Neural Side-By-Side: Predicting Human Preferences for No-Reference Super-Resolution Evaluation

Valentin Khrulkov
Yandex, Russia
khrulkov.v@gmail.com

Artem Babenko
Yandex, Russia
National Research University Higher School of Economics
Moscow, Russia
artem.babenko@phystech.edu

Abstract

Super-resolution based on deep convolutional networks is currently gaining much attention from both academia and industry. However, lack of proper evaluation measures makes it difficult to compare approaches, hampering progress in the field. Traditional measures, such as PSNR or SSIM, are known to poorly correlate with the human perception of image quality. Therefore, in existing works common practice is also to report Mean-Opinion-Score (MOS) — the results of human evaluation of super-resolved images. Unfortunately, the MOS values from different papers are not directly comparable, due to the varying number of raters, their subjectivity, etc. By this paper, we introduce Neural Side-By-Side — a new measure that allows super-resolution models to be compared automatically, effectively approximating human preferences. Namely, we collect a large dataset of aligned image pairs, which were produced by different super-resolution models. Then each pair is annotated by several raters, who were instructed to choose a more visually appealing image. Given the dataset and the labels, we trained a CNN model that obtains a pair of images and for each image predicts a probability of being more preferable than its counterpart. In this work, we show that Neural Side-By-Side generalizes across both new models and new data. Hence, it can serve as a natural approximation of human preferences, which can be used to compare models or tune hyperparameters without raters' assistance. We open-source the dataset and the pretrained model[1] and expect that it will become a handy tool for researchers and practitioners.

1. Introduction

Image super-resolution (SR) is a long-standing task in image processing, which aims to recover high-resolution images from low-resolution ones. During the last decades, this task has attracted ever-growing research attention, since super-resolution is a critical component of many computer vision pipelines, including surveillance[15, 31], video enhancement[1], medical imaging[4] and others. Over the years, a plethora of super-resolution methods has been developed, from simple interpolation techniques to more powerful models, employing deep architectures, which currently dominate the super-resolution landscape. Overall, these days SR is an active research direction, and the community constantly develops new model architectures, optimization objectives, regularization, and normalization techniques.

Given a large number of different SR models appearing in the literature, practitioners need an instrument to select the model that is the most effective on the particular data. Furthermore, in many papers model hyperparameters are chosen based on the performance on the academical benchmarks and can be suboptimal for the real task in hand. In practice, model selection and hyperparameter tuning are even more challenging, since "ground truth" high-resolution images can be absent for real data. This implies that the established full-reference measures (such as PSNR, SSIM[29], LPIPS[32]) cannot be used, and human evaluation is often the only option to choose the model. The human evaluation is typically referred to as Mean-Opinion-Score (MOS) and denotes the average rating that human raters assigned to images produced by the particular SR model. Being an adequate evaluation measure for SR, MOS, however, is both time-consuming and expensive, which prevents its usage, e.g. to tune hyperparameters.

In this paper, we introduce *Neural Side-By-Side* — a new instrument for no-reference super-resolution evaluation that can be used for model comparison or hy-

[1] https://github.com/KhrulkovV/NeuralSBS

perparameter tuning. In a nutshell, we collect a large number of aligned image pairs, where each image is produced by some super-resolution model. Then each pair is labeled by humans, who are instructed to select a more visually appealing image from each pair. Importantly, the pairs consist of aligned images, i.e. they correspond to the same low-resolution image. Hence, the rater labels reflect only image quality and do not relate to their content. We refer to this dataset as **SBS180K** and release both images and labels for further research. Given the **SBS180K** dataset, we train a CNN model that obtains an image pair and predicts probabilities of each image being more attractive compared to its counterpart. These probabilities, averaged over a large number of pairs, can then be used as a quantitative measure to compare two super-resolution models. In the experimental section, we demonstrate that Neural Side-By-Side generalizes to both new models and new image sets, hence it can be used as an effective "approximation" of human evaluation. Indeed, there were several attempts to learn no-reference measures of general image quality[16, 8, 17, 11, 7, 23, 10, 13, 27, 5]. However, these measures are not tailored specifically for image super-resolution and we demonstrate their inferiority to Neural Side-By-Side in the experimental section.

We summarize the contributions of our paper as follows:

1. We introduce Neural Side-By-Side — a new no-reference technique to compare super-resolution models. The Neural Side-By-Side outperforms the existing no-reference measures in terms of approximating human evaluation and can serve as a handy tool for both academicians and practitioners.

2. We release **SBS180K** — a dataset of aligned image pairs, produced by different super-resolution models. The dataset is needed to reproduce the results from our paper and can be used to train and evaluate new models.

3. We evaluate several established models with Neural Side-By-Side on common benchmarks. We expect that the obtained numbers can be useful for further development of more advanced super-resolution models.

2. Related work

In this section, we aim to put our work in context with existing literature.

Super-resolution evaluation. The two most common measures for image super-resolution are currently peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM)[29]. Since both PSNR and SSIM are known to only loosely correlate with visual perception[12], DNN-based metrics, such as LPIPS[32], are currently gaining popularity. In many practical scenarios, PSNR/SSIM/LPIPS cannot be used, since they are full-reference, i.e., require ground truth high-resolution images, which can be absent for real data. Therefore, a reliable no-reference measure is needed for super-resolution practitioners.

No-reference image quality evaluation. No-Reference Image Quality Assessment (NR-IQA) is a well-known problem of predicting the "quality" of individual images by themselves, without a reference image. The state-of-the-art NR-IQA models are usually trained on large datasets of images with assigned labels of visual quality, produced by human raters. The first methods of this family were based on hand-engineered image descriptors[20, 19], while more recent methods use deep neural networks[16, 8, 17, 11, 7, 23, 10, 13, 27, 5]. These methods typically differ in architecture details and optimization objectives. In the context of super-resolution the common weakness of NR-IQA measures, however, is that all of them are trained on datasets of natural images, which do include images produced by different super-resolution methods. It is this weakness of the existing measures we address by our paper.

NR-IQA for image super-resolution. Probably, the closest work to ours is [18], which also aims to learn NR-IQA measure for super-resolution. [18] introduces a dataset of 1620 super-resolved color images from 30 source images. Specifically, each source image is first processed by six different combinations of downsampling and blurring to generate six low-resolution images. Then nine super-resolution image reconstruction algorithms are adopted to generate the super-resolved images. All 1620 images are rated by humans and a simple model based on handcrafted visual features is trained to approximate the rates. While the measure from [18] has received certain interest from the community (e.g. it is used in the PIRM challenge[2]), this measure has three important drawbacks:

- It is based only on 30 source images, which can be insufficient to generalize to unseen data.

- It is based on human rates that were obtained only for early super-resolution models (up to 2014), hence it can perform inadequately on the results of more advanced models.

- It is based on handcrafted image features, which can be suboptimal in terms of capturing the signal needed for NR-IQA.

In contrast, the proposed SBS180K dataset is much larger, contains the human rates for the recent SR models, and uses learnable deep features for the corresponding NR-IQA model.

Prior datasets for learnable NR-IQA. We summarize the main existing datasets that can be potentially used to learn NR-IQA measures for super-resolution in Table 1. Compared to existing alternatives, our SBS180K dataset is both large-scale and is tailored to the super-resolution task.

Dataset	Size	SR specific	Comment
Ma[18]	1680	✓	SR before 2014
AVA[24]	255K	✗	Natural images
SBS180K	180K	✓	SR models before 2020

Table 1: Comparison of the existing datasets that can potentially be used to train NR-IQA measure for super-resolution.

3. The SBS180K dataset

Our dataset consists of 176440 aligned image pairs (split into 167019 train pairs and 9421 test pairs), labeled according to human aesthetic preferences. The labeling is provided with a single number — *score* in the range $[0, 1]$, reflecting the aesthetic appeal of the second image in the pair with respect to the first one. Formally, this score equals the percentage of raters that prefer the second image to the first image. Each pair of images corresponds to two variants of the same low-resolution image, upsampled via two different super-resolution algorithms (including SRGANs, MSE upsampling networks, bicubic interpolation, etc.). As a source of images, we used random frames from a diverse set of video fragments, described below. To avoid very similar images, we sampled only one frame from each chunk of 200 consecutive frames. An average length of fragments was about 30 seconds, and 3–5 frames were sampled from each video.

Now we describe the specifics of video collection and the annotation protocol.

3.1. Video selection

The video fragments were gathered to cover a variety of possible practical super-resolution scenarios, such as old classic movies, cartoons, TV shows and sporting events. In total there were 2071 unique video fragments of 30-second length gathered from various sources: an old B&W movie, three colored movies from before 1970, three colored movies from 1970–2000s and three from 2000–2010s, one old animated cartoon, one modern animated cartoon, and one anime series, HD TV shows from the following categories each: crime, adventures, news, musical clips and concerts, one soccer and one hockey sporting events from 1981 and 1986, broadcasted TV channels, auxilliary videos including Vimeo clips, UltraHD concert clips, HDR youtube videos.

3.2. Model selection

Our dataset was collected during a continuing effort to find the best possible super-resolution model. We have experimented with roughly 170 unique models based on various tweaks and adjustments of popular super-resolution models, including models based on Generative Adversarial Networks (GANs) and super-resolution convolutional networks trained with MSE loss. More concretely, most of our models were based on the following SR algorithms: SRGAN [12], DRCAN [9], SRResNet [12], ESRGAN [28]. These algorithms usually include a large number of hyperparameters, e.g., weight coefficients of various loss components, with no obvious and straightforward effect on the performance. To measure these effects with human evaluation, we have extensively experimented with many possible tweaks and architecture designs, proposed in the literature on generative models and super-resolution models. These tricks include various data preparation aspects (e.g., adjusting noise/constrast), architecture specifics, auxuliary losses, and various loss weighting. Our models were trained on a specifically constructed dataset consisting of 75 proprietary Ultra HD movie trailers.

Image resolution. Most of the images from our dataset have resolution 1280×720, with a small number of 1920×1080 images. Most of SR models were trained to perform x2 upscaling with an exception of several models performing x1.25, x1.5 or x3 upscaling.

3.3. Annotation

A comparison of different SR models was performed using a human evaluation protocol, described below. For each evaluation session, termed *experiment*, we prepared a batch of roughly 100–200 video fragments and selected several (3–6) SR models from our pool. Each video fragment was upscaled (frame-wise) by each model. Additionally, we have always included the bicubic upscaling as the baseline model in the experiment. Human annotators were presented with a pair of video fragments, and were asked a question: "*which one do you like more?*". The total *score* of the pair (i, j) is then defined as the ratio of clicks on the image j and the total number of annotators of this pair. On av-

Figure 1: An example of pairs from our dataset. According to human evaluators, images from the bottom row are significantly (score 0.9 or more) better than their counterparts from the top row.

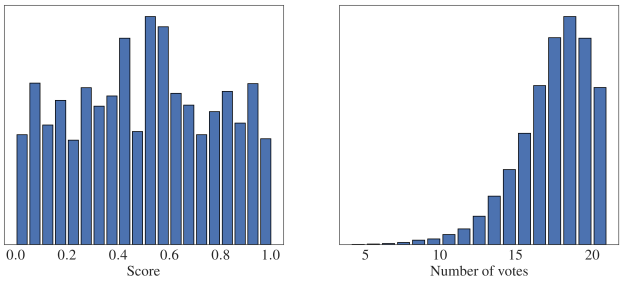

Figure 2: [left] Distribution of human evaluation scores in the collected dataset. [right] Distribution of the number of evaluations of each video pair.

erage, there were about 20 annotators for each pair (see Figure 2 for the exact distribution). During this process, each pair was randomly shuffled to determine which video to show on the left, and which one on the right. The annotators, participating in all experiments, do not have a computer vision background and represent typical users in the Web.

Postprocessing In order to collect a more tractable dataset for side-by-side evaluation, we have assumed that human preferences on videos are sufficiently well approximated on the frame level. I.e, if for two videos (i, j) we have obtained a score s, then for any aligned frames extracted from these videos the resulting score would also be s. We hypothesize that this is a reasonable assumption since our models operated frame-wise. Based on this observation, we simply extracted every 200th frame from each video from a pair and assigned the same score for each resulting image pair. A small sample of frames from our dataset are presented on Figure 1.

3.4. Statistics

In this subsection, we provide some key statistical information about our dataset.

Score distribution Figure 2 demonstrates the obtained distribution of scores in our dataset. We observe that the resulting distribution is quite close to uniform (even though it has a peak at roughly 0.5), which suggests that our dataset is quite diverse and contains useful information about human perceptual quality evaluation. In particular, there is a large number of "difficult" pairs, where both images are preferable for a large number of annotators.

Number of annotations Figure 2 also summarizes the number of annotators for each video pair. Note that 98% of the pairs were evaluated by at least 10 people and 76% had at least 16 votes.

3.5. Train and test sets

In order to verify the generalization of super-resolution evaluation models, we have constructed a split based both on separating *models* and *videos* with a goal to prevent overfitting of NeuralSBS models to particular SR algorithms or image types. To achieve this, we constructed a matrix $\mathcal{M}$ of size $N \times M$, with N=2071 being the number of unique *original* video fragments and M=169 being the total number of models. We then set $\mathcal{M}_{ij} = 1$ if video i was processed by the model j. Based on $\mathcal{M}$ we partitioned all the labeled pairs into two non-overlapping subsets (so no model/video from the test subset appeared in the training subset).

3.6. Neural SBS model

In this section, we describe our approach to train a CNN based model to predict outcomes of side-by-side image comparisons. For a given pair of images, such a model should output a single number — expected result of side-by-side comparisons by human annotators.

Formally, given a dataset of aligned pairs of images and assigned score labels $X = \{(x_1^i, x_2^i), s_i\}_{i=1}^N$, we model s_i as a parameter of the Bernoulli distribution conditioned on the input (x_1^i, x_2^i). Thus, our goal is to train a neural network $F((x_1, x_2); \theta)$ such that

$$F((x_1^i, x_2^i); \theta) \sim \text{Bernoulli}(s_i). \tag{1}$$

Additionally, by construction we have the following constraint on the network F:

$$F((x_1^i, x_2^i); \theta) = 1 - F((x_2^i, x_1^i); \theta) \tag{2}$$

In order to build a neural network for the task at hand we have used a *Siamese network* [3] (as commonly done in the cases of tasks with dual input) and as the backbone we chose Inception v3 [26], with the last fully connected layer of size 2048×1000 replaced by a fully connected layer of size 2048×2048. Each image from the pair is passed through the Inception network, resulting in two feature vectors, which in turn are normalized to be of length 1. In order to produce the output score satisfying Equation (2) we have applied the *antisymmetric bilinear pooling*. Namely, for two vectors f_1 and f_2 of size N this operation is defined as:

$$s(f_1, f_2) = \sigma(\langle f_1, \frac{1}{2}(\Omega - \Omega^\top) f_2 \rangle), \tag{3}$$

where Ω is a trainable matrix of size $N \times N$ and $\sigma(\cdot)$ is the sigmoidal activation function. Note that due to antisymmetricity of the matrix $\frac{1}{2}(\Omega - \Omega^\top)$ and property $\sigma(-x) = 1 - \sigma(x)$ we obtain that

$$s(f_2, f_1) = 1 - s(f_1, f_2),$$

as desired. Our model is illustrated on Figure 3. We use the standard cross entropy loss for training.

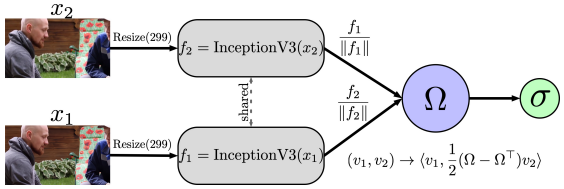

Figure 3: Our NeuralSBS model. Two images are passed through the InceptionV3 network, normalized and fed into the antisymmetric bilinear unit.

The constructed network is then trained in an end-to-end manner via backpropagation. The Inception V3 network was initialized with the network pretrained on the ImageNet. For training and testing, we rescale both input images to 299×299 (similarly with [27, 11]). We have also tried training on random aligned crops, however, the obtained results were not compelling, probably due to significant changes in scene composition.

4. Experiments

The goal of this section is to confirm that the NeuralSBS model outperforms the existing NR-IQA baselines in terms of approximating human evaluation of super-resolved images. First, we analyze the performance of our model on the **SBS180K** test set, which was discussed in Section 3. This experiment aims to demonstrate that NeuralSBS successfully generalizes to unseen images and super-resolution models. Second, we prove that NeuralSBS can be used as a no-reference evaluation measure on the established benchmark datasets for super-resolution. We start with the Ma[18] dataset containing human evaluations of 9 different SR models applied to 180 natural images. Compared to existing baselines, we demonstrate that the NeuralSBS model most accurately predicts outcomes of comparisons of various models compared with the ranking given by human rates. Finally, we test the baselines and the NeuralSBS model on several popular SR benchmarks.

4.1. Baselines

In order to verify the benefits of the model trained to evaluate the relative visual quality of image pairs, we compare NeuralSBS with other approaches for estimating the *no–reference* perceptual image quality. Below we list the baselines, which source-code and pretrained models are available online.

Neural Image Assessment: The first approach proposed in [27] evaluates the perceptual quality of an image by predicting a distribution of human opinion scores using a convolutional neural network trained on the AVA dataset. More concretely, for each possible opinion score (represented by an integer in $[1, 10]$) the model predicts a probability of this score to be assigned to an image by a human evaluator. For assigning a score to a single image, we simply computed the expected value of scores: $\text{NIMA(x)} = \sum_{i=1}^{10} i s(x)_i$.

Photo Aesthetics Ranking Network with Attributes and Content Adaptation: The second approach [11] (termed PARNAC later in the text) similarly uses a CNN to predict the aesthetic score of a single image (as a number in the interval $[0, 1]$).

Naturalness Image Quality Evaluator (NIQE): NIQE [22] estimates the *naturalness* of an image using the Natural Scene Statistics(NSS)-based features. By construction, lower values of NIQE correspond to more 'natural' images. Using this score we can also directly compare two images.

NeuralSBS trained on AVA: To highlight the importance of training side-by-side models on datasets of aligned image pairs, we train a model (termed NeuralSBS$^-$ later in the text) to predict the relative image quality of pairs containing images of different visual content. We employ the Aesthetic Visual Analysis (AVA) dataset containing 255K images. For each image, a number of opinion scores (on a range from 1 to 10) given by human evaluators are provided. For training, we randomly sample a pair of images from the dataset and convert the vectors of these opinion scores to a single relative score using 4. In other words, we simply find the probability of a human evaluator assigning a higher label to the second image. The model and training settings are identical to those of the standard NeuralSBS model.

$$\text{NeuralSBS}^-(x_1, x_2) = \sum_{j>i} s(x_2)_j s(x_1)_i + \frac{1}{2} \sum_{j=i} s(x_2)_j s(x_1)_i. \quad (4)$$

For NIMA we have used the open-source implementation[2] in PyTorch, and for PARNAC we took the original implementation[3] in Caffe. For NIQE we used the open-source implementation found in `scikit-video`[4].

4.2. SBS180K results prediction

The design of our model and our dataset is motivated by the side-by-side human evaluation, used to evaluate the relative performance of two super-resolution approaches in typical development pipelines. Hence, in the first experiment we verify that the NeuralSBS model is an effective "approximator" of human evaluation scores on the test subset of **SBS180K**.

4.2.1 Results

Our results are summarized in Table 2 and Figure 4. First, both NeuralSBS and the baselines are able to identify the better model, as demonstrated in Table 2. However, the baselines do not predict that certain images are drastically better to a human eye than their counterparts. This observation is quite intuitive since both baseline models were trained to estimate the quality of a *single image*, and while two images may look quite good to a person, during the task of paired comparison it may be obvious that one is clearly more appealing. In contrast, our NeuralSBS model decently predicts the shape of the distribution as shown on Figure 4.

[2] https://github.com/kentsyx/Neural-IMage-Assessment
[3] https://github.com/aimerykong/deepImageAestheticsAnalysis
[4] http://www.scikit-video.org

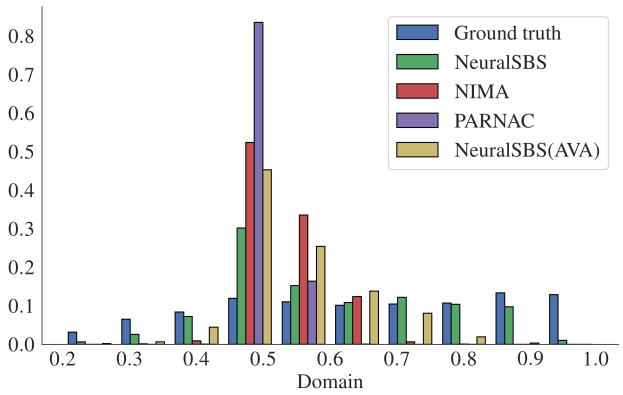

Figure 4: Distribution of the ground truth score and scores produced by various algorithms on the **SBS180K** test set. We observe that NeuralSBS most closely follows the scores distribution, while the baselines do not capture significant relative perceptual differences.

4.3. Scoring of SR models

In this subsection, we argue that NeuralSBS can serve as a research tool for adequate no-reference evaluation of SR models on the common benchmarks. **Ma dataset.** This experiment is organized as follows. For each image in the dataset, we select all the possible pairs from 9 available SR methods (resulting in 36 comparisons) and rank them with NeuralSBS and each of the baselines. We report the accuracy of the obtained predictions compared to the ground truth human rankings present in the dataset. Results are summarized in Table 3. We observe that the NeuralSBS model significantly outperforms its competitors on this task. We hypothesize that one possible reason for this is the presence of a large number of blurry/low-quality images present in **SBS180K**, which were not available for other methods.

SR benchmarking. The final experiment is organized as follows. We select four popular SR models — MSRResNet[12] and MSRGAN[12] using the modified ResNet architecture from [28] as the backbone, ESRGAN [28] in two versions: ESRGAN(PSNR) optimized to achieve the highest PSNR and ESRGAN(GAN) fine-tuned in the GAN regime, and one of the recently proposed models SRFBN [14]. We used the pretrained models and code from the authors of [28] available at github[5]. For SRFBN we used the code and weights available at github[6].

Additionally, we also use two standard upscaling

[5] https://github.com/xinntao/BasicSR
[6] https://github.com/Paper99/SRFBN_CVPR19

	NeuralSBS	NeuralSBS$^-$	NIMA	PARNAC	NIQE
Accuracy, %	**81.2**	74.1	80.3	76.1	42.1

Table 2: The accuracy of no-reference evaluation methods described in Section 4.1 on the **SBS180K** test set.

	NeuralSBS	NeuralSBS$^-$	NIMA	PARNAC	NIQE
Accuracy, %	**62.4**	47.0	48.4	49.0	57.4

Table 3: The accuracy of no-reference evaluation methods described in Section 4.1 on the Ma [18] dataset.

approaches — nearest neighbor and bicubic interpolation. For three standard SR datasets — Urban100[6], Set14 [30], BSD100[21] we apply previously described SR algorithms, and evaluate the obtained samples using NeuralSBS and other baselines. For Set14 and BSD100 we have also included Mean Opinion Scores (MOS) available from [12]. To evaluate the accuracy of the obtained predictions, for each model and dataset we also compute the average LPIPS[32] scores utilizing the ground truth images, and compare the no–reference predictions against these values.

As the proposed NeuralSBS (and NeuralSBS$^-$) model is trained to perform the pairwise comparison, we choose MSRResNet to be a "reference" model for both NeuralSBS and the NeuralSBS$^-$ baselines. Our choice of the reference model is the following — traditional methods like nearest neighbor interpolation might be too weak and too easy to beat by other models, so we might not get enough signal on the power of the models. On the other hand, MSSResNet optimized for the pixelwise MSE loss serves as a good starting point for many SR models, which typically involve an MSE pretraining stage. With MSRResNet as a reference model, we expect elementary interpolation techniques to be much worse compared to all the DL based approaches, and models optimized for MSE (or, equivalently, PSNR) to be inferior to models trained with additional adversarial objectives. For comparison reasons, we also add the MOS values from [12]. For other baselines, we simply compute its predicted score for each image in the dataset and then average across all the images.

The results are summarized in Table 4 and in Table 5, where for each pair of models (resulting in 21 comparisons) for each dataset we find the number of models correctly ranked by each of the no–reference methods (where we assume correct ranking to be given by LPIPS values computed using the ground truth images). We highlight several key observations below:

1. The ranking of models, provided by NeuralSBS, is the same as provided by human evaluation (MOS). Meanwhile, other measures often disagree with LPIPS values.

2. The NIMA and PARNAC baselines often rank simple interpolation techniques higher than the recent MSRResNet model, see the BSD100 dataset. In contrast, for all datasets, NeuralSBS scores of deep models are much higher compared to shallow counterparts.

3. The range of NeuralSBS values is much wider compared to the competitors, i.e. it is a more sensitive measure, which is useful for the development process.

Overall, among all the measures, NeuralSBS is the most consistent with the human evaluation results, reported in the previous works. We argue that it justifies its usage as an evaluation measure of SR models in scenarios when ground truth data is absent.

As a demonstrative example, on Figure 5 we provide a gradient saliency map [25], produced with the NeuralSBS model, for two images of significantly different visual quality. Interestingly, the model is focused more on minor details of the main object on the image (a tiger), than on the background. We conjecture that this behavior is quite similar to how humans assess the relative perceptual quality of two images.

5. Discussion

In this paper, we have introduced Neural Side-By-Side (NeuralSBS) — a new no-reference measure for image super-resolution, which allows us to compare different SR models or to tune their hyperparameters. By extensive experiments, we show that NeuralSBS outperforms existing no-reference measures in terms of approximating human preferences for super-resolved images. The NeuralSBS design is motivated by a realistic practical scenario, where developers currently have to use expensive human evaluation to understand if certain tweaks or model changes result in higher percep-

Dataset	Model	Metric						
		NIMA	PARNAC	**NeuralSBS**	NeuralSBS$^-$	NIQE	MOS	LPIPS
Set14	nearest	5.018	0.437	0.117	0.408	21.33	1.28	0.405
	bicubic	4.575	0.372	0.115	0.357	22.06	1.97	0.439
	MSRResNet	5.102	0.425	0.500	0.500	22.36	3.37	0.284
	SRFBN	5.106	0.426	0.511	0.501	22.24	-	0.280
	MSRGAN	5.326	0.467	0.587	0.604	20.30	2.29	0.145
	ESRGAN (PSNR)	5.149	0.429	0.543	0.526	22.28	3.58	0.271
	ESRGAN (GAN)	5.355	0.464	0.618	0.592	19.50	-	0.133
Urban100	nearest	5.416	0.473	0.167	0.372	20.03	-	0.411
	bicubic	5.335	0.460	0.142	0.318	19.96	-	0.473
	MSRResNet	5.722	0.513	0.500	0.500	20.24	-	0.227
	SRFBN	5.729	0.514	0.514	0.502	20.03	-	0.214
	MSRGAN	5.752	0.515	0.530	0.516	18.81	-	0.143
	ESRGAN (PSNR)	5.734	0.514	0.520	0.503	19.87	-	0.196
	ESRGAN (GAN)	5.757	0.516	0.552	0.521	18.02	-	0.123
BSD100	nearest	5.292	0.426	0.107	0.449	21.53	1.11	0.475
	bicubic	4.521	0.348	0.118	0.366	21.73	1.47	0.526
	MSRResNet	4.873	0.395	0.500	0.500	22.66	2.29	0.371
	SRFBN	4.898	0.397	0.517	0.503	22.58	-	0.367
	MSRGAN	5.420	0.452	0.640	0.668	19.57	3.56	0.178
	ESRGAN (PSNR)	4.956	0.403	0.568	0.531	22.62	-	0.357
	ESRGAN (GAN)	5.449	0.455	0.642	0.662	19.10	-	0.161

Table 4: Evaluation results of various SR models on popular SR datasets. For NeuralSBS and NeuralSBS$^-$ evaluation was performed against the images produced by MSRResNet [28, 12]. Mean Opinion Scores (MOS) were taken from [12]. For LPIPS and NIQE a lower score indicates a better quality, for other algorithms — a higher score is better.

	Set14	Urban100	BSD100	Total (of 63)
NIQE	14	17	14	45
PARNAC	17	21	18	56
NIMA	21	21	18	60
NeuralSBS$^-$	20	21	20	61
NeuralSBS	21	21	20	**62**

Table 5: The number of pairs of models correctly compared by each method.

Figure 5: Visualization of the saliency map of the NeuralSBS model. Images were produced by MSRResNet and MSRGAN respectively. The score assigned by NeuralSBS to this pair is 0.89. Notably, the model is focused on minor details, such as tiger whiskers and stripes.

tual quality. As demonstrated in the experiments, the usage of NeuralSBS allows to automate the evaluation, thereby substantially expediting the development process. Given that our dataset is large and diverse, we expect that it can serve as a useful benchmark for new super-resolution methods, which comes with a "build-in" human evaluation, provided by our model. Furthermore, **SBS180K** is a natural fit to train image enhancement models. Finally, since NeuralSBS per se is a differentiable computational unit, it can be used as an additional optimization objective in the future models for super-resolution or general image generation.

References

[1] Bishop, C.M., Blake, A., Marthi, B.: Super-resolution enhancement of video. In: AISTATS (2003) 4321

[2] Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: The 2018 pirm challenge on perceptual image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018) 4322

[3] Chopra, S., Hadsell, R., LeCun, Y., et al.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR (1). pp. 539–546 (2005) 4325

[4] Greenspan, H.: Super-resolution in medical imaging. The Computer Journal **52**(1), 43–63 (2008) 4321

[5] Hosu, V., Goldlucke, B., Saupe, D.: Effective aesthetics prediction with multi-level spatially pooled features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9375–9383 (2019) 4322

[6] Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5197–5206 (2015) 4327

[7] Jin, X., Wu, L., Li, X., Chen, S., Peng, S., Chi, J., Ge, S., Song, C., Zhao, G.: Predicting aesthetic score distribution through cumulative jensen-shannon divergence. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018) 4322

[8] Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1733–1740 (2014) 4322

[9] Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1637–1645 (2016) 4323

[10] Ko, K., Lee, J.T., Kim, C.S.: Pac-net: pairwise aesthetic comparison network for image aesthetic assessment. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 2491–2495. IEEE (2018) 4322

[11] Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: European Conference on Computer Vision. pp. 662–679. Springer (2016) 4322, 4325

[12] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017) 4322, 4323, 4326, 4327, 4328

[13] Lee, J.T., Kim, C.S.: Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1191–1200 (2019) 4322

[14] Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3867–3876 (2019) 4326

[15] Lin, F., Fookes, C., Chandran, V., Sridharan, S.: Super-resolved faces for improved face recognition from surveillance video. In: International Conference on Biometrics. pp. 1–10. Springer (2007) 4321

[16] Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rapid: Rating pictorial aesthetics using deep learning. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 457–466. ACM (2014) 4322

[17] Lu, X., Lin, Z., Shen, X., Mech, R., Wang, J.Z.: Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 990–998 (2015) 4322

[18] Ma, C., Yang, C.Y., Yang, X., Yang, M.H.: Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding **158**, 1–16 (2017) 4322, 4323, 4325, 4327

[19] Marchesotti, L., Murray, N., Perronnin, F.: Discovering beautiful attributes for aesthetic image analysis. International journal of computer vision **113**(3), 246–266 (2015) 4322

[20] Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: 2011 International Conference on Computer Vision. pp. 1784–1791. IEEE (2011) 4322

[21] Martin, D., Fowlkes, C., Tal, D., Malik, J., et al.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Iccv Vancouver: (2001) 4327

[22] Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2012) 4325

[23] Murray, N., Gordo, A.: A deep architecture for unified aesthetic prediction. arXiv preprint arXiv:1708.04890 (2017) 4322

[24] Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2408–2415. IEEE (2012) 4323

[25] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013) 4327

[26] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015) 4325

[27] Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE Transactions on Image Processing **27**(8), 3998–4011 (2018) 4322, 4325

[28] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018) 4323, 4326, 4328

[29] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 4321, 4322

[30] Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International conference on curves and surfaces. pp. 711–730. Springer (2010) 4327

[31] Zhang, L., Zhang, H., Shen, H., Li, P.: A super-resolution reconstruction algorithm for surveillance images. Signal Processing **90**(3), 848–859 (2010) 4321

[32] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018) 4321, 4322, 4327