SR+Codec: a Benchmark of Super-Resolution for Video Compression Bitrate Reduction

Evgeney Bogatyrev[1]
evgeney.zimin@graphics.cs.msu.ru
Ivan Molodetskikh[1]
ivan.molodetskikh@graphics.cs.msu.ru
Dmitriy Vatolin[1,2]
dmitriy@graphics.cs.msu.ru

[1] Lomonosov Moscow State University, Moscow, Russia
[2] MSU Institute for Artificial Intelligence, Moscow, Russia

Abstract

In recent years, there has been significant interest in Super-Resolution (SR), which focuses on generating a high-resolution image from a low-resolution input. Deep learning-based methods for super-resolution have been particularly popular and have shown impressive results on various benchmarks. However, research indicates that these methods may not perform as well on strongly compressed videos.

We developed a super-resolution benchmark to analyze SR's capacity to upscale compressed videos. Our dataset employed video codecs based on five widely-used compression standards: H.264, H.265, H.266, AV1, and AVS3. We assessed 19 popular SR models using our benchmark and evaluated their ability to restore details and their susceptibility to compression artifacts. To get an accurate perceptual ranking of SR models, we conducted a crowd-sourced side-by-side comparison of their outputs. We found that some SR models, combined with compression, allow us to reduce the video bitrate without significant loss of quality. We also compared a range of image and video quality metrics with subjective scores to evaluate their accuracy on super-resolved compressed videos. The benchmark is publicly available at videoprocessing.ai/benchmarks/super-resolution-for-video-compression.html.

1 Introduction

Super-resolution (SR) is the task of increasing the resolution of images and videos, with potential use ranging from detail restoration to quality enhancement [30, 39]. Some state-of-the-art SR methods can restore details not clearly visible in the original (lower-resolution) clip while working in real time [46]. Neighboring frames can help fill gaps when upscaling, because small movements caused by camera tremor may provide enough information to accurately increase the resolution, as demonstrated using a Google Pixel 3 camera [45]. This ability of SR methods to restore video details and enhance video quality suggests their use for improving video compression efficiency.

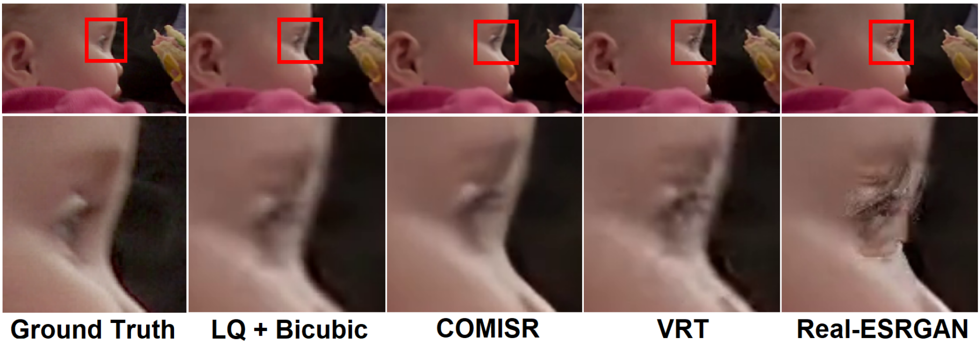

Figure 1: The comparison between 3 super-resolution models on compressed video sequence. COMISR eliminates the compression artifact because it is designed to work with compressed video. On the other hand, VRT and Real-ESRGAN fail to remove this artifact.

Video traffic had accounted for more than 80% of all consumer Internet traffic by 2022 [3] and that number is continuing to rise. Video compression that can lower bandwidth consumption with minimal changes to the visual quality of the video is more critical than ever. There have been some breakthroughs in video compression techniques, such as the recently developed MPEG compression standard MPEG-5 Part 2 Low Complexity Enhancement Video Coding (LCEVC) [32]. The core idea of LCEVC is to use a conventional video codec at a lower resolution and reconstruct a full-resolution video after decoding using enhancement layers, leading to a significant decrease to bandwidth consumption.

Some recent codecs [23, 32] downscale a video before compression to cut the bitrate and then upscale it to its original resolution using SR. Not all SR methods are suitable for such downscale-based video compression, however, since few real-time SR models can generate acceptable-quality video. Our research shows that many SR models are unable to deal with compression artifacts, as shown in Figure 1. Several existing benchmarks assess SR methods' ability to upscale compressed videos [49, 50], focusing on perceptual quality. We improve upon this work by additionally considering video bitrate reduction.

To analyze which SR models work better with each compression standard, and to help researchers find the best models for their codecs, we present our Super-Resolution for Video Compression benchmark. To develop it we selected 19 popular SR models with different architectures and assessed their compressed-video-restoration capabilities on our dataset, which includes videos compressed using five codecs. Our effort employed objective metrics and subjective evaluation to assess model quality. In addition, we analysed the correlation between objective-quality metrics and subjective scores and calculated bitrate reduction that can be achieved using each SR model during the compression.

Our main contributions are as follows:

1. We present a comprehensive SR benchmark to test the ability of 19 SR models to upscale and restore videos compressed by five video codecs of different standards. We evaluate the perceptual quality of the restored videos by conducting a crowd-sourced subjective comparison with 5397 subjects. The benchmark is publicly available at https://videoprocessing.ai/benchmarks/super-resolution-for-video-compression.html.

2. For every tested codec, we show which SR methods provide the most video bitrate

reduction with the least quality loss. We find bitrate improvements of up to 65% compared to using the base codec directly without SR.

3. We analyse six video quality metrics by their correlation with subjective scores on our dataset. Based on their performance on individual video clips, we construct a simple metric combination with improved results.

2 Related Work

In this section we provide an overview of existing SR methods, downscale-based video codecs, and SR benchmarks.

2.1 Super-resolution methods

SR has received extensive attention since neural networks were first applied to this area, resulting in many approaches.

Some SR algorithms rely on the temporal redundancy of video frames, allowing them to restore a single high-resolution frame from a series of low-resolution ones. **RBPN** [17] integrates spatial and temporal contexts from a video using a recurrent encoder-decoder module. **COMISR** [27] upscales compressed videos; it employs bidirectional recurrent warping for detail-preserving flow estimation, and it applies Laplacian enhancement. **BasicVSR++** [15] also adopts bidirectional propagation and spatial alignment. **VRT** [29] extracts video features, upscales them, and then reconstructs HQ frames on the basis of these features using a transformer network. **RVRT** [30] divides videos into multiple clips and uses previously inferred clip features to estimate subsequent clip features. Additionally, a guided deformable attention mechanism facilitates clip-to-clip alignment by predicting and aggregating multiple relevant locations across different clips. **Swin2SR** [16] explores Swin Transformer V2 to improve **SwinIR** [28] compressed image super-resolution.

Generative adversarial networks (GANs) serve widely in deep learning and especially in SR. **ESRGAN** [38] modifies the SRGAN architecture by adding residual-in-residual dense block as well as improved adversarial loss and perceptual loss. **Real-ESRGAN** [39] enhances this approach by incorporating high-order degradation modeling to simulate real-world degradation.

Lately, diffusion models have demonstrated impressive abilities to generate high-quality results in various applications, and SR is no exception [33, 35, 41, 51]. Although diffusion-based models have shown promising results, their drawback lies in the long inference process. These models usually need multiple inference steps to produce a final output, which hinders their practical use. Since they are unable to work in real time, they can't be used in the video decoding process. Therefore, in this study we are not considering diffusion models.

Given the limited number of SR models designed to work with compressed video, assessing the performance of existing SR models on compressed video remains a critical task.

2.2 Downscale-based video codecs

Recently, some video codecs have been designed to downscale the video before compression to reduce the bitrate, and then upscale it to the original resolution on the decoder side. Researchers are exploring many approaches to the upscaling module, ranging from simple filters to extensive neural networks.

The core idea of **LCEVC** [32] is to use a conventional video codec at a lower resolution and reconstruct a full-resolution video by combining the decoded low-resolution video with up to two enhancement sub-layers of residuals encoded with specialized low-complexity coding tools. Some video codecs use the similar idea of implementing a super-resolution network on the decoder side.

SRVC [23] encodes video into two bitstreams: a content stream and a model stream. The content stream is produced by compressing downsampled low-resolution video with the existing codec. The model stream encodes updates to the lightweight super-resolution network, which is used to upscale video at the decoder side. The SR network is trained on local segments of the video during the encoding process.

RR-DnCNN [19] addresses the problem of removing compression artifacts during the downscale-based video coding. A straightforward approach of applying compression artifact removal techniques before SR may result in detail loss. The paper proposes an end-to-end restoration-reconstruction deep neural network using the degradation-aware technique.

ViSRTA [11] takes one step further and dynamically resamples the video not only spatially, but also temporally during the encoding process. CNN-based architecture is used to restore spatial resolution, and temporal upsampling is performed by frame repetition.

Lin *et al.* [31] proposes a CNN-based SR method for resample-based video coding on the basis of the VTM codec. The authors designed separate networks for the luma and chroma components which exploit the cross-component correlation by using the luma reconstruction as the auxiliary information for the chroma network.

In our benchmark we take the idea of these codecs and try to apply it to different combinations of widely-used codecs and SR methods. Downscale-based video codecs are usually designed to transmit extra features of the original video alongside the low-resolution video data, to aid in the upscaling process during decoding. In our benchmark, codecs and SR methods work independently of each other, as our main goal is to evaluate how different codec and SR methods fit together.

2.3 Super-resolution benchmarks

Many SR benchmarks and challenges have appeared recently. We focus our attention on the ones prioritizing super-resolving compressed videos.

NTIRE 2022 Challenge on Super-Resolution and Quality Enhancement of Compressed Video [50] presents evaluation results of quality enhancement of compressed videos (track 1), quality enhancement with 2× and 4× SR (track 2 and track 3 respectively). More than 600 participants have registered on all tracks in total. This challenge proposes LDV 2.0 dataset with 335 videos and uses PSNR to evaluate participants.

The **AIM 2022** Challenge on Super-Resolution of Compressed Image and Video [49] presents two tracks: restoration of compressed images and compressed video. PSNR is used to evaluate the results, and an extension of the previously mentioned LDV 2.0 dataset, the LDV 3.0 dataset of 365 videos, is proposed. The second track of this task requires participants to enhance and 4× super-resolve HEVC compressed videos.

Our benchmark improves upon all the benchmarks presented above by a wider set of video quality metrics and a variety of video codecs used to create distorted videos. We also measure the bitrate reduction that SR models can provide when used during decompression.

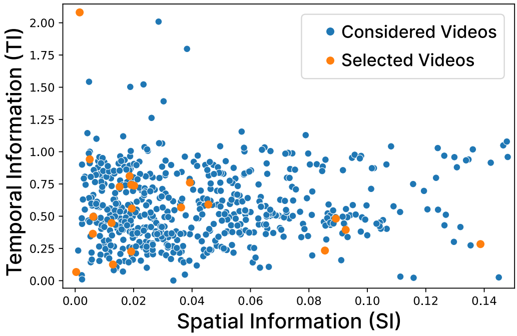

Figure 2: Distribution of Google Spatial and Temporal [40] information for videos we considered when creating our training dataset. Chosen videos appear in orange, others in blue.

Codec	Standard
x264	H.264
x265	H.265
aomenc	AV1
vvenc [44]	H.266
uavs3e [9]	AVS3

Table 1: Video codecs used to compress videos from benchmark dataset. We used *medium* preset for all of these codecs.

3 Benchmark Methodology

In this section, we describe the process of preparing the dataset for our benchmark, including the methodology for selecting videos and preprocessing them. We also describe our benchmark evaluation procedure and quality assessment.

3.1 Dataset preparation

To ensure the benchmark dataset is diverse enough to test various aspects of SR models, we collected 1920 × 1080 videos from multiple sources:

- **Vimeo**: We gathered 50 sequences, including both real world and animation. We split them into scenes using the Scene Change Detector plugin for VQMT [6].
- **Camera**: We shot several videos using a Canon EOS 7D. The settings aimed to minimize blur and achieve the appropriate brightness — the ISO was 4000 and the aperture 400. Those settings provided clear ground-truth (GT) videos without blur or noise. We shot 20 indoor videos and 30 outdoor videos. The indoor ones consist of synthetically crafted scenes containing objects from everyday life. Each scene includes either moving objects or horizontal camera motion.
- **Games**: We recorded 20 clips from various 2D and 3D videogames.

We then obtained the following features for each video: Google Spatial and Temporal features [40], frames per second (FPS), colorfulness [18], and maximal number of faces [5] throughout the video. On the basis of these features we separated all videos into 20 clusters using the K-Means clustering and selected one video from each cluster, as shown in Figure 2. We refer to these selections as *source videos*. A preview of them appears in Figure 3.

To ensure that important fine details of the video are not completely lost after downscaling and compression, we considered only videos with low space and time complexity, and no major blur or noise. However, camera motion was required, as it can help video SR models restore each frame more faithfully using neighboring frames.

Figure 3: Example videos from the dataset. The dataset includes real-world sequences, animation, and clips from games.

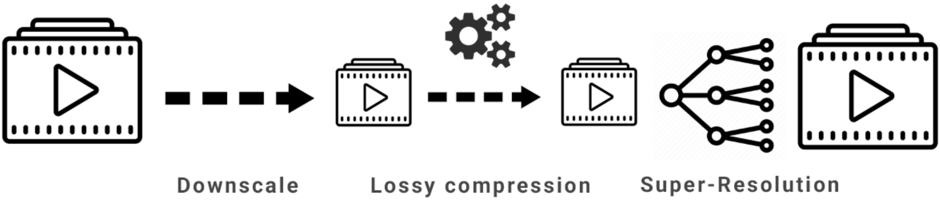

Figure 4: The evaluation pipeline of our benchmark. The pipeline consists of three steps: 4× bicubic downscaling, compression, and 4× SR upscaling.

3.2 Benchmark pipeline

To select SR models for evaluation we used SR benchmarks that target two tasks: detail restoration [4] and perceptual-quality improvement [10]. We excluded similar SR methods based on their performance in these benchmarks.

We selected the following 19 methods: BasicVSR++ [15], COMISR [27], DBVSR [34], EGVSR [14], LGFN [36], RBPN [17], Real-ESRGAN [39], RealSR [21], RSDN [20], SOF-VSR-BD [37], SOF-VSR-BI [37], SwinIR [28], TMNet [47], VRT [29], RVRT [30], IART[48], AnimeSR[46], Topaz Video AI [8], and bicubic interpolation. For all models we used pretrained weights provided by authors. Given the scarcity of high-quality video SR models, we decided to include image SR in our comparison.

We employed the standard video codecs for five widely-used compression standards, as detailed in Table 1.

A brief visualization of the benchmark pipeline can be seen in Figure 4. First, we downscaled the source video to 480 × 270 resolution using FFmpeg with the `flags=bicubic` option. We then compressed the low-resolution video using each of the five video codecs at three target bitrates: 0.6, 1.0, and 2.0 Mbps. To get more accurate objective metric scores, we also compressed videos at 0.1, 0.3, and 4.0 Mbps; however, these bitrates were not used in the subjective evaluation. We chose these bitrates to be relatively low and to form a logarithmic curve. All codecs employed the *medium* preset during compression. Compressed videos underwent transcoding to PNG sequences using FFmpeg, which were then passed as inputs to an SR model. We applied image SR models to each frame individually; video SR models received the path to the directory containing frames in the correct order. We tested

Codec	SR	Subj. score ↑	ERQA ↑	LPIPS ↓	PSNR ↑	MDTVSFA ↑
x264	SwinIR [28]	**3.695**	0.698	**0.206**	24.895	**0.564**
x264	Real-ESRGAN [39]	3.092	0.740	0.267	26.983	0.501
x264	*No SR*	2.895	**0.802**	0.306	**27.125**	0.547
x265	RVRT [30]	**2.768**	0.751	0.281	27.021	**0.506**
x265	*No SR*	2.011	**0.825**	**0.221**	**27.152**	0.496
x265	COMISR [27]	1.831	0.726	0.285	26.967	0.506
vvenc	RVRT [30]	**2.421**	0.755	**0.295**	**27.425**	0.501
vvenc	RBPN [17]	1.393	0.752	0.331	27.424	**0.509**
vvenc	*No SR*	0.944	**0.835**	0.325	26.467	0.500
uavs3e	*No SR*	**2.313**	**0.833**	**0.270**	**27.174**	**0.505**
uavs3e	RVRT [30]	1.682	0.761	0.291	27.040	0.503
uavs3e	RBPN [17]	1.404	0.737	0.294	26.998	0.499
aomenc	*No SR*	**2.884**	**0.856**	**0.235**	**27.238**	**0.509**
aomenc	RVRT [30]	1.752	0.755	0.283	27.046	0.501
aomenc	RBPN [17]	1.599	0.730	0.288	27.018	0.499

Table 2: Comparison of SR+codec pairs by subjective score and objective metrics for "Restaurant" sequence.[1] The best result appears in **bold**.

a 4× upscale using our benchmark, but some SR models can only handle 2×. In this latter case, we applied the model twice.

3.3 Quality estimation and subjective study

After super-resolving videos, we calculated the following objective-video-quality metrics on the results: PSNR, MS-SSIM [43], VMAF [2], LPIPS [52], MDTVSFA [26] and ERQA [24]. We considered mainly full-reference metrics since we prioritized detail restoration over perceptual quality. PSNR, MS-SSIM, and VMAF are the standard for compressed video quality evaluation [1], while LPIPS and ERQA are well-suited for super-resolved images [4]. The only no-reference metric we used is MDTVSFA, since it shows promising results when evaluating compressed sequences [12]. To rank bitrate reduction, we calculated BSQ-rate (bitrate-for-the-same-quality rate) [53] for each SR+codec pair relative to base codec performance, where the base codec is the one we used to compress low-resolution video. A lower BSQ-rate means more bitrate saving for the same quality.

To subjectively rank SR models, we conducted a crowd-sourced comparison through Subjectify.us [7] service. Because detail loss and compression artifacts can be difficult to notice in a full frame, the subjective evaluation employed crops. We needed pairs of crops to fit on the assessors' screens during the comparison, so we choose the resolution of the crops to be 480×270. First, we generated saliency maps for each source video using a method proposed in Kroner *et al.* [25]. Second, we averaged the saliency maps over all frames and applied a Gaussian-blur kernel to the result in order to determine the video's most salient region. Third, we took distorted videos from the benchmark dataset and cut one 480×270 crop from each one, with the most salient area at the center of the crop. We evaluated objective metrics on these crops to determine the correlation with the subjective scores.

We split our comparison into five sections by codec, using only the 10 best SR models as

[1] Full results are available on the benchmark page: https://videoprocessing.ai/benchmarks/super-resolution-for-video-compression.html

codec SR	x264	x265	aomenc	vvenc	uavs3e
No SR	1.000	1.000	**1.000**	1.000	*1.000*
RealSR[21]	**0.196**	0.502	*1.513*	4.470	**0.639**
RVRT[30]	0.271	0.724	2.806	*1.665*	1.750
SwinIR[28]	*0.304*	**0.346**	1.505	2.502	0.640
Real-ESRGAN[39]	0.335	*0.640*	2.411	4.368	2.430
COMISR[27]	0.367	0.741	2.799	**0.701**	1.603
VRT[29]	1.245	3.175	4.157	4.185	3.663
BasicVSR++[15]	1.971	3.390	4.199	4.185	4.250
RBPN[17]	1.979	3.434	4.226	4.246	4.470

Table 3: Average BSQ-rate [53] over subjective scores for each SR+codec pair. Lower BSQ-rate is better. The best result appears in **bold**, the second best is underlined, and the third best is in *italics*.

determined by the LPIPS value. Also, from each group of models with similar architectures, like SOF-VSR-BI and SOF-VSR-BD, we selected only one model showing the best LPIPS values.

During the experiment we showed each participant a pair of videos from two random SR models and asked them to choose the video that looks more realistic and has fewer compression artifacts ("indistinguishable" was also an option). Every video pair was viewed by 15 participants. Each participant compared 25 pairs total.

Among the 25 questions were three verification ones, which had obvious predefined answers. We excluded the results from any participant who failed to correctly answer one or more of the verification questions. A total of 5662 people participated in our subjective evaluation. We excluded the results from 265 of them because they failed to correctly answer verification questions. Our calculation of the final subjective scores, using the Bradley-Terry model [13], employed the remaining 120,316 responses.

4 Evaluation Results

In this section we present the results of our Super-Resolution for Video Compression benchmark. We discuss two different ways of comparing SR models in our benchmark: by evaluating the quality of the results and by the bitrate reduction compared to the baseline codec. We also analyze existing video quality metrics based on subjective evaluation.

4.1 Comparison by resulting video quality

Table 2 lists two best SR methods for each codec and *"No SR"* method, which applies the video codec to the source video without downscaling or super-resolving. The best methods are chosen based on subjective scores on the *"Restaurant"* sequence compressed at an approximate bitrate of 2 Mbps.

We can see that RBPN and RVRT appear to be the best methods for various advanced codecs, mainly due to their video restoration capabilities. However, this is not the case for x264 codec, where SwinIR and Real-ESRGAN come on top. Videos compressed with x264 at lower bitrates have many compression artifacts that make it hard to faithfully restore the

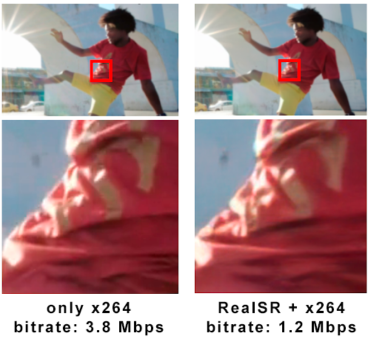

Figure 5: RealSR applied to video compressed with x264 codec at 1.2 Mbps can achieve the same visual quality as plain x264 codec at 3.8 Mbps.

Metric	PLCC	SRCC
MS-SSIM [42]	0.146	0.151
PSNR	0.187	0.285
VMAF [2]	0.344	0.448
LPIPS [52]	0.414	0.431
ERQA [24]	0.582	0.624
MDTVSFA [26]	0.634	0.644
ERQA×MDTVSFA	**0.770**	**0.801**

Table 4: Mean Pearson (PLCC) and Spearman (SRCC) rank correlation coefficients between metrics and subjective-comparison results.

original sequence. Thus, generative SR models such as Real-ESRGAN perform subjectively better by generating realistic frames.

We also see that *"No SR"* exhibits the best results for the aomenc codec. This is likely because at low bitrates, AV1 codecs use a special mode that encodes frames at a low resolution and applies an upsampler when decoding [22], making the additional use of SR redundant.

4.2 Comparison by bitrate reduction

The results of BSQ-rate calculation over subjective scores of each SR+codec pair appears in Table 3. As the table shows, the best SR model differs by codec, proving that no single SR model can handle distortion from all compression standards with equal effectiveness. Although RealSR shows significantly less accurate results at high bitrates than other SR models, it gives more visually pleasing results at low bitrates, which are subjectively comparable to *"No SR"* results at high bitrates. We can see that *"No SR"* shows better results with aomenc codec for the same reason we discussed in the previous subsection.

4.3 Video quality metrics assessment

Judging by the results in Section 4.1, some video quality metrics cannot reproduce the result of the subjective evaluation obtained by crowd-sourcing. To analyze the objective metrics, we calculated them on the crops that were used for the crowd-sourced evaluation. Pearson (PLCC) and Spearman (SRCC) correlations of objective metrics with subjective scores are presented in Table 4.

We analyzed the performance of the best metrics and noticed that ERQA shows a high correlation on the video crops where MDTVSFA has a low correlation, and vice versa. We believe that these two metrics take into account different features of the input video and can complement each other. Since ERQA values range from 0 to 1, ERQA can be used as a correction value for MDTVSFA. We attempted to combine the ERQA and MDTVSFA metrics by multiplying their results. We called this method ERQA×MDTVSFA. This approach yielded a significant increase in correlation, as shown in Table 4.

5 Conclusion

In this paper we proposed a new benchmark for super-resolution (SR) compression restoration. Our work assessed 19 SR models applied to five video codecs using both objective-quality metrics and subjective evaluation.

Our research shows that SR models, such as RealSR [21] and RVRT [30], can serve in downscale-based codecs on the decoder side to enhance the subjectively perceived quality of videos with low bitrates. RVRT can improve the quality of videos compressed by x265 and vvenc codecs. RealSR can be used with x264 to reduce the video bitrate by more than 65% without significant quality loss, as shown in Figure 5.

We find that existing video-quality metrics correlate poorly with subjective scores and are therefore unsuitable for assessing the results of downscale-based video coding. MDTVSFA and ERQA show better results, and their combination ERQA×MDTVSFA achieves Spearman's rank correlation of 0.801 on our dataset.

Our benchmark is publicly available at videoprocessing.ai/benchmarks/super-resolution-for-video-compression.html. We welcome SR researchers to contribute to it by submitting SR models.

References

[1] MSU Video Codecs Comparison 2022 report. http://www.compression.ru/video/codec_comparison/2022/main_report.html. Online; accessed 2024-05-03.

[2] VMAF - Video Multi-Method Assessment Fusion. https://github.com/Netflix/vmaf. Online; accessed 2024-05-03.

[3] VNI Complete Forecast Highlights. https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_Device_Growth_Traffic_Profiles.pdf. Online; accessed 2024-05-03.

[4] MSU Video Super-Resolution Benchmark: Detail Restoration. https://videoprocessing.ai/benchmarks/video-super-resolution.html. Online; accessed 2024-05-03.

[5] Face Recognition by ageitgey. https://github.com/ageitgey/face_recognition. Online; accessed 2024-05-03.

[6] MSU Scene Change Detector. https://www.compression.ru/video/quality_measure/metric_plugins/scd_en.htm. Online; accessed 2024-05-03.

[7] Subjectify.us: Crowd-sourced subjective quality evaluation platform. https://subjectify.us/. Online; accessed 2024-05-03.

[8] Topaz Video Enhance AI. https://www.topazlabs.com/topaz-video-ai. Online; accessed 2024-05-03.

[9] uavs3e video codec. https://github.com/uavs3/uavs3e. Online; accessed 2024-05-03.

[10] MSU Video Upscalers Benchmark: Quality Enhancement. https://videoprocessing.ai/benchmarks/video-upscalers.html. Online; accessed 2024-05-03.

[11] Mariana Afonso, Fan Zhang, and David R Bull. Video compression based on spatio-temporal resolution adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1):275–280, 2018.

[12] Anastasia Antsiferova, Sergey Lavrushkin, Maksim Smirnov, Aleksandr Gushchin, Dmitriy Vatolin, and Dmitriy Kulikov. Video compression dataset and benchmark of learning-based video-quality metrics. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 13814–13825. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/59ac9f01ea2f701310f3d42037546e4a-Paper-Datasets_and_Benchmarks.pdf.

[13] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[14] Yanpeng Cao, Chengcheng Wang, Changjun Song, Yongming Tang, and He Li. Real-time super-resolution system of 4k-video based on deep learning. In *2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pages 69–76. IEEE, 2021.

[15] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022.

[16] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *European Conference on Computer Vision*, pages 669–687. Springer, 2022.

[17] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019.

[18] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–95. SPIE, 2003.

[19] Man M Ho, Jinjia Zhou, and Gang He. RR-DnCNN v2. 0: enhanced restoration-reconstruction deep neural network for down-sampling-based video coding. *IEEE Transactions on Image Processing*, 30:1702–1715, 2021.

[20] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *European conference on computer vision*, pages 645–660. Springer, 2020.

[21] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 466–467, 2020.

[22] Urvang Joshi, Debargha Mukherjee, Yue Chen, Sarah Parker, and Adrian Grange. In-loop frame super-resolution in av1. In *2019 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2019.

[23] Mehrdad Khani, Vibhaalakshmi Sivaraman, and Mohammad Alizadeh. Efficient Video Compression via Content-Adaptive Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4521–4530, 2021.

[24] Anastasia Kirillova, Eugene Lyapustin, Anastasia Antsiferova, and Dmitry Vatolin. ERQA: Edge-restoration quality assessment for video super-resolution. *arXiv preprint arXiv:2110.09992*, 2021.

[25] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129: 261–270, 2020. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2020.05.004. URL https://www.sciencedirect.com/science/article/pii/S0893608020301660.

[26] Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4):1238–1257, 2021.

[27] Yinxiao Li, Pengchong Jin, Feng Yang, Ce Liu, Ming-Hsuan Yang, and Peyman Milanfar. Comisr: Compression-informed video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2543–2552, 2021.

[28] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.

[29] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022.

[30] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022.

[31] Chaoyi Lin, Yue Li, Kai Zhang, Zhaobin Zhang, and Li Zhang. Cnn-based super resolution for video coding using decoded information. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021.

[32] Guido Meardi, Simone Ferrara, Lorenzo Ciccarelli, Guendalina Cobianchi, Stergios Poularakis, Florian Maurer, Stefano Battista, and Ahmad Byagowi. MPEG-5 part 2: Low complexity enhancement video coding (LCEVC): Overview and performance evaluation. *Applications of Digital Image Processing XLIII*, 11510:238–257, 2020.

[33] Mehdi Noroozi, Isma Hadji, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. You only need one step: Fast super-resolution with stable diffusion via scale distillation. *arXiv preprint arXiv:2401.17258*, 2024.

[34] Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4811–4820, 2021.

[35] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[36] Dewei Su, Hua Wang, Longcun Jin, Xianfang Sun, and Xinyi Peng. Local-global fusion network for video super-resolution. *IEEE Access*, 8:172443–172456, 2020.

[37] Longguang Wang, Yulan Guo, Zaiping Lin, Xinpu Deng, and Wei An. Learning for video super-resolution through HR optical flow estimation. In *Asian Conference on Computer Vision*, pages 514–529. Springer, 2018.

[38] 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 *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018.

[39] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[40] Yilin Wang, Sasi Inguva, and Balu Adsumilli. YouTube UGC dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019.

[41] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. *arXiv preprint arXiv:2311.14760*, 2023.

[42] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[44] Adam Wieckowski, Jens Brandenburg, Tobias Hinz, Christian Bartnik, Valeri George, Gabriel Hege, Christian Helmrich, Anastasia Henkel, Christian Lehmann, Christian Stoffers, Ivan Zupancic, Benjamin Bross, and Detlev Marpe. VVenC: An Open And Optimized VVC Encoder Implementation. In *Proc. IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–2. doi: 10.1109/ICMEW53276.2021.9455944.

[45] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld Multi-Frame Super-Resolution. *ACM Trans. Graph.*, 38(4), July 2019. ISSN 0730-0301. doi: 10.1145/3306346.3323024. URL https://doi.org/10.1145/3306346.3323024.

[46] 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.

[47] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2021.

[48] Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. Enhancing Video Super-Resolution via Implicit Resampling-based Alignment. *arXiv preprint arXiv:2305.00163*, 2024.

[49] Ren Yang, Radu Timofte, Xin Li, Qi Zhang, Lin Zhang, Fanglong Liu, Dongliang He, He Zheng, Weihang Yuan, Pavel Ostyakov, et al. Aim 2022 challenge on super-resolution of compressed image and video: Dataset, methods and results. *arXiv preprint arXiv:2208.11184*, 2022.

[50] Ren Yang, Radu Timofte, Meisong Zheng, Qunliang Xing, Minglang Qiao, Mai Xu, Lai Jiang, Huaida Liu, Ying Chen, Youcheng Ben, et al. NTIRE 2022 challenge on super-resolution and quality enhancement of compressed video: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1221–1238, 2022.

[51] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024.

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[53] Anastasia V Zvezdakova, Dmitriy L Kulikov, Sergey V Zvezdakov, and Dmitriy S Vatolin. BSQ-rate: a new approach for video-codec performance comparison and drawbacks of current solutions. *Programming and computer software*, 46(3):183–194, 2020.