

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

Zixi Tuo^{1*}, Huan Yang², Jianlong Fu², Yujie Dun¹, Xueming Qian¹

¹Xi'an Jiaotong University ²Microsoft Research Asia

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 **Vector-Quantized Degradation** model for animation video **Super-Resolution (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 **Real Animation Low-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.

1. Introduction

Animation video super-resolution (VSR) [34], which is a subdiscipline of video super-resolution (VSR) [1, 2, 3, 19, 28, 35], 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 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.

*This work was done while Zixi Tuo was a research intern at Microsoft Research Asia.

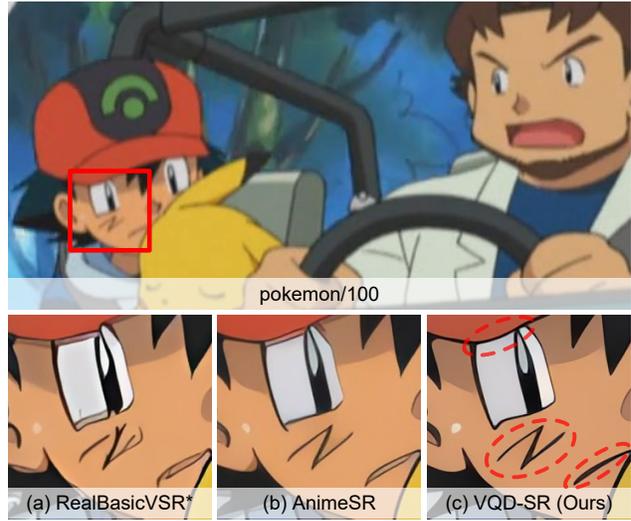


Figure 1. A comparison between our proposed VQD-SR and other SOTA VSR methods in real scenario. As indicated by the red dotted circles, our VQD-SR could restore more appealing details by considering the degradation priors in real animation videos. (* denotes fine-tune on animation videos).

Dedicated to real applications, animation VSR is a blind SR [21, 22, 31, 42, 33] 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 [10, 20, 31, 42] methods try to combine basic degradation operators like noise, blur and rescaling to imitate the real-world degradation pipeline. In-

stead, implicit modeling methods employ a neural network, which is usually in the manner of generative adversarial networks (GAN), 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 [34] 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 [34], 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 **Vector-Quantized Degradation** model for animation video **Super-Resolution (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) [7, 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 **Real Animation Low-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, 28, 30, 36]. 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, 30, 36, 39] and recurrent structure [1, 2, 8, 13, 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, 35], 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 [10, 43] 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 [31] for ringing artifacts and strategies like high-order modeling [31], random shuffling [42], 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, 33, 40] 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, 5, 6], animation video colorization [18, 27, 41], animation video generation [12] and animation video interpolation [26]. AnimeSR [34] 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 [7]. Recently in VQFR [11], 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 **Real Animation Low-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 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 sub-

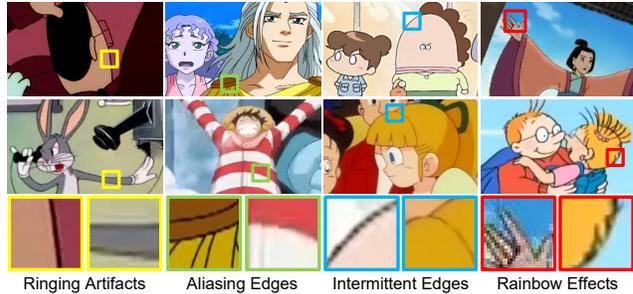


Figure 2. **Samples of RAL and typical degradation phenomena in real-world LR animation videos.** RAL is a real-world animation video dataset contains over 10K LR frames. Ringing artifacts, aliasing edges, intermittent edges and rainbow effects are four typical degradation phenomena in real-world animation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

tle 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 [31, 42], we propose to explore the data intrinsic characteristics further and utilize degradation priors of real-world animation videos. As illustrated in Sec. 3.1, 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 pat-

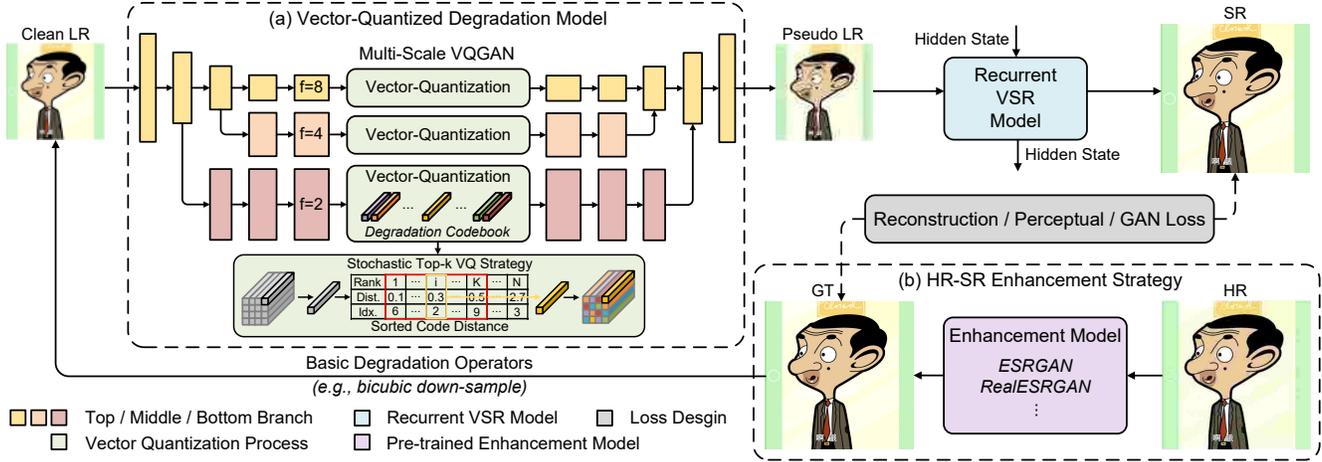


Figure 3. **The overview of VQD-SR for training the animation VSR model.** (a) Vector-quantized degradation model is a multi-scale VQGAN which transfers the real-world degradations from the learned VQ codebook to clean LR video frames to synthesize pseudo LR videos for VSR training, with the stochastic 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.

terns 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 [7] 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 [7], 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 be 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 [7], as these two kinds of components are entangled. 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, 37] 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 encoder modules and codebooks 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 convergence. 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 VQGAN, 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 responsible 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) :

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

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.

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 [34] 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 [31]) 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. For the {top, middle, bottom} branch in our multi-scale VQGAN, we set the compression factor f to {8,4,2} and use {1024,256,128} codebook entries with 256 channels. 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 [7] 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} [44] in the first stage, while adding the adversarial loss L_{adv} [7] in the second stage. The full objective function for encoder E , decoder G and hierarchical VQ codebook $Z_{\{t,m,b\}}$ is:

$$L(E, G, Z_{\{t,m,b\}}) = L_{vq} + L_{per} + L_{adv}, \quad (3)$$

in which L_{vq} is calculate by:

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

Where x, \hat{x} are the input image and reconstructed image, $sg[\cdot]$ 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 [34] 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 adding 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 appending 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 [34], 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-RealQ, which contains 46 real-world low-quality clips for evaluation. Each clip consists of 100 frames. We adopt MANIQA [38], the winner in NTIRE 2022 NR-IQA challenge [9], as the no-reference image quality evaluation metric, as it in the previous animation VSR work [34].

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

We compare our VQD-SR with six SOTA SR methods. Among them, AnimeSR [34] and RealBasicVSR [3] are blind video super-resolution (VSR) methods. Real-ESRGAN [31] and BSRGAN [42] 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 non-blind VSR method and blind SISR with implicit degradation modeling. Apart from AnimeSR [34], 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-RealQ [34] for quantitative comparisons. Among them, Real-ESRGAN [31] and BSRGAN [42] 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 [32] 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 [34] only utilizes a small number of real data (three human-annotated animation videos), which hinders 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 [34] by 0.0264 in MANIQA on AVC-RealQ [34]. 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

Table 1. Quantitative comparison on AVC-RealQ for 4× animation VSR. ‘*’ denotes fine-tune on animation dataset AVC-Train [34]. **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* [31]	16.70	438	0.3825
BSRGAN* [42]	16.70	438	0.3836
RealBasicVSR*[3]	6.29	252	0.3504
AnimeSR [34]	1.50	26	0.3832
VQD-SR (Ours)	1.47	25	0.4096

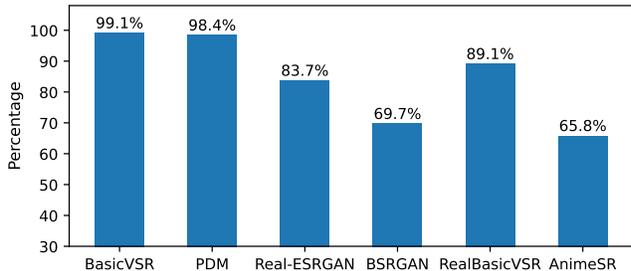


Figure 4. **Results of User Study.** The values on Y-axis represent the percentage of users that prefer VQD-SR over other approaches.

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-RealQ video clips. It can be observed that VQD-SR greatly improves the visual quality, especially for the recover of clear and visually natural lines. Additionally, VQD-SR is also capable of handling the background blur (the 5th row) or the blur effect in some intended scenarios (e.g. underwater scene in the 6th row), while most of the other methods suffer from over-sharp artifacts like irregular lines and patterns. More visual comparisons 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-RealLR [34] 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 [34].

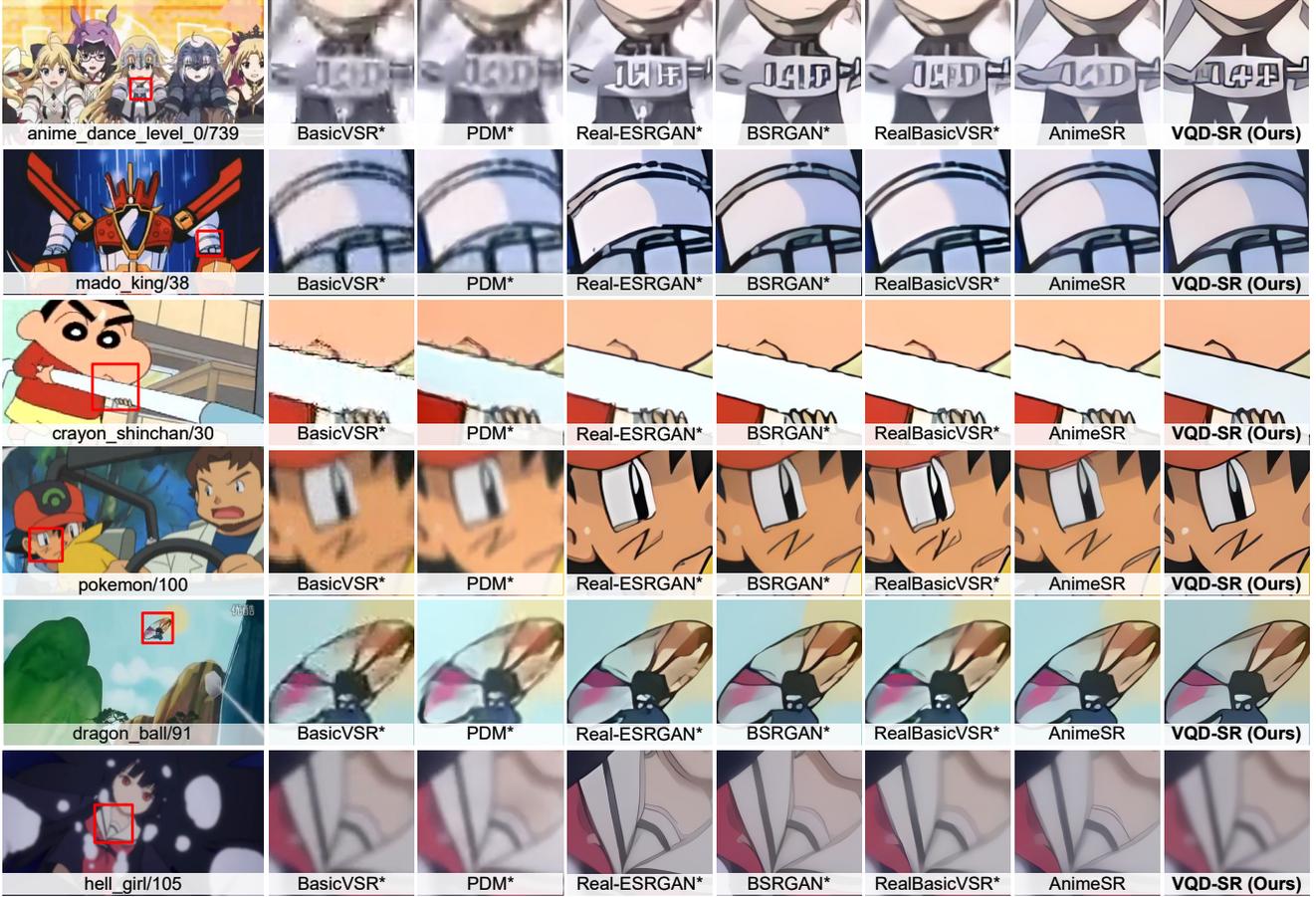


Figure 5. **Qualitative comparisons on AVC-RealLQ [34] for 4x scaling factor.** ‘*’ denotes fine-tune on animation domain dataset. The frame name is shown at the bottom of each case. Zoom in to see better visualization.

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, 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 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. Referring 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 multi-scale model trained in single-stage, the performances dropped by



Figure 6. Ablation study on VQ degradation model.

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 top-1 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.

HR Enhancement. Because 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 ef-

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↑	Top-1	VQD-SR (base)
	0.3770	0.3857

Table 3. Ablation study of different SR models used in HR-SR enhancement.

Enhancement Method	MANIQA↑
No Enhancement	0.3857
AnimeSR [34]	0.4146
Real-ESRGAN (L) [31]	0.4211
Real-ESRGAN (S) [31]	0.4037

fective training ground truths. We first compare three different HR enhancement models by experiments on VQD-SR (base). The quantitative results are shown in Tab. 3. 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) [31] with relatively small model size, as the HR-SR model. Visual examples of the enhanced HR frames can be found in Fig. 7.

We further utilize the enhanced HR animation videos processed by Real-ESRGAN(S) [31] 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 [34] is an VSR model specified for animation domain. As shown in Tab. 4, 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. 8 also verify the advantages of the proposed HR-SR enhancement strategy for animation VSR.

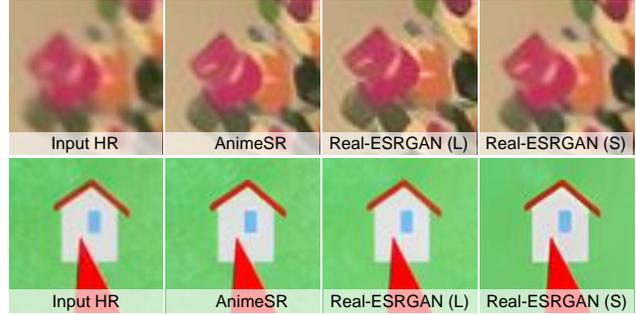


Figure 7. Visual comparison of different SR models for HR-SR enhancement.



Figure 8. Ablation study of HR-SR enhancement strategy for different VSR models.

Table 4. Ablation study of HR-SR enhancement for different VSR models on AVC-RealLQ. We report MANIQA for comparison.

MANIQA↑	wo /HR-SR	w /HR-SR
RealBasicVSR [3]	0.3504	0.3855
AnimeSR [34]	0.3832	0.4079
VQD-SR	0.3857	0.4096

5. 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 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. [1](#), [2](#), [6](#), [12](#)
- [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. [1](#), [2](#)
- [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. [1](#), [2](#), [6](#), [8](#), [11](#), [12](#)
- [4] Jie Chen, Gang Liu, and Xin Chen. AnimeGAN: A novel lightweight gan for photo animation. In *ISICA*, pages 242–256, 2020. [3](#)
- [5] Yugang Chen, Muchun Chen, Chaoyue Song, and Bingbing Ni. CartoonRenderer: An instance-based multi-style cartoon image translator. In *ACM MM*, pages 176–187, 2020. [3](#)
- [6] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. CartoonGAN: Generative adversarial networks for photo cartoonization. In *CVPR*, pages 9465–9474, 2018. [3](#)
- [7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. [2](#), [3](#), [4](#), [5](#)
- [8] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *ICCVW*, pages 3476–3485, 2019. [2](#)
- [9] 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. [6](#)
- [10] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, pages 1604–1613, 2019. [1](#), [2](#)
- [11] 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. [3](#)
- [12] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *ECCV*, pages 598–613, 2018. [3](#)
- [13] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, pages 3897–3906, 2019. [2](#)
- [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. [1](#), [2](#)
- [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. [5](#)
- [16] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [5](#)
- [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. [5](#)
- [18] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *CVPR*, pages 3753–3761, 2019. [3](#)
- [19] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *CVPR*, pages 5687–5696, 2022. [1](#)
- [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. [1](#), [2](#)
- [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. [1](#), [2](#), [6](#), [12](#)
- [22] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *CVPR*, pages 291–300, 2020. [1](#), [2](#)
- [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. [4](#)
- [24] 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*, pages 0–0, 2019. [13](#)
- [25] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 32, 2019. [2](#), [3](#)
- [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. [3](#)
- [27] Harrish Thasarathan, Kamyar Nazeri, and Mehran Ebrahimi. Automatic temporally coherent video colorization. In *CRV*, pages 189–194, 2019. [3](#)
- [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. [1](#), [2](#)
- [29] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. [2](#), [3](#), [5](#)
- [30] 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](#)
- [31] 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. [1](#), [2](#), [3](#), [5](#), [6](#), [8](#), [12](#)
- [32] 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. [6](#)

- [33] 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. [1](#), [2](#)
- [34] 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. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#), [11](#), [12](#), [17](#), [18](#), [19](#)
- [35] 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. [1](#), [2](#)
- [36] 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. [1](#), [2](#)
- [37] 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. [4](#)
- [38] 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. [6](#), [12](#)
- [39] 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. [2](#)
- [40] 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. [2](#)
- [41] 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. [3](#)
- [42] 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. [1](#), [2](#), [3](#), [6](#), [12](#)
- [43] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, pages 3262–3271, 2018. [2](#)
- [44] 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. [5](#)

Supplementary Material

In this supplementary material, Sec. A first illustrates the implementation details of VQD-SR, which includes the implementation of VQ degradation model in Sec. A.1, the details of VSR model in Sec. A.2, and the experiment details in Sec. A.3. Then Sec. B introduces our RAL dataset with statistics and representative samples. Sec. C shows more comparison results. Finally, Sec. D discusses the limitations of VQD-SR.

A. Implementation Details

A.1. VQ Degradation Model

Network Architecture. The architecture of our multi-scale VQGAN for degradation modeling is described in Tab. 6. The inputs of (b) Middle Branch and (c) Bottom Branch are the intermediate outputs of (a) Top Branch, represented by x_m and x_b respectively, and the output features \hat{x}_m and \hat{x}_b are further added back to the top branch when decoding. The vector-quantization is conducted pixel by pixel on the encoded outputs z_t , z_m and z_b in latent space with independent VQ codebooks, sized $\{1024, 256, 128\}$ with 256 channels. The compression factor f , which denotes the patch size when projecting each entry in the VQ codebook from latent space to the original image space, is set to $\{8, 4, 2\}$ controlled by the number of downsample steps in these three branches. During training, the middle branch and the bottom branch only take effect in the second-stage. All the training procedures are performed on eight NVIDIA 32G V100 GPUs.

Degradation Pipeline. The whole degradation pipeline with multi-scale VQGAN to transfer the real-world degradation priors can be formulated as: $x = D^n(y) = (\text{FFmpeg} \circ \text{VQD} \circ \text{Down} \circ \text{Noise} \circ \text{Blur})(y)$. Where x denotes the degraded LR clips and y denotes the HR clips. For basic operators (blur, noise, and FFmpeg), we follow the settings and hyperparameters in AnimeSR [34]. For VQ degradation, we keep the nearest neighbor search on the top and the middle branch while adopting the stochastic top-k VQ strategy on the bottom branch.

Stochastic Top-k VQ strategy. Specifically, for each training clip in every iteration, we uniformly sample an integer k from $[1, K]$ as the degradation level, and utilize the k_{th} nearest codebook entry to conduct the element-wise quantization on the encoded output in the bottom branch. A larger k denotes a more severe degradation, and K denotes the max degradation level in training. As shown in Tab. 5 and Fig. 9, we compare the results of animation VSR when $K = \{1, 30, 50, 100\}$ on VQD-SR (base). When trained with a small K (e.g. $K = 1$, also the nearest neighbor search), the animation VSR model has limited generalization ability due to the rigid degradation level in training. However, when K is too large (e.g. $K = 100$), the degrada-

Table 5. Ablation Study of the value K in stochastic Top-k VQ strategy.

	Top-1	Top-30	Top-50	Top-100
MANIQA↑	0.3770	0.3846	0.3857	0.3756

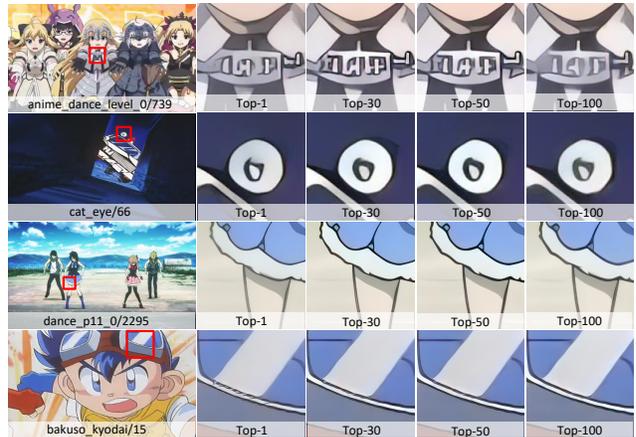


Figure 9. Ablation Study of the value K in stochastic Top-k VQ strategy. We choose $K = 50$ in our VQD-SR.

tions are so severe that contaminate the original image structures and disturb the training of VSR model. Based on the results, we finally choose $K = 50$ for our VQ degradation model, which leads to better results with sharper lines in smooth shapes. Referring to Fig. 18-19, we also show some visual examples of LR video frames degraded by multi-scale VQGAN in multi-levels with different k .

A.2. Video Super-Resolution Model

We follow the VSR model in AnimeSR [34] because of its efficiency but remove the SR feedback in the recurrent block as shown in Fig. 10. Different from natural domain videos, the continuities between animation video frames are relatively poor which causes difficulties for explicit alignment modules and further impacts the final VSR results. Thus, the explicit alignment module is left out in the architecture of animation VSR model, where the misaligned recurrent features are directly adopted, which also greatly shrinks the computation costs. However, as is studied by Chan *et al.* [3] that although long-term information is beneficial for VSR, it may suffer from error accumulation during propagation. For VSR models without explicit alignment modules, this problem could be more severe. As shown in Fig. 11, we find that too many misaligned recurrent features make the animation VSR model [34] susceptible to error accumulation, and sometimes cause a collapse of GAN loss in training, leading to corrupt VSR models. Based on this observation, we remove the SR feedback in the recurrent block to ensure more stable training. We train the VSR model on eight NVIDIA 32G V100 GPUs.

Table 6. **Architecture of Multi-scale VQGAN.** The proposed multi-scale VQGAN is composed of three parallel branches: (a) Top Branch, (b) Middle Branch, and (c) Bottom Branch. The downsample block is realized by a 3×3 convolution layer with stride 2. The upsample block is composed of a 3×3 convolution layer with stride 1 and a pixelshuffle layer. $C = 128$ is the number of base channel, $n_z = 256$ is the embedding dimension of VQ codebook.

(a) Top Branch	
Encoder	Decoder
$x \in \mathbb{R}^{H \times W \times 3}$	$z_q \in \mathbb{R}^{H/8 \times W/8 \times n_z}$
Conv2D $\rightarrow \mathbb{R}^{H \times W \times C}$	Conv2D $\rightarrow \mathbb{R}^{H/8 \times W/8 \times 4C}$
Downsample Block, $2 \times$ Residual Block $\rightarrow x_b \in \mathbb{R}^{H/2 \times W/2 \times C}$	Non-Local Block $\rightarrow \mathbb{R}^{H/8 \times W/8 \times 4C}$
Downsample Block, $2 \times$ Residual Block $\rightarrow x_m \in \mathbb{R}^{H/4 \times W/4 \times 2C}$	$2 \times$ Residual Block, Conv2D $\rightarrow \mathbb{R}^{H/8 \times W/8 \times 2C}$
Downsample Block, $2 \times$ Residual Block $\rightarrow \mathbb{R}^{H/8 \times W/8 \times 2C}$	$(+ \hat{x}_m)$, $2 \times$ Residual Block, Upsample Block $\rightarrow \mathbb{R}^{H/4 \times W/4 \times 2C}$
Conv2D, $2 \times$ Residual Block $\rightarrow \mathbb{R}^{H/8 \times W/8 \times 4C}$	$(+ \hat{x}_b)$, $2 \times$ Residual Block, Upsample Block $\rightarrow \mathbb{R}^{H/2 \times W/2 \times C}$
Non-Local Block $\rightarrow \mathbb{R}^{H/8 \times W/8 \times 4C}$	$2 \times$ Residual Block, Upsample Block $\rightarrow \mathbb{R}^{H \times W \times C}$
GroupNorm, Conv2D $\rightarrow z_t \in \mathbb{R}^{H/8 \times W/8 \times n_z}$	$2 \times$ Residual Block $\rightarrow \hat{x} \in \mathbb{R}^{H \times W \times 3}$
(b) Middle Branch	
Encoder	Decoder
$x_m \in \mathbb{R}^{H/4 \times W/4 \times 2C}$	$z_{q(m)} \in \mathbb{R}^{H/4 \times W/4 \times n_z}$
Conv2D, $2 \times$ Residual Block $\rightarrow \mathbb{R}^{H/4 \times W/4 \times 2C}$	Conv2D $\rightarrow \mathbb{R}^{H/4 \times W/4 \times 4C}$
Conv2D, $2 \times$ Residual Block $\rightarrow \mathbb{R}^{H/4 \times W/4 \times 4C}$	$2 \times$ Residual Block, Conv2D $\rightarrow \mathbb{R}^{H/4 \times W/4 \times 2C}$
GroupNorm, Conv2D $\rightarrow z_m \in \mathbb{R}^{H/4 \times W/4 \times n_z}$	$2 \times$ Residual Block, Conv2D $\rightarrow \hat{x}_m \in \mathbb{R}^{H/4 \times W/4 \times 2C}$
(c) Bottom Branch	
Encoder	Decoder
$x_b \in \mathbb{R}^{H/2 \times W/2 \times C}$	$z_{q(b)} \in \mathbb{R}^{H/2 \times W/2 \times n_z}$
Conv2D, $2 \times$ Residual Block $\rightarrow \mathbb{R}^{H/2 \times W/2 \times 2C}$	Conv2D $\rightarrow \mathbb{R}^{H/2 \times W/2 \times 4C}$
Conv2D, $2 \times$ Residual Block $\rightarrow \mathbb{R}^{H/2 \times W/2 \times 2C}$	$2 \times$ Residual Block, Conv2D $\rightarrow \mathbb{R}^{H/2 \times W/2 \times 2C}$
Conv2D, $2 \times$ Residual Block $\rightarrow \mathbb{R}^{H/2 \times W/2 \times 4C}$	$2 \times$ Residual Block, Conv2D $\rightarrow \mathbb{R}^{H/2 \times W/2 \times 2C}$
GroupNorm, Conv2D $\rightarrow z_b \in \mathbb{R}^{H/2 \times W/2 \times n_z}$	$2 \times$ Residual Block, Conv2D $\rightarrow \hat{x}_b \in \mathbb{R}^{H/2 \times W/2 \times C}$

A.3. Experiment Details

Evaluation with MANIQA. We test the super-resolution methods on AVC-ReallQ [34], which is a real-world animation video dataset containing 46 low-quality clips with 100 frames per clip. As the lack of ground truths for testing, we follow AnimeSR [34] and adopt the no-reference image quality evaluation metric MANIQA [38] to evaluate the final SR results. MANIQA is designed for rating GAN-based distorted images, which is suitable for evaluating GAN-based image restoration algorithms. Following [38] and [34], we measure MANIQA every 10 frames in each video. For each testing frame, we randomly crop 224×224 sized images 20 times and calculate the average score. We run the evaluation process 3 times and report the mean metrics for all of the reported results.

User Study on Visual Quality. We conduct A-B tests to further compare the visual quality of VQD-SR with other six SOTA methods (BasicVSR [1], PDM [21], RealESRGAN [31], BSRGAN [42], RealBasicVSR [3], AnimeSR [34]) successively. For each comparison with one of the six methods, there are 20 subjects involved in the tests on the SR results of AVC-ReallQ [34] with 46 animation video clips. Considering the subtle qualitative differences between frames, we uniformly sample 4 frames for testing in each clip. The user interface for the A-B test, as shown in

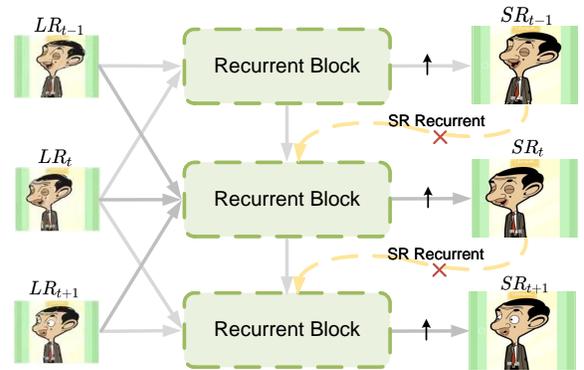


Figure 10. **Architecture of VSR model.** We follow the VSR model in AnimeSR [34] but remove the SR feedback in the recurrent block for stable training.

Fig. 12, provides the users with two images in random order which include one VQD-SR frame and one frame from the other method. Users are asked to select one with higher visual quality. As the resolutions of SR results are too large (e.g. 5760×4320) to fit the screen and display the details at the same time if only providing the complete images, we further show two side-by-side zoom-in windows controlled by mouse with adjustable positions and sizes. The final results are the percentages of votes which prefer VQD-SR to



Figure 11. Too many recurrent features cause unstable training leading to corrupt animation VSR models.



Figure 12. UI of the A-B test. The main window shows the complete image with the current testing progress and frame name beneath. Two sub-windows on the right display the zoom-in details of two methods where the mouse hovers, and the size of the viewport can also be adjusted by scrolling the mouse wheel. Users are asked to select the one with higher visual quality by pressing ‘1’ or ‘2’ on the keyboard.

other methods.

B. Statistics and Samples of RAL

Our Real Animation Low-quality (RAL) video dataset contains over 10K LR frames extracted from 441 real-world low-quality animation videos and contains rich real-world degradations in animation domain. The statistics of RAL is shown in Fig. 13. In Fig. 14, we also show some representative samples and typical real-world degradations in RAL.

C. More Qualitative Comparisons

In this section, we show more qualitative results to verify the effectiveness of our proposed methods.

Animation VSR Results. In Fig. 20 - 22, we compare our VQD-SR with six SOTA SR methods. Our VQD-SR is capable to recover visually natural and sharper lines (Fig. 20, Fig.22 row 1-2) with fewer artifacts, restore clear details (Fig. 21), handle intense scenarios (*e.g.* background blur) without over-sharp artifacts (Fig. 22 row 3-4).

HR-SR Enhancement. In Fig. 16, we show the enhanced HR animation video frames with different SR models. Our HR-SR enhancement strategy could alleviate the compression artifacts and sharpen the edges without contaminating

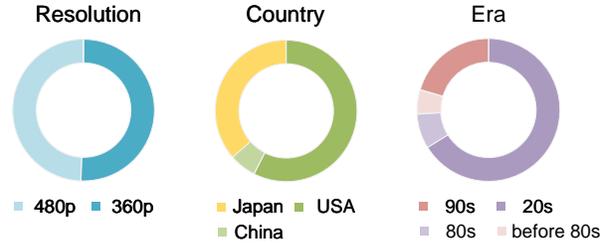


Figure 13. Statistics of RAL.

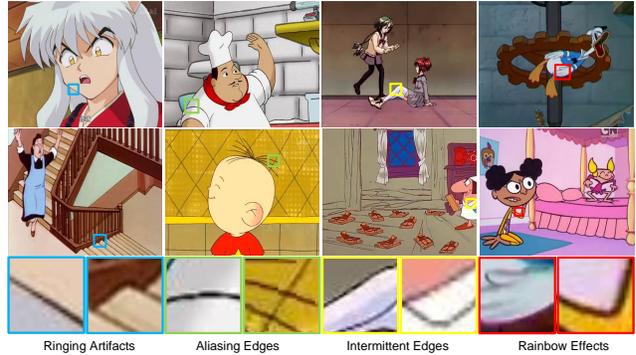


Figure 14. Samples of RAL and typical degradation phenomena in real-world LR animation videos.

the original details in animation HR frames. As shown in Fig. 23, the proposed HR-SR strategy is valid to improve the results of animation VSR, regardless of the specific VSR model, with the help of more effective ground truths for training.

We further extend the HR-SR enhancement strategy from animation videos to natural videos (REDS [24]) in Fig. 17. Because of the complex textures and irregular illumination conditions, directly adopting HR-SR enhancement strategy to natural videos would cause amplified illumination artifacts (row 1), contaminated details (row 2), and over-sharp textures (row 3), leading to unappealing results.

D. Limitation and Discussion

In this section, we visualize the failure cases of VQD-SR in Fig. 15. 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.

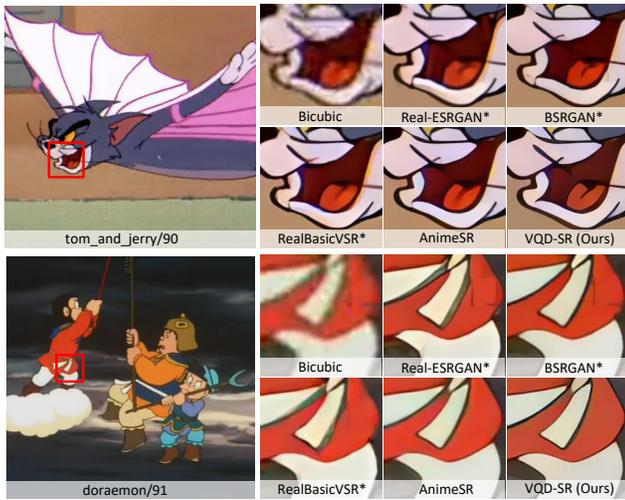


Figure 15. **Failure cases** when LR frames endure severe color distortions around edges.

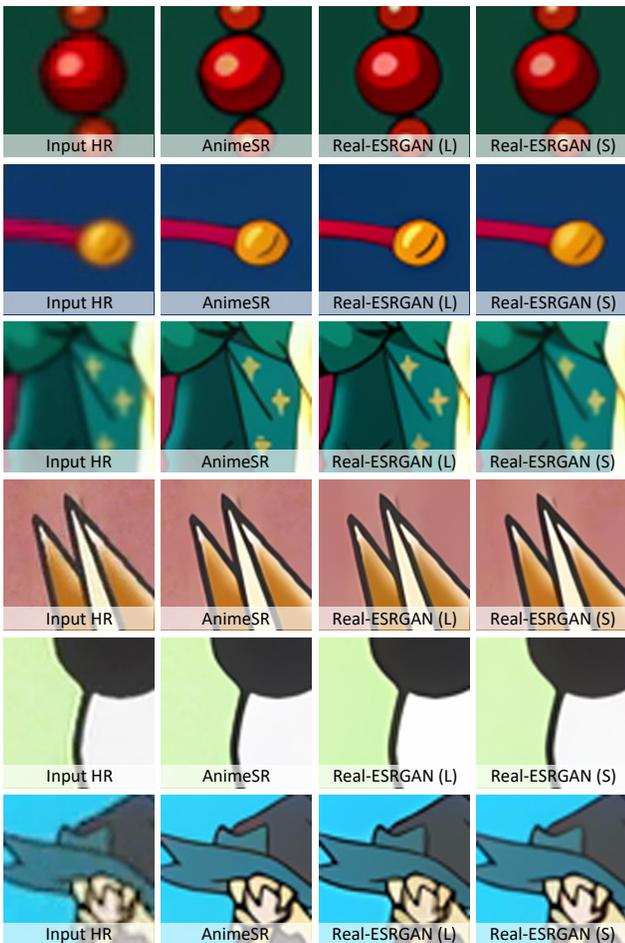


Figure 16. Visual comparison of different SR models for HR-SR enhancement in animation domain.

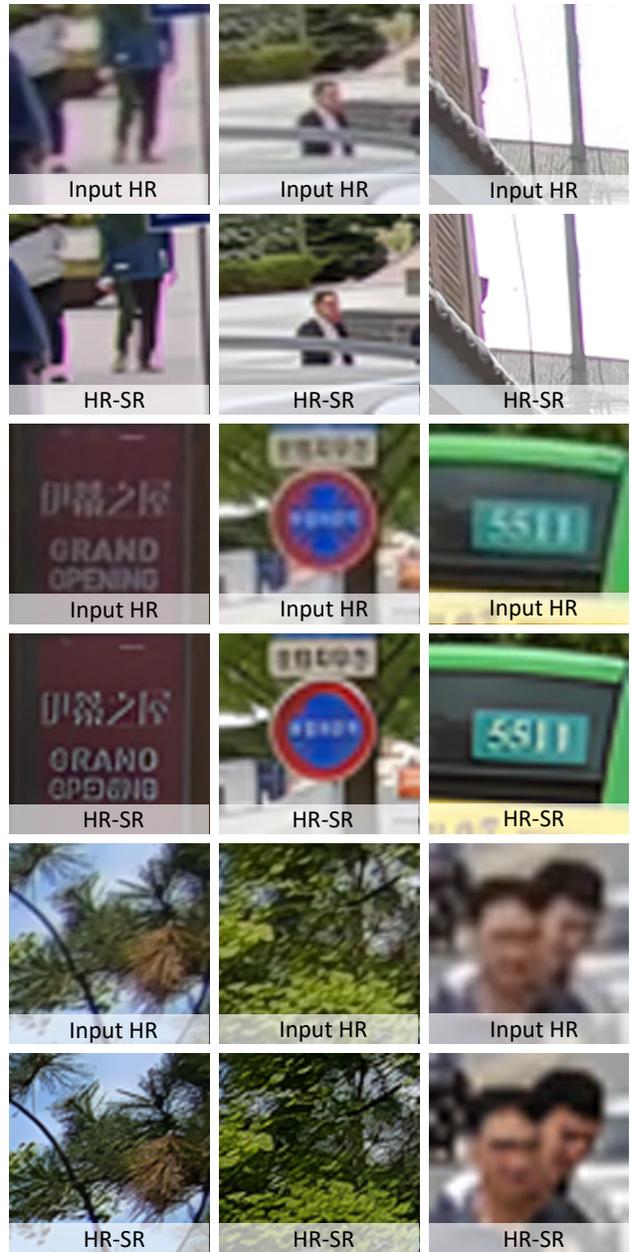


Figure 17. Extend HR-SR enhancement strategy to natural videos. Complex textures and irregular illumination conditions lead to amplified illumination artifacts (row 1), contaminated details (row 2), and over-sharp textures (row 3).

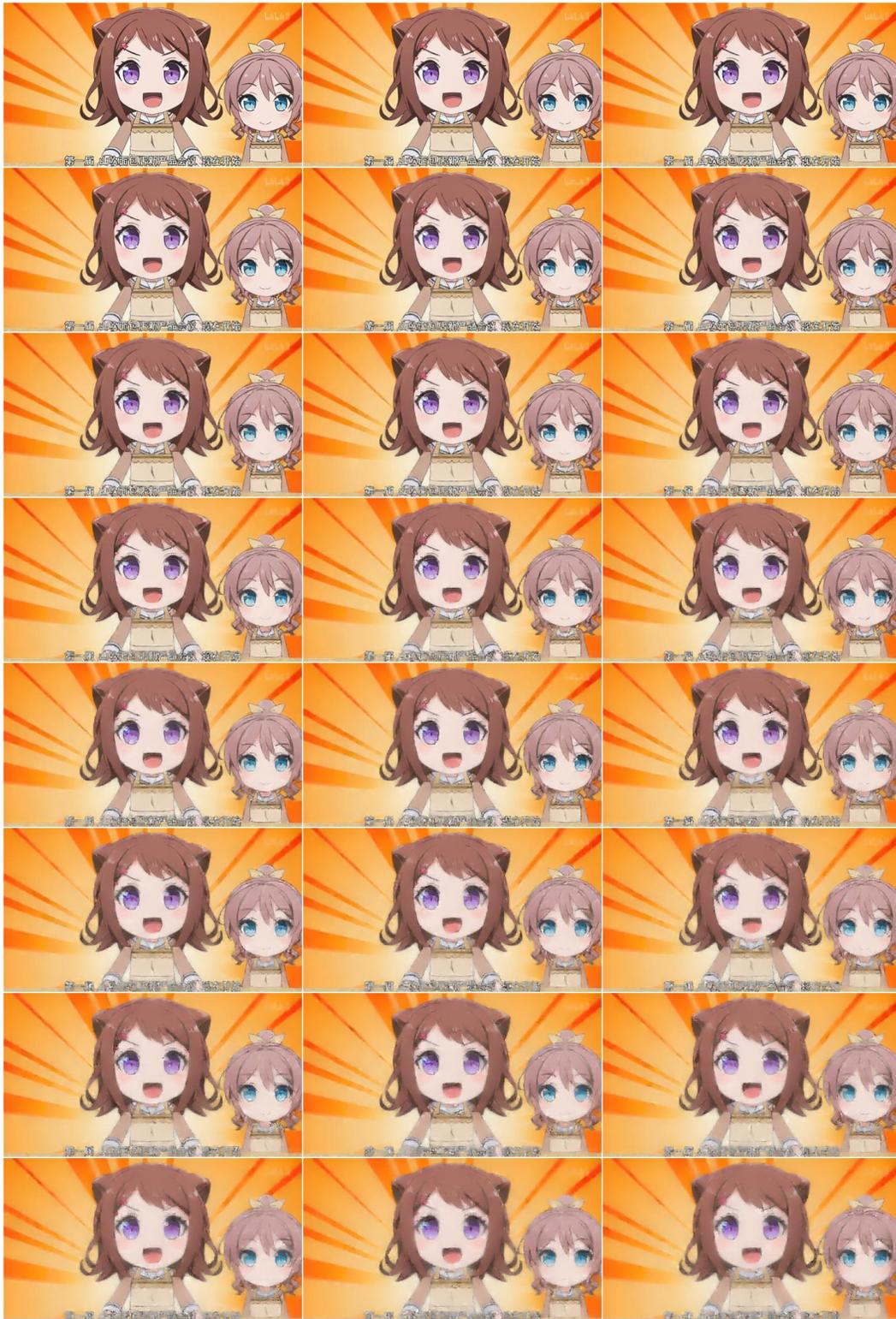


Figure 18. **Examples of LR frames** degraded by multi-scale VQGAN in multi-levels. From top to bottom, left to right, we show the degradation levels in ascending order. For the sake of clear comparisons, we show the results every 3 levels in the first 70 levels ($k = 1, 4, 7, \dots, 70$) here. Input HR: [AVC-Train/bang_dream.p1_0/00000049.png](#)

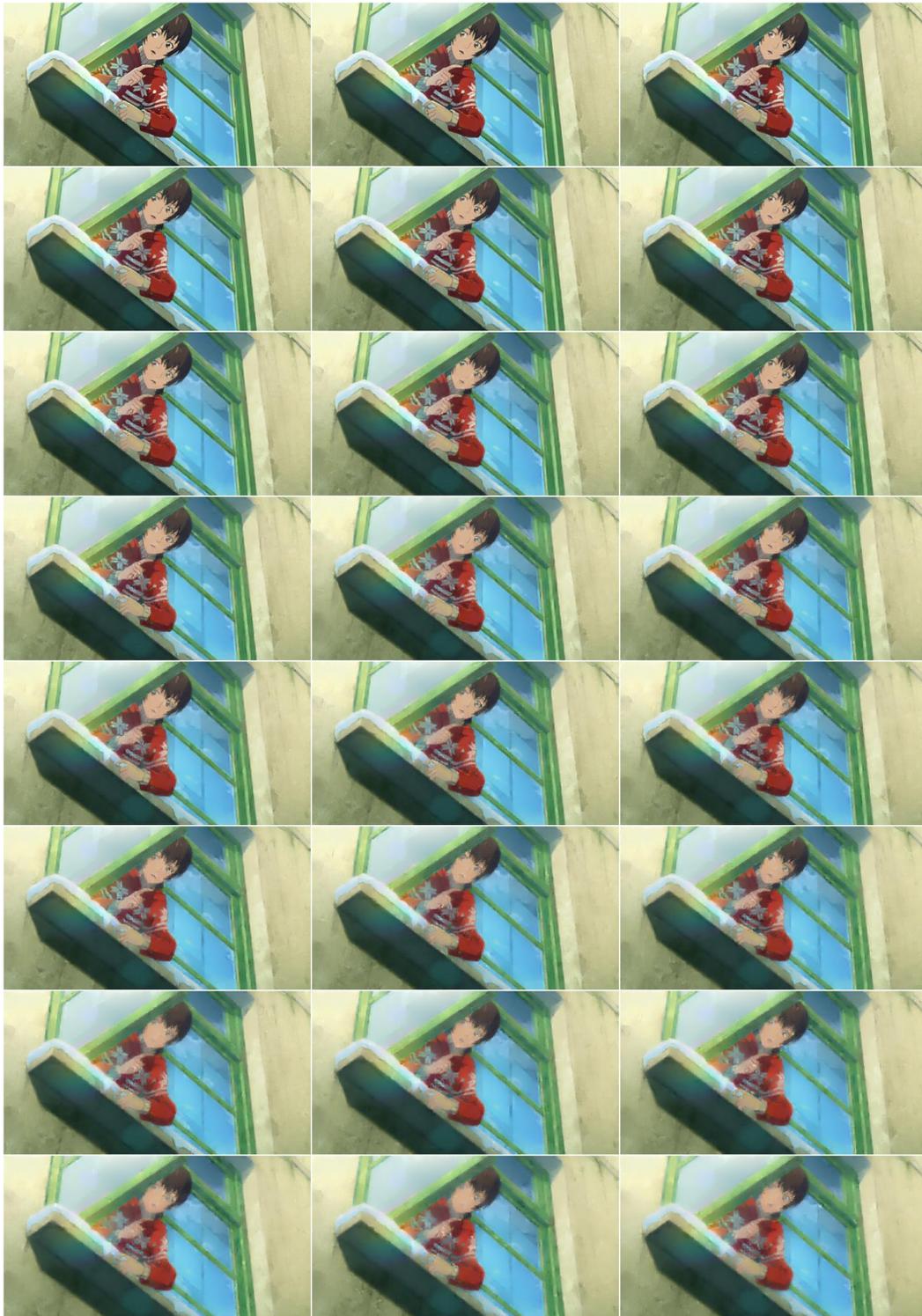


Figure 19. **Examples of LR frames** degraded by multi-scale VQGAN in multi-levels. From top to bottom, left to right, we show the degradation levels in ascending order. For the sake of clear comparisons, we show the results every 3 levels ($k = 1, 4, 7, \dots, 70$) here. Input HR: [AVC-Train/b0034m9uleq.10005_movie001_0/00000049.png](#)

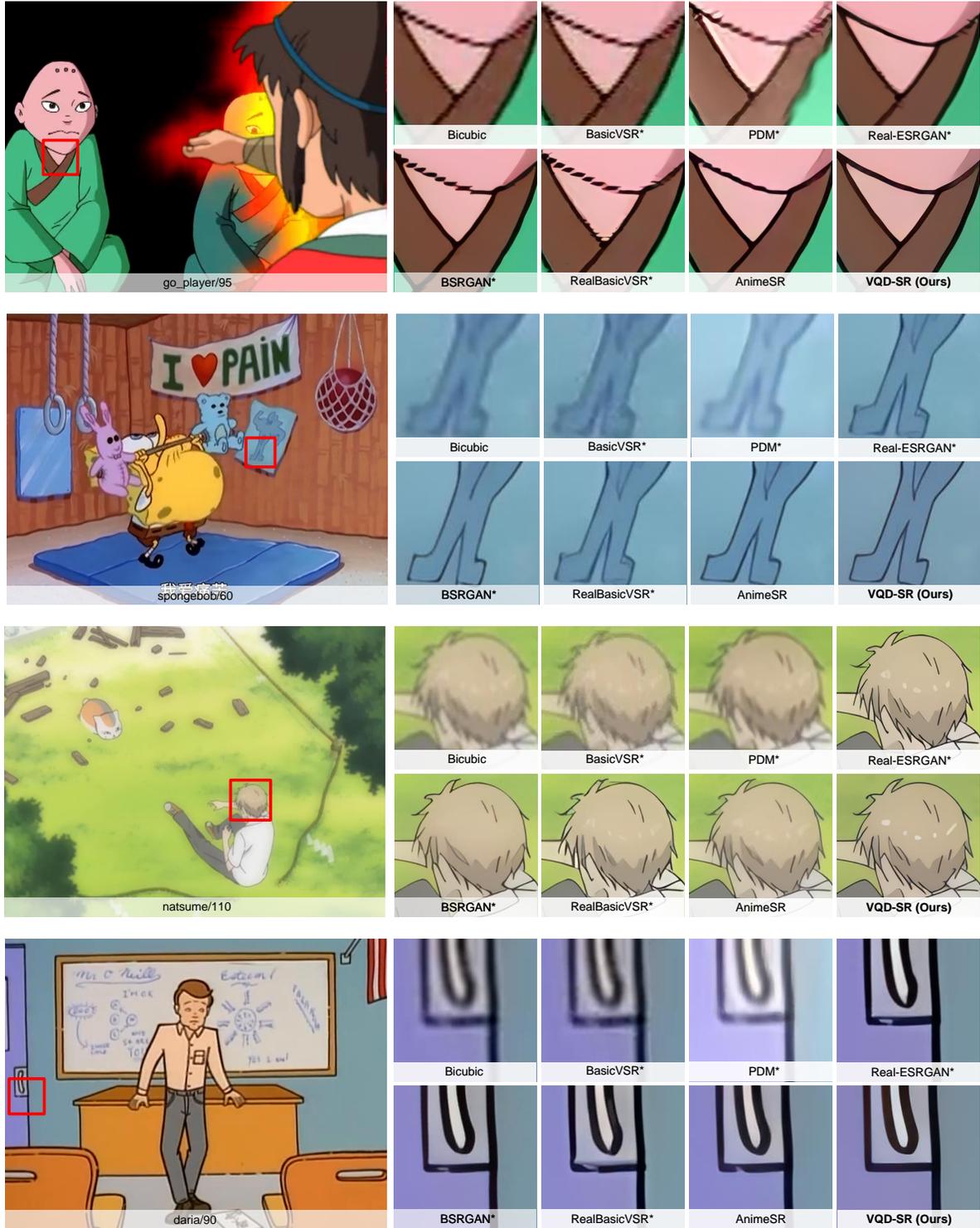


Figure 20. **Qualitative comparisons with SOTA methods.** ‘*’ denotes fine-tune on animation dataset AVC-Train [34]. Our VQD-SR is capable to recover visually natural and sharper lines with fewer artifacts.

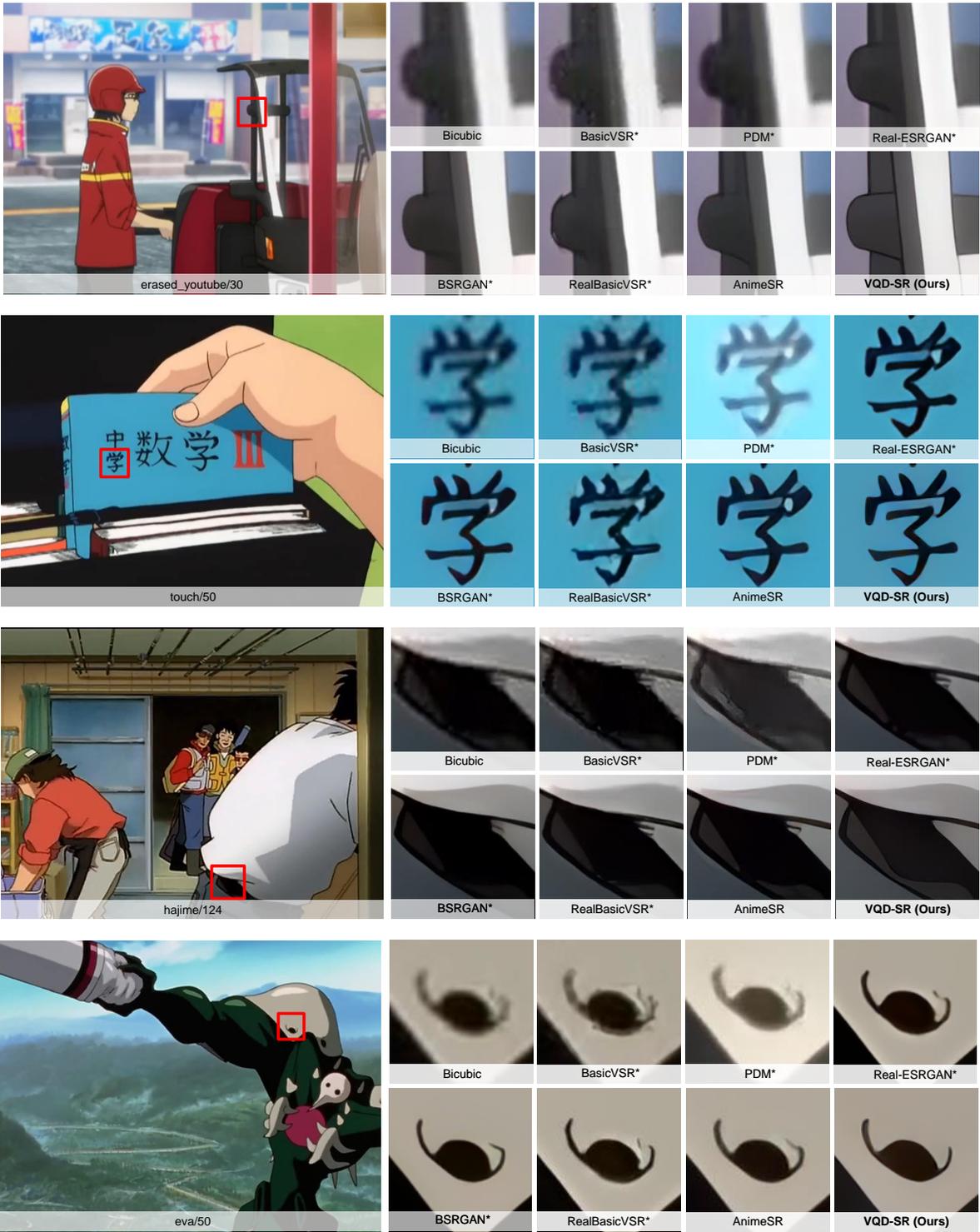


Figure 21. **Qualitative comparisons with SOTA methods.** ‘*’ denotes fine-tune on animation dataset AVC-Train [34]. Our VQD-SR is capable to restore clear details.



Figure 22. **Qualitative comparisons with SOTA methods.** ‘*’ denotes fine-tune on animation dataset AVC-Train [34]. Our VQD-SR is capable to recover visually natural and sharper lines with fewer artifacts (row 1-2) and handle intense scenarios (e.g. background blur) without over-sharp artifacts (row 3-4)

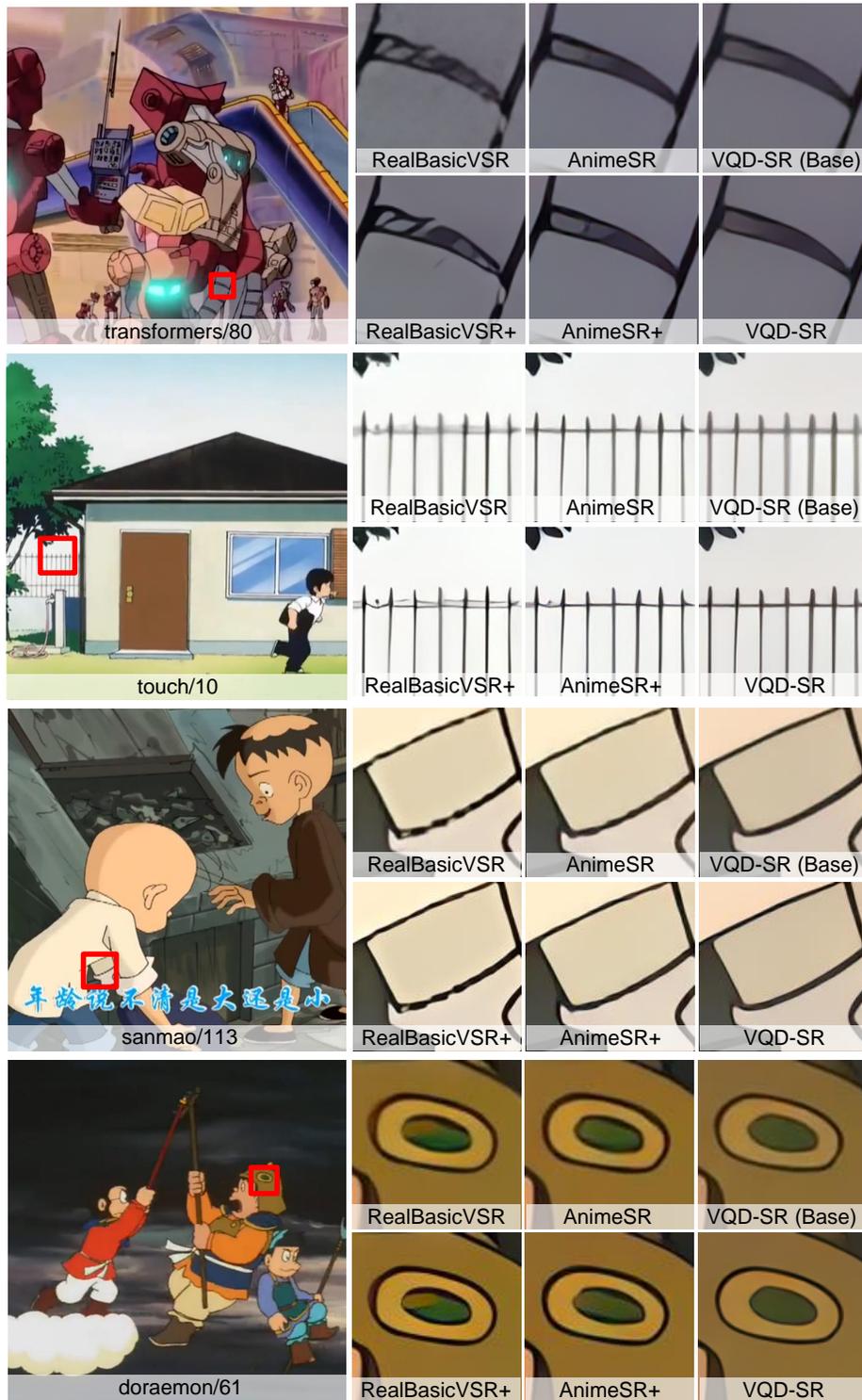


Figure 23. Ablation study of HR-SR enhancement for different VSR methods.