in HR animation training videos with an existing SR model to generate more ideal ground truths for supervising the VSR results.

animation videos are roughly composed of *smooth color patches and clear lines*. Compared to open-domain videos, which contain complex textures and irregular illumination conditions, it is possible to encode such animation video frames into a codebook of finite basic visual patterns (i.e., color patches and lines). More importantly, any animation video frame can be further decoded with corresponding patterns with high fidelity. If such a process is restricted to the real LR domain (*e.g.*, the proposed RAL dataset), the degradation priors of real animation videos will be encapsulated into the aforementioned codebook simultaneously and could be further used for synthesizing pseudo LR data for the VSR training.

VQGAN [5] is a reconstruction network composed of an encoder E, a decoder G and a vector-quantized codebook Z. Motivated by the perceptually rich codebook, we resort to VQGAN [5], trained on the proposed RAL dataset with rich degradations of animation videos, to realize the decomposition of low-quality frames and encapsulate degradation priors into a learned codebook. When given a clean LR frame, the real-world degradation priors can be transferred by replacing the original clean patches with the corresponding degraded versions by looking up the learned degradation codebook. In this section, we first introduce our modified multi-scale VQGAN, then illustrate the two-stage training pipeline, and finally use the trained model to synthesize pseudo LR-HR pairs in diverse levels of degradations by a stochastic top-k vector-quantization strategy.

Multi-Scale VQGAN. As discussed in Sec. 3.1, the degradations in animation videos mostly happen around the edges, which are fine-grained local details. A degradation model should capable of processing these local details while

keeping the global structures (*e.g.*, colors and shapes) at the same time. However, these two goals are usually conflicted with each other under the scope of the original single-scale VQGAN [5], as these two kinds of components are entangled together. Disentangling local details from global structures is a requisite for degradation modeling with VQGAN.

The multi-scale design has been widely explored in low-level vision [23, 39] and achieved great success, especially in the image restoration task. Inspired by this, we propose a multi-scale VQGAN which contains three parallel encoding branches. As shown in Fig. 3 a, these three branches share a shallow feature extraction module but have independent deeper encoding modules with different compression factors f (we choose $f = \{8, 4, 2\}$ for the top, middle and bottom branches respectively). Benefiting from our multi-scale design, image information of different levels could be effectively disentangled and learned by each encoding branch in one VQ-GAN model.

Two-Stage Training Pipeline. However, it is non-trivial to achieve the feature disentanglement as we expected by directly training the proposed model from scratch. This is because the bottom branch with the smallest scale will dominate the model training since it provides the easiest way to model converge. We call this phenomenon "short-cut". Such a problem prevents the top and the middle branches from learning image information for different scales, and the learned image representations are still entangled with each other. To take full advantage of our multi-scale VQ-GAN, we propose a two-stage training pipeline. We first remove the middle and the bottom branches, and only train the top branch with the largest compression factor f. In this stage, the top branch is forced to take effect and be re-

sponsible to reconstruct the global structures. After the convergence of the top branch, we add the other two branches back to fine-tune the overall multi-scale VQGAN. In the second stage, the two new branches only need to recover the remaining smaller structures and fine-grained details, thus avoiding the "short-cut" problem.

Stochastic Top-k VQ Strategy. After training our multi-scale VQGAN on a real-world low-quality animation video dataset (*e.g.*, the proposed RAL dataset in Sec. 3.1), the learned codebooks will contain rich degradation information and are capable of transferring such degradations into unseen animation data. Specifically, as shown in Fig. 3, the learned multi-scale VQGAN will be used as the degradation model to synthesize pseudo LR images from clean LR images for further VSR model training. Although the degradation priors in the codebooks are diverse, there is only one fixed degradation output for each input as the quantization process takes the nearest neighbor search strategy to reconstruct the input, leading to poor generalization capability of the final VSR model. To expand the degradation space and introduce more randomness during training, we adopt a stochastic top-k search strategy when looking up the codebook for the VQ process. More precisely, given the encoded output $\hat{z} = E(x) \in \mathbb{R}^{h \times w \times n_z}$ from the input image x, the former vector quantization methods search the nearest entry in the codebook $\mathcal{Z} = \{z_l\}_{l=1}^L$ for each vector $\hat{z}_{ij}$ in position (i, j) to get the quantized output z_q:

$$z_q = q(\hat{z}) := \left(\arg \min_{z_l \in \mathcal{Z}} \|\hat{z}_{ij} - z_l\| \right) \in \mathbb{R}^{h \times w \times n_z}. \quad (1)$$

We expand the nearest search strategy to a stochastic top-k search strategy. That is, we obtain the k_{th} closest entry in the codebook for each vector $\hat{z}_{ij}$ in position (i, j):

$$z_q = q(\hat{z}) := \left(\arg \, k_{th} \atop z_l \in \mathcal{Z} \right. \left. \|\hat{z}_{ij} - z_l\| \right) \in \mathbb{R}^{h \times w \times n_z}. \quad (2)$$

For each sample in every training iteration, the value of k is randomly chosen from a predefined range $[1, K]$ and controls the level of degradation. We chose K=50 in VQD-SR and the experiment of the choice of K can be found in the supplementary.

A key point here is that we adopt the stochastic top-k strategy only for the bottom branch while keeping the nearest neighbor search for the top and middle branches. This is because the larger-scale structures like colors and shapes are encoded in these two branches, and we do not want them to be changed when augmenting the local details for degradation. That is also the reason why we make a great effort to design the multi-scale VQGAN that encodes the different components in input images separately. The idea of stochastic top-k VQ strategy also corresponds with the fact that one HR image could have multiple degraded versions.

3.3. HR-SR Enhancement Strategy

AVC [35] is the latest large-scale high-quality animation video dataset, specified for animation VSR tasks. As opposed to images that could be recorded in lossless format, most of the animation videos available on the Web all went through compression to reduce the network traffic. As demonstrated in Fig. 3 b, the HR animation video frames still suffered from the compression artifacts, and therefore are not completely ideal as the training ground truths. Based on this observation, we suggest enhancing the HR animation videos to further lift the upper bound for the existing animation VSR model. Since the contents of animation videos are relatively simple and the HR video frames are relatively clean, general pre-trained SR models (*e.g.*, Real-ESRGAN [32]) could be used to reduce the artifacts while not contaminating image details. We thus enhance all the HR video frames in the AVC dataset by first applying an SR model, then downsampling to align the original sizes.

3.4. Training Details

Degradation Model. In our multi-scale VQGAN, we set the compression factor f to {8,4,2} for the {top, middle, bottom} branch respectively and use 1024 codebook entries with 256 channels in total. The input images are randomly cropped to patches of 256×256. The training of our multi-scale VQGAN is a two-stage procedure. We follow [5] to adopt Adam [16] optimizer with the base learning rate of 4.5×10^{-6} for both stages. The learning rate in top branch is decreased by 4 times in the second stage. We set the total batch size to 32 and 24 for the two stages, respectively. We only apply VQ loss L_{vq} [29] and perceptual loss L_{per} [48] in the first stage, while adding the adversarial loss L_{adv} [5] in the second stage. The full objective function for encoder E, decoder G and hierarchical VQ codebook Z is:

$$L(E, G, Z) = L_{vq} + L_{per} + L_{adv}, \quad (3)$$

in which L_{vq} is calculate by:

$$\begin{aligned} L_{vq} = &\|x - \hat{x}\| + \beta \|sg[E(x)_{(s)} - z_{q(s)}]\|_2^2 \\ &+ \|sg[z_{q(s)}] - E(x)_{(s)}\|_2^2. \end{aligned} \quad (4)$$

Where $x, \hat{x}$ are the input and reconstructed images, $sg[\cdot]$ denotes the stop-gradient operation, and $s \in \{t, m, b\}$ means the top, middle or bottom branches. More implementation details can be found in the supplementary.

Video Super-Resolution Model. We adopt the unidirectional recurrent network in AnimeSR [35] as the VSR model because of its efficiency, and further remove the SR feedback in the recurrent block to ensure a more stable training according to our observation that too many recurrent features would sometimes cause a collapse of GAN loss as the omitting of explicit alignment module. We train the

VSR model in two stages. In the first stage, we pre-train the VSR model with $L1$ loss for 300K iterations, with batch size 16 and learning rate 10^{-4}. In the second stage, we further add perceptual loss [15] and GAN loss [17] to fine-tune the network for another 300K iterations, with batch size 16 and learning rate 5×10^{-5}. As the promotion of perceptual quality is mainly during the second stage, we only append our multi-scale VQGAN for more realistic degradations in the second stage to reduce the training cost while only adopting basic operators in the first stage.

4. Experiments

4.1. Datasets and Metrics

We evaluate our method and compare its performance with other SOTA SR methods on AVC [35], the latest proposed large-scale animation video clip dataset. AVC is composed of three subsets. The training set AVC-Train contains 553 high-quality clips, and the testing set AVC-Test contains 30 high-quality clips. There is also a AVC-RealLQ , which contains 46 real-world low-quality clips for evaluation. Each clip consists of 100 frames. We adopt MANIQA [41], the winner in NTIRE 2022 NR-IQA challenge [8], as the no-reference image quality evaluation metric, as it in the previous animation VSR work [35].

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

We compare our VQD-SR with six SOTA SR methods. Among them, AnimeSR [35] and RealBasicVSR [3] are blind video super-resolution (VSR) methods. Real-ESRGAN [32] and BSRGAN [46] are blind single-image super-resolution (SISR) methods with explicit degradation modeling. We also conduct experiments on BasicVSR [1] and PDM [21] as two representative methods for none-blind VSR method and blind SISR with implicit degradation modeling. Apart from AnimeSR [35], for which we report the result in the original paper, all the others are general SR methods for open-domain images or videos. For fair comparisons, we fine-tune their officially released models on animation dataset AVC-Train.

Quantitative Comparison. As shown in Tab. 1, we evaluate all the methods on AVC-RealLQ [35] for quantitative comparisons. Among them, Real-ESRGAN [32] and BSRGAN [46] expand the former explicit degradation models by introducing high-order degradations and random shuffling respectively, which greatly improve the synthetic ability. And with the help of powerful SR backbone (RRDB-Net [33] with 16.70M parameters), their results are quite remarkable. Nonetheless, ignoring the intrinsic characteristics of animation data limits their performance when applied to animation domain. Although specialized for animation videos, AnimeSR [35] only utilizes a small number of real data (three human-annotated animation videos), which hin-

Table 1. Quantitative comparison on AVC-RealLQ for $4\times$ animation VSR. '*' denotes fine-tune on animation dataset AVC-Train [35]. **Bold** text indicates the best performance.

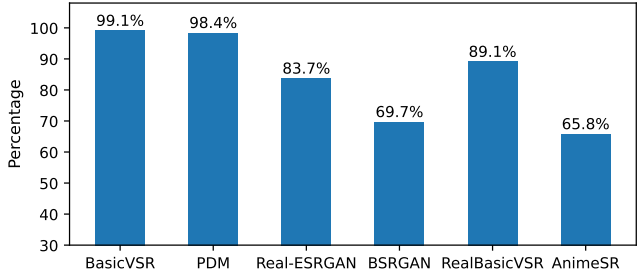| Method	Params (M)	Runtime (ms)	MANIQA↑
BasicVSR* [1]	6.29	175	0.2829
PDM* [21]	16.70	438	0.2574
Real-ESRGAN* [32]	16.70	438	0.3825
BSRGAN* [46]	16.70	438	0.3836
RealBasicVSR*[3]	6.29	252	0.3504
AnimeSR [35]	1.50	26	0.3832
VQD-SR (Ours)	**1.47**	**25**	**0.4096**

Figure 4. Results of User Study. The values on Y-axis represent the percentage of users that prefer VQD-SR over other approaches.

ders the performance of VSR model in real scenarios. Different from them, our VQD-SR considers the intrinsic characteristics of animation videos and leverages the enormous degradation priors contained in rich-content real animation videos. VQD-SR also adopts the HR enhancement strategy for more effective SR supervision. Due to these advantages, VQD-SR achieves a result of 0.4096 and significantly outperforms the SOTA animation VSR model [35] by 0.0264 in MANIQA on AVC-RealLQ [35]. Meanwhile, as the removal of SR recurrent feedback in VSR model, the number of parameters in VQD-SR are further reduced to 1.47 M with the runtime of 25 ms computed on a NVIDIA Tesla M40 GPU with the LR size of 180×320, which demonstrates the superiority of our VQD-SR in real applications.

Qualitative Comparison. As shown in Fig. 5, we compare visual qualities of different approaches on real-world AVC-RealLQ video clips. It can be observed that VQD-SR greatly improves the visual quality, especially for the recovery of clear lines. Additionally, the results of VQD-SR are more visually natural with fewer artifacts in some intended scenarios (e.g., the out-focus background blur and smooth underwater scene in row 5-6), while most of the other methods suffer from over-sharp artifacts like irregular lines and patterns. More visual comparisons with original inputs for reference and result analysis are in the supplementary materials.

We also conduct a user study to further evaluate the visual quality of our approach. There are 20 subjects involved in this user study and 22080 votes are collected on AVC-
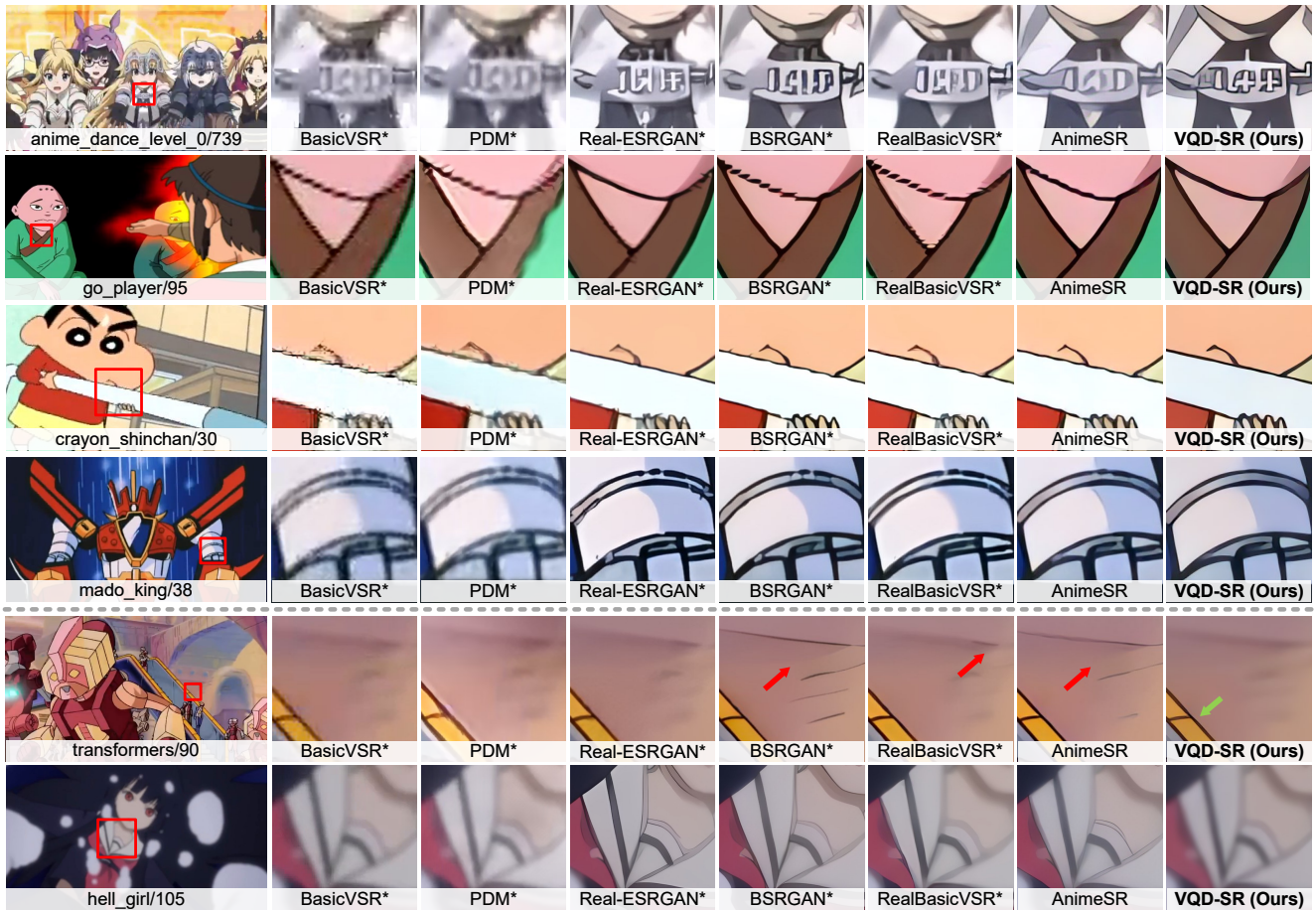
Figure 5. Qualitative comparisons on AVC-RealLQ [35] for $4\times$ scaling factor. '$*$' denotes fine-tune on animation domain dataset. The frame name is shown at the bottom of each case. Our VQD-SR recovers visually natural and sharper lines with fewer artifacts (the first 4 rows) and is also capable to handle some intended scenarios with fewer over-sharp artifacts (row 5th with smooth background and clear foreground, row 6th in underwater scene). Zoom in to see better visualization.

RealLQ [35] with 46 video clips. For each comparison, we adopt an A-B test that provides the users with two images in random order which include one VQD-SR frame. Users are asked to select the one with higher visual quality. As shown in Fig. 4, comparing with the SR methods designed for open-domain, over 80% of users voted for our VQD-SR. By considering the data characteristic of animation, there are still over 65% users who prefer our method to AnimeSR [35].

4.3. Ablation study

In this section, we conduct ablation studies on each significant component of our VQ degradation model and HR-SR enhancement strategy in VQD-SR.

VQ Degradation Model. In the ablation study of VQ degradation model design, we conduct all the experiments based on VQD-SR (base), which denotes a plain version of VQD-SR trained without HR-SR enhancement. As shown

Figure 6. Ablation study on VQ degradation model.

in Tab. 2 a, we first compare the degradation model which is simply composed of basic operators (blur, noise and FFmpeg) to our VQD-SR (base). With the help of real-world degradation priors contained in VQ codebook, the result can be improved from 0.3768 to 0.3857 in MANIQA. Refering to Fig. 6 a, VSR model trained with only basic operators fails to handle some real-world degradations, leading to artifacts around the edges. We further compare the structure and training strategy of VQ degradation model in Tab. 2 b and Tab. 2 c. As the global structures and local details are entangled together in single-scale VQ model and the

Table 2. Ablation study results of degradation models, structures of VQ degradation model, training pipelines of VQ degradation model and VQ codebook search strategies. The MANIQA scores are reported for comparisons.

	(a) Degradation Model	
MANIQA↑	BasicOP-Only	VQD-SR (base)
	0.3768	0.3857
	(b) VQGAN Structure	
MANIQA↑	Single-scale	VQD-SR (base)
	0.3466	0.3857
	(c) Training Strategy	
MANIQA↑	Single-stage	VQD-SR (base)
	0.3430	0.3857
	(d) Codebook Search	
MANIQA↑	Nearest	VQD-SR (base)
	0.3770	0.3857

Table 3. Applying VQ degradation to widely used SR backbones. For each backbone, settings are consistent apart from degradation methods.

Backbone	Real-ESRGAN*	RealBasicVSR*	AnimeSR
Params (M)	16.7	6.3	1.5
Original degradation	0.3825	0.3504	0.3832
VQ degradation	**0.3967**	**0.3849**	**0.3857**

multi-scale model trained in single-stage, the performances dropped by 0.0391 and 0.0427, respectively. As shown in Fig. 6 b and Fig. 6 c, some fine-grained details are missing. For VQ strategies, as shown in Tab. 2 d and Fig. 6 d, our stochastic top-k VQ strategy can improve MANIQA by 0.0087 compared with the original nearest neighbor search strategy, as the top-k VQ strategy enables multi-level degradations which improves the generalization ability of the trained animation model, leading to more appealing results.

To further illustrate the general applicability, we also apply our VQ degradation model to other SOTA SR backbones, Real-ESRGAN and RealBasicVSR, fine-tuned with animation data in the same setting. Referring to Tab. 3, SR models with VQ degradation achieve better results comparing with their original degradation methods, which indicates that the proposed VQ degradation is generally effective for animation SR models. It is noteworthy that the performance gains could be more significant for larger backbones. The visual examples in Fig. 7 also show the superiority of our VQ degradation method, especially to recover clear and visual natural lines which are the most import patterns deciding the visual quality of animation videos.

HR Enhancement. The training HR animation videos collected from the Web also contain artifacts which are not ideal for ground truths, hindering the performances of animation VSR models. We propose the HR-SR strategy to enhance the quality of HR animation videos for more effective training ground truths. We first compare three different HR enhancement models by experiments on VQD-SR (base).
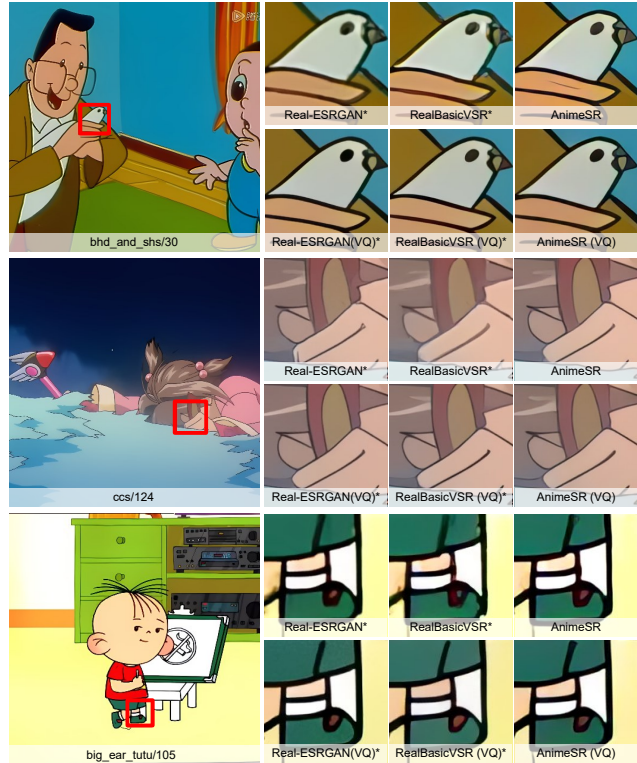

Figure 7. Applying VQ degradation to widely used SR backbones. Settings are consistent apart from degradation methods. Models trained with VQ degradation could produce more clear and visual natural lines with fewer artifacts.

The quantitative results are shown in Tab. 4. All three models improve the performance of animation VSR model, which verifies that the proposed HR-SR strategy is suitable for the training of animation VSR model, and improvements are not subject to the specific model for enhancement. However, as the SR process for data enhancement is conducted in HR domain, which takes a high demand on computing resources, we finally choose SISR model, Real-ESRGAN (S) [32] with relatively small model size. Visual examples of the enhanced HR frames can be found in Fig. 8.

We further utilize the enhanced HR animation videos processed by Real-ESRGAN(S) [32] to fine-tune different VSR models. Apart from the proposed VQD-SR, RealBasicVSR [3] is a blind VSR model designed for open-domain videos and AnimeSR [35] is an VSR model specified for animation domain. As shown in Tab. 5, with the help of HR enhancement, all three VSR models get improved, which illustrates that the proposed HR enhancement strategy is effective for animation VSR task, regardless of the specific VSR model. The visual results of RealBasicVSR+, AnimeSR+, and VQD-SR in Fig. 9 also verify the advantages of the HR-SR enhancement strategy for animation VSR.

Table 4. Ablation study of different SR models for HR-SR enhancement. All three models improve the performance. We choose Real-ESRGAN (S) considering the computing cost.

Enhancement Method	MANIQA↑
No Enhancement	0.3857
AnimeSR [35]	0.4146
Real-ESRGAN (L) [32]	0.4211
Real-ESRGAN (S) [32]	0.4096

Table 5. Ablation study of HR-SR enhancement for different animation VSR models on AVC-RealLQ. We report MANIQA for comparison.

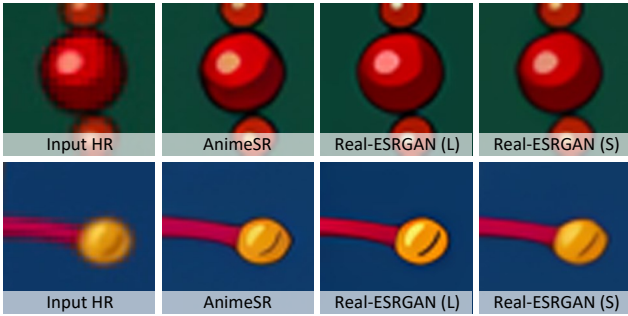| MANIQA↑	wo /HR-SR	w /HR-SR
RealBasicVSR [3]	0.3504	0.3855
AnimeSR [35]	0.3832	0.4079
VQD-SR	0.3857	0.4096

Figure 8. Enhanced HR data with HR-SR strategy by different enhancement models.

5. Limitation and Discussion

In this section, we visualize the failure cases of VQD-SR in Fig. 10. Our VQD-SR is capable to restore sharper lines and suppress diverse artifacts in animation videos. However, in some extreme circumstances (*e.g.*, severe lossy compressions, long passing ages), there still gets room for improvement, for example the remaining color distortions around edges. Nevertheless, comparing with other SOTA methods, our VQD-SR still achieves better performances.

6. Conclusion

In this paper, we thoroughly study the characteristics of animation videos and fully leverage the rich prior knowledge contained in real data for a more practical animation VSR model. To be specific, we propose a novel degradation model which utilize a multi-scale VQGAN, self-supervised on purely real low-quality animation data (RAL), to decompose animation frames and establish a codebook containing diverse real-world degradation priors. We also propose a
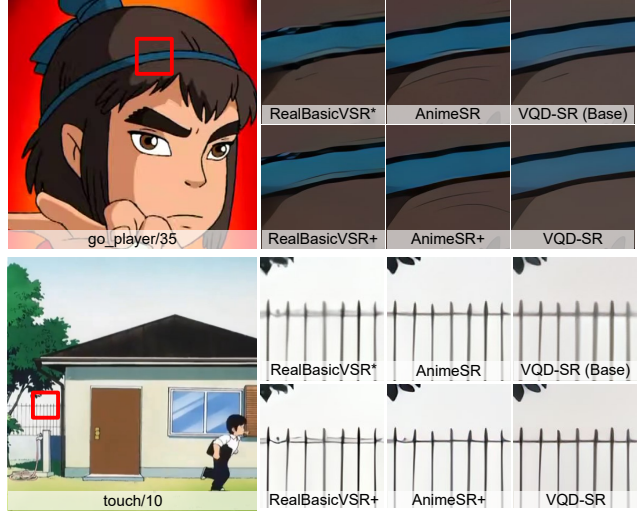

Figure 9. Ablation study of HR-SR enhancement strategy for different VSR models. HR-SR enhancement is generally effective task, regardless of the specific VSR model.
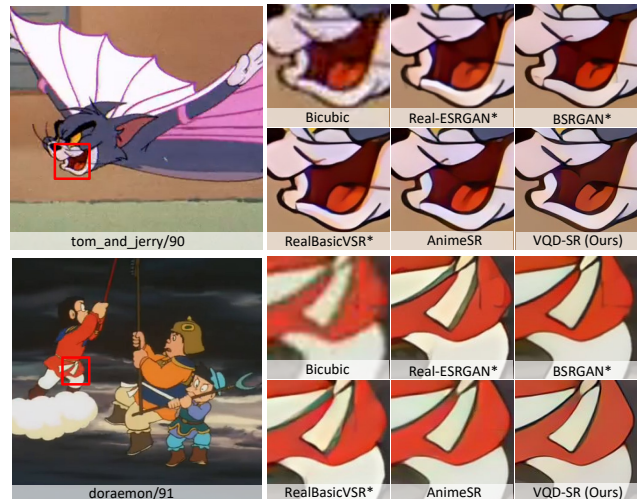

Figure 10. Failure cases when LR frames endure severe color distortions around edges.

two-stage training pipeline and a stochastic top-k VQ strategy to match up with the VQ degradation model. We further propose to enhance the HR training videos to lift the performance ceiling for animation VSR tasks. Experiment results show the superiority of the proposed VQD-SR over existing SOTA models, as our methods could generate more effective LR-HR training pairs in view of the characteristics in real animation videos.

References

[1] 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*, pages 4947–4956, 2021.

[2] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, pages 5972–5981, 2022.

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, pages 5962–5971, 2022.

[4] Jie Chen, Gang Liu, and Xin Chen. AnimeGAN: A novel lightweight gan for photo animation. In *ISICA*, pages 242–256, 2020.

[5] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021.

[6] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *ICCVW*, pages 3476–3485, 2019.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[8] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S Ren, Radu Timofte, Yuan Gong, Shanshan Lao, Shuwei Shi, Jiahao Wang, Sidi Yang, et al. NTIRE: 2022 challenge on perceptual image quality assessment. In *CVPR*, pages 951–967, 2022.

[9] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, pages 1604–1613, 2019.

[10] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, pages 126–143, 2022.

[11] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *ECCV*, pages 598–613, 2018.

[12] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, pages 3897–3906, 2019.

[13] Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. Domain-aware universal style transfer. In *ICCV*, pages 14609–14617, 2021.

[14] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *ECCV*, pages 645–660, 2020.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016.

[16] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.

[18] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *CVPR*, pages 3753–3761, 2019.

[19] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *CVPR*, pages 5687–5696, 2022.

[20] Pengju Liu, Hongzhi Zhang, Yue Cao, Shigang Liu, Dongwei Ren, and Wangmeng Zuo. Learning cascaded convolutional networks for blind single image super-resolution. *Neurocomputing*, 417:371–383, 2020.

[21] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Learning the degradation distribution for blind image super-resolution. In *CVPR*, pages 6063–6072, 2022.

[22] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *CVPR*, pages 291–300, 2020.

[23] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, pages 5690–5699, 2020.

[24] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *ECCV*, pages 257–273, 2022.

[25] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *NeurIPS*, 32, 2019.

[26] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep animation video interpolation in the wild. In *CVPR*, pages 6587–6595, 2021.

[27] Harrish Thasarathan, Kamyar Nazeri, and Mehran Ebrahimi. Automatic temporally coherent video colorization. In *CRV*, pages 189–194, 2019.

[28] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. TDAN: Temporally-deformable alignment network for video super-resolution. In *CVPR*, pages 3360–3369, 2020.

[29] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017.

[30] Jianbo Wang, Huan Yang, Jianlong Fu, Toshihiko Yamasaki, and Baining Guo. Fine-grained image style transfer with visual transformers. In *ACCV*, pages 841–857, 2022.

[31] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019.

[32] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021.

[33] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018.

[34] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *CVPR*, pages 13385–13394, 2021.

[35] Yanze Wu, Xintao Wang, Gen Li, and Ying Shan. AnimeSR: Learning real-world super-resolution models for animation videos. *arXiv preprint arXiv:2206.07038*, 2022.

[36] Jun Xiao, Xinyang Jiang, Ningxin Zheng, Huan Yang, Yifan Yang, Yuqing Yang, Dongsheng Li, and Kin-Man Lam. Online video super-resolution with convolutional kernel bypass grafts. *TMM*, 2023.

[37] Liangbin Xie, Xintao Wang, Shuwei Shi, Jinjin Gu, Chao Dong, and Ying Shan. Mitigating artifacts in real-world video super-resolution models. *arXiv preprint arXiv:2212.07339*, 2022.

[38] 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*, pages 6388–6397, 2021.

[39] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020.

[40] Fuzhi Yang, Huan Yang, Yanhong Zeng, Jianlong Fu, and Hongtao Lu. Degradation-guided meta-restoration network for blind super-resolution. *arXiv preprint arXiv:2207.00943*, 2022.

[41] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, pages 1191–1200, 2022.

[42] 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*, pages 3106–3115, 2019.

[43] Guanghao Yin, Xinyang Jiang, Shan Jiang, Zhenhua Han, Ningxin Zheng, Huan Yang, Donglin Bai, Haisheng Tan, Shouqian Sun, Yuqing Yang, et al. Online video streaming super-resolution with adaptive look-up table fusion. *arXiv preprint arXiv:2303.00334*, 2023.

[44] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPRW*, pages 701–710, 2018.

[45] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *CVPR*, pages 8052–8061, 2019.

[46] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021.

[47] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, pages 3262–3271, 2018.

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.