Towards True Detail Restoration for Super-Resolution: A Benchmark and a Quality Metric

Eugene Lyapustin, Anastasia Kirillova, Viacheslav Meshchaninov, Evgeny Zimin, Nikolai Karetin, and Dmitriy Vatolin
Lomonosov Moscow State University, Moscow, Russia
\{evgeny.lyapustin,anastasia.kirillova,vyacheslav.meshchaninov,evgeney.zimin,nikolai.karetin,dmitriy\}@graphics.cs.msu.ru

Abstract. Super-resolution (SR) has become a widely researched topic in recent years. SR methods can improve overall image and video quality and create new possibilities for further content analysis. But the SR mainstream focuses primarily on increasing the naturalness of the resulting image despite potentially losing context accuracy. Such methods may produce an incorrect digit, character, face, or other structural object even though they otherwise yield good visual quality. Incorrect detail restoration can cause errors when detecting and identifying objects both manually and automatically.

To analyze the detail-restoration capabilities of image and video SR models, we developed a benchmark based on our own video dataset, which contains complex patterns that SR models generally fail to correctly restore. We assessed 32 recent SR models using our benchmark and compared their ability to preserve scene context. We also conducted a crowd-sourced comparison of restored details and developed an objective assessment metric that outperforms other quality metrics by correlation with subjective scores for this task. In conclusion, we provide a deep analysis of benchmark results that yields insights for future SR-based work.

Keywords: Super-Resolution, detail restoration, quality metrics, benchmark

1 Introduction

Super-resolution (SR) involves increasing the spatial resolution of images and videos. Potential uses include image and video enhancement, restoration, and compression [5] through increases in perceptual quality and through more-accurate identification and recognition of objects that are unclear in the original (lower-resolution) image. The topic is under considerable research, with new works appearing monthly.

Although image-based methods, by design, lack additional information when attempting to super-resolve images on the basis of spatial information and
learned distributions, video- and burst-based methods employ a sequence of images, allowing them to restore details using temporal information. Because digital images and videos are discrete, the scene is undersampled, but neighboring frames can help fill in the gap when upscaling. Even small movements caused by camera tremor may yield enough information to accurately increase the resolution 2–3 times, as demonstrated in a Google Pixel 3 camera [37].

Recent advances in SR take advantage of deep learning [8,4,22], and new works regularly set higher quality standards for existing datasets [28]. Most of these works target subjective beauty in the resulting image, which is often fine when an appealing image is helpful or necessary. But no such works to our knowledge have analyzed how those models will restore actual details and context from the original scene. For example, recent GAN methods likely produce results similar to the distribution of their training dataset rather than restoring reference details [13]. But preserving context is necessary for video surveillance, dashboard cameras, and other tasks that require scene interpretation. An incorrectly restored digit or character (Fig. 2) can lead to incorrect conclusions when analyzing images or videos. Likewise, an improperly reproduced face will cause misidentification during both manual and automatic classification. But, as far as we know, there are no existing benchmarks for SR detail restoration.

Analyzing detail restoration requires objective quality metrics that can automatically detect misrepresented details compared with ground truth and that can validate SR models. The most common image-quality metrics, PSNR and SSIM [36], are unable to detect mistaken details. Prior works reveal that both PSNR and SSIM, on which existing benchmarks rely heavily, are unsuitable for assessing SR methods [13,6]. An objective-quality metric for detail restoration should consider how well object structure is restored relative to ground-truth im-

Fig. 1: Left: ranking of subjective score and runtime among 25 Super-Resolution models. Every model was tested on the same video sequence of 100 frames of 480 × 320 resolution. Subjective score is calculated using the Bradley-Terry model. Dotted line shows Pareto front. Right: Visual comparison between the ground truth (GT) images and the results of super-resolution models.
Towards True Detail Restoration for Super-Resolution

Fig. 2: Example of characters changing after video upscaling: two characters from source frame (GT, leftmost) merge into a single one for one method (Real-ESRGAN, central) but not for the other (BasicSR, rightmost).

agery. Because some works use edge maps to prevent distortion of super-resolved faces \[17\], we argue that a similar method will help in assessing SR models and find promising SR approaches.

With detail restoration in mind, our main contributions are the following:

1. We developed a comprehensive benchmark\(^1\) for restoring details when performing super-resolution. It includes objective and subjective assessments.
2. We gathered and analyzed 32 SR models and provided a deep analysis as well as insights for future work in this area.
3. We developed a new quality metric, ERQA\(^2\)v2.0, to assess detail restoration; it outperforms other quality metrics by correlation with subjective scores for SR assessment.

2 Related work

Super-resolution methods. The SR problem gets a lot of attention recently with new different approaches pushing state-of-the-art visual quality and performance.

Earlier deep learning methods like ESPCN \[27\] extracted feature maps in the low-resolution space and used a sub-pixel convolution layer which learned an array of upscaling filters to produce high-resolution output. Several approaches were proposed that predict degradation kernels for super-resolution \[12,20\].

Crucial part for video super-resolution is usage of temporal information. DB-VSR \[26\], for example, estimates a motion blur for the particular video and compensates the motion between frames explicitly. SOF-VSR \[32\] explicitly compensates motion by high-resolution optical flow, estimated from the low-resolution video in a coarse-to-fine manner. Several works \[14,41,30\] use a deformable 3D convolution to compensate the motion between frames. LGFN \[29\] extends deformable convolution approach with decreased multi-dilation convolution units (DMDCUs) for explicit frames alignment. Another approach to dealing with temporal information is using recurrent architectures. Harris et al. \[8\] treated each

\(^1\) https://videoprocessing.ai/vsr/
\(^2\) https://github.com/msu-video-group/erqa
context frame as a separate source of information. These sources are combined in an iterative refinement framework which is aided with explicit inter-frame motion estimation. Isobe et al. [9] divided the input into structure and detail components which are fed to a recurrent unit composed of several proposed two-stream structure-detail blocks.

Generative adversarial networks (GAN) became very popular in deep learning and particularly in super-resolution. Wang et al. [35] has improved network architecture, adversarial loss and perceptual loss of SRGAN [19]. They use RRDB without batch normalization as the basic network building unit and let the discriminator predict relative realness instead of the absolute value. They also use loss features before activation. Later it was extended [34] by a high-order degradation modeling process to better simulate complex real-world degradations. iSeeBetter [3] combines GAN and recurrent approaches. It extracts spatial and temporal information from the current and neighboring frames. Recurrent back-projection network is used as generator and the discriminator is the same as for SRGAN [19].

Some works propose other unique approaches. TGA [10] reorganizes input sequence into several groups of subsequences with different frame rates. Groups allow to extract spatio-temporal information in a hierarchical manner. They are followed by an intra-group and inter-group fusion modules. TMNet [38] is trained for space-time video super-resolution. The temporal information is integrated by deformable convolution with the multi-frame input.

Super-resolution quality metrics. Various video quality metrics are used to measure the performance of super-resolution methods.

PSNR is a traditional quality metric. It uses mean square error and maximum of pixel values to calculate peak to noise ratio. SSIM [36] is another traditional quality metric which calculates average, variance and covariance pixel values on windows of an image. Despite the poor correlation of these metrics with subjective score, they are still commonly used in super-resolution papers.

Recently, deep learning approaches to video quality metrics are gaining popularity. LPIPS [43] is a metric which utilizes deep features of various neural networks (for example, VGG trained on ImageNet) for image comparison. It performs well for its general task and now it is gaining popularity among researchers. Another example is DISTS [6], which uses a convolutional neural network to extract texture representations from images and then combines spatial features of these texture representations with the traditional structure similarity.

Some metrics take more unique approaches. SFSN [45] considers super-resolved images in a two-dimensional space of structural fidelity versus statistical naturalness to account for the behavior of specific super-resolution approaches, such as generative adversarial networks. ERQA [18] is a metric which validates super-resolved images in terms of real detail restoration. In order to achieve this, object edges are detected to be matched with their counterparts on the reference image as edge fidelity is crucial for truthful restoration.

While most super-resolution methods are evaluated by full-reference metrics, they require ground-truth images, which are not always available in prac-
Towards True Detail Restoration for Super-Resolution

tice. NeuralSBS [16] is a no-reference metric designed specifically for the super-resolution task. This metric uses Siamese neural network trained on aligned image pairs. Labels for the training are pick rates for each image in a pair. After training model is able to predict which one of two super-resolved images will be preferred by humans.

Super-resolution benchmarks. A few super-resolution benchmarks were organized by companies and conferences with the purposes to compare diverse algorithms and to find the state-of-the-art for the task.

Vimeo 90k [39] is a large high-quality video dataset for low-level video processing. Authors use this dataset to evaluate their TOFlow algorithm in three video processing tasks: frame interpolation, video denoising and super-resolution. PSNR and SSIM are used for evaluation.

NTIRE 2021 Challenge on Video Super-resolution [28] presents evaluation results of quality restoration competition on full (track 1) and half (track 2) framerate, 247 and 223 participants have registered, respectively. This challenge uses REDS [25] dataset with 30000 images divided in 100-frame sequences. PSNR and SSIM are used to score participants with additional LPIPS values available.

RealVSR [40] dataset consists of 500 paired LR-HR videos captured using the multi-camera system of iPhone 11 Pro Max. To combat misalignment and color differences between the images caused by using two separate cameras, Laplacian pyramid with different loss functions is used. PSNR, SSIM, NIQE [24] and BRISQUE [23] are used for evaluation.

Super-resolution assessment. There is a number of datasets which contain super-resolved images aligned with some subjective scores.

QADS [44] includes 20 reference images and 980 super-resolved images by 21 image SR methods. Mean opinion score for these images were then collected from 100 participants.

SBS180K [16] consists of aligned pairs of images with human pick rates of each image in a pair. Images are evenly extracted from large video database. It was used for training NeuralSBS [16] metric.

PIPAL [13] dataset consists of 250 image patches from two high-quality image datasets DIV2K [1] and Flickr2K [31]. Selected areas are hard to restore due to the presence of high-frequency textures. More than 1.13 million human judgments are collected to assign subjective scores for PIPAL images. PIPAL dataset includes the results of GAN-based methods, which are missing in previous datasets.

3 Benchmark

In this section, we describe our dataset and the benchmark.
3.1 Scene

The scene for our dataset consists of eight patterns separated by X marks, which we used later for processing. Each pattern aims to challenge SR models and, preferably, represent something from everyday life. Scene can be seen in Fig. 3. Here are descriptions for every pattern:

- **Board** is a set of several objects placed on a cork board. Objects present volumetric and textural complexity for SR models. Moreover, there are 6 face images generated using StyleGAN2 [15] for face restoration assessment.

- **License-plate numbers** include several car plates from different regions: USA, China and Russia. It is becoming common to use deep learning for investigations where it is crucial for SR methods to work reliably or explicitly limited.

- **QR codes** can be used to approximate information loss after consecutive image degradation and restoration. There are several QR codes in scene with varying size and similar coded content complexity.

- **Text** is common thing to see in images and videos. Incorrect text restoration may cause corruption of sentences and overall scene context. There are multiple characters of different fonts and a set of handwritten text.

- **Noise** pattern provides a unique “footprint” for each SR method and can be used to detect overfitting.

- **Resolution test chart** is a commonly used for camera calibration and analysis. The chart used in a scene is compliant with ISO 12233 standard.

- **Metal paper** is a crumpled foil which provide unique light reflections for each video frame that complicates overall restoration.

- **Color lines** add complexity to restoration because of high-frequency color changes.
3.2 Dataset preparation

Capture. We shot the videos using a Canon EOS 7D camera in daylight conditions to ensure naturalness. The camera settings aimed to minimize blur and obtain the appropriate brightness: the ISO was 4000 and the aperture 400. Each video is a sequence of 100 photos taken in burst mode with a maximum resolution of $5184 \times 3456$ because the camera video resolution and compression quality are limited. We refer to these videos as source videos.

High- and low-resolution frames. To prevent undesirable motion blur and noise in the ground-truth frames, and to avoid running afoul of memory restrictions when applying models to high-resolution images, we downscaled the source videos to $1920 \times 1080$ using bicubic interpolation. We refer to these downscaled videos as the ground-truth or high-resolution (HR) data. The low-resolution (LR) frames are the HR frames further downscaled by $4 \times$ to $480 \times 270$.

Realistic noise. Because real-world imagery generally contains noise, we generated additional LR videos using noise from the source videos. We use a noise model from CBDNet [7]. To make it natural, we conducted an experiment to measure the camera’s noise parameters. We recorded 100 frames using the Canon EOS 7D on a tripod, calculated the average frame, and subtracted it from the actual frames to obtain noise samples. Our next step was to tune the noise-model parameters to match the real noise distribution in the camera. The results were $\sigma_s = 0.001$ and $\sigma_c = 0.035$ (see Fig. 4).

3.3 Subjective comparison

To calculate accurate ranks for detail restoration we conduct crowd-source subjective comparison on Subjectify.us service. From each video we extract a few video-crops with the most difficult patterns for detail-restoration. For comparison we randomly show each participant a pair of frame patches from two super-resolution models and ask them to choose the patch on which details are restored better, contrary to the most beautiful one. Each pair of patches was shown to 10–15 participants until confidence interval stopped changing. We decided to work with patches, because it is very difficult to visually compare small details in a full frame. Each participant compared 25 pairs total with 3 of them for verification to filter out untruthful participants. 1400 participants total participated in comparison and their answer were used for final subjective scores calculation with Bradley-Terry model [2].

4 Quality Metric

Inspired by ERQA metric [18], we propose sequential refinement of restored boundaries ERQAv2.0, following the naming from original authors (ERQAv1.0 and ERQAv1.1). Unlike ERQAv1.1, however, our approach is using a lower-level technique instead of an explicit edge detector.
First, we compute the gradients of the ground-truth and input frames using the simple convolutional kernels $[-0.5, 0, 0.5]$ and $[-0.5, 0, 0.5]^T$. Since outliers may degrade the results, we filter out gradients with magnitude greater than the 85th percentile of the frame-gradient distribution. Experimental observations show that employing such filter, increases the algorithm’s performance in both quality and speed. Next, we consider gradients in the ground-truth and input frames matched, if the cosine of the angle between them exceeds 0.85.

As in Kirillova et al. [18] was shown, some SR methods can generate frames with small pixel shifts relative to ground-truth (see Fig. 5). Authors of original ERQA addressed the issue by grid-searching pixel shifts to find one that maximizes PSNR. Our experiments shown that pixel shifts may vary across frame so we developed an iterative approach. ERQAv2.0 first finds measure of similarity — the number of matching gradients between the input and ground truth frames with 5-pixel radius. Next we sort the shifts in descending order by this measure of similarity. We then iteratively take the first $N$ shifts, find matches between gradient maps, add these matches to the true positive mask and remove them from the gradient maps of the input and ground-truth frames. We found that the true positive mask stops changing after $N = 35$.

Our approach yields three gradient maps: true positive, false positive and false negative. False positive and false negative masks are the rest of the gradient maps of the input and ground-truth frames accordingly. Then we compute the $F_\beta$-score with $\beta = 0.5$. Super-resolution methods tend to generate wide edges which means high false positive rate, so we use $\beta = 0.5$. Experimental results are presented in Sec. 5.

The metric can automatically assess detail-restoration quality and, on the basis of an error map, find frame areas that contain artifacts produced by SR, as shown in Fig. 6.

5 Experiments

In this section we show evaluation results of SR methods and quality metrics on our dataset.

5.1 Super-resolution models

For our experiments we chose models so that various approaches were included in our benchmark and compared with each other: alignment by optical flow [33,26] and 3D convolutions [14,41], deformable convolutions [30,29], recurrent models [11,8,9], meta-learning [20], and generative adversarial networks [3,34]. We mostly used the latest methods, because they generally show better results.

We evaluated 32 different Super-resolution models by applying them to our dataset on a server with Intel(R) Xeon(R) Silver 4216 CPU and NVIDIA Titan RTX GPU. The execution time is calculated as a full runtime divided by the number of frames in a sequence with $480 \times 320$ resolution. FPS is an inverse value to the execution time. You can see top-10 leaderboard in Tab. 1 and full leaderboard in supplementary materials.
Towards True Detail Restoration for Super-Resolution

Fig. 5: The distribution of global pixel shifts of some methods relative to ground-truth. Some models produce non-constant shifts which can be explained by training on a different downsampling method.

Fig. 6: Heat map of Super-Resolution artifacts showcasing the difference in text restoration.

| Rank | Model         | Subjective ↑ | ERQA v2.0 ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FPS ↑ |
|------|---------------|--------------|-------------|--------|--------|---------|------|
| 1    | VRT [22]      | 7.627        | 0.851       | 31.669 | 0.902  | 0.241   | 0.166|
| 2    | BasicVSR [4]  | 7.186        | 0.846       | 31.443 | 0.900  | 0.240   | 2.128|
| 3    | RBPN [8]      | 7.068        | 0.841       | 31.407 | 0.899  | 0.260   | 0.043|
| 4    | DBVSR [26]    | 6.947        | 0.835       | 31.071 | 0.894  | 0.274   | 0.241|
| 5    | iSeeBetter [3]| 6.809        | 0.839       | 31.104 | 0.896  | 0.259   | 0.044|
| 6    | LGFN [29]     | 6.505        | 0.831       | 31.291 | 0.898  | 0.275   | 0.667|
| 7    | DynaVSR-R [20]| 6.135        | 0.802       | 28.37  | 0.865  | 0.274   | 0.177|
| 8    | TMNet [38]    | 6.000        | 0.821       | 30.364 | 0.885  | 0.270   | 1.136|
| 9    | COMISR [21]   | 5.636        | 0.794       | 26.708 | 0.840  | 0.271   | 1.613|
| 10   | RSDN [9]      | 5.565        | 0.764       | 25.321 | 0.826  | 0.333   | 1.961|

Table 1: A comparison of top-10 Super-Resolution methods by subjective score and objective metrics.
5.2 Quality metrics

On top of super-resolution results on our dataset we calculated several quality metrics that are usually used for super-resolution assessment. You can see metrics correlation with subjective scores in Tab. 2.

| Metric     | PLCC | SRCC |
|------------|------|------|
| ERQA v2.0  | 0.899| 0.805|
| ERQA v1.1  | 0.878| 0.787|
| LPIPS      | 0.835| 0.738|
| DIoTS      | 0.828| 0.719|
| SSIM       | 0.670| 0.575|
| MS-SSIM    | 0.599| 0.548|
| SFSN       | 0.474| 0.512|
| PSNR       | 0.431| 0.471|
| VMAF       | 0.264| 0.459|
| NeuralSBS  | 0.066| 0.029|

Table 2: Mean Pearson (PLCC) and Spearman (SRCC) correlation coefficients of used metrics with subjective comparison.

Fig. 7: Clustering of algorithms by a few objective metrics (ERQA v2.0, LPIPS, PSNR, and SSIM). Centers of clusters were chosen as representative methods for the analysis of subjective comparison. The mean values of ERQA v2.0 and LPIPS are chosen for axes on this plot.

5.3 Analysis of subjective-comparison method

As subjective assessment is expensive and time consuming with a number of comparisons grows quadratic to the number of compared methods. Therefore we sampled a small number of SR methods to analyze several hypotheses. Conducted a subjective evaluation of all models on tests with hand tremor motion and bicubic degradation (Tab. 1), we were able to choose a few metrics that have high correlation with subjective assessment (Tab. 2). Then we used two traditional metrics (PSNR and SSIM) and two metrics more consistent with subjective assessment (ERQA v1.1 [18] and LPIPS [43]). We also selected three patches from the scene for this experiment so scores between every patch are distinct. We used 12 features (4 different metrics for 3 patches) to group 21 SR models using KMedoids with 6 clusters. We consider the resulting six group centers to be representative methods for later subjective experiments. (Fig. 7).
Finally, we conducted four subjective comparisons of representative methods under different test settings:

1. Hand tremor and Bicubic Interpolation (BI)
2. Parallel motion and BI
3. Rotation and BI
4. Hand tremor and Blur Downsampling (BD) — direct downsampling after Gaussian smoothing

The experiment showed that subjective scores on tests with different types of motion correlate highly in both the Pearson and Spearman coefficients (Tab. 3) and are greatly consistent with each other. So, it is not necessary to conduct a subjective assessment on all types of camera motion. On the contrary, it is important to collect subjective scores for models on both BI and BD degradation as the results strongly differ. For the top two models, the subjective score on a test with BD degradation differs from the score on the same test with BI degradation. Finally, we conduct subjective comparison on both BI and BD degradation.

| Test                | Hand tremor + BI | SRCC | PLCC |
|---------------------|------------------|------|------|
| Parallel motion + BI| 0.943            | 0.946|
| Rotation + BI       | 1.000            | 0.955|
| Hand tremor + BD    | 0.943            | 0.685|

Table 3: Pearson (PLCC) and Spearman (SRCC) correlation coefficients between subjective scores on Hand tremor + BI test and other tests. The correlation was calculated on 6 representative algorithms.

| Metric      | SRCC | PLCC |
|-------------|------|------|
| ERQA v2.0   | 0.721| 0.797|
| ERQA v1.1   | 0.548| 0.677|
| PSNR        | 0.868| 0.861|
| SSIM        | 0.877| 0.941|
| NeuralSBS   | 0.879| 0.645|
| SFSN        | 0.546| 0.662|

Table 4: Pearson (PLCC) and Spearman (SRCC) correlation coefficients of metrics between tests with BI and BD input degradation types.

5.4 Analysis of values of objective metrics for VSR models

While there is no one true way to downsample images for SR models training, the two most popular ones are Bicubic Interpolation (BI) and Blur Downsampling (BD) — direct downsampling after Gaussian smoothing. As we considered in Sec. 5.3, subjective scores differ between different degradations. ERQA v1.1 and ERQA v2.0 also exhibit low correlation between tests with BI and BD (Tab. 4), which shows that all models vary in detail restoration quality between BI and BD degradation. ERQA values on BI almost always exceed values of the same
metric on BD, because Gaussian smoothing generally reduce the quality of edge restoration. PSNR and SSIM exhibit high correlation between BI and BD, which proves that these metrics are less sensitive to changes of the input degradation. NeuralSBS achieves a high SRCC between BI and BD, but its BI value are always greater than its BD value. On the other hand, SFSN values for BI are less than or equal to its values for BD.

We notice that models generally yield better results with degradation which they used during training which strongly suggest that SR models may overfit to the specific downsampling method. So a model’s rank can change greatly depending on the input degradation and a model produce much better result on the same degradation type it was trained on (Tab. 5), but some stable models, such as LGFN [29], provide similar results on both tests. To our knowledge, some authors use several types of degradation chosen randomly to neglect such overfitting [42].

Fig. 8: Average difference of ERQAv1.1 values, calculated on tests with parallel motion and rotation.

Fig. 9: Average difference of ERQAv2.0 values calculated on tests with and without noise.

In Sec. 5.3 we showed that subjective scores on different types of motion are highly correlated. The same is true for metric values, but the more complicated the motion, the greater challenge for a given model. Thus, we considered how these values change with motion type for each model. We compared results using clean data to exclude noise effects. ERQAv2.0 values for parallel camera motion are generally higher than those for rotation. PSNR and SSIM indicate a much smaller difference between motion types, but for nearly all models, the metric values for hand tremor average slightly better than those for parallel motion, and those for parallel motion average slightly better than those for rotation. The dependence between metric values for rotation and parallel motion is more difficult because different models work better on different motion types (Fig. 8). We showed that the ranking can change depending on camera motion (Tab. 7),

In Sec. 5.3 we showed that subjective scores on different types of motion are highly correlated. The same is true for metric values, but the more complicated the motion, the greater challenge for a given model. Thus, we considered how these values change with motion type for each model. We compared results using clean data to exclude noise effects. ERQAv2.0 values for parallel camera motion are generally higher than those for rotation. PSNR and SSIM indicate a much smaller difference between motion types, but for nearly all models, the metric values for hand tremor average slightly better than those for parallel motion, and those for parallel motion average slightly better than those for rotation. The dependence between metric values for rotation and parallel motion is more difficult because different models work better on different motion types (Fig. 8). We showed that the ranking can change depending on camera motion (Tab. 7),
but these changes are much smaller than they are with input-degradation type (Tab. 5). For example, iSeeBetter [3] declines considerably when forceful motion is added, whereas other top models are less sensitive to changes in camera motion. DBVSR [26] and ESRGAN [35] appear more stable to motion because their metric values decrease, but that decrease is less than for other top models (RBPN [8], LGFN [29], and iSeeBetter [3]).

| Model           | BI | BD | Mean | Trained |
|-----------------|----|----|------|---------|
| DynaVSR-R       | 10 | 1  | BD   |         |
| Real-ESRGAN     | 9  | 2  | 2    | both    |
| LGFN            | 5  | 6  | 3    | BI      |
| DBVSR           | 1  | 11 | 4    | BI      |
| RBPN            | 2  | 10 | 5    | BI      |
| iSeeBetter      | 4  | 8  | 6    | BI      |
| HCFlow          | 6  | 7  | 7    | —       |
| RealSR          | 8  | 9  | 8    | —       |
| DynaVSR-V       | 18 | 3  | 9    | BD      |
| ESRGAN          | 3  | 16 | 10   | BI      |
| SwinIR          | 16 | 4  | 11   | —       |
| TMNet           | 7  | 14 | 12   | BI      |
| Real-ESRnet     | 21 | 5  | 13   | both    |

Table 5: Ranking of models by ERQAv2.0 on BI and BD input down-sampling.

| Model            | w/o noise | w/ noise | Mean |
|------------------|-----------|----------|------|
| Real-ESRGAN      | 8         | 1        | 1    |
| RealSR           | 9         | 3        | 2    |
| SwinIR           | 15        | 2        | 3    |
| RBPN             | 2         | 9        | 4    |
| DBVSR            | 1         | 12       | 5    |
| iSeeBetter       | 4         | 8        | 6    |
| Real-ESRnet      | 21        | 4        | 7    |
| LGFN             | 5         | 19       | 10   |
| GFPGAN           | 25        | 5        | 19   |
| ESRGAN           | 3         | 25       | 23   |

Table 6: Ranking of models by ERQAv2.0 on tests with and without noise, along with ranking by mean ERQAv2.0 on both tests. Models with a rank not worse than 5 on any test (with or without noise, or by mean ERQAv2.0) were visualized. Table is sorted by mean ERQAv2.0.

We also compared tests with and without noise. Generally, metric values for tests with noise are lower than those without noise (Fig. 9), but all metrics correlate poorly between these types of tests (Tab. 8). We showed that although models generally works worse with noise input data, rank of models can change substantially when noise is added to the input video (Tab. 6), because metric values for some models decrease more significantly than for others. Real-ESRGAN [34], D3Dnet [41], and SOF-VSR-BI are more stable to noise than others, while RBPN [8], DBVSR [26], and ESRGAN [35] are superior to other models on clean data but fail to perform well with noise.

6 Conclusion

In this paper we propose a new benchmark for super-resolution detail restoration. We assess 32 super-resolution models both objectively and subjectively. We propose a new quality metric ERQAv2.0 that has better correlations with subjective scores than other metrics and more closely resembles human ranking.
| Model        | Motion 1 | Motion 2 | Motion 3 | Mean |
|-------------|----------|----------|----------|------|
| RBPN        | 2        | 1        | 3        | 1    |
| DBVSF       | 4        | 2        | 2        | 2    |
| ESRGAN      | 5        | 4        | 1        | 3    |
| LGFN        | 3        | 3        | 4        | 4    |
| iSeeBetter  | 1        | 5        | 6        | 5    |
| HCFlow      | 6        | 7        | 5        | 6    |
| DynaVSR-R   | 8        | 6        | 7        | 7    |
| TMNet       | 7        | 8        | 8        | 8    |
| RealSR      | 10       | 9        | 9        | 9    |
| TDAN        | 9        | 10       | 11       | 10   |
| Real-ESRGAN | 14       | 11       | 10       | 11   |

Table 7: Ranking of models by ERQAv2.0 on tests with different camera motion (Motion 1 = Hand tremor, Motion 2 = Parallel motion, Motion 3 = Rotation), along with ranking by mean ERQAv2.0 on these three tests. Models with a rank not worse than 10 on any motion type were visualized. Table is sorted by mean ERQAv2.0.

It was improved and then used to study the impact of input degradations on the behavior of different super-resolution models.

Experiments show that modern SR models may overfit to downsampling method used during training, LGFN [29] being the exception. Camera movement also affects model behavior but not as much, however performance of some models declines with presence of motion, iSeeBetter [3] being an example. Adding noise to the image generally lowers metric values and substantially changes model rank: models with poor performance on clean images become relatively better with noise appearance, and on the contrary algorithms achieving high metric values before injection noticeably decrease in performance on noisy images. And in total video-based methods perform better for the detail restoration task than image-based method, confirming the earlier hypotheses.

We hope that our work can help to gain insights into detail restoration for Super-Resolution methods and develop this research area in the future. We will continue to update the benchmark with new promising methods as they appear.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017) 5
2. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39(3/4), 324–345 (1952) 7
3. Chadha, A., Britto, J., Roja, M.M.: iseebetter: Spatio-temporal video super-resolution using recurrent generative back-projection networks. Computational Visual Media 6(3), 307–317 (2020) 4, 8, 9, 13, 14
4. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4947–4956 (2021) 2, 9
5. Chen, Y., Li, Q., Zhang, A., Zou, L., Jiang, Y., Xu, Z., Li, J., Yuan, Z.: Higher quality live streaming under lower uplink bandwidth: An approach of super-resolution based video coding. In: Proceedings of the 31st ACM Workshop on Network and Operating Systems Support for Digital Audio and Video. p. 74–81. NOSSDAV ’21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3458306.3458874, https://doi.org/10.1145/3458306.3458874
6. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. CoRR abs/2004.07728 (2020), https://arxiv.org/abs/2004.07728 2, 4
7. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 7
8. Haris, M., Shakharovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3897–3906 (2019) 2, 3, 8, 9, 13
9. Isobe, T., Jia, X., Gu, S., Li, S., Wang, S., Tian, Q.: Video super-resolution with recurrent structure-detail network. In: European Conference on Computer Vision. pp. 645–660. Springer (2020) 4, 8, 9
10. Isobe, T., Li, S., Jia, X., Yuan, S., Slabaugh, G., Xu, C., Li, Y.L., Wang, S., Tian, Q.: Video super-resolution with temporal group attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8008–8017 (2020) 4
11. Isobe, T., Zhu, F., Jia, X., Wang, S.: Revisiting temporal modeling for video super-resolution. arXiv preprint arXiv:2008.05765 (2020) 8
12. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 466–467 (2020) 3
13. Jinjin, G., Haoming, C., Haoyu, C., Xiaoxing, Y., Ren, J.S., Chao, D.: Pipal: A large-scale image quality assessment dataset for perceptual image restoration. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 633–651. Springer International Publishing, Cham (2020) 2, 5
14. Jo, Y., Oh, S.W., Kang, J., Kim, S.J.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3224–3232 (2018) 3, 8
15. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 6
16. Khrulkov, V., Babenko, A.: Neural side-by-side: Predicting human preferences for no-reference super-resolution evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4988–4997 (June 2021) 5
17. Kim, J., Li, G., Yun, I., Jung, C., Kim, J.: Edge and identity preserving network for face super-resolution. Neurocomputing 446, 11–22 (2021). https://doi.org/https://doi.org/10.1016/j.neucom.2021.03.048, https://www.sciencedirect.com/science/article/pii/S0925231221004227

18. Kirillova, A., Lyapustin, E., Antsiferova, A., Vatolin, D.: Erqa: Edge-restoration quality assessment for video super-resolution. arXiv preprint arXiv:2110.09992 (2021) 4, 7, 8, 10

19. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017) 4

20. Lee, S., Choi, M., Lee, K.M.: Dynavsr: Dynamic adaptive blind video super-resolution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2093–2102 (2021) 3, 8, 9

21. Li, Y., Jin, P., Yang, F., Liu, C., Yang, M.H., Milanfar, P.: Comisr: Compression-informed video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2543–2552 (2021) 9

22. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. arXiv preprint arXiv:2201.12288 (2022) 2, 9

23. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing 21(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050

24. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. IEEE Signal Processing Letters 20(3), 209–212 (2013). https://doi.org/10.1109/LSP.2012.2227726

25. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019) 5

26. Pan, J., Bai, H., Dong, J., Zhang, J., Tang, J.: Deep blind video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4811–4820 (2021) 3, 8, 9, 13

27. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016) 3

28. Son, S., Lee, S., Nah, S., Timofte, R., Lee, K.M.: Ntire 2021 challenge on video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 166–181 (June 2021) 2, 5

29. Su, D., Wang, H., Jin, L., Sun, X., Peng, X.: Local-global fusion network for video super-resolution. IEEE Access 8, 172443–172456 (2020) 3, 8, 9, 12, 13, 14

30. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3360–3369 (2020) 3, 8

31. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 114–125 (2017) 5
32. Wang, L., Guo, Y., Lin, Z., Deng, X., An, W.: Learning for video super-resolution through HR optical flow estimation. In: ACCV (2018) 3
33. Wang, L., Guo, Y., Liu, L., Lin, Z., Deng, X., An, W.: Deep video super-resolution using hr optical flow estimation. IEEE Transactions on Image Processing 29, 4323–4336 (2020) 8
34. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: International Conference on Computer Vision Workshops (ICCVW) (2021) 4, 8, 13
35. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: The European Conference on Computer Vision Workshops (ECCVW) (September 2018) 4, 13
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13(4), 600–612 (2004) 2, 4
37. Wronski, B., Garcia-Dorado, I., Ernst, M., Kelly, D., Krainin, M., Liang, C.K., Levoy, M., Milanfar, P.: Handheld multi-frame super-resolution. ACM Trans. Graph. 38(4) (Jul 2019). https://doi.org/10.1145/3306346.3323024, https://doi.org/10.1145/3306346.3323024 2
38. Xu, G., Xu, J., Li, Z., Wang, L., Sun, X., Cheng, M.M.: Temporal modulation network for controllable space-time video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6388–6397 (2021) 4, 9
39. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision 127(8), 1106–1125 (2019) 5
40. Yang, X., Xiang, W., Zeng, H., Zhang, L.: Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4781–4790 (October 2021) 5
41. Ying, X., Wang, L., Wang, Y., Sheng, W., An, W., Guo, Y.: Deformable 3d convolution for video super-resolution. IEEE Signal Processing Letters 27, 1500–1504 (2020) 3, 8, 13
42. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4791–4800 (October 2021) 12
43. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 4, 10
44. Zhou, F., Yao, R., Liu, B., Qiu, G.: Visual quality assessment for super-resolved images: Database and method. IEEE Transactions on Image Processing 28(7), 3528–3541 (2019). https://doi.org/10.1109/TIP.2019.2898638 5
45. Zhou, W., Wang, Z., Chen, Z.: Image super-resolution quality assessment: Structural fidelity versus statistical naturalness. arXiv preprint arXiv:2105.07139 (2021) 4