APISR: Anime Production Inspired Real-World Anime Super-Resolution

Supplementary Material

In this supplementary material, Sec. A first presents more statistics and details of our proposed anime image SR training dataset. Then, Sec. B shows details about our implementations in super-resolution (SR) network training. Specifically, Sec. B.1 presents the image SR network we used in our training. Sec. B.2 presents details of post-processing techniques we use on the pseudo-GT preparation for hand-drawn line enhancement. Sec. B.3 presents figures and details of the ResNet50 [10] perceptual loss for our proposed balanced twin perceptual loss. Sec. B.4 provides the hyperparameter setting for our proposed prediction-oriented compression and shuffled resize module in the degradation model. Finally, Sec. C provides more visual results of comparisons among SOTA methods and ablation studies.

A. API Dataset Details

Our **A**nime **P**roduction-oriented **I**mage (API) SR dataset contains 3,740 high-quality and informative images. This quantity is roughly the same quantity as the previous photorealistic SR training dataset size [26, 29], which includes DIV2K [1], Flickr2K [20], and OutdoorSceneTraining [24]. The aspect ratio and resolution information before scaling are shown in Fig. 1.

B. Implementation Details

B.1. Training Network Details

The generator network we deploy is GRL [14], a SOTA image SR network (CVPR 2023). GRL leverages interconnected relationships within various layers of image structures through a Transformer-based framework, attaining improvement in multiple tasks of SR and image restoration. The model we chose is its tiny version, which has 0.91M parameters. To better adapt the real-world SR task, we changed its upsampler module from the default pixel shuffle strategy to the nearest neighbor interpolation with the convolution layer approach, which is used for the base model version but not for the tiny version in their proposed methods. We change the upsampler because the nearest neighbor interpolation with the convolution layer is claimed to show fewer artifacts in the upsampling process than the pixel shuffle strategy. The final network parameter is 1.03M, which is the smallest network among all image and video-based SOTA methods that we compare.

B.2. Hand-drawn line enhancement Details

In the hand-drawn line enhancement, we have proposed outlier filter and passive dilate techniques to obtain a clean XDoG-extracted [27] hand-drawn line edge map. XDoG is widely used in paired dataset preparation in anime colorization [4, 5, 11, 23]. The extracted edge map by XDoG is a binary output, where the white pixel stands for the active edge map region and the black pixel stands for the unrelated region.

For the outlier filter, we use breadth-first search in eight directions to recursively detect the surrounding pixels of all white pixels and turn white pixel regions into black pixels if the total quantity of connected white pixels is less than the threshold. We empirically set the threshold as 32.

For the dilation, we passively replace the black pixel with the white pixel if it has more than 3 white pixel neighbors, which is different from independent kernel-based active dilation methods in [7, 9, 13] that directly spread the surrounding neighbors to be white pixels if the central pixel is white. Compared to active dilation methods, our proposed passive dilation is more concentrated on the hand-drawn lines region instead of covering unrelated pixel information (see Fig. 3). Thus, we name our methods as passive dilatation.

In the implementation, we will do an unsharp mask for the whole image first to increase overall visualization sharpness and then apply two extra turns of sharpening to the hand-drawn lines specifically based on the pipeline design mentioned above. More implementation details can be found in our released code.

B.3. Balanced Twin Perceptual Loss Details

As shown in Fig. 2, our proposed middle-layer output comparisons for ResNet50 [10] follow the idea proposed by ESRGAN [25] which compares feature map outputs before the activation layer. Following VGG-based perceptual loss [12], we compare the last convolution layer of each stage. There are five middle-layer output comparisons, which are the same quantity as VGG-based perceptual loss [12]. Thus, our proposed twin perpetual loss reaches a mutual balance in training.

B.4. Degradation Details

For the prediction-oriented compression module of the degradation model, we deploy both the image compression with prediction mechanism (*i.e.*, WebP [17] and AVIF [8]) and single-frame video compression. Meanwhile, for the robustness of the degradation model, we keep the

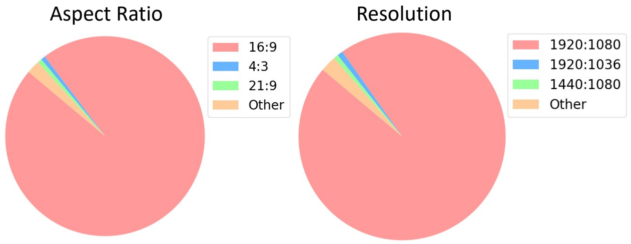

Figure 1. **API dataset extra statistics.**

JPEG [22]. The quality factor range of JPEG, WebP, and AVIF is $[20, 95]$ with encoding speed in the range of $[0, 6]$ for WebP and AVIF. The probability of fetching the value in the range is equal.

For the stability of video compression processing, we choose the widely-used video processing tools, *ffmpeg*, to perform the proposed single-frame compression of MPEG2 [15], MPEG4 [2], H.264 [16], and H.265 [19]. In *ffmpeg*, CRF is an engineering system to control the quantization level, and preset is a speed control mechanism whose setting is directly related to compression distortions. For MPEG2 and MPEG4, we empirically find that the quality factor control (*-qscale:v*) is a better way to control single-frame compression, but for H.264 and H.265, CRF is a better way to control. For MPEG2 and MPEG4, we set the quality factor in the range $[8, 31]$. For H.264 and H.265, we set the CRF in the range $[23, 38]$ and $[28, 42]$ respectively. The preset for all of them is $\{slow, medium, fast, faster, superfast\}$ with probability $\{0.05, 0.35, 0.3, 0.2, 0.1\}$.

The first prediction-oriented compression includes JPEG [22] and WebP [17] with a probability of $\{0.4, 0.6\}$ respectively. The second prediction-oriented compression includes JPEG [22], WebP [17], AVIF [8], and single-frame compression of MPEG2 [15], MPEG4 [2], H.264 [16], and H.265 [19] with probability of $\{0.06, 0.1, 0.1, 0.12, 0.12, 0.3, 0.2\}$ respectively. For the first resize module, we set the scaling in the range of $[0.1, 1.2]$ with probability $\{0.2, 0.7, 0.1\}$ to scale up, scale down, or remain current resolution. For the second resize module, we choose the range of $[0.15, 1.2]$ with probability $\{0.2, 0.7, 0.1\}$. More implementation details can be found in our released code.

C. More Qualitative Comparisons

In this section, we present more qualitative results to verify the effectiveness of our APISR among SOTA methods. Moreover, we provide visual comparisons for the ablation studies.

Extra Qualitative Comparisons with SOTA methods. Fig. 4 and Fig. 5 show extra qualitative comparisons on AVC-RealLQ [28] datasets for $4\times$ scaling. This includes image-based Real-ESRGAN [26] and BSRGAN [29], and video-based RealBasicVSR [6], AnimeSR [28], and VQD-SR [21]. Our APISR presents clearer and sharper hand-drawn lines (first example of Fig. 4, first and second examples of Fig. 5, and third example of Fig. 6), better restoration with more natural details (second and third examples of Fig. 4, and third example of Fig. 5), and does not present unwanted color artifacts (first and second examples of Fig. 6).

Qualitative Comparisons of Ablation Studies. Fig. 7, Fig. 8, and Fig. 9 shows the qualitative comparisons of ablations studies.

As shown in Fig. 7, the network trained with AVC-Train [28] over-sharpens the grid texture and produces annoying artifacts as denoted by the arrows in the figure. Similarly, the network trained with the random sampled or IQA-based sampled dataset can alleviate this artifact but is still hard to completely remove it. However, when we introduce the ICA-based selection method with I-Frame dataset collection, this artifact is greatly removed and the generated image shows more natural details. This is thanks to versatile complex scenes included in the dataset due to ICA-based selection. With 720P rescaling, fewer ringing artifacts appear.

As shown in Fig. 8, the network trained with high-order [26] and random order [29] degradation model presents ringing artifacts, rainbow effects, and color distortions as denoted by the arrows in the figure. Nevertheless, introducing our proposed prediction-oriented compression module in the degradation model promotes the network to greatly restore these problems and generate more natural details with less distorted hand-drawn lines. Moreover, with the shuffled resize module in the degradation model, more distortions are restored and present natural shadow details.

As shown in Fig. 9, the network trained with the plain version presents unwanted color pixel artifacts and sparse hand-drawn line information as denoted by the arrows in the figure. With the hand-drawn line enhancement, the hand-drawn line around the eyes of the character is greatly intensified and more details are generated. However, the unwanted color pixels still exist and they are presented as an annoying artifact. With the twin perceptual loss, the unwanted color pixels are greatly alleviated. Further, with the scaling to early layers in ResNet perceptual loss, more shadow artifacts and distortions are restored.

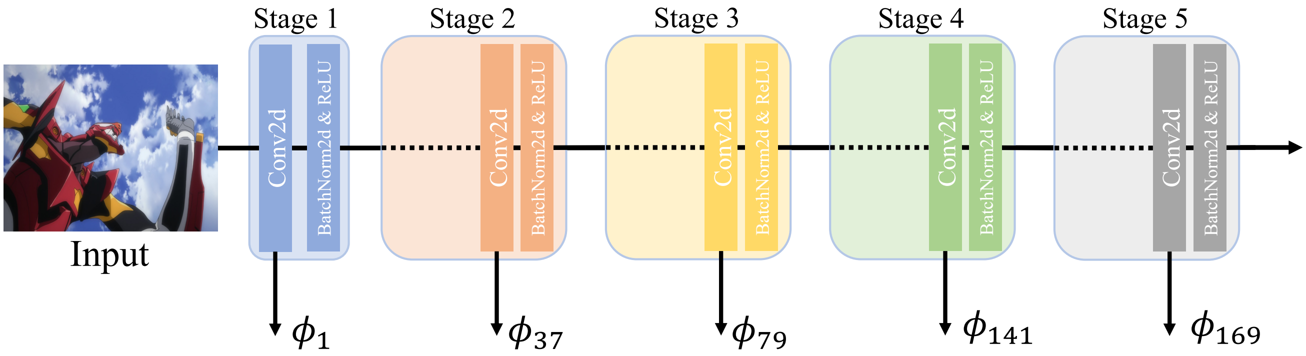

Figure 2. The overview of our proposed middle-layer outputs of ResNet50 [10] perceptual loss trained by Danbooru dataset [3]. Overall, ResNet50 can be summarized into five stages which is similar to VGG [18]. ϕ_j represents the perceptual function that returns jth layer output of ResNet50.

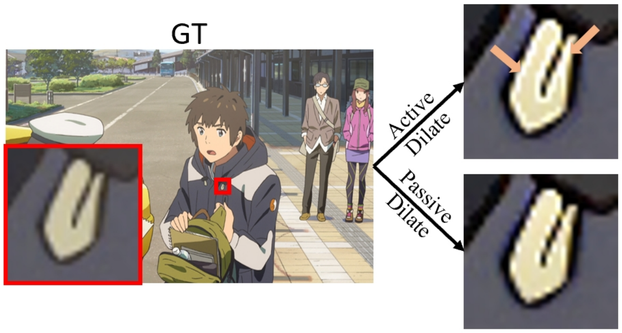

Figure 3. **Comparisons between active and passive dilation.** Our proposed passive dilation is more concentrated on the hand-drawn line region without producing over-sharpened pseudo-GT images as in active dilation methods.

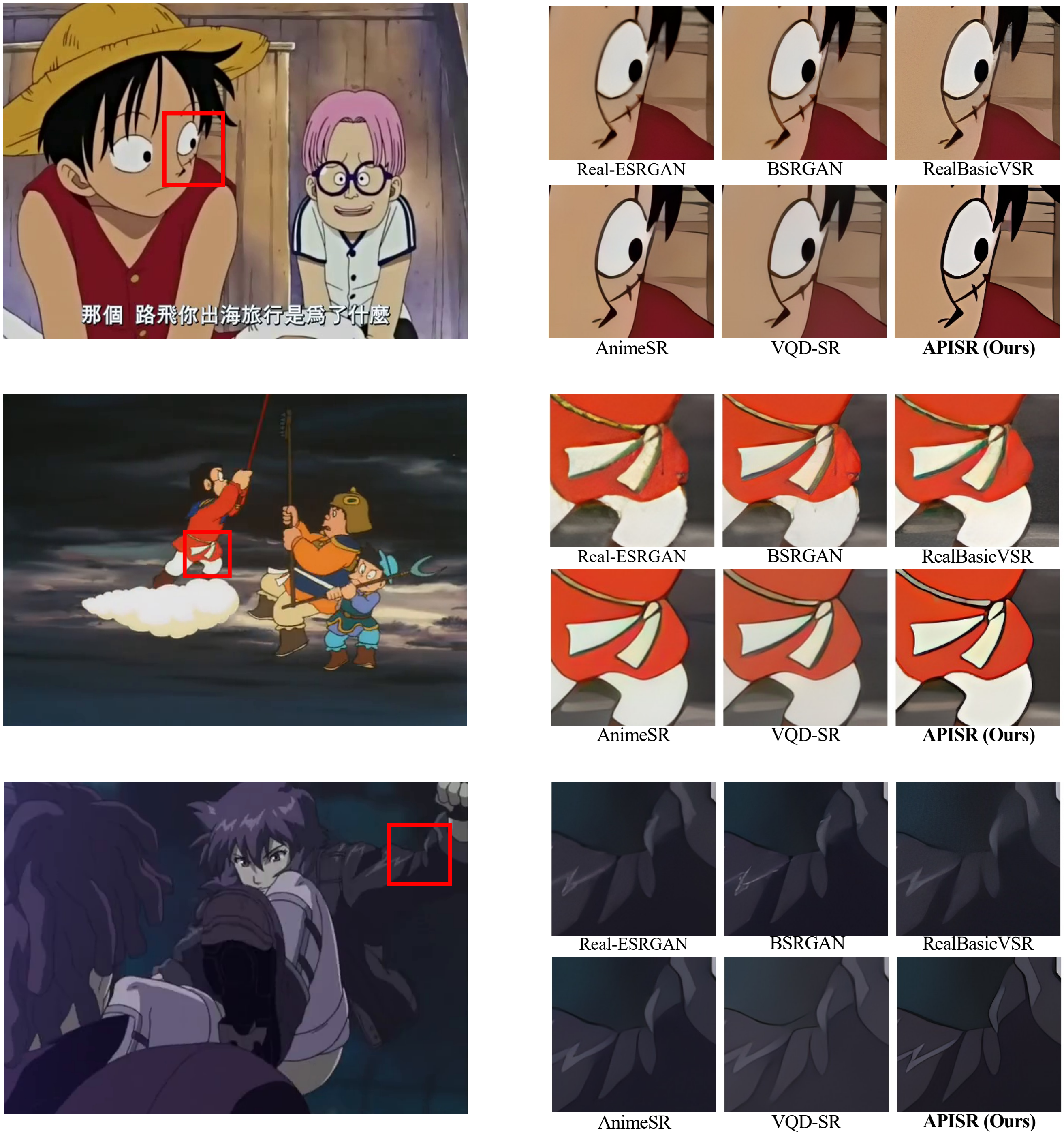

Figure 4. Qualitative comparisons on AVC-RealLQ [28] for 4× scaling. Our APISR presents clearer and sharper hand-drawn lines, better restoration with more natural details, and does not present unwanted color artifacts. **Zoom in for the best view.**

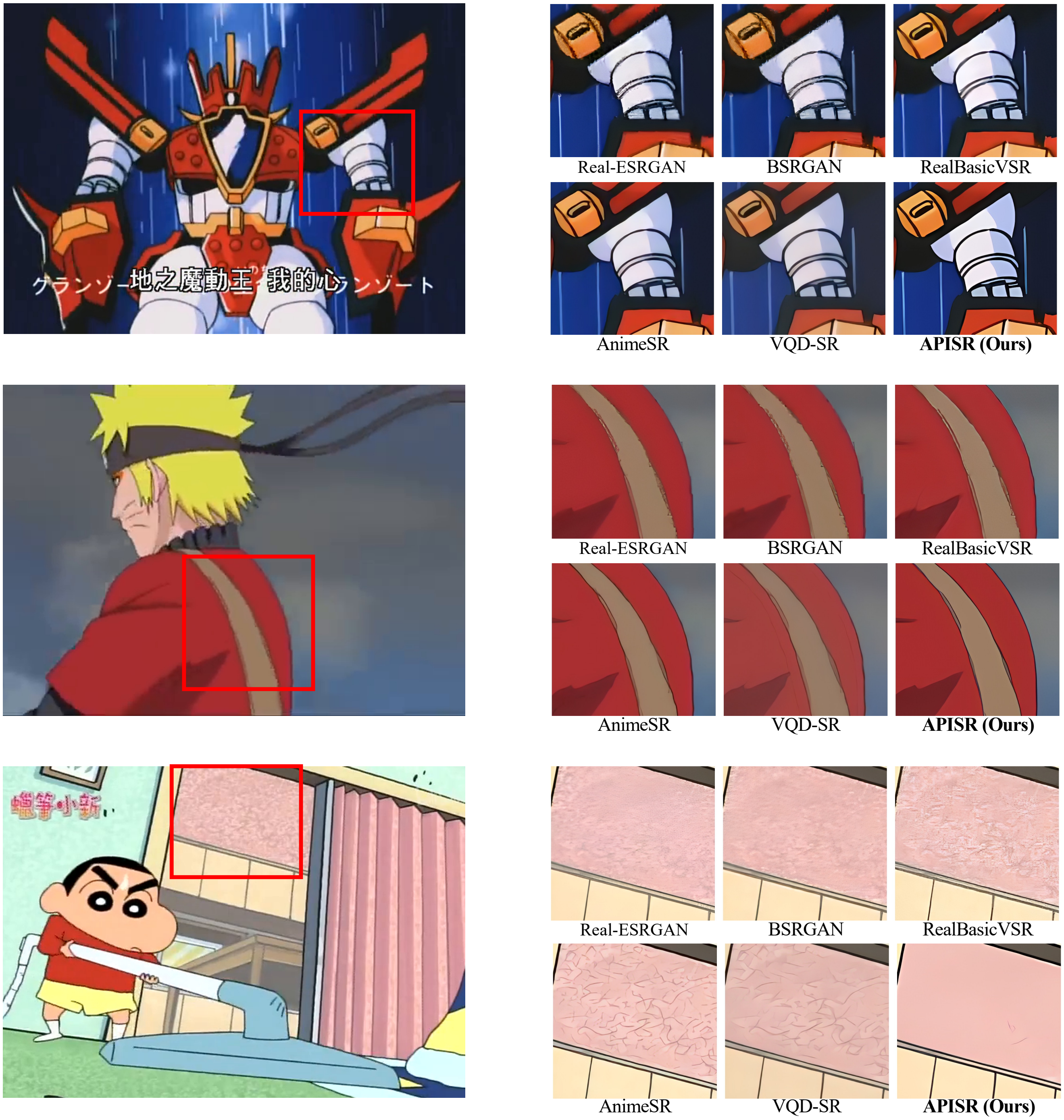

Figure 5. Qualitative comparisons on AVC-RealLQ [28] for 4× scaling. Our APISR presents clearer and sharper hand-drawn lines, better restoration with more natural details, and does not present unwanted color artifacts. **Zoom in for the best view.**

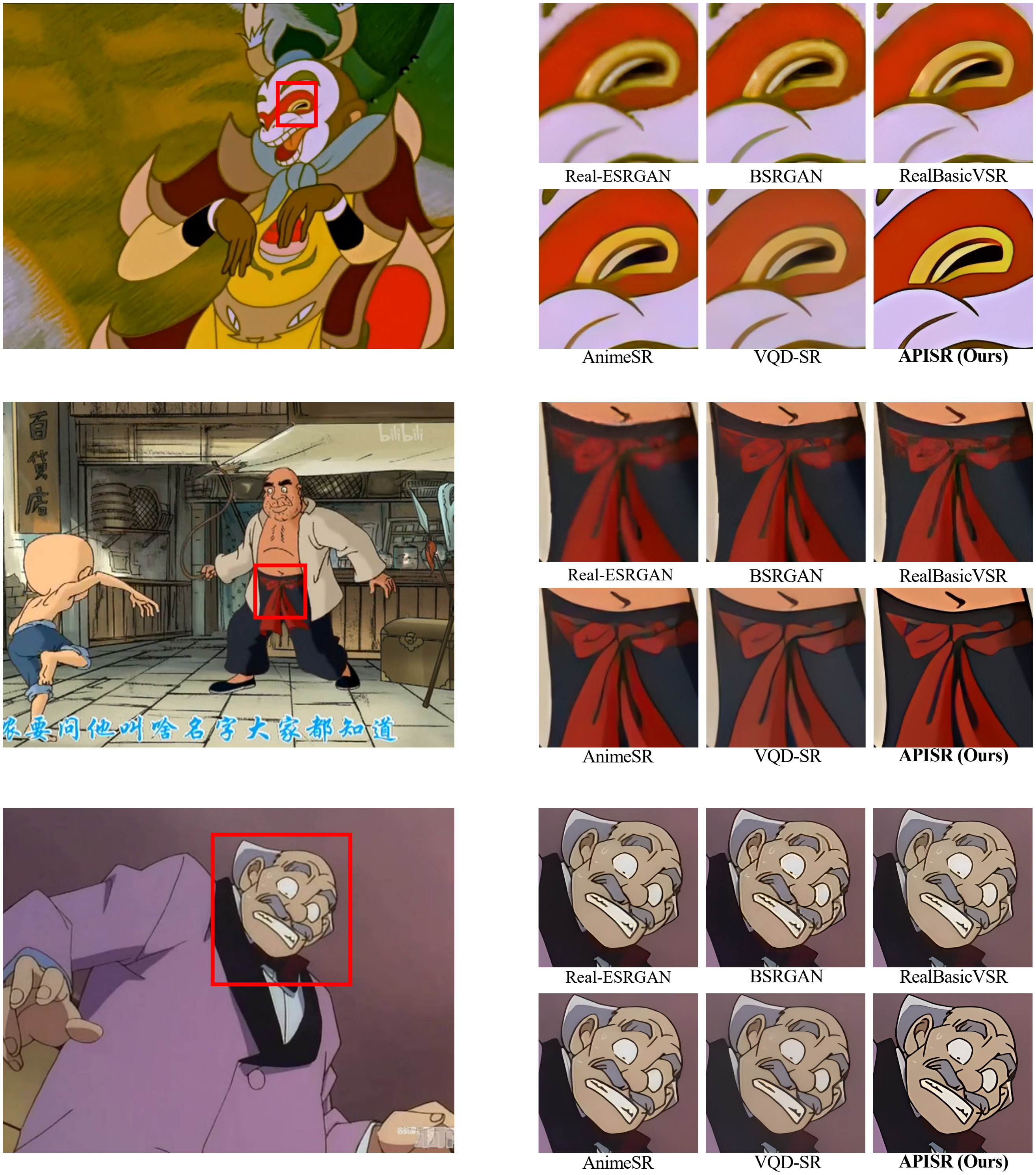

Figure 6. Qualitative comparisons on AVC-RealLQ [28] for 4× scaling. Our APISR presents clearer and sharper hand-drawn lines, better restoration with more natural details, and does not present unwanted color artifacts. **Zoom in for the best view.**

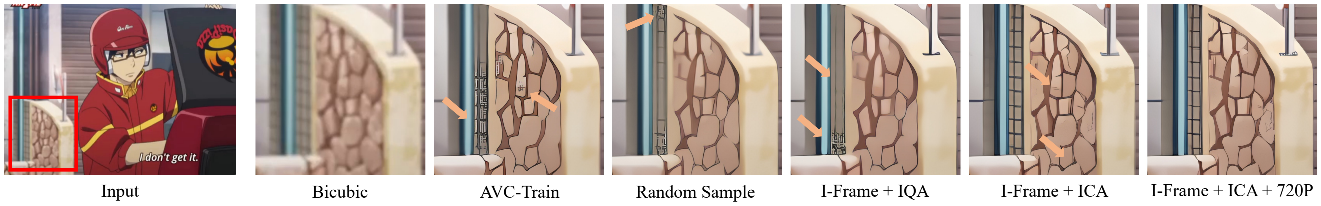

Figure 7. **Qualitative comparisons of the first ablation study.** IQA stands for image quality assessment. ICA stands for image complexity assessment. 720P stands for our proposed 720P rescaling. **Zoom in for the best view.**

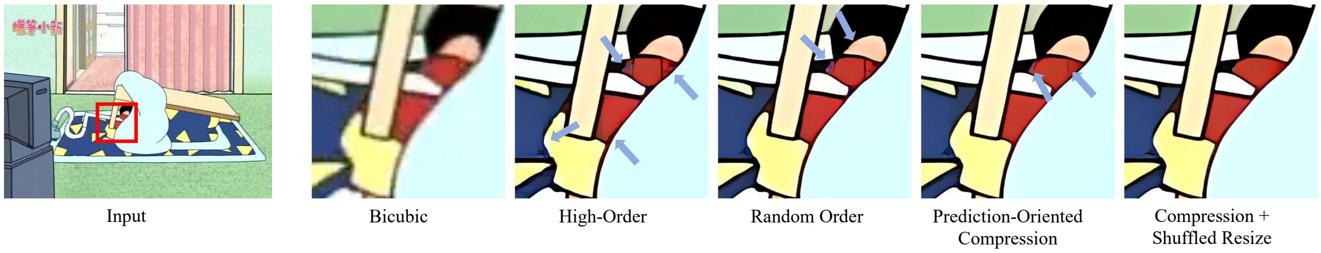
Input
Bicubic
High-Order
Random Order
Prediction-Oriented Compression
Compression + Shuffled Resize

Figure 8. **Qualitative comparisons of the second ablation study. Zoom in for the best view.**

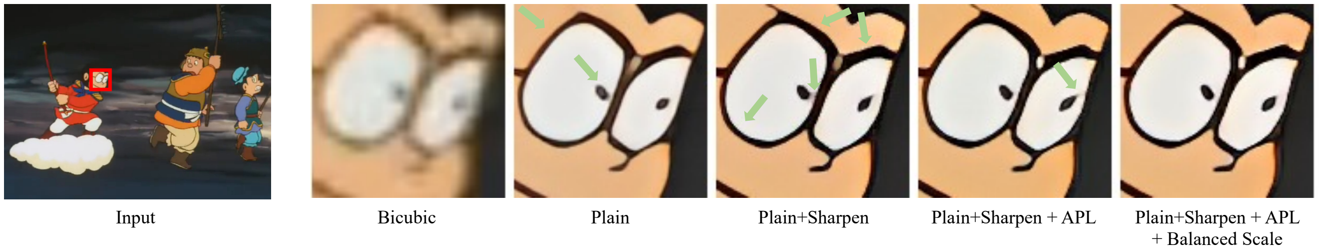

Figure 9. **Qualitative comparisons of the third ablation study.** Hand-drawn lines enhancement is denoted as **Sharpen** and twin perceptual loss is denoted as **APL**. **Balanced Scale** presents the early layer scaling to ResNet perceptual loss. **Zoom in for the best view.**

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 1

[2] Olivier Avaro, Alexandros Eleftheriadis, Carsten Herpel, Ganesh Rajan, and Liam Ward. Mpeg-4 systems: overview. *Signal Processing: Image Communication*, 15(4-5):281–298, 2000. 2

[3] Matthew Baas. Danbooru2018 pretrained resnet models for pytorch. https://rf5.github.io, 2019. Accessed: DATE. 3

[4] Yu Cao, Xiangqiao Meng, PY Mok, Xueting Liu, Tong-Yee Lee, and Ping Li. Animediffusion: Anime face line drawing colorization via diffusion models. *arXiv preprint arXiv:2303.11137*, 2023. 1

[5] Hernan Carrillo, Michaël Clément, Aurélie Bugeau, and Edgar Simo-Serra. Diffusart: Enhancing line art colorization with conditional diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3485–3489, 2023. 1

[6] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 2

[7] Yuanzheng Ci, Xinzhu Ma, Zhihui Wang, Haojie Li, and Zhongxuan Luo. User-guided deep anime line art colorization with conditional adversarial networks. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1536–1544, 2018. 1

[8] Jingning Han, Bohan Li, Debargha Mukherjee, Ching-Han Chiang, Adrian Grange, Cheng Chen, Hui Su, Sarah Parker, Sai Deng, Urvang Joshi, et al. A technical overview of av1. *Proceedings of the IEEE*, 109(9):1435–1462, 2021. 1, 2

[9] Yliess Hati, Gregor Jouet, Francis Rousseaux, and Clément Duhart. Paintstorch: a user-guided anime line art colorization tool with double generator conditional adversarial network. In *Proceedings of the 16th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–10, 2019. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 1, 3

[11] Zhengyu Huang, Haoran Xie, Tsukasa Fukusato, and Kazunori Miyata. Anifacedrawing: Anime portrait exploration during your sketching. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 1

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 1

[13] Yeongseop Lee and Seongjin Lee. Automatic colorization of anime style illustrations using a two-stage generator. *Applied Sciences*, 10(23):8699, 2020. 1

[14] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023. 1

[15] Joan L Mitchell, William B Pennebaker, Chad E Fogg, Didier J LeGall, Joan L Mitchell, William B Pennebaker, Chad E Fogg, and Didier J LeGall. Mpeg-2 overview. *MPEG Video Compression Standard*, pages 171–186, 1996. 2

[16] Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. Overview of the scalable video coding extension of the h. 264/avc standard. *IEEE Transactions on circuits and systems for video technology*, 17(9):1103–1120, 2007. 2

[17] Zhanjun Si and Ke Shen. Research on the webp image format. In *Advanced graphic communications, packaging technology and materials*, pages 271–277. Springer, 2016. 1, 2

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[19] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 2

[20] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 1

[21] Zixi Tuo, Huan Yang, Jianlong Fu, Yujie Dun, and Xueming Qian. Learning data-driven vector-quantized degradation model for animation video super-resolution. *arXiv preprint arXiv:2303.09826*, 2023. 2

[22] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1): xviii–xxxiv, 1992. 2

[23] Ning Wang, Muyao Niu, Zhi Dou, Zhihui Wang, Zhiyong Wang, Zhaoyan Ming, Bin Liu, and Haojie Li. Coloring anime line art videos with transformation region enhancement network. *Pattern Recognition*, 141:109562, 2023. 1

[24] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 1

[25] 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 *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1

[26] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 1, 2

[27] Holger Winnemöller, Jan Eric Kyprianidis, and Sven C Olsen. Xdog: An extended difference-of-gaussians compendium including advanced image stylization. *Computers & Graphics*, 36(6):740–753, 2012. 1

[28] 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. 2, 4, 5, 6

[29] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1, 2