ERQA: Edge-Restoration Quality Assessment for Video Super-Resolution

Anastasia Kirillova[1][a], Eugene Lyapustin[1][b], Anastasia Antsiferova[1][c] and Dmitry Vatolin[1][d]
[1]*Lomonosov Moscow State University, Moscow, Russia*
{*anastasia.kirillova, evgeny.lyapustin, aantsiferova, dmitriy*}*@graphics.cs.msu.ru*

Keywords: Video Super-Resolution, Quality Assessment, Video Restoration

Abstract: Despite the growing popularity of video super-resolution (VSR), there is still no good way to assess the quality of the restored details in upscaled frames. Some SR methods may produce the wrong digit or an entirely different face. Whether a method's results are trustworthy depends on how well it restores truthful details. Image super-resolution can use natural distributions to produce a high-resolution image that is only somewhat similar to the real one. VSR enables exploration of additional information in neighboring frames to restore details from the original scene. The ERQA metric, which we propose in this paper, aims to estimate a model's ability to restore real details using VSR. On the assumption that edges are significant for detail and character recognition, we chose edge fidelity as the foundation for this metric. Experimental validation of our work is based on the MSU Video Super-Resolution Benchmark, which includes the most difficult patterns for detail restoration and verifies the fidelity of details from the original frame. Code for the proposed metric is publicly available at https://github.com/msu-video-group/ERQA.

1 INTRODUCTION

As a fundamental image- and video-processing task, super-resolution remains a popular research topic. It has a wide range of applications, from low-complexity encoding[1] to old-film restoration and medical-image enhancement. Trends in quality assessment of upscaled videos and images are favoring estimation of statistical naturalness in combination with fidelity. But restoration fidelity is much more important than statistical naturalness for some tasks: small-object recognition (e.g., license-plate numbers) in CCTV recordings, text recognition, and medical-image reconstruction.

With the development of deep-learning-based approaches, many super-resolution models produce visually natural frames but lose important details. For example, the rightmost image in Figure 1 is perceptually better than the leftmost one, but the shape of the shiny thread in the leftmost image is closer to ground truth (GT, center). Occasionally, such models can even change the context in an image by, for example, producing an incorrect number, character, or even human face without decreasing traditional-metric values. In Figure 2, RPBN (Haris et al., 2019) added horizontal lines to the bottom-right character, but all three models score the same on traditional metrics. In Figure 3, Real-ESRGAN (Wang et al., 2021) mixed two letters from low-resolution images to form a completely different letter. In Figure 4, Real-ESRGAN and RealSR (Ji et al., 2020) produced unnatural faces that greatly differ from the source one.

Figure 1: Example of upscaled images that vary in detail-restoration quality. The rightmost image is visually more natural, but the shape of the details in the leftmost image is closer to the original.

The examples in Figures 1–4 demonstrate that assessment of detail-restoration quality for image and video super-resolution is difficult. The best way to estimate restoration fidelity is to conduct a subjective comparison; it's the most precise approach but is time

[a] https://orcid.org/0000-0002-0799-3135
[b] https://orcid.org/0000-0002-2515-9478
[c] https://orcid.org/0000-0002-1272-5135
[d] https://orcid.org/0000-0002-8893-9340
[1] https://www.lcevc.org/

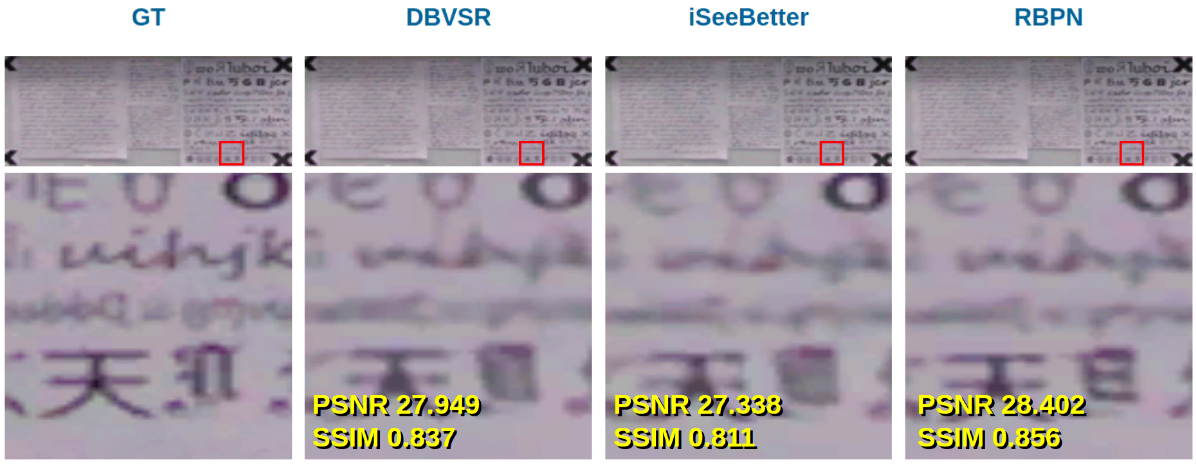

Figure 2: Example of changing context in an upscaled video: text restoration has changed a character in the rightmost image.

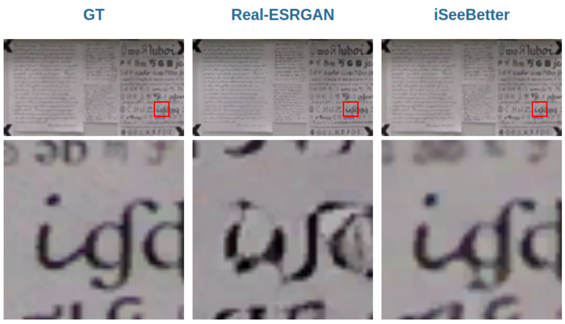

Figure 3: Example of changing context in an upscaled video: two characters (left) mix to yield a new one (center) during text restoration.

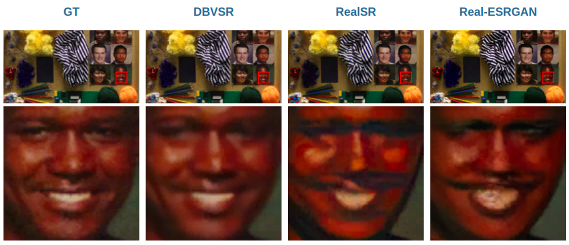

Figure 4: Example of changing context in an upscaled video: unnatural faces are the result here, differing considerably from the source one (GT).

consuming and expensive. Another way involves reference quality metrics. Traditional similarity metrics such as PSNR and SSIM (Wang et al., 2004) are often used to evaluate super-resolution models, but they yield poor results and are unstable when dealing with shifts and other common super-resolution artifacts. LPIPS (Zhang et al., 2018) is increasingly popular for this task, but it originally aimed to assess perceptual similarity rather than fidelity. The new DISTS (Ding et al., 2020a) metric is an improvement on LPIPS, but it also focuses on perceptual similarity.

Our research focuses on analyzing super-resolution algorithms, particularly their restoration fidelity. When we started working on a benchmark for video super-resolution, including a test for restoration-quality assessment, we discovered that existing metrics work fine for other tests (restoration naturalness and beauty) but have a low correlation with subjective detail-quality estimation. In this paper, therefore, we introduce a new method for evaluating information fidelity. Experiments reveal that our metric outperforms other super-resolution quality metrics in assessing detail restoration. [2]

The main contributions of our work are the following:

1. A video-super-resolution benchmark based on a new dataset containing the most difficult patterns for detail restoration.
2. A subjective comparison examining the fidelity of details from the original frame, instead of traditional statistical naturalness and beauty.
3. A new metric for assessing the detail-restoration quality of video super-resolution.

2 RELATED WORK

PSNR and SSIM (Wang et al., 2004) are common metrics for assessing super-resolution quality. We analyzed 378 papers that propose super-resolution methods and found that since 2008, PSNR and SSIM have remained the most popular metrics. But both have been shown to exhibit a low correlation with subjective scores. Only LPIPS (Zhang et al., 2018) has

[2] https://videoprocessing.ai/benchmarks/video-super-resolution.html

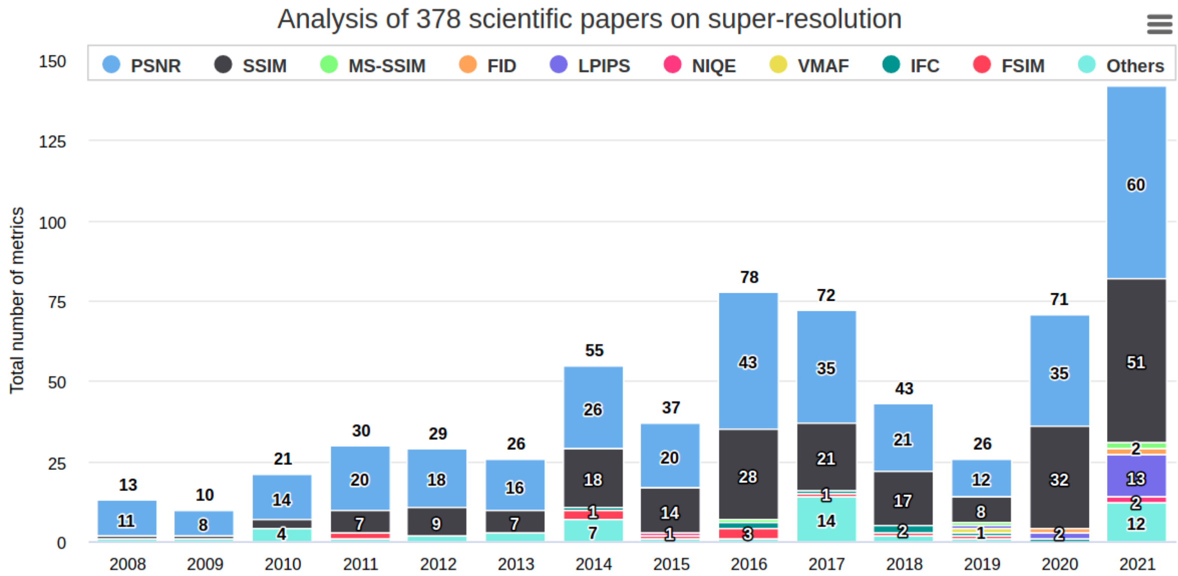

Figure 5: Metrics for estimating super-resolution quality cited in papers proposing new methods, by year. PSNR and SSIM (Wang et al., 2004) are the most popular; LPIPS saw wide use in 2020 and 2021.

grown in popularity over the last two years; other metrics remain less popular (Figure 5).

Several full-reference metrics for assessing super-resolution visual quality have emerged. (Wan et al., 2018) used four features (gradient magnitude, phase congruency, anisotropy, and directionality complexity) to calculate the perceptual structure measurement (PFSM) in both the upscaled and original high-resolution frames. The authors' comparison of PFSMs using a similarity function showed more-consistent results than previous approaches with regard to visual perception on the SRSED dataset. (Zhou et al., 2021) calculated structural fidelity and statistical naturalness, fused these coefficients into a weighted sum, and achieved good correlation on the QADS image database. (Zhou et al., 2019) compared distorted and original images by separately estimating textural, structural, and high-frequency similarity. The final score revealed an even better correlation on QADS.

In a few papers, the authors proposed reduced-reference metrics, which use low-resolution (LR) images as a reference. A popular approach is to extract structure or texture features from LR and upscaled (SR) images, compare them separately, and fuse the resulting similarity indices (Yeganeh et al., 2015; Tang et al., 2019; Fang et al., 2019). Metrics based on this idea achieve a Pearson correlation coefficient of 0.79 to 0.83 and a Spearman correlation coefficient of 0.69 to 0.85 on various datasets, depending on the implementation. (Yang et al., 2019a) trained a regression model using statistical features extracted from LR and SR images, obtaining a correlation similar to that of other top metrics on the dataset from (Ma et al., 2017). (Shi et al., 2019) proposed another approach for reduced-reference assessment that uses the visual-content-prediction model to measure the structure of the reference HR and SR images. This method outperforms previous ones on the SISRSet dataset.

A number of no-reference metrics are also used for video super-resolution. Most of them train regression models on statistical features extracted from upscaled frames (Ma et al., 2017; Zhang et al., 2019; Beron et al., 2020), achieving a Spearman correlation of 0.740 to 0.939 and a Pearson correlation of 0.728 to 0.9463, depending on the implementation and test dataset. Also, many no-reference metrics are based on features extracted using a pretrained neural network—VGGNet, for example (Zhang et al., 2021). In a few papers, the authors trained SVM using extracted features (Qian et al., 2019; Wang et al., 2018) and obtained results similar to those of other metrics. (Greeshma and Bindu, 2018) calculated acutance and spatial-discontinuity features in the gradient and wavelet domains and then pooled them in the so-called super-resolution entropy metric (SREM). They later proposed the SRQC metric (Greeshma and Bindu, 2020), which estimates structure changes and quality-aware features by considering fuzzy gradient points and multiscale energy bands. Their metrics exhibited good results, but they consider only a few images and four SR methods for the test dataset.

Edges have a strong influence on the human visual system. Furthermore, edge fidelity is a base criterion for assessing detail-restoration quality. Several methods thus consider edge features as the basis for quality assessment (Table 1). Some calculate edge features, including number, length, direction, strength, contrast, and width, and compare them using the similarity measure to estimate image or video quality (Attar et al., 2016; Ni et al., 2017; Yang et al., 2019b). Nevertheless, these metrics achieve on their datasets almost the same correlation as traditional PSNR and SSIM. In (Xue and Mou, 2011), the authors detected edges in both reference and distorted images and compared them by calculating recall. (Chen et al., 2011) used histogram analysis for edge comparison. These metrics deliver a slightly greater correlation than PSNR and SSIM. Liu et al. (Liu et al., 2019) proposed using the F1 score to evaluate edge fidelity, but they declined to conduct a comparison with other metrics and kept their code under wraps. Our method is based on the same edge-comparison idea, but it's robust for small local and global edge shifts, which appear during super-resolution but are unessential for detail recognition. It yielded much better results than other quality-assessment approaches.

A number of datasets are used for testing super-resolution quality assessment (Table 2), but not for detail restoration, and they lack difficult patterns for that task. Therefore, we built a dataset for assessing super-resolution quality that includes the most challenging content for detail restoration.

Summarizing the above analysis, few metrics aim to assess and compare detail-restoration quality. Some that use edge features have emerged, but no one uses them for super-resolution, which involves peculiar artifacts. Therefore, it's important to obtain an objective metric that correlates highly with human estimation of detail-restoration quality and that allows comparison of super-resolution models, not only for naturalness but also for information fidelity.

3 PROPOSED METHOD

3.1 Dataset Preparation

To analyze a VSR model's ability to restore real details, we built a test stand containing patterns that are difficult for video restoration (Figure 6).

To calculate metrics for particular content types and to verify how a model works with different inputs, we divide each output frame into parts by detecting crosses:

Part 1 "Board" includes a few small objects and photos of human faces[3]. Our goal is to obtain results for the model operating on textures with small details. The striped fabric and balls of yarn may produce a Moire pattern (Figure 7). Restoration of human faces is important for video surveillance.

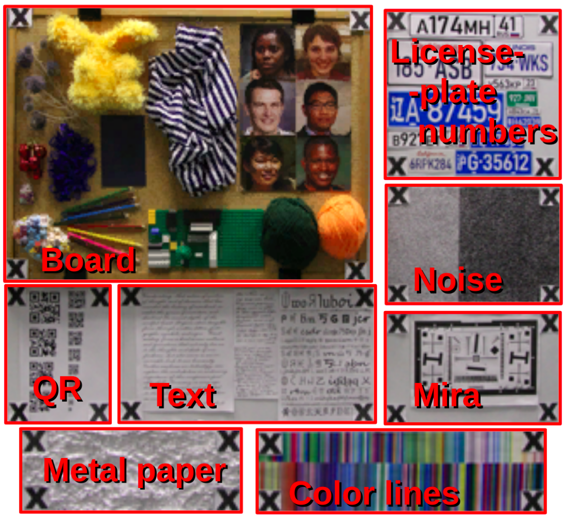

Figure 6: Test stand for the proposed VSR benchmark.

Figure 7: Example of a Moire pattern on the "Board."

Part 2 "QR" comprises multiple QR codes of differing sizes; the aim is to find the size of the smallest recognizable one in the model's output frame. A low-resolution frame may blend QR-code patterns, so models may have difficulty restoring them.

Part 3 "Text" includes two kinds: handwritten and typed. Packing all these difficult elements into the training dataset is a challenge, so they are each new to the model as it attempts to restore them.

Part 4 "Metal paper" contains foil that was vigorously crumpled. It's an interesting example because of the reflections, which change periodically between frames.

Part 5 "Color lines" is a printed image with numerous thin color stripes. This image is difficult be-

[3]Photos were generated by https://thispersondoesnotexist.com/

Table 1: A comparison of metrics based on edge features consideration. The results are taken from the original papers.

Metric	Edge comparison	Test dataset	PLCC	SRCC
EBIQA (Attar et al., 2016)	Similarity measure of length, number, orientation	A57 WIQ	0.8786 0.8113	0.8603 0.7582
ESIM (Ni et al., 2017)	Similarity measure of contrast, width, direction	SIQAD SCID	0.8788 0.8630	0.8632 0.8478
MSEA (Yang et al., 2019b)	Similarity measure of contrast, structure	SIQAD	0.8867	0.8773
rNSE (Xue and Mou, 2011)	Recall	A57, CSIQ, LIVE, IVC MICT, TID2008	0.8747	0.8649
EDHSSIM (Chen et al., 2011)	Histogram analysis	LIVE	—	0.9660
S. Liu et al. (Liu et al., 2019)	F1-score	Custom dataset	—	—

Table 2: A comparison of datasets using for testing super-resolution quality assessment approaches.

Dataset	# references	# SR images	# SR algorithms	Subjective type
C. Ma et al.'s (Ma et al., 2017)	30	1620	9	MOS
QADS (Zhou et al., 2019)	20	980	21	Pairwise comparison
SupER (Köhler et al., 2019)	14	3024	20	Pairwise comparison
SRIJ (Beron et al., 2020)	32	608	7	MOS
SISRSet (Shi et al., 2019)	15	360	8	MOS
ECCV (Yang et al., 2014)	10	540	6	MOS
SRID (Wang et al., 2017)	20	480	8	MOS

cause thin lines of similar colors end up mixing in low-resolution frames.

Part 6 "License-plate numbers" consists of a set of car license plates of varying sizes from different countries [4]. This content is important for video surveillance and dashcam development.

Part 7 "Noise" includes difficult noise patterns. Models cannot restore real ground-truth noise, and each one produces a unique pattern.

Part 8 "Mira" contains a resolution test chart with patterns that are difficult to restore: a set of straight and curved lines of differing thicknesses and directions.

We captured the dataset using a Canon EOS 7D camera. We quickly took a series of 100 photos and used them as a video sequence. The shots were from a fixed point without a tripod, so the video contains a small amount of random motion. We stored the video as a sequence of frames in PNG format, converted from JPG. The camera's settings were ISO 4000, aperture 400, and resolution 5,184x3,456.

The source video also has a resolution of 5,184x3,456 and was stored in the sRGB color space.

[4]The license-plate numbers are generated randomly and printed on paper.

We degraded it using bicubic interpolation to generate a ground truth of resolution 1,920x1,280. This step is essential because many open-source models lack the code to process a large frame; processing large frames is also time consuming. We further degraded the input video from ground truth, again using bicubic interpolation, to 480x320 to test the models for 4x upscaling. The output of each model is also a sequence of frames, which we compare with the ground-truth sequence to verify the model's performance.

3.2 Subjective Comparison

We used 21 super-resolution algorithms in our quality assessment. We also added a ground-truth video, so the experimental validation involves 22 videos. We cut the sequences to 30 frames and converted them to 8 frames per second (fps). This length allows subjects to easily consider details and decide which video is better. We then cropped from each video 10 snippets that cover the most difficult patterns for restoration and conducted a side-by-side pairwise subjective evaluation using the Subjectify.us service, which enables crowd-sourced comparisons.

To estimate information fidelity, we asked participants in the subjective comparison to avoid choosing the most beautiful video, but instead choose the one that shows better detail restoration. Each participant was shown 25 paired videos and in each case had to choose the best video ("indistinguishable" was also an option). Three of these pairs are for verification, so the final results exclude their answers. All other responses from successful participants are used to predict subjective scores using the Bradley-Terry model.

3.3 Edge Restoration Quality Assessment Method

On the basis of the hypothesis that edges are significant for detail restoration, we developed the edge-restoration quality assessment (ERQA) metric, which estimates how well a model can restore edges in a high-resolution frame. Our metric compensates for small global and local edge shifts, assuming they don't complicate detail recognition.

First, we find edges in both the output and ground-truth frames. Our approach uses an OpenCV implementation[5] of the Canny algorithm (Canny, 1986). The threshold for initially identifying strong edges is 200, and the threshold for linking edges is 100. These coefficients allow us to highlight the edges of all objects, even small ones, while skipping lines, which are unimportant (Figure 8).

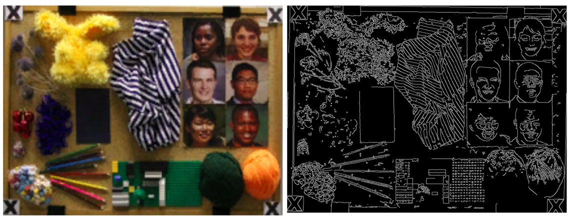

Figure 8: Example of edges highlighted by Canny algorithm (Canny, 1986) with chosen parameters.

Having found the edges in the ground-truth and distorted frames as binary masks, we compare them using the F1 score. Some models can generate frames with a global pixel shift relative to ground truth, so we checked the integer pixel shifts $[-3,3]$ along both axes and chose the one with the maximum PSNR value. Compensating for this global shift aids our metric considerably (Table 3).

During an upscaling, models may also shift edge pixels locally, which in many cases is insignificant to human perception of information. To compensate for local single-pixel edge shifts, we consider as true positive any pixels on the output edges, which are not on the ground-truth edges but are near (on the difference of one pixel) with the edge of GT (Figure 9).

We then noticed that some models produce a wider edge compared with the ground truth, and our method with local compensation (ERQAv1.0) marks these edges as fully true positive. To correct this shortcoming, ERQAv1.1 considers each point on a ground-truth edge as corresponding to true positive only once (Figure 10).

Figure 9: Visualization of ERQA metric.

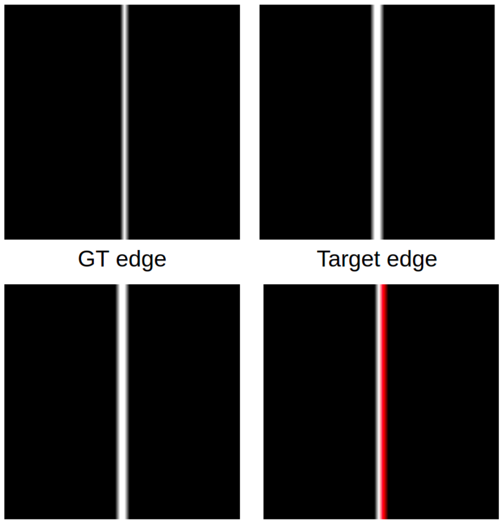

Figure 10: A comparison of ERQAv1.0 and ERQAv1.1. White = true positive, red = false positive.

4 EXPERIMENTAL VALIDATION

4.1 Ablation Study

To verify the significance of the global- and local-shift compensation, we conducted a basic edge comparison without compensation, with only global compensation, with both global and local compensation (v1.0), and with penalization of wide edges (v1.1).

[5]https://docs.opencv.org/3.4/dd/d1a/group__imgproc__feature.html#ga04723e007ed888ddf11d9ba04e2232de

All consistently increased both the Pearson (PLCC) and Spearman (SRCC) correlation coefficients (Table 3).

4.2 Comparison with Other Metrics

We conducted a study of existing metrics for video-quality assessment and found that some work well for naturalness and beauty, but none works well for restoration. We calculated several well-known metrics on a new dataset: PSNR, SSIM (Wang et al., 2004), MS-SSIM (Wang et al., 2003), VMAF[6], the recently developed LPIPS (Zhang et al., 2018), which showed good results when assessing super-resolution imaging, its improvement DISTS (Ding et al., 2020b) and metric for SR assessment (Ma et al., 2017). Our metric outperforms all others in both the Pearson and Spearman correlation coefficients (Figure 13). LPIPS places second. A popular metric for video-quality assessment, VMAF, exhibits poor results even compared with the traditional SSIM for this case. Multi-scale structural similarity (MS-SSIM), which usually delivers better results than simple structural similarity (SSIM), ranked last on super-resolution.

We tried our global-shift-compensation scheme in an attempt to improve the performance of these metrics. Nearly all metrics (except VMAF) were better as a result (Table 4).

Because metrics can work differently on different content types, we separately considered the correlation of metric values with subjective assessment on all crops and then calculated the mean correlation. Despite its simple and straightforward construction, ERQA delivers more-consistent results with subjective assessment (Figure 14) and outperforms all other metrics in both the Pearson (Table 5) and Spearman (Table 6) coefficients when assessing information fidelity.

4.3 QADS Dataset

We also verified our metric on the QADS dataset (Zhou et al., 2019). Although the mean correlation is lower than that of a few other metrics, the reason is that this dataset was developed for another test case (visual perception). In some situations, an image with lower visual perception looks more like the original one than does an image with higher visual perception (Figure 12). At the same time, working with images closer to our test-case ERQA yields good results (Figure 11).

[6]https://github.com/Netflix/vmaf

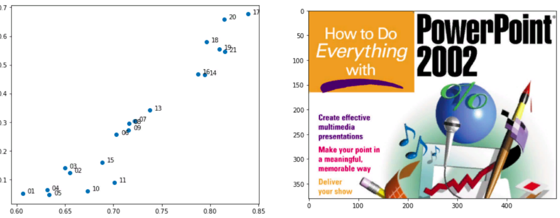

Figure 11: Scatterplot of ERQA and subjective scores on a test using the QADS dataset along with a corresponding ground-truth image.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed a new full-reference ERQA metric for assessing detail restoration by video super-resolution. It compares edges in reference and target videos to analyze how well a VSR model restores the source structure and details. We also created a special dataset for assessing VSR quality and used it to analyze our metric through subjective comparisons. ERQA shows a high correlation with human detail perception and overall better results than traditional as well as state-of-the-art VQA methods. The concept underlying our metric allows it to serve for similar restoration tasks, such as deblurring, deinterlacing, and denoising.

ACKNOWLEDGEMENTS

This work was partially supported by Russian Foundation for Basic Research under Grant 19-01-00785 a and by Foundation for Assistance to Small Innovative Enterprises under Grant UMNIK 16310GU/2021.

REFERENCES

Attar, A., Shahbahrami, A., and Rad, R. M. (2016). Image quality assessment using edge based features. *Multimedia Tools and Applications*, 75(12):7407–7422.

Beron, J., Benitez-Restrepo, H. D., and Bovik, A. C. (2020). Blind image quality assessment for super resolution via optimal feature selection. *IEEE Access*, 8:143201–143218.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698.

Chen, X., Zhang, R., and Zheng, S. (2011). Image quality assessment based on local edge direction histogram. In *2011 International Conference on Image Analysis and Signal Processing*, pages 108–112. IEEE.

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. (2020a).

Table 3: An ablation study of the proposed method.

Stage	PLCC	SRCC
Without compensation (baseline)	0.5035	0.4745
+ Compensation of global shift	0.7395 (+0.2360)	0.6342 (+0.1597)
+ Compensation of local shift (v1.0)	0.8243 (+0.0848)	0.7383 (+0.1041)
+ Penalize false wide edges (v1.1)	0.8316 (+0.0540)	0.7519 (+0.0486)

Table 4: Performance comparison of all metrics with and without global compensation shifts.

Metric	Without compensation		With global pixel shift compensation	
	PLCC	SRCC	PLCC	SRCC
LPIPS	0.8103	0.7077	0.8352 (+0.0249)	0.7377 (+0.0300)
DISTS	0.8094	0.6513	0.8278 (+0.0184)	0.6931 (+0.0418)
MS-SSIM	0.2796	0.4282	0.5992 (+0.3196)	0.5484 (+0.1202)
VMAF	0.2998	0.4692	0.2644 (-0.0354)	0.4572 (-0.012)
VMAF(not clipped)	0.3428	0.4706	0.2999 (-0.0429)	0.4586 (-0.012)

Figure 12: Example of visual perception correlating poorly with restoration quality. The leftmost image looks better, but the rightmost is closer to the original (center).

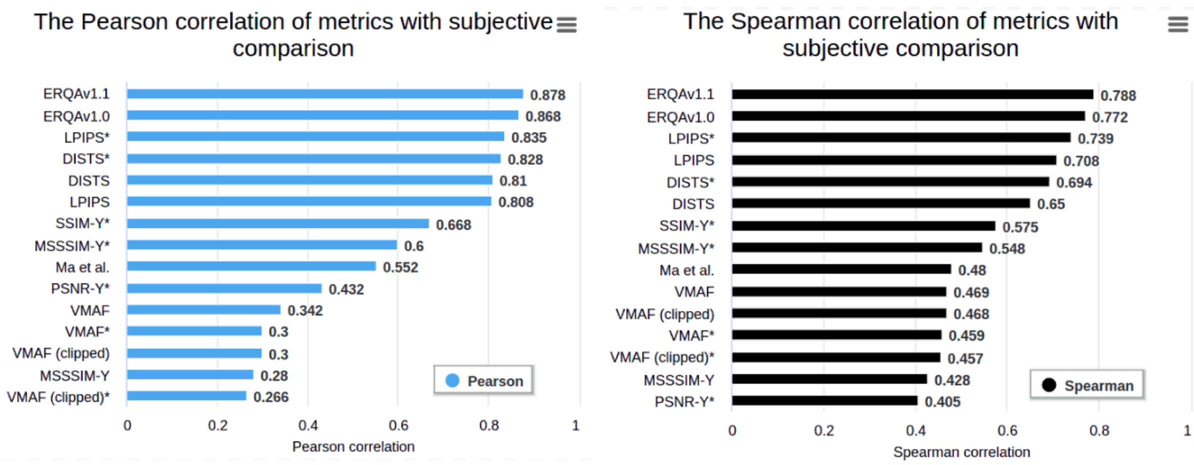

Figure 13: Mean Pearson (PLCC) and Spearman (SRCC) correlations between metric values and subjective assessment. *Denotes metrics calculated using the global-shift compensation scheme.

Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*.

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. (2020b). Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728.

Fang, Y., Liu, J., Zhang, Y., Lin, W., and Guo, Z. (2019). Reduced-reference quality assessment of image super-resolution by energy change and texture variation. *Journal of Visual Communication and Image Representation*, 60:140–148.

Greeshma, M. and Bindu, V. (2018). Novel quality metric for image super resolution algorithms-super resolution entropy metric (srem). In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 151–168. Springer.

Greeshma, M. and Bindu, V. (2020). Super-resolution qual-

Table 5: Pearson correlation of metrics with subjective assessment on all test cases.

Metric	Lego	Toy	Faces	Yarn	QRs	Text-1	Text-2	Car-1	Car-2	Mira	Mean
ERQAv1.0	0.88	0.91	0.84	0.90	0.88	0.87	0.91	0.92	0.87	0.70	0.87
ERQAv1.1	0.87	0.92	0.87	0.90	0.91	0.85	0.88	0.95	0.93	0.70	0.88
SSIM*	0.74	0.37	0.76	0.51	0.71	0.62	0.72	0.89	0.88	0.48	0.67
PSNR*	0.45	0.07	0.64	0.05	0.48	0.29	0.57	0.82	0.75	0.20	0.43
LPIPS	0.85	0.87	0.55	0.88	0.85	0.89	0.79	0.85	0.76	0.79	0.81
LPIPS*	0.87	0.89	0.60	0.90	0.87	0.89	0.79	0.88	0.81	0.85	0.84
DISTS	0.77	0.7	0.71	0.81	0.8	0.91	0.93	0.89	0.83	0.75	0.81
DISTS*	0.77	0.7	0.72	0.82	0.83	0.91	0.95	0.93	0.89	0.76	0.83
MS-SSIM	0.32	0.21	0.19	0.13	0.24	0.30	0.43	0.40	0.31	0.27	0.28
MS-SSIM*	0.76	0.27	0.72	0.44	0.59	0.56	0.57	0.86	0.88	0.35	0.60
VMAF	0.19	0.25	0.25	0.14	0.43	0.42	0.54	0.52	0.40	0.28	0.34
VMAF*	0.12	0.22	0.17	0.17	0.42	0.38	0.48	0.46	0.32	0.26	0.30
VMAF (clip)	0.15	0.21	0.20	0.12	0.4	0.4	0.52	0.45	0.31	0.23	0.30
VMAF (clip)*	0.10	0.20	0.13	0.15	0.39	0.36	0.46	0.40	0.25	0.22	0.27
Ma et al.	0.54	0.85	-0.16	0.8	0.73	—	—	—	—	—	0.55

Table 6: Spearman correlation of metrics with subjective assessment on all test cases.

Metric	Lego	Toy	Faces	Yarn	QRs	Text-1	Text-2	Car-1	Car-2	Mira	Mean
ERQAv1.0	0.87	0.72	0.85	0.85	0.66	0.85	0.89	0.86	0.79	0.38	0.77
ERQAv1.1	0.87	0.66	0.89	0.84	0.65	0.85	0.91	0.92	0.88	0.41	0.79
SSIM*	0.68	0.20	0.81	0.33	0.52	0.57	0.63	0.86	0.86	0.29	0.58
PSNR*	0.36	-0.05	0.66	0.14	0.40	0.40	0.54	0.82	0.72	0.06	0.41
LPIPS	0.70	0.79	0.52	0.79	0.68	0.88	0.63	0.67	0.67	0.75	0.71
LPIPS*	0.78	0.81	0.56	0.84	0.69	0.88	0.65	0.72	0.71	0.75	0.74
DISTS	0.6	0.35	0.69	0.65	0.54	0.84	0.79	0.74	0.72	0.58	0.65
DISTS*	0.6	0.35	0.72	0.71	0.58	0.86	0.88	0.87	0.81	0.56	0.694
MS-SSIM	0.38	0.30	0.59	0.20	0.32	0.48	0.47	0.56	0.59	0.39	0.43
MS-SSIM*	0.77	0.19	0.68	0.35	0.48	0.53	0.6	0.82	0.81	0.25	0.55
VMAF	0.36	0.35	0.61	0.33	0.36	0.52	0.53	0.55	0.60	0.48	0.47
VMAF*	0.33	0.36	0.57	0.38	0.34	0.52	0.48	0.56	0.58	0.47	0.46
VMAF (clip)	0.36	0.35	0.60	0.33	0.36	0.52	0.53	0.55	0.59	0.49	0.47
VMAF (clip)*	0.33	0.36	0.56	0.38	0.34	0.52	0.48	0.56	0.57	0.47	0.46
Ma et al.	0.47	0.88	-0.28	0.62	0.71	—	—	—	—	—	0.48

ity criterion (srqc): a super-resolution image quality assessment metric. *Multimedia Tools and Applications*, 79(47):35125–35146.

Haris, M., Shakhnarovich, G., and Ukita, N. (2019). Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906.

Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., and Huang, F. (2020). 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.

Köhler, T., Bätz, M., Naderi, F., Kaup, A., Maier, A., and Riess, C. (2019). Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2944–2959.

Liu, S., Peng, X., and Liu, Z. (2019). Image quality assessment through contour detection. In *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*, pages 1413–1417. IEEE.

Ma, C., Yang, C.-Y., Yang, X., and Yang, M.-H. (2017). Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16.

Ni, Z., Ma, L., Zeng, H., Chen, J., Cai, C., and Ma, K.-K. (2017). Esim: Edge similarity for screen content image quality assessment. *IEEE Transactions on Image Processing*, 26(10):4818–4831.

Qian, J., Hu, B., Tang, L., Zhang, J., and Liang, S. (2019). Blind quality index for super resolution reconstructed images using first-and second-order structural degradation. *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, 102(11):1533–1541.

Shi, G., Wan, W., Wu, J., Xie, X., Dong, W., and Wu, H. R. (2019). Sisrset: Single image super-resolution subjec-

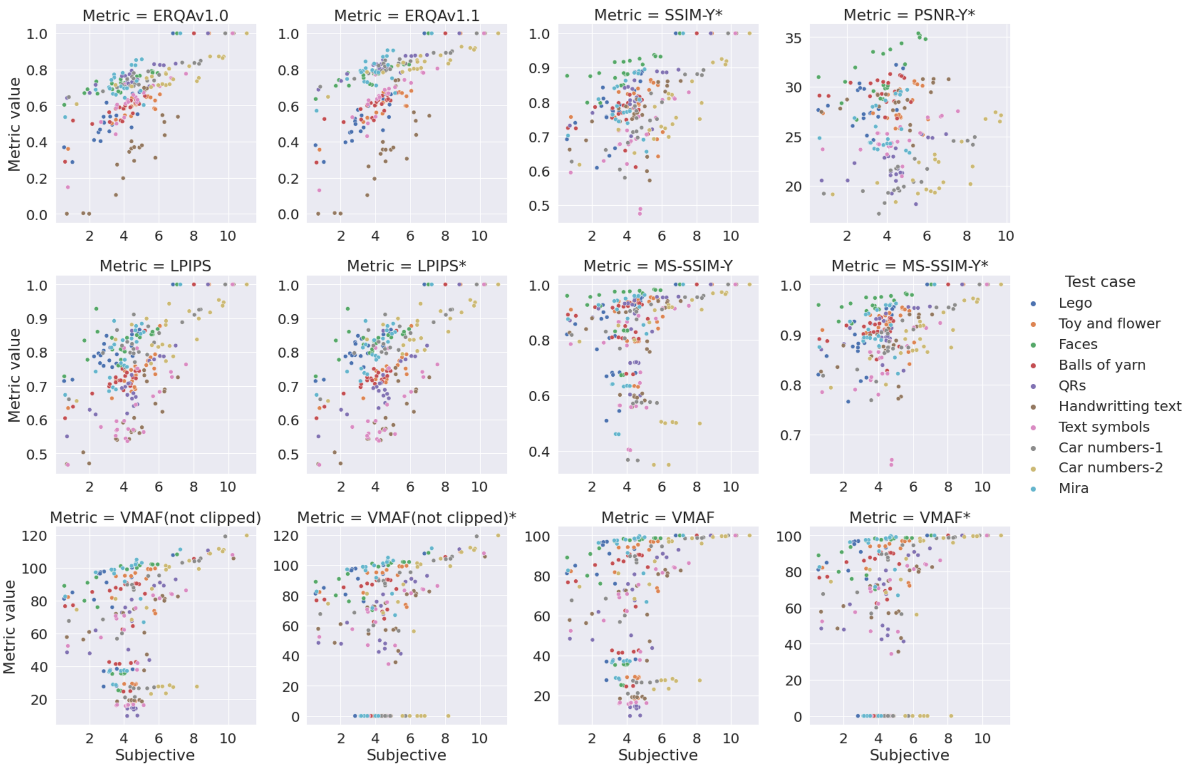

Figure 14: Scatterplots for all metrics in all test cases.

tive evaluation test and objective quality assessment. *Neurocomputing*, 360:37–51.

Tang, L., Sun, K., Liu, L., Wang, G., and Liu, Y. (2019). A reduced-reference quality assessment metric for super-resolution reconstructed images with information gain and texture similarity. *Signal Processing: Image Communication*, 79:32–39.

Wan, W., Wu, J., Shi, G., Li, Y., and Dong, W. (2018). Super-resolution quality assessment: Subjective evaluation database and quality index based on perceptual structure measurement. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Wang, G., Li, L., Li, Q., Gu, K., Lu, Z., and Qian, J. (2017). Perceptual evaluation of single-image super-resolution reconstruction. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3145–3149. IEEE.

Wang, G., Zhu, F., Lu, Z., Yuan, X., and Li, L. (2018). No-reference quality assessment of super-resolution reconstructed images by incorporating domain knowledge. *J. Inf. Hiding Multim. Signal Process.*, 9(2):496–505.

Wang, X., Xie, L., Dong, C., and Shan, Y. (2021). Real-esrgan: Training real-world blind super-resolution with pure synthetic data. *arXiv preprint arXiv:2107.10833*.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee.

Xue, W. and Mou, X. (2011). An image quality assessment metric based on non-shift edge. In *2011 18th IEEE International Conference on Image Processing*, pages 3309–3312. IEEE.

Yang, C.-Y., Ma, C., and Yang, M.-H. (2014). Single-image super-resolution: A benchmark. In *European conference on computer vision*, pages 372–386. Springer.

Yang, L., Sheng, Y., and Chai, L. (2019a). A machine learning based reduced-reference image quality assessment method for single-image super-resolution. In *2019 Chinese Control Conference (CCC)*, pages 3571–3576. IEEE.

Yang, Q., Ma, Z., Xu, Y., Yang, L., Zhang, W., and Sun, J. (2019b). Modeling the screen content image quality via multiscale edge attention similarity. *IEEE Transactions on Broadcasting*, 66(2):310–321.

Yeganeh, H., Rostami, M., and Wang, Z. (2015). Objective quality assessment of interpolated natural images. *IEEE Transactions on Image Processing*, 24(11):4651–4663.

Zhang, K., Zhu, D., Jing, J., and Gao, X. (2019). Learning a cascade regression for no-reference super-resolution image quality assessment. In *2019 IEEE International*

Conference on Image Processing (ICIP), pages 450–453. IEEE.

Zhang, K., Zhu, D., Li, J., Gao, X., Gao, F., and Lu, J. (2021). Learning stacking regression for no-reference super-resolution image quality assessment. *Signal Processing*, 178:107771.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). 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.

Zhou, F., Yao, R., Liu, B., and Qiu, G. (2019). Visual quality assessment for super-resolved images: Database and method. *IEEE Transactions on Image Processing*, 28(7):3528–3541.

Zhou, W., Wang, Z., and Chen, Z. (2021). Image super-resolution quality assessment: Structural fidelity versus statistical naturalness. *arXiv preprint arXiv:2105.07139*.