Bridging the Simulated-to-Real Gap: Benchmarking Super-Resolution on Real Data

Thomas Köhler, Michel Bätz, Farzad Naderi, André Kaup, *Fellow, IEEE*, Andreas Maier, *Member, IEEE*, and Christian Riess, *Member, IEEE*

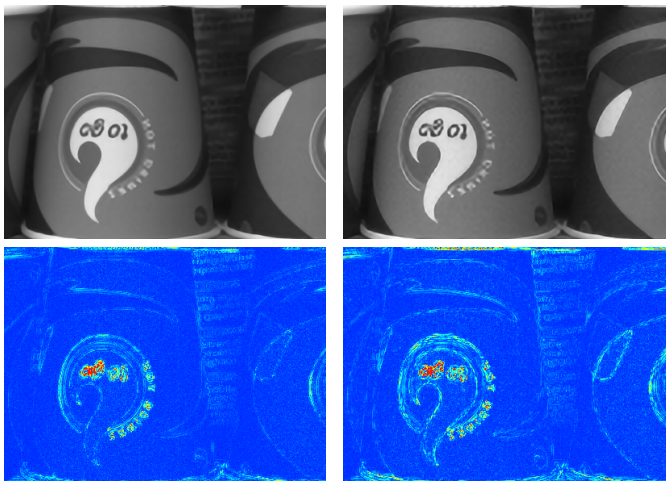
Abstract—Capturing ground truth data to benchmark super-resolution (SR) is challenging. Therefore, current quantitative studies are mainly evaluated on simulated data artificially sampled from ground truth images. We argue that such evaluations overestimate the actual performance of SR methods compared to their behavior on real images. To bridge this *simulated-to-real gap*, we introduce the *Super-Resolution Erlangen* (SupER) database, the first comprehensive laboratory SR database of all-real acquisitions with pixel-wise ground truth. It consists of more than 80k images of 14 scenes combining different facets: CMOS sensor noise, real sampling at four resolution levels, nine scene motion types, two photometric conditions, and lossy video coding at five levels. As such, the database exceeds existing benchmarks by an order of magnitude in quality and quantity. This paper also benchmarks 19 popular single-image and multi-frame algorithms on our data. The benchmark comprises a quantitative study by exploiting ground truth data and qualitative evaluations in a large-scale observer study. We also rigorously investigate agreements between both evaluations from a statistical perspective. One interesting result is that top-performing methods on simulated data may be surpassed by others on real data. Our insights can spur further algorithm development, and the publicy available dataset can foster future evaluations.

Index Terms—Super-resolution, ground truth, simulated-to-real gap, benchmark, quantitative evaluation, observer study

1 Introduction

Super-resolution (SR) [1] enhances the spatial resolution of digital images without modifying camera hardware. This facilitates low-cost high-resolution (HR) imagery to improve vision tasks, e. g. in surveillance [2], remote sensing [3], 3D imaging [4], or healthcare [5], [6]. Single-image SR (SISR) infers HR details from a low-resolution (LR) image using self-similarities [7], [8] or example data via classical regression [9], [10], [11], [12] or deep learning [13], [14], [15], [16], [17]. Multi-frame SR (MFSR) fuses LR frames with relative motion via interpolation [18], [19], iterative reconstruction [20], [21], [22], [23], [24], or deep learning [25], [26]. Such methods complement related resolution enhancement strategies like frame rate up-conversion (temporal resolution enhancement) [27] or high dynamic range imaging (radiometric resolution enhancement) [28].

SR performance guarantees have also been subject to much research. Examples are seminal works on inherent performance guarantees, like algebraic studies of maximum resolution gains [29], [30] or statistical validations [31]. These works use approximations like linearity and shift invariance of imaging systems, hence they can only roughly predict upper or lower performance bounds. In contrast to that, performance on real data is still widely unexplored, due to the lack of quantitative benchmarks. This is surprising, also in comparison to other computer vision areas like motion analysis [32] or deblurring [33]. Evaluations on real LR data are performed visually or with no-reference measures due to the absence of ground truth data [34], [35]. However, this is inappropriate in applications that require evidence that SR has high fidelity w. r. t. a ground truth (e. g. medical imaging) as this property could be anti-correlated to perceptual quality [36]. Experiments on simulated data with a ground truth to quantity SR quality are by far the



(a) VDSR [14] on simulated data (PSNR: 34.46, IFC: 3.93)

(b) VDSR [14] on real data (PSNR: 32.46, IFC: 2.63)

Fig. 1: Can experiments on simulated data predict the behavior of SR on real data? Our study reveals that this is not the case. (a) VDSR [14] and its color-coded error w. r. t. the ground truth on simulated low-resolution data. (b) VDSR on our real acquisitions of the same scene. The simulation considerably overestimates the performance both visually and quantitatively. We benchmark SR algorithms on captured data to overcome shortcomings of simplistic simulations.

T. Köhler, F. Naderi, and A. Maier are with the Pattern Recognition Lab, Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Erlangen, Germany. (e-mail: {thomas.koehler, farzad.naderi, andreas.maier}@fau.de)
M. Bätz and A. Kaup are with the Chair of Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Erlangen, Germany. (e-mail: {michel.baetz, andre.kaup}@fau.de)
C. Riess is with the IT Security Infrastructures Lab, FAU Erlangen-Nürnberg, Erlangen, Germany. (e-mail: christian.riess@fau.de)

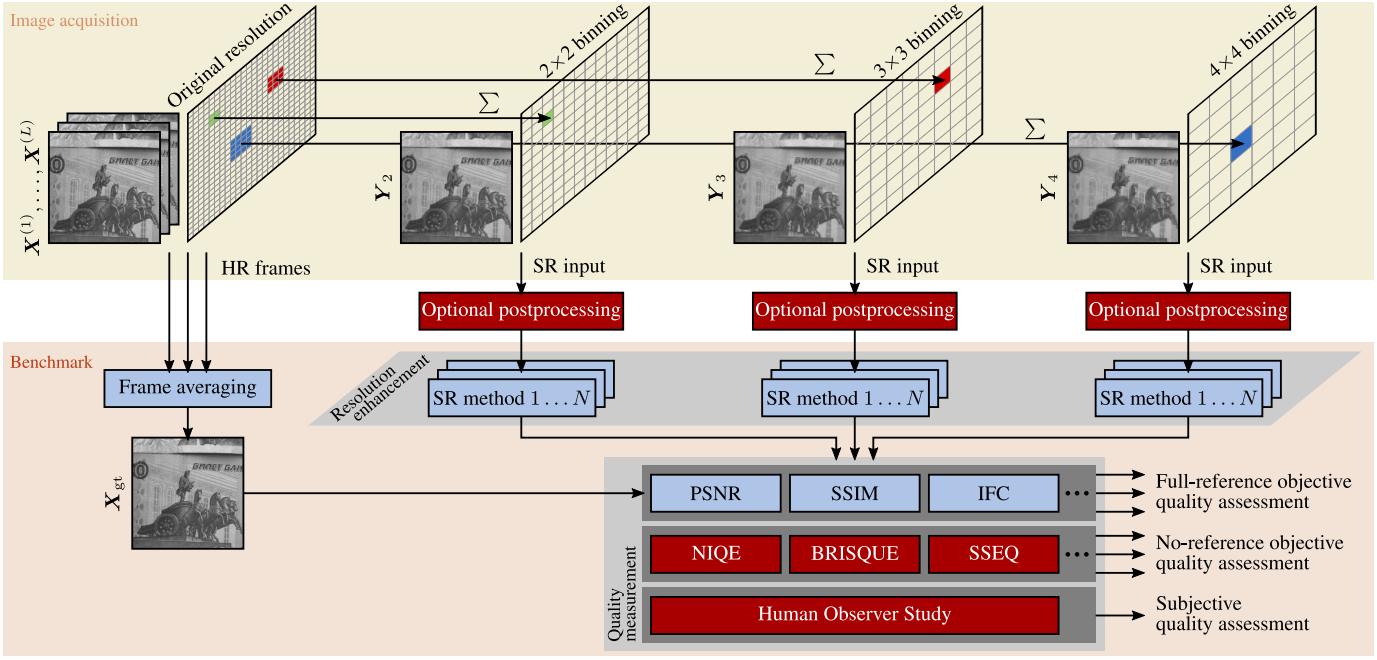


Fig. 2: Overview of our data collection and benchmark. In our data collection, we capture multiple frames at the actual pixel resolution to obtain ground truth high-resolution images via frame averaging. We employ hardware binning on the sensor to gain captured low-resolution images and include postprocessing (e. g. image/video compression) to collect multiple versions of this low-resolution data (see Sect. 3). In the benchmark, we use 1) full-reference and no-reference measures to quantitatively assess super-resolved data with and without exploiting the ground truth (see Sects. 4 and 5), and 2) observer studies to evaluate image quality according to human perception (see Sect. 6).

most common evaluation strategy. Unfortunately, simulations only partially address practical constraints like physically true image noise, low-light exposures, or photometric variations. This is due to simplifying assumptions on LR image formation, such as bicubic downsampling of HR images [37].

To demonstrate limitations of simulated data, Fig. 1 compares the popular VDSR method [14] on simulated images from bicubic downsampling of the ground truth to real images on the same scene using the hardware-based acquisition proposed here. It shows that simulation is a weak indicator for the performance on real data with degradations like non-Gaussian noise. SR on real data suffers from noise breakthroughs and artifacts at fine structures, e. g. text, resulting in considerably lower image quality expressed by PSNR and IFC. In this paper, we show that there is indeed a *simulated-to-real gap* in today's SR benchmarks. More specifically, we reveal that simulation studies often overestimate the actual performance of SR on real images as quantitatively shown in Fig. 5. Similar conclusions have also been drawn for image denoising [38] and motivate the use of real acquisitions to benchmark SR.

This paper introduces the *Super-Resolution Erlangen (SupER)* database – a large database of captured LR images at *multiple levels of spatial resolution* and ground truth HR data – to close the simulated-to-real gap. Data is obtained via *hardware binning*, and covers difficult conditions like local object motion and photometric variations. It also comes with *postprocessed* images with different levels of H.265/HEVC coding to investigate video compression as shown in Fig. 2 (top). To our knowledge, this is the first comprehensive dataset of *all-real* image sequences to allow quantitative evaluations. It comprises more than 80k images of 14 scenes at 4 resolution and 5 compression levels. This size also opens the opportunity to use it for fine-tuning learning-based methods to real data.

We benchmark 19 SR algorithms on the SupER database from two perspectives as shown in Fig. 2 (bottom). First, this includes a *quantitative evaluation* using full-reference and no-reference quality assessment. This is by far the most comprehensive SR comparison, which particularly cross-compares SISR and MFSR. Second, we present a *large-scale observer study* to benchmark SR algorithms according to human visual perception. Our experiments quantitatively reveal on real data some unexpected results as manifestations of the simulated-to-real gap: for instance, several classical methods like shallow regression or sparsity priors for reconstruction-based SR compare very well to much more elaborated deep learning methods. They also show so far unexplored mismatches between quantitative evaluations and human perception: for example, correlations between quantitative and perceptual quality are higher for larger SR factors but deteriorates under real conditions like photometric variations. To maximize the use for the community and to foster benchmarks on real images, we publish all data and source code implementing the evaluation protocols[1].

The remainder of this paper is organized as follows. In Sect. 2, we review existing SR datasets and evaluation strategies. In Sect. 3, we introduce the proposed benchmark database. In Sect. 4, we present the underlying evaluation protocol. In Sect. 5 and Sect. 6, we evaluate current SR approaches quantitatively and in the observer study. In Sect. 7, we draw conclusions for future algorithm developments. Section 8 concludes this work.

1. https://superresolution.tf.fau.de/

2 Related Work

In comparison to the great number of algorithmic contributions, there is only few prior work on their comparative experimental evaluations. Most papers follow two evaluation strategies.

2.1 Benchmarking on Simulated Data

The most common evaluation strategy is the use of simulated data. Yang *et. al.* [39] and Timofte *et. al.* [37] have evaluated current SISR methods, where LR images are obtained by artificial downsampling of HR ground truth data. In [24], Liu and Sun evaluated MFSR on video datasets by artificially sampling HR videos. All of these benchmarks have in common that only simplistic sampling kernels that are known a priori (e. g. bicubic [37] or Gaussian kernels [24], [39]) are simulated but SR in case of more general kernels is unexplored.

One considerable step forward is the DIVerse 2K resolution (DIV2K) database [40] with LR data generated under bicubic downsampling and more difficult kernels that are hidden to a user. This provides valuable insights to SR performance and the associated NTIRE 2017 challenge [41] revealed that overall very deep neural networks [14], [15], [17] and methods evolved thereof [42], [43] are currently the top performing ones. DIV2K covers challenging situations in terms of sampling but the simulations lack MFSR facets, i. e. sequences with motion or photometric variations. The LR data is also not the outcome of physical imaging system with effects like non-Gaussian noise or image/video compression.

In general, the use of simulated data enables comparisons to a ground truth by full-reference quality measures but limits the significance to study SR under realistic constraints. The neglection of physically meaningful sampling kernels, realistic noise models, or environmental conditions manifest crucial limitations and is no realistic depiction of real-world applications. Surprisingly, the impact of realistic image formation models for SR evaluations compared to such simplifying conditions is still widely unexplored. In Sect. 3.4, we address this question and show the overall weak correlations between benchmarks on simulated and real data termed simulated-to-real gap.

2.2 Benchmarking on Real Data

Existing real-world image databases [34], [35] lack of ground truth data and are hence designed for perceptual evaluations. This requires no-reference quality measures as for example proposed in [44], [45]. In the perception-distortion plane as proposed by Blau and Michaeli [36], such methods are used in conjunction with full-reference measures to jointly analyze perceptual quality and fidelity to a ground truth. However, finding appropriate no-reference measures to assess SR on general scenes is difficult. In Sect. 6.4, we show that on real data popular no-reference measures show lower correlations to human visual perception than measures with a reference. Another strategy are large-scale observer studies, as previously conducted for deblurring [33] or SISR [39]. This ensures high agreement to human perception but is cumbersome and the results are difficult to reproduce. Our work aims at constructing a database with ground truth HR data to circumvent the use of no-reference measures. We further statistically analyze the agreement between quantitative and observer studies.

Fig. 3: Overview of the scenes covered by our SupER database.

Other studies [46], [47] also validated SR in specific vision tasks. However, such evaluations have limited informative value for general benchmarks on natural scenes. Our work aims at broadly benchmarking SR algorithms on real captured images.

In a work closely related to ours, Qu *et. al.* [48] have collected a database of real face images with corresponding ground truth data. Their setup utilizes two cameras combined with a beam splitter to capture LR and HR images at the same time. However, the required LR/HR alignment in this *multi-camera* setup is potentially affected by error-prone calibrations and image registrations. This makes the use of full-reference quality measures for pixel-wise comparisons unreliable. Furthermore, the data of [48] comprises only single images, which excludes MFSR. We propose a *single-camera* setup that avoids these limitations and also allows us to acquire more than two resolution levels.

3 SupER Database

We collect sets of LR and HR images at multiple resolutions with a single camera by capturing *stop-motion* videos. At each time step of a stop-motion video, the underlying scene, environmental conditions, and the camera pose are kept static. For consecutive time steps, the scene undergoes changes related to camera and/or object movements and/or environmental variations. One time step is represented by the $(n+1)$-tuple $(\boldsymbol{X}_{\mathrm{gt}}, \boldsymbol{Y}_{b_1}, \boldsymbol{Y}_{b_2} \ldots, \boldsymbol{Y}_{b_n})$, where $\boldsymbol{X}_{\mathrm{gt}}$ denotes a ground truth HR image of size $N_u \times N_v$ and $\boldsymbol{Y}_{b_i}$, $i = 1, \ldots, n$ are LR frames of size $N_u/b_i \times N_v/b_i$ at n different hardware binning factors b_i. Our database covers 14 lab scenes including text, emulated surveillance scenes, and various objects, see Fig. 3.

3.1 Image Formation

In order to gain the ground truth $\boldsymbol{X}_{\mathrm{gt}}$, we capture L frames $\boldsymbol{X}^{(l)}$, $l = 1, \ldots, L$ at each time step of a stop-motion video using the actual pixel resolution of the camera. These frames are acquired under constant illumination and without inter-frame motion. To reduce sensor noise, the ground truth is computed by averaging over L ($L = 10$) consecutive frames:

$$\boldsymbol{X}_{\mathrm{gt}} = \frac{1}{L} \sum_{l=1}^{L} \boldsymbol{X}^{(l)} \quad . \tag{1}$$

To obtain the LR data $\boldsymbol{Y}_{b_i}$ associated with $\boldsymbol{X}_{\mathrm{gt}}$, we use camera hardware binning. This aggregates adjacent pixels on the sensor array as depicted in Fig. 2 (top). Let $x(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbb{R}^2$ be an irradiance light field [30]. Hardware binning links $x(\boldsymbol{u})$ to the image $\boldsymbol{Y}_b$ according to:

$$\boldsymbol{Y}_b = \mathcal{Q}\left\{\mathcal{D}_b\left\{x(\boldsymbol{u})\right\} + \boldsymbol{\epsilon}\right\} \quad , \tag{2}$$

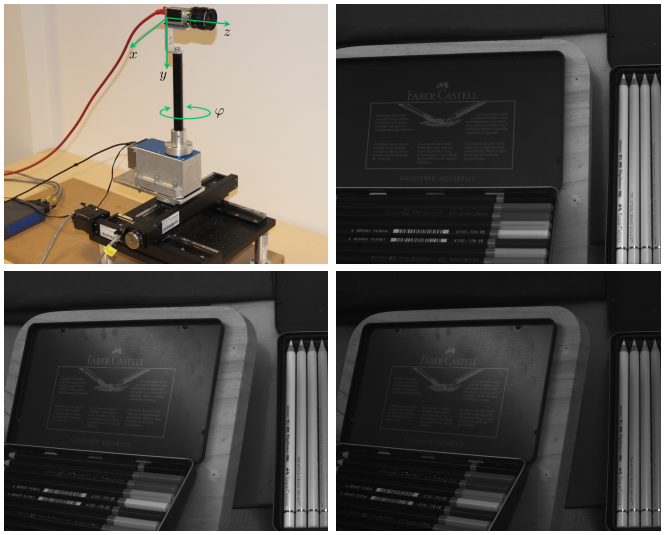


Fig. 4: Proposed hardware setup and an example scene. Our data comprise challenging conditions for SR like local object motion and photometric variations.

where $\mathcal{D}_b\{\cdot\}$ denotes sampling according to the binning factor b, $\mathcal{Q}\{\cdot\}$ denotes quantization to capture image intensities, and $\boldsymbol{\epsilon}$ is additive noise. The sampling $\mathcal{D}_b\{\cdot\}$ is described by:

$$\mathcal{D}_b\left\{x(\boldsymbol{u})\right\} = \left(\boldsymbol{H}_{\text{sensor},b} \star \boldsymbol{H}_{\text{optics}} \star x\right)(\boldsymbol{u}) \quad , \tag{3}$$

where $\boldsymbol{H}_{\text{optics}}$ denotes the optical point spread function (PSF), $\boldsymbol{H}_{\text{sensor},b}$ models the spatial integration over $b \times b$ pixels on the sensor array, and $\star$ is the convolution operator [30]. As we use a single optical system to capture HR and LR data, $\boldsymbol{H}_{\text{sensor},b}$ is determined by the binning factor b while $\boldsymbol{H}_{\text{optics}}$ does not depend on the binning. Our data collection is based on high-quality optical equipment, such that $\boldsymbol{H}_{\text{sensor},b}$ is the main limiting factor for resolution and signal degradations. We use $n = 3$ binning factors $b \in \{2, 3, 4\}$ to acquire data at different resolution levels.

Most SR algorithms deal either with grayscale or a single luminance channel, while the chrominance channels are simply interpolated [13], [25], [37], [49]. Thus, we limited ourselves to monochromatic acquisitions to compromise between hardware requirements and practical applicability. In order to study full color SR, the setup can be generalized to provide multiple channels, e. g. using color filters or a full RGB camera.

3.2 Image Postprocessing

We capture raw LR and ground truth data in the proposed multi-resolution scheme while camera internal processing is avoided. This enables to explicitly investigate SR under different types of *postprocessing* (or *preprocessing* from the SR perspective) like white balancing or gamma correction. Raw LR data forms an ideal base for SR while additional (latent) postprocessing steps might deteriorate the performance of SR algorithms. This provides a new testbed for quantitative benchmarks.

Lossy image and video compression is a type of postprocessing that is of high practical relevance. We compute an additional compressed version $\boldsymbol{Y}_{b,c}$ of the LR image $\boldsymbol{Y}_b$ according to:

$$\boldsymbol{Y}_{b,c} = \mathcal{C}_c\left\{\boldsymbol{Y}_b\right\} \quad , \tag{4}$$

TABLE 1: Overview of motion types captured in our database.

Motion type	Camera trajectory
Translation z	Linear
Translation x,z	Sinusoidal
Panning	Circular
Translation x,y,z and panning	Joint sinusoidal and circular

where $\mathcal{C}_c\{\cdot\}$ denotes compression at compression level c.

We employ H.265/HEVC video coding [50] on our raw data using the *main* profile and *random access* mode (version HM-16.2). Uncoded data and the quantization parameters (QP) 10, 20, 30, and 40 were chosen so as to cover the full range from a raw data without compression artifacts to a strong compression with heavy artifacts. When combined with the multi-resolution scheme, this approach leads to 15 versions of LR data associated with a given ground truth, i. e. three resolutions levels and five compression levels.

3.3 Motion Types and Environmental Conditions

For our data collection, we use a Basler acA2000-50gm CMOS camera [51] equipped with a f/1.8, 16 mm fixed focus lens [52] on a positioning stage as shown in Fig. 4. The camera pose is controlled by a stepper motor and a height-adjustable table. This enables camera panning in one dimension (in-plane rotation) and translations in three dimensions. We consider four basic motion types that were described by different camera trajectories, see Tab. 1. The photometric conditions are controlled by artificial lighting and we consider bright (*daylight*) and low-light illumination (*nightlight*). In conjunction with movements of objects in a scene, this forms five dataset categories with different levels of difficulty.

Global motion: This baseline category consists of static scenes with constant daylight conditions. All inter-frame motion is global camera motion using the trajectories in Tab. 1 with uniformly distributed camera positions.

Local motion: This category consists of dynamic scenes captured under daylight conditions with a static camera but moving objects, see Fig. 4. All inter-frame motion is translational and/or rotational object motion.

Mixed motion: This category combines global and local motion. To this end, each camera trajectory is combined with translational and/or rotational object motion.

Photometric variation: This category comprises sequences of K frames, where the first $K - K_{\text{night}}$ frames are taken from the global, local, and mixed motion data and the remaining K_{night} *outliers* frames are obtained under nightlight conditions, see Fig. 4.

Video compression: This category further extends the aforementioned datasets by five H.265/HEVC compression levels, i. e. all LR images are provided as uncompressed and compressed versions.

Overall, our database comprises 56 global, 56 mixed, and 14 local motion image sequences with $K = 40$ frames each captured from 14 scenes. The photometric variation datasets augment each sequence by $K_{\text{night}} = 5$ nightlight images.

TABLE 2: Comparison of our SupER database to other publicly available benchmark datasets. Unlike existing datasets, we provide captured image sequences at four spatial resolution levels (ground truth HR images plus three levels for LR images) without involving simulation. In addition, our LR data is provided at five compression levels (uncoded plus four quantization levels) using H.265/HEVC video coding. All quantitative properties refer to the original versions of the datasets. We excluded datasets without separate LR data.

Dataset	Real/Simulated	# Res. levels	# Comp. levels	# Sequences	# LR images	# HR images
MDSP [34]	Real	1	1	21	915	✗
Vandewalle [35]	Real	1	1	3	12	✗
Liu and Sun [24]	Simulated	2	1	4	171	171
Yang *et. al.* [39]	Simulated	4	1	Single images	2,061	229
DVI2K [40]	Simulated	4	1	Single images	6,000	1,000
Qu *et. al.* [48]	Real	2	1	Single images	93	93
SupER (ours)	Real	4	5	254	85,050	5,670

3.4 Comparison to Existing Datasets

Our data acquisition scheme goes beyond existing real-world databases [34], [35] by providing 1) real LR acquisitions, 2) compressed LR data using H.265/HEVC video coding, and 3) corresponding HR ground truth data as summarized in Tab. 2. Its size in terms of LR/HR exemplars also exceeds the state-of-the-art by an order of magnitude. This fosters both large-scale benchmarks and training of learning-based methods.

3.4.1 Hardware Binning for Realistic Image Artifacts

In contrast to our image formation in (3), the widely used *software binning* [24], [39], [40] is based based on the model:

$$\boldsymbol{Y}_b = \mathcal{D}_b\{\boldsymbol{X}\} + \boldsymbol{\eta} \ , \tag{5}$$

where $\boldsymbol{X}$ is a discretization of $x(\boldsymbol{u})$ and $\boldsymbol{\eta}$ is additive noise. Typically $\boldsymbol{X}$ is a reference image from an existing database, e. g. LIVE [53], Set5, Set14, B100, or L20 [37], or from HR videos [24]. Note that a simulation cannot model the true physics of image formation since it does not have access to the original irradiance $x(\boldsymbol{u})$ as used in (2). For instance, simulated data is based on simplified models for $\boldsymbol{\eta}$ like ideal quantization noise [39] or Gaussian noise [54]. The sampling $\mathcal{D}_b\{\cdot\}$ is modeled by bicubic downsampling or other artificial operators [24], [40]. The LR images in our database are degraded by noise and subsampling from actual physical processes.

We validate the importance of real acquisitions by analyzing correlations between SR benchmarks on simulated and real data. Following prior work [40], we simulate LR images from our ground truth data using bicubic downsampling as a counterpart to LR images that were captured by hardware binning. Figure 5 shows the performance of various algorithms of our benchmark (see Sect. 4) for different binning factors on simulated versus real data. The performance is depicted w. r. t. different full-reference quality measures (PSNR, SSIM, and IFC) averaged over 14 scenes with global motion. We can observe that the performance of most algorithms is higher on simulated data but drops on real data as depicted in Fig. 1. For instance, two recent deep networks (VDSR [14], DRCN [15]) outperform most classical algorithms on simulated data. Interestingly, these deep networks are outperformed by several other approaches on real data. Overall, there are Spearman rank correlations of $\rho < 0.6$ between the quality measures on simulated and real data. These correlations are even negative for all measures in case of small magnification factors corresponding to small binning factors (2×2) in the image formation. This demonstrates that benchmarks on simulated data are weak indicators for an algorithm ranking on real data especially for small magnifications. We argue that this simulated-to-real gap appears because simulations follows a relatively simple model, while real data can be affected by more realistic artifacts like non-Gaussian noise or oversaturated pixels, among others. Such observations have also been reported in related areas like blind deblurring [33] or denoising [38], which underlines the importance of real data for thorough SR evaluations.

Some prior works also use the same models for simulations and SR. This can be seen as *inverse crime* [55] and limits the significance of experiments. For instance, many deep learning methods are trained and validated for LR/HR exemplars related to each other by bicubic downsampling. This explains their lower performance on real data that follows a more complicated image formation. It also adds another merit of using real LR data: beyond the aspect of evaluation, our datasets can also complement the fine-tuning of learning-based methods.

3.4.2 Single-Camera Setup for Registered Ground Truth

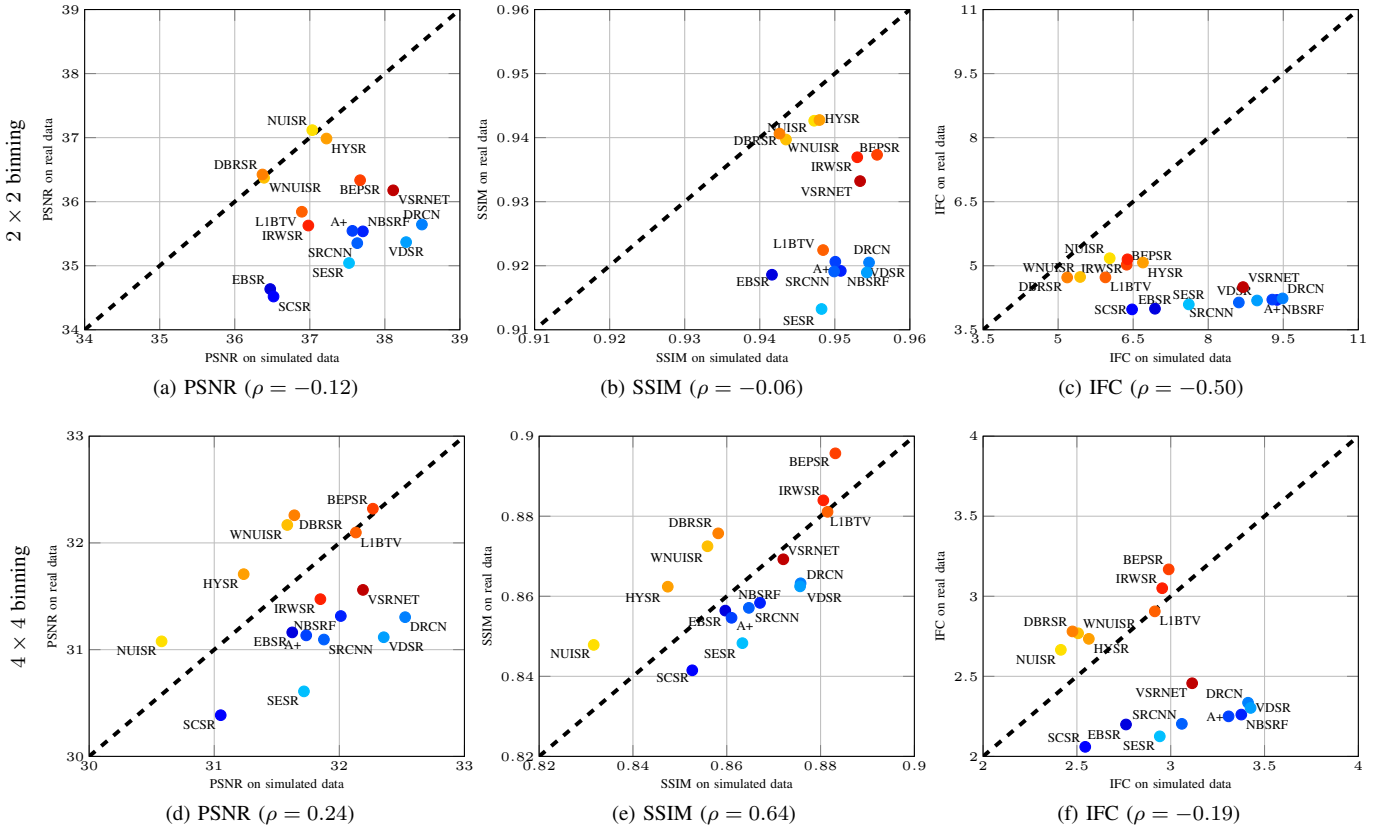
The used single-camera setup overcomes several limitations of related multi-camera setups as proposed in [48]. First and foremost, it guarantees by design a perfect alignment between LR and ground truth data. This allows pixel-wise comparisons among super-resolved images and the ground truth by full-reference quality measures. In contrast to data captured with multi-camera setups, our ground truth is not the outcome of a potentially error-prone registration procedure.

The proposed single-camera setup further extends the scope of our database in comparison to [48]: one important benefit is the existence of multiple resolution levels. Additionally, we collect image sequences instead of single images. This makes the data usable for both, SISR and MFSR.

4 Benchmark Setup

4.1 Evaluation Protocol

We perform SR on $K = 2M + 1$ consecutive LR frames $\boldsymbol{Y}^{(-M)}, \ldots, \boldsymbol{Y}^{(0)}, \ldots, \boldsymbol{Y}^{(M)}$ in a sliding window scheme. $\boldsymbol{Y}^{(0)}$ is referred to as the reference frame. For SISR, $\boldsymbol{Y}^{(0)}$ serves as input to determine the corresponding HR image $\boldsymbol{X}_{\mathrm{sr}}$. In case of MFSR, $\boldsymbol{Y}^{(-M)}, \ldots, \boldsymbol{Y}^{(M)}$ is exploited to obtain $\boldsymbol{X}_{\mathrm{sr}}$ using variational optical flow [56] to estimate subpixel motion towards $\boldsymbol{Y}^{(0)}$. For MFSR with customized motion



(a) PSNR ($\rho = -0.12$) (b) SSIM ($\rho = -0.06$) (c) IFC ($\rho = -0.50$)

(d) PSNR ($\rho = 0.24$) (e) SSIM ($\rho = 0.64$) (f) IFC ($\rho = -0.19$)

Fig. 5: Correlation between the benchmarks of SR algorithms on simulated data and our captured LR data in terms of different full-reference quality measures for 2×2 (top row) and 4×4 binning (bottom row). The individual algorithms are categorized either as SISR (shown with blue color map) or MFSR (shown with red color map). Algorithms located below the line of equal image quality perform worse on real data compared to simulated data. For each quality measure, the corresponding Spearman rank correlation ρ is shown.

compensation [24], [25], [57], we employ the optical flow estimation used in the original versions of the algorithms.

We study the magnification factors 2, 3, and 4 to super-resolve LR images at the respective binning factors to the resolution of the ground truth. The number of input frames for MFSR is chosen according to the desired magnification and therefore we use 5, 11, and 17 frames for the factors 2, 3, and 4, respectively. Our benchmark comprises ten sliding windows per dataset. We crop central regions of size 1200×960 pixels with sufficient details in the ground truth of each dataset to increase the significance of evaluating SR quality.

4.2 Quantitative Quality Measures

We use nine full-reference and no-reference quality measures $Q(\boldsymbol{X}_{\mathrm{sr}})$. The full-reference methods comprise the PSNR, SSIM [58], MS-SSIM [59], and IFC [60]. These measures focus on different features, namely intensities (PSNR), structural information (SSIM, MS-SSIM), and natural scene statistics (IFC) to assess the fidelity of super-resolved data $\boldsymbol{X}_{\mathrm{sr}}$ w. r. t. the ground truth $\boldsymbol{X}_{\mathrm{gt}}$ assuming that both are aligned. A three-pixel boundary is cropped to provide space for compensating alignment differences by some algorithms. We additionally use five popular no-reference measures, namely S3 [61], BRISQUE [62], SSEQ [63], NIQE [64], and SRM [65], which is a recent measure trained from SISR examples. Higher S3 and SRM measure express higher perceptual quality of the assessed SR image. For BRISQUE, SSEQ and NIQE, we used the negated scores such that higher measures express higher quality.

Note that scene content can considerably influence the absolute values of these measures [39]. To reduce dependencies from scene content and to analyze the improvement of SR over the input data, we also evaluate normalized quality measures:

$$\tilde{Q}(\boldsymbol{X}_{\mathrm{sr}}) = (Q(\boldsymbol{X}_{\mathrm{sr}}) - Q(\tilde{\boldsymbol{Y}}))/Q(\tilde{\boldsymbol{Y}}) \ , \qquad (6)$$

where $\tilde{\boldsymbol{Y}}$ denotes the nearest-neighbor interpolation of the reference frame $\boldsymbol{Y}^{(0)}$ on the target HR grid.

4.3 Perceptual Pairwise Comparisons

A large-scale observer study is set up to assess image quality according to human visual perception. The study adopts *forced-choice pairwise comparisons* [66], where two images obtained by different SR algorithms from the same data are presented side-by-side. An observer is requested to choose from each pair the image with higher quality. The pairs are randomly sampled without replacement out of $n_{\mathrm{data}}\binom{n_{\mathrm{sr}}}{2}$ pairs with n_{sr} algorithms and n_{data} datasets. Within a pair, both images are exposed to identical conditions, i. e. identical binning, motion, lighting, and compression. Thus the only differences in the presented images are due to the applied SR algorithms. Observers are guided interactively through sessions of n_{pairs} pairs by a dynamic

webpage. Among the n_{pairs} pairs, n_{sanity} pairs are randomly mixed in as sanity checks to identify careless observers. Sanity checks comprise ground truth images and results with severe artifacts (aliasing, noise, or motion artifacts). We discard a session if the sanity check is failed more than once.

We denote by $\boldsymbol{M} \in \mathbb{Z}^{n_{\text{sr}} \times n_{\text{sr}}}$ the winning matrix for a session, where M_{ij}, $i \neq j$ denotes the number of times that the i-th method is preferred over the j-th method. To globally rank algorithms from pairwise votes, we use $\boldsymbol{M}$ to fit a Bradley-Terry (B-T) model [33]. This describes the probability $P(i \succ j)$ that the i-th method is ranked over the j-th method:

$$P(i \succ j) = \frac{e^{\delta_i}}{e^{\delta_i} + e^{\delta_j}} \ , \tag{7}$$

where δ_i and δ_j are quality scores associated with the i-th and the j-th method, respectively. Then, based on (7), the negative log-likelihood for the B-T scores $\boldsymbol{\delta} \in \mathbb{R}^{n_{\text{sr}}}$ is given by:

$$\mathcal{L}(\boldsymbol{\delta}) = -\log \left(\prod_{i=1}^{n_{\text{sr}}} \prod_{j=1, j \neq i}^{n_{\text{sr}}} P(i \succ j)^{M_{ij}} \right) \ . \tag{8}$$

The B-T scores $\boldsymbol{\delta}$ are obtained by minimizing (8) using expectation-maximization [33].

To analyze the agreement among the votes from n_{observer} sessions, we employ the Kendall coefficient of agreement [67]:

$$u = \frac{2W}{\binom{n_{\text{sr}}}{2}\binom{n_{\text{observer}}}{2}} - 1, \quad W = \sum_{i=1}^{n_{\text{sr}}} \sum_{j=1, j \neq i}^{n_{\text{sr}}} \binom{M_{ij}}{2} \ . \tag{9}$$

This describes inter-observer variances and perfect agreement leads to $u = 1$, while uniformly random votes lead to $u = -\frac{1}{n_{\text{sr}}}$.

4.4 Evaluated Algorithms

We investigate 18 classical and state-of-the-art SR methods and bicubic interpolation as categorized in Tab. 3.

For MFSR, we study interpolation, reconstruction, and deep learning methods. Interpolation MFSR comprises non-uniform interpolation (NUISR) [68], NUISR with outlier weighting (WNUISR) [18], and denoising-based refinement (DBRSR) [69]. The reconstruction methods are non-blind L_1 norm minimization with bilateral total variation prior (L1BTV) [22], bilateral edge preserving prior (BEPSR), and iteratively re-weighted minimization (IRWSR) [23]. We also evaluated blind Bayesian video SR (BVSR) [24] and SR with motion blur handling (SRB) [57]. We further use the video SR network (VSRNET) of [25] and the hybrid approach (HYSR) of [70].

For SISR, we study dictionary and deep learning methods. The dictionary methods are example-based ridge regression (EBSR) [9], sparse coding (SCSR) [71], Naive Bayes SR forests (NBSRF) [10], and adjusted anchored neighborhood regression (A+) [12]. The deep learning methods comprise CNNs (SRCNN) [13], very deep networks (VDSR) [14], and deeply-recursive networks (DRCN) [15]. As an internal method, we evaluate transformed self-exemplars (SESR) [8].

We used published reference implementations if available. For L1BTV and BEPSR, we used the publicly available MATLAB SR toolbox [23]. For NUISR, we adopted the method in [70]. BVSR is based on source code provided by the authors of [25]. All learning-based methods use their original pretrained models wherever possible. NBSRF and SCSR were retrained for $3\times$ and $4\times$ magnification on the original training data as pretrained models were unavailable. For VSRNET, we used the pretrained network for $K = 5$ frames for all magnifications. We selected free parameters following the guidelines in the source codes. For methods that require knowledge on the camera PSF, an isotropic Gaussian kernel of size $\lceil 6\sigma_{\text{PSF}} \rceil \times \lceil 6\sigma_{\text{PSF}} \rceil$ pixels was used, where $\sigma_{\text{PSF}} = b\sigma_0$, b is the binning factor, and $\sigma_0 = 0.4$ is the standard deviation on the LR grid.

TABLE 3: Categorization of the SR algorithms in our benchmark.

Category	Single-image (SISR)	Multi-frame (MFSR)
Self-exemplars	SESR [8]	
Deep learning architectures	SRCNN [13],VDSR [14] DRCN [15]	VSRNET [25]
Shallow architectures	SCSR [71], EBSR [9] NBSRF [10], A+ [12]	HYSR [70]
Interpolation	BICUBIC	NUISR [68], WNUISR [18] DBRSR [69]
Non-blind reconstruction		L1BTV [22], BEPSR [72] IRWSR [23]
Blind reconstruction		BVSR [24], SRB [57]

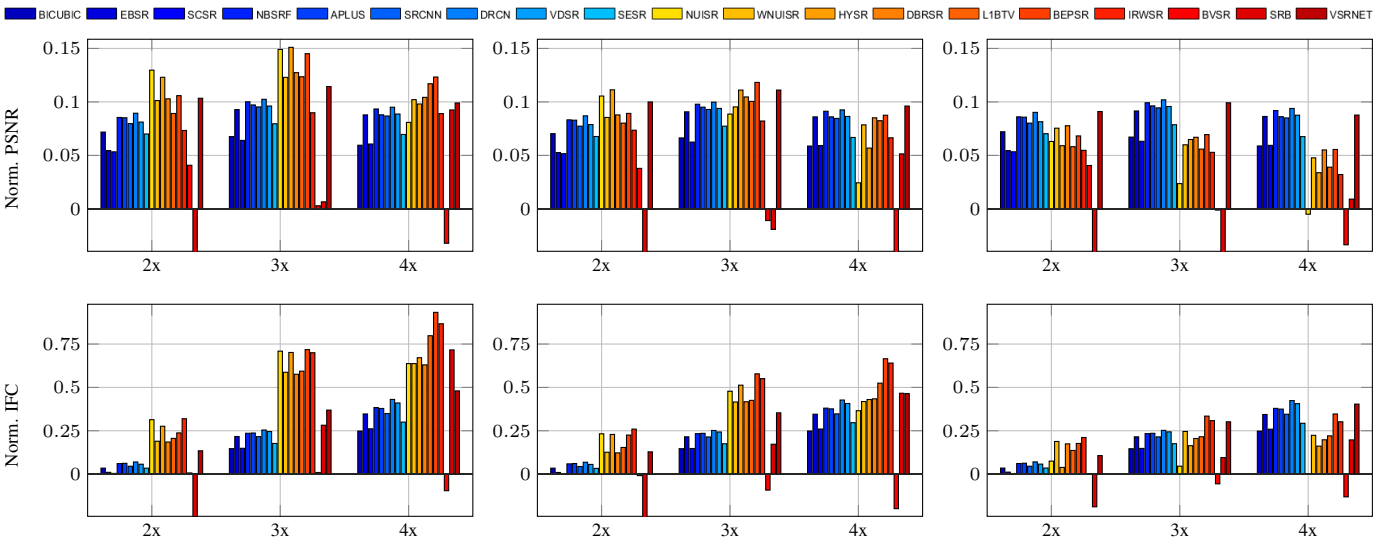
5 Quantitative Study

In this section, we quantitatively evaluate SR on our database. This benchmark aims at evaluating the fidelity of SR images w. r. t. the ground truth data using normalized full-reference quality measures. Results for other measures including no-reference assessment for a perceptual benchmark are provided on our project website. We analyze different motion and environmental conditions as well as video compression.

5.1 Super-Resolution on Static Scenes

Figure 6a shows a comparison of the SR methods with different magnification factors on the global motion datasets using the normalized PSNR and IFC. Overall, we found that relative performances of algorithms depend on the utilized quality measure. Except for large magnifications, interpolation-based MFSR (NUISR, HYSR) performed best in terms of PSNR. In case of IFC, reconstruction-based methods (BEPSR, IRWSR) achieved better scores. This can be explained by two properties. 1) PSNR weighs deviations to the ground truth in homogeneous and textured regions uniformly. We observed that PSNR tends to prefer slightly oversmoothed images, which is consistent with evaluations of full-reference quality assessment [73]. As interpolation-based SR tends to introduce blur, these methods are ranked higher by PSNR. 2) IFC puts the emphasis on high frequencies [39]. Reconstruction-based SR use statistical priors on natural images, e. g. sparsity [22], [23], which leads to a better recovery of high frequencies and thus a higher IFC score.

Interestingly, blind SR (SRB, BVSR) did not perform better than computationally more efficient non-blind methods. SRB was prone to ringing artifacts while BVSR was affected by oversmoothing due to inaccurate PSF estimation, which partly led to negative normalized measures. Figure 7 (top) shows a



(a) Global motion datasets (b) Mixed motion datasets (c) Local motion datasets

Fig. 6: Benchmark of the SR algorithms for global (a), mixed (b), and local motion (c) using the mean normalized PSNR and IFC for $2\times$, $3\times$ and $4\times$ magnification. The algorithms are categorized either as SISR (shown with blue color map) or MFSR (shown with red color map).

(a) LR input (b) A+ [12] (c) VDSR [14] (d) WNUISR [18] (e) IRWSR [23] (f) Ground truth

Fig. 7: SR methods under global motion on the *newspapers* dataset ($3\times$ magnification). Top: SR on raw data. MFSR (e. g. , WNUISR and IRWSR) outperforms SISR algorithms (e. g. , A+ and VDSR) w.r.t. the recovery of fine structures like text and reconstruction algorithms with sparsity priors (e. g. , IRWSR) enhanced the recovery of HR details compared to interpolation-based methods (e. g. , WNUISR). Bottom: SR under H.265/HEVC coding (quantization QP30). All methods are affected by compression artifacts and become indistinguishable.

comparison among different methodologies. Here, the sparsity priors contributed to the recovery of the printed text.

The benchmark also reveals that under pure global motion MFSR outperforms SISR. This is because MFSR exploits complementary information across multiple images to recover HR details, while SISR "hallucinates" these details.

In SISR, it is worth noting that the use of external data outperformed the self-exemplar approach (SESR). However, even some of the sophisticated deep nets (VDSR, DRCN) performed only comparable to classical shallow architectures. This is quite surprising as results on simulated data [14], [15] indicate that deep learning surpasses such classical approaches. In MFSR, interpolation-based algorithms were suitable for small magnification ($2\times$) while reconstruction and deep learning approaches (VSRNET) performed better for larger factors ($3\times$ and $4\times$). We explain this behavior by the use of statistical priors or the use of LR/HR exemplars, which guide the recovery of fine structures.

5.2 Super-Resolution under Video Compression

We investigated the influence of video compression using H.265/HEVC coding. Figure 8 benchmarks the different SR methods at five compression levels for $3\times$ magnification. As expected, video compression considerably affects the overall performance of SR. In particular, we found that at large compression levels the algorithms become indistinguishable as shown in Fig. 7 (bottom). This is related to the deficiency of current SR methodologies to handle video compression. It is interesting to note that most of the current deep learning (SRCNN, VDSR, DRCN) or dictionary learning techniques (SCSR, EBSR, NBSRF, A+) hold the potential to consider video compression in their underlying models. However, this aspect is often ignored within the image formation and training of these methods, which explains their behavior on real data affected by video compression.

Interestingly, video compression has more influence to MFSR. We can explain this observation by antialiasing that

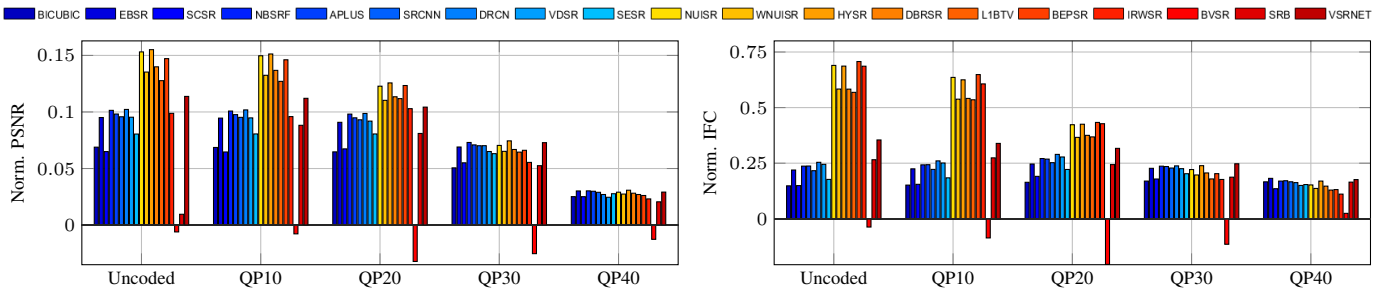


Fig. 8: Robustness analysis of SR w. r. t. video compression. The x-axis denotes the compression level in terms of H.265/HEVC quantization. The y-axis depicts the normalized PSNR and IFC averaged over 14 scenes with global motion and $3\times$ magnification. The individual algorithms are categorized either as SISR (shown with blue color map) or MFSR (shown with red color map).

(a) A+ [12] (b) NUISR [18] (c) IRWSR [23] (d) VSRNET [25] (e) Ground truth

Fig. 9: SR on the dynamic *coffee* dataset ($3\times$ magnification). Top: Mixed motion due movements of a coffee cup. In contrast to SISR (e. g. A+ [12]), simple interpolation-based MFSR (e. g. NUISR [68]) is prone to erroneous optical flow caused by occlusions near object boundaries, while robust reconstruction (e. g. IRWSR [23]) and deep learning methods (e. g. VSRNET [25]) partly compensate for uncertainties. Bottom: Local object motion without camera motion. Except VSRNET [25] that exploits training data, MFSR cannot effectively enhance the resolution.

is implicitly performed by H.265/HEVC coding. Since MFSR exploits aliased signal components to recover HR details, the performance of these algorithms is inherently limited.

5.3 Super-Resolution on Dynamic Scenes

Figure 6b benchmarks SR under mixed motion. Here, the performance of most MFSR algorithms considerably deteriorated compared to static scenes while SISR was unaffected. Motion artifacts were more significant for more input frames at larger magnifications. That is because optical flow estimation becomes more difficult for large displacements related to local motion over longer input sequences. We found that algorithms building on simple interpolation (NUISR, HYSR) were most sensitive. Interpolation-based SR with proper outlier weighting (WNUISR) or refinement (DBRSR) as well as reconstruction-based SR with outlier-insensitive models showed higher robustness. Interestingly, VSRNET was only slightly affected by local motion. We explain this observation by the neural network architecture that was trained for a fixed number of frames and the underlying adaptive motion compensation. Figure 9 (top) depicts some representative methods, where local motion is related to translational movements of a cup. The insufficient optical flow estimation leads to motion artifacts.

Figure 6c depicts our benchmark under local motion. The absence of global motion inherently affected MFSR algorithms as complementary information across LR frames does not exist. Thus, they effectively perform multi-frame deblurring/denoising but cannot overcome undersampling. In our benchmark, SISR partly outperformed MFSR. Among the MFSR algorithms, VSRNET performed best. This can be explained by the external training data used for VSRNET. Without global motion, it drops back to SISR and better recovers HR details than other MFSR methods, as shown in Fig. 9 (bottom).

5.4 Super-Resolution under Photometric Variations

We also studied SR under photometric variations over the input frames. This situation appears if input frames are collected over a longer period of time with environmental changes, e. g. in remote sensing, and is crucial for MFSR. An exact handling requires a photometric registration [74], which is omitted by most state-of-the-art algorithms.

Figure 10 evaluates MFSR for an increasing number of photometric outlier within $K = 11$ consecutive frames with global motion and $3\times$ magnification. We found that even for a few outliers most methods performed worse than simple bicubic interpolation as photometric variations are neither considered implicitly by generative models nor explicitly by proper correction methods. Reconstruction-based algorithms with robust and adaptive models (IRWSR, BEPSR) were less sensitive and adaptively handled photometric variations. Figure 11 depicts this behavior on the *games* dataset. The photometric variations resulted in intensity distortions and

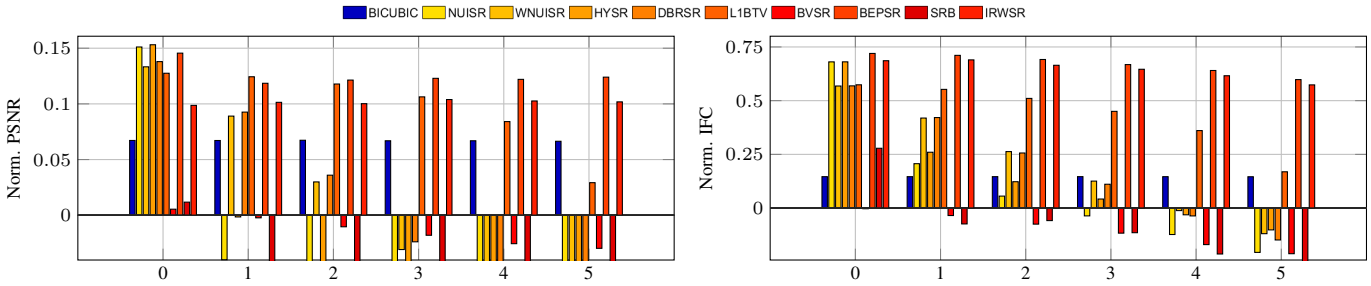
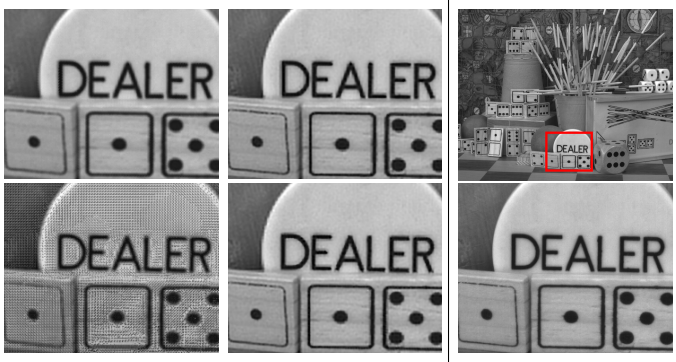


Fig. 10: Robustness analysis of MFSR w. r. t. photometric variations. The x-axis denote the number of photometric outliers within a set of $K = 11$ frames. The y-axis depicts the normalized PSNR and IFC averaged over 14 scenes with global motion and $3\times$ magnification.



(a) NUISR [68] (b) BEPSR [72] (c) Ground truth

Fig. 11: MFSR without (top) and with photometric variations across multiple LR frames (bottom) on the *games* dataset ($3\times$ magnification).

noise in interpolation-based SR (NUISR) while adaptive reconstruction-based SR (BEPSR) was unaffected.

6 Human Observer Study

We conducted an observer study for global motion, mixed motion, local motion, and photometric variations. The data comprises uncompressed images, one camera trajectory (translation x, y, z and pan), and one sliding window per image sequence. Overall, the study was conducted with 3,024 images obtained from 19 SR algorithms at three resolution levels, which yields 26,712 image pairs. The observers comprise paid participants from Amazon Mechanical Turk as well as volunteers including experts in computer vision. We collected 292,400 votes in 5,848 sessions[2]. Each session comprises 50 image pairs, where 8 pairs served as sanity checks. The median time to complete one session was 14.8 minutes. We discarded 16.8% of the votes due to observers who failed the sanity checks.

6.1 Ranking of the Super-Resolution Methods

The algorithm benchmark in our observer study is based on the B-T model. In Fig. 12, we rank the competing SR methods based on their mean B-T scores $\bar{\delta}$ on the global motion, mixed motion, local motion, and photometric variation datasets. One can observe that the ranking heavily depends on the underlying motion and photometric conditions. We make the following observations that partly agree with our quantitative evaluation:

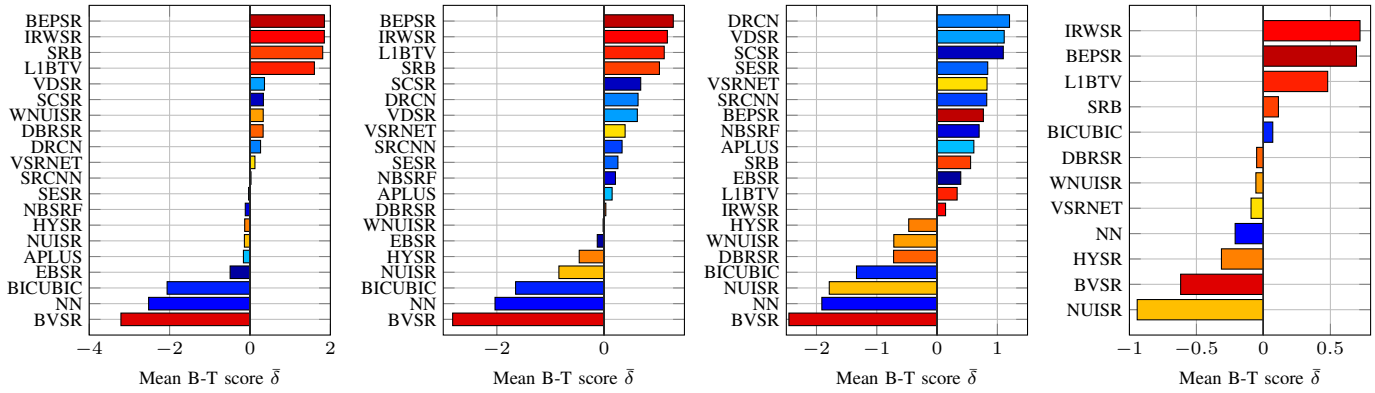
- In case of global motion, reconstruction-based MFSR (BEPSR, IRWSR, SRB, L1BTV) ranked highest.
- In case of mixed motion, MFSR is partly outperformed by SISR. We explain this behavior by the sensitivity of MFSR against inaccurate local motion estimates.
- In case of local motion, SISR algorithms are ranked highest. Particularly, recent deep learning (DRCN, VDSR) and dictionary methods (SCSR) performed best.
- In case of photometric variations, robust reconstruction methods (IRWSR, BEPSR) are ranked highest.

6.2 Inter-Observer Variance

The inter-observer variance is analyzed by the Kendall coefficient of agreement u for different datasets and magnification factors in Fig. 13. We found that the agreement among different observers increases with the magnification factor. This can be explained by a higher perceptual consensus at large factors, where differences among SR algorithms are often clearly visible. For smaller factors, observers are more often controversial about algorithm performance. Moreover, the agreement is higher under global motion compared to local and mixed motion. This is because motion artifacts in MFSR are more likely under local motion. Therefore, some observers subjectively prefer SISR that hallucinates HR details without motion artifacts, while others tolerate slight motion artifacts but prefer the recovery of true HR details as achieved by MFSR. Photometric variations lead to the highest agreement, since here quality differences are clearly visible.

Figure 14 depicts Kendall's u for different numbers of sessions included in the evaluation. Furthermore, Kendall's u is shown for different error levels $n_f \in \{0, 1, 8\}$ tolerated for the sanity checks. We performed a Monte-Carlo simulation similar to [33] and randomly sampled the respective number of sessions from the entire set of sessions. The mean and standard deviation of Kendall's u was determined over 1,000 samples for each number of sessions. We found that Kendall's u converges in terms of the standard deviation and becomes stable beyond 2,000 sessions with lower values for higher error levels n_f. Notice that the inter-observer variance is considerably lower if we omit the sanity checks (i. e. $n_f = 8$). This result justifies the sanity checks that aim at removing outliers from the paired comparisons. We use $n_f = 1$ as a tradeoff between outlier removal and tolerating mistakenly entered votes of observers.

2. We tolerated repeated participations for the observers, but ensured that each session consists of randomly selected image pairs.



(a) Global motion datasets (b) Mixed motion datasets (c) Local motion datasets (d) Photometric variation datasets

Fig. 12: Ranking of the competing SR algorithms on the different datasets in our observer study. We ranked the algorithms w. r. t. their mean B-T scores $\bar{\delta}$, where higher, positive scores express better image quality according to human visual perception. The individual algorithms are categorized either as SISR (shown with blue color map) or MFSR (shown with red color map).

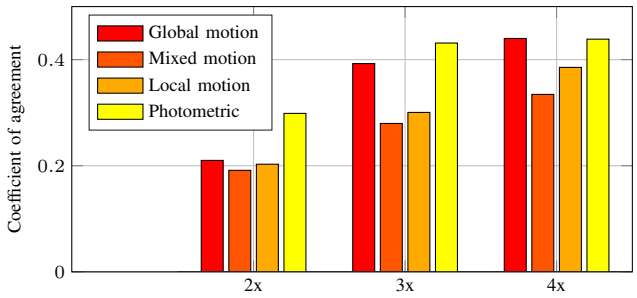


Fig. 13: Kendall coefficients of agreement among the observers in our study for different datasets and magnification factors.

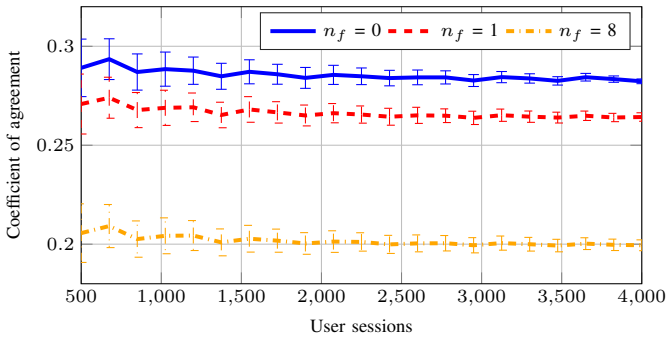


Fig. 14: Convergence of the Kendall coefficient of agreement. We determined mean $\pm$ standard deviation of the agreement in a Monte-Carlo simulation at different error levels n_f for our sanity checks.

6.3 Image Quality and Computation Time Tradeoff

Computation time is relevant to many practical SR applications. For all methods, MATLAB sources on the CPU are used, where some modules are accelerated by C++. SR is computed to the 1200×960 pixels resolution of the ground truth images and computation times for MFSR include optical flow estimation. Figure 15 shows the tradeoff between image quality in the B-T model and computation time for $2\times$ and $4\times$ magnification. We identify three classes of algorithms:

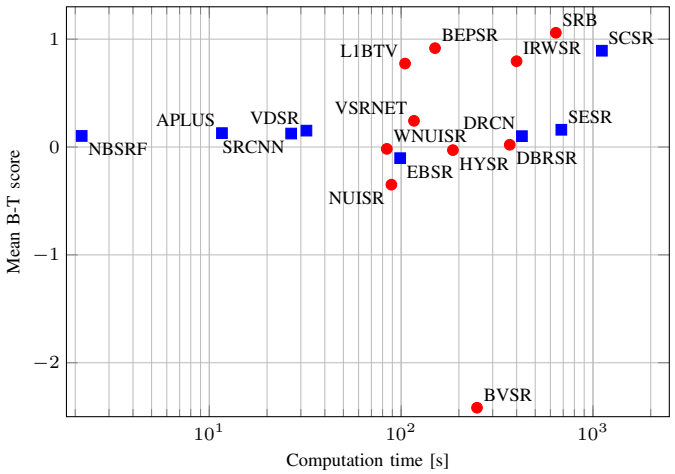
- Five SISR (EBSR, NBSRF, SRCNN, DRCN, VDSR) and six MFSR methods (HYSR, DBSR, SRB, L1BTV, BEPSR, BVSR) have higher complexities for larger magnifications. The computational time of MFSR methods additionally depends on the number of input images.
- The computation time of three SISR methods (SCSR, A+, SESR) are mainly influenced by the input image resolution, but not by magnification.
- The computational time of four MFSR methods (VSRNET, NUISR, WNUISR, IRWSR) is unaffected by these factors.

Overall, VDSR and SRCNN – two recent deep learning methods – have excellent quality/time tradeoffs. For small magnifications, SCSR yields higher quality but is much slower. NBSRF and A+ are faster but achieve lower quality. In terms of MFSR, non-blind reconstruction (L1BTV, IRWSR, BEPSR) as well as VSRNET show good quality/time tradeoffs. The interpolation-based NUISR and WNUISR are slightly faster but achieve lower image quality. However, interpolation-based SR doubles the complexity of motion estimation as optical flow is computed in forward and backward direction [18].

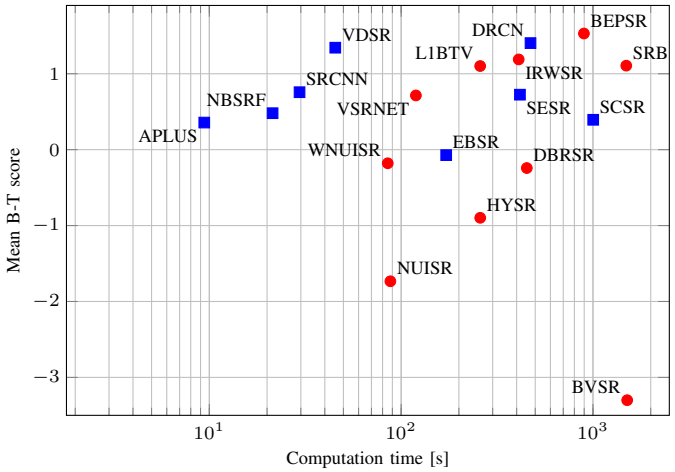
6.4 Correlation to Quantitative Study

To analyze the correlation between quantitative quality assessment and human visual perception expressed by the B-T model, we employ the weighted Kendall τ distance [75]. This approach compares rankings obtained by the quantitative measures to rankings in the B-T model, where lower distances express stronger correlations. Figure 16 shows the mean weighted Kendall τ (normalized to $[0, 1]$) of nine measures for the different datasets and magnification factors. We found that for most measures their respective weighted Kendall τ depends on the magnification factor to a certain extent. In particular, the higher the magnification the lower the weighted Kendall τ. This is consistent with prior studies [36], [65] and suggests that quantitative evaluations should focus on higher magnifications if one desires fair correlations to human visual perception.

For no-reference quality assessment, BRISQUE shows the strongest correlations for most datasets and magnification factors. For the full-reference measures, IFC shows strong correlations while the commonly used PSNR is often a



(a) Mean B-T score vs. computation time (2× magnification)



(b) Mean B-T score vs. computation time (4× magnification)

Fig. 15: Tradeoff between image quality in terms of the B-T scores and the corresponding computation time for 2× and 4× magnification. We compare SISR (shown in blue) and MFSR method (shown in red).

weaker indicator for human visual perception, especially for small magnifications. Furthermore, IFC often shows higher correlations than no-reference approaches, especially for larger magnifications. This is remarkable as prior work [36] proposed to use no-reference measures to quantify perceptual quality. We explain this contradiction by the fact that current no-reference approaches are either generic (e. g. NIQE) or customized to SR but developed under simplified conditions (e. g. SRM). For instance, SRM was trained from SISR results on simulated data, which explains the lower performance on real data with MFSR facets. The weak correlations of no-reference assessment are particularly noticeable under photometric variations, where we observe high weighted Kendall τ distances. This is another occurrence of the simulated-to-real gap and also underlines the importance of ground truth data for SR benchmarking.

7 Discussion

The benchmark reveals several interesting conclusions related to the simulated-to-real gap and provides guidelines for future research.

7.1 Remarks on the Quantitative Study

In SISR, the use of external training data outperforms self-exemplar approaches [8]. Surprisingly, depending on the quality measure, popular deep nets [13], [14], [15] do not clearly outperform classical methods [9], [10], [12], [71]. This contrasts benchmarks on simulated data, where deep nets perform best.

We also found that MFSR surpasses SISR in baseline experiments with pure global motion. Despite the success of learning-based methods, classical sparsity priors [22], [23], [72] are still invaluable in this field. We consider the further development of such priors as well as their combination with learning-based architectures as a promising way.

SR quality heavily depends on environmental conditions. Generally, MFSR is sensitive to failures in optical flow estimation. This explains the weaker performance on mixed or local motion as well as photometric variations, while SISR is unaffected by these factors. Within MFSR, reconstruction [22], [23], [72] or deep learning [25] methods are more robust than interpolation-based approaches [18], [68], [69].

Video compression, e. g. H.265/HEVC, challenges all methods and MFSR in particular. The proposed dataset can be useful to create compressed training data for future learning-based methods. Also for future work, hybrid approaches to combine strengths of SISR and MFSR appear to be highly relevant, like [70] or architectures like VSRNET [25].

7.2 Remarks on the Human Observer Study

Visual perception of SR image quality in our observer study shows reasonable agreements to the algorithm rankings in the quantitative study. Specifically, classical reconstruction-based algorithms like [22], [23], [72] (in MFSR) and learning-based methods like [14], [15], [71] (in SISR) are ranked highest in a B-T model derived from pair-wise comparisons.

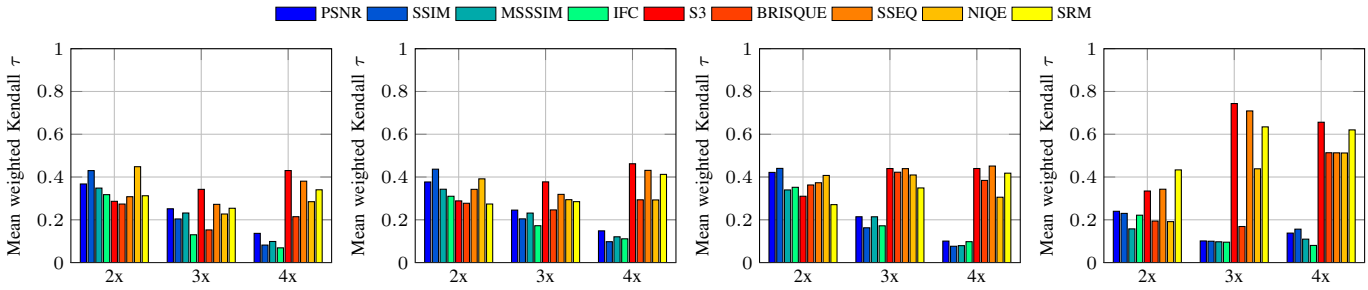
Inter-observer variances heavily depend on SR parameters. Most importantly, our analysis shows that larger magnification factors lead to a higher consensus between different observers. In view of this finding, reliable evaluations should conduct SR with large magnifications if agreements to human visual perception are desired.

If low computational complexities are important, non-blind reconstruction algorithms [22], [23], [72] and VSRNET [25] (in MFSR) as well as deep learning [13], [14] (in SISR) feature good tradeoffs between image quality and computation time.

The actual correlation between human perception and quantitative benchmarks depends on the quality measure, as also reported in [33], [39]. The popular PSNR shows weak correlations compared to the information-theoretic IFC. For future work, we encourage the use of such elaborated measures to reliably benchmark SR. Beyond that full-reference measures have stronger correlations compared to no-reference measures on real images, which underlines the importance of ground truth data and the need to improve current no-reference methods.

8 Conclusion

This paper presented a SR database to conduct the largest SR benchmark to date, across both SISR and MFSR algorithms.



(a) Global motion datasets (b) Mixed motion datasets (c) Local motion datasets (d) Photometric variation datasets

Fig. 16: Mean weighted Kendall τ distance among the B-T scores and different quantitative quality measures. The lower the weighted Kendall τ the higher the correlation to human visual perception expressed by the B-T model. We compared four full-reference measures (shown with blue color map) as well as five no-reference measures (shown with red color map) for different datasets and magnification factors.

This is the first dataset to combine image sequences of real LR acquisitions with ground truth data. Additionally, the size of the database is a magnitude larger than currently existing benchmarks. We also demonstrated that evaluations on simulated data do not necessarily reflect the performance of SR on real data. This simulated-to-real gap is closed by our data that was captured by means of hardware binning with challenging effects like non-Gaussian noise. Future works can also use the data to move model training from simulated to real images, and to perform thorough quantitative evaluations.

Our database considers various types of environmental aspects (local motion and photometric variation), technological aspects (hardware binning) as well as software aspects (video coding) without involving simulations. We plan to also acquire data with sensor-specific artifacts, e. g. from the color-filter array (CFA) or rolling shutter [76]. Also, some application-specific artifacts are left for future work. For instance, motion blur is important for SR in the wild and could be addressed by *recording-and-playback* of motion similar to [77].

Acknowledgments

We would like to thank all observers of our human observer study for their participation. We also thank the authors of prior works for providing the source codes for our benchmark.

References

[1] P. Milanfar, *Super-Resolution Imaging*. CRC Press, 2010.
[2] L. Zhang, H. Zhang, H. Shen, and P. Li, "A Super-Resolution Reconstruction Algorithm for Surveillance Images," *Signal Processing*, vol. 90, no. 3, pp. 848–859, 2010.
[3] H. Zhang, L. Zhang, and H. Shen, "A Super-Resolution Reconstruction Algorithm for Hyperspectral Images," *Signal Processing*, vol. 92, no. 9, pp. 2082–2096, 2012.
[4] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "LidarBoost: Depth Superresolution for ToF 3D Shape Scanning," *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 343–350, 2009.
[5] T. Köhler, A. Brost, K. Mogalle, Q. Zhang, C. Köhler, G. Michelson, J. Hornegger, and R. P. Tornow, "Multi-Frame Super-Resolution with Quality Self-Assessment for Retinal Fundus Videos," in *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2014, pp. 650–657, LNCS Vol. 8673, Part I.
[6] T. Köhler, S. Haase, S. Bauer, J. Wasza, T. Kilgus, L. Maier-Hein, C. Stock, J. Hornegger, and H. Feussner, "Multi-Sensor Super-Resolution for Hybrid Range Imaging with Application to 3-D Endoscopy and Open Surgery," *Medical Image Analysis*, vol. 24, no. 1, pp. 220–234, 2015.
[7] D. Glasner, S. Bagon, and M. Irani, "Super-Resolution from a Single Image," in *Proc. International Conference on Computer Vision (ICCV)*, 2009, pp. 349–356.
[8] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5197–5206.
[9] K. I. Kim and Y. Kwon, "Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.
[10] J. Salvador and E. Pérez-Pellitero, "Naive Bayes Super-Resolution Forest," in *Proc. International Conference on Computer Vision (ICCV)*, 2015, pp. 325–333.
[11] S. Schulter, C. Leistner, and H. Bischof, "Fast and Accurate Image Upscaling with Super-Resolution Forests," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3791–3799.
[12] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. Asian Conference on Computer Vision (ACCV)*, vol. 9006, 2015, pp. 111–126.
[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a Deep Convolutional Network for Image Super-Resolution," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 184–199.
[14] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1646–1654.
[15] ——, "Deeply-Recursive Convolutional Network for Image Super-Resolution," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1637–1645.
[16] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[17] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[18] M. Bätz, A. Eichenseer, and A. Kaup, "Multi-Image Super-Resolution using a Dual Weighting Scheme based on Voronoi Tessellation," in *Proc. International Conference on Image Processing (ICIP)*, 2016, pp. 2822–2826.
[19] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel Regression for Image Processing and Reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349–366, 2007.
[20] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Variational Bayesian Super Resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984–999, 2011.
[21] C. Bercea, A. Maier, and T. Köhler, "Confidence-Aware Levenberg-Marquardt Optimization for Joint Motion Estimation and Super-Resolution," in *Proc. International Conference on Image Processing (ICIP)*, 2016, pp. 1136–1140.
[22] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.
[23] T. Köhler, X. Huang, F. Schebesch, A. Aichert, A. Maier, and J. Hornegger, "Robust Multiframe Super-Resolution Employing Iteratively Re-Weighted Minimization," *IEEE Transactions on Computational Imaging*, vol. 2, no. 1, pp. 42–58, 2016.
[24] C. Liu and D. Sun, "On Bayesian Adaptive Video Super Resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 346–360, 2014.

[25] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video Super-Resolution With Convolutional Neural Networks," *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, 2016.

[26] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jiaya, "Video Super-Resolution via Deep Draft-Ensemble Learning," in *Proc. International Conference on Computer Vision (ICCV)*, 2015, pp. 531–539.

[27] U. S. Kim and M. H. Sunwoo, "New Frame Rate Up-Conversion Algorithms With Low Computational Complexity," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 3, pp. 384–393, 2014.

[28] M. Bätz, T. Richter, G. Jens-Uwe, A. Papst, J. Seiler, and A. Kaup, "High Dynamic Range Video Reconstruction from a Stereo Camera Setup," *Signal Processing: Image Communication*, vol. 29, no. 2, pp. 191–202, 2014, special Issue on Advances in High Dynamic Range Video Research.

[29] S. Baker and T. Kanade, "Limits on Super-Resolution and How to Break Them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, 2002.

[30] Z. Lin and H.-Y. Shum, "Fundamental Limits of Reconstruction-Based Superresolution Algorithms under Local Translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 83–97, 2004.

[31] D. Robinson and P. Milanfar, "Statistical performance analysis of super-resolution," *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1413–1428, 2006.

[32] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[33] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang, "A Comparative Study for Single Image Blind Deblurring," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] S. Farsiu, D. Robinson, and P. Milanfar, "Multi-Dimensional Signal Processing Dataset, last accessed 03/17/17," https://users.soe.ucsc.edu/ milanfar/software/sr-datasets.html, 2016.

[35] P. Vandewalle, "LCAV Super-Resolution Datasets, last accessed 03/17/17," http://lcav.epfl.ch/software/superresolution, 2016.

[36] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] R. Timofte, R. Rothe, and L. V. Gool, "Seven Ways to Improve Example-Based Single Image Super Resolution," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1865–1873.

[38] T. Plotz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1586–1595.

[39] C. Y. Yang, C. Ma, and M. H. Yang, "Single-Image Super-Resolution: A Benchmark," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 372–386.

[40] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[41] R. Timofte, E. Agustsson, L. V. Gool, M. H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee et al., "NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1110–1121.

[42] W. Bae, J. Yoo, and J. C. Ye, "Beyond Deep Residual Learning for Image Restoration: Persistent Homology-Guided Manifold Simplification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1141–1149.

[43] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1132–1140.

[44] H. Yeganeh, M. Rostami, and Z. Wang, "Objective quality assessment for image super-resolution: A natural scene statistics approach," in *Proc. International Conference on Image Processing (ICIP)*, 2012, pp. 1481–1484.

[45] Q. Yuan, L. Zhang, and H. Shen, "Multiframe Super-Resolution Employing a Spatially Weighted Total Variation Model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 3, pp. 379–392, 2012.

[46] D. Dai, Y. Wang, Y. Chen, and L. Van Gool, "Is image super-resolution helpful for other vision tasks?" in *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–9.

[47] R. Raghavendra, K. Raja, B. Yang, and C. Busch, "Comparative evaluation of super-resolution techniques for multi-face recognition using light-field camera," in *International Conference on Digital Signal Processing*, 2013, pp. 1–6.

[48] C. Qu, D. Luo, E. Monari, T. Schuchert, and J. Beyerer, "Capturing Ground Truth Super-Resolution Data," in *Proc. International Conference on Image Processing (ICIP)*, 2016, pp. 2812–2816.

[49] H. Dirks, J. Geiping, D. Cremers, and M. Moeller, "Multiframe Motion Coupling via Infimal Convolution Regularization for Video Super Resolution," *arXiv preprint 1611.07767v1*, 2016.

[50] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[51] "Basler ace data sheet," Basler AG, 2016.

[52] Edmund Optics Ltd, "High resolution fixed focal length lens #85-866," https://www.edmundoptics.com/imaging-lenses/, 2016.

[53] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE Image Quality Assessment Database Release 2, last accessed 03/17/17," http://live.ece.utexas.edu/research/quality, 2016.

[54] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[55] A. Wirgin, "The inverse crime," *arXiv preprint arXiv:math-ph/0401050*, 2004.

[56] C. Liu, "Beyond Pixels: Exploring New Representations and Applications for Motion Analysis," PhD thesis, Massachusetts Institute of Technology, 2009.

[57] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu, "Handling Motion Blur in Multi-Frame Super-Resolution," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5224–5232.

[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[59] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-Scale Structural Similarity for Image Quality Assessment," in *Proc. IEEE Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems, and Computers*, 2003, pp. 1398–1402.

[60] H. R. Sheikh, A. Bovik, and G. de Veciana, "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[61] C. T. Vu, T. D. Phan, and D. M. Chandler, "S3: A Spectral and Spatial Measure of Local Perceived Sharpness in Natural Images," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 934–45, 2012.

[62] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–708, 2012.

[63] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014.

[64] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[65] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[66] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Computer Graphics Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[67] M. G. Kendall and B. B. Smith, "On the Method of Paired Comparisons," *Biometrika*, vol. 31, p. 324, 1940.

[68] S. C. Park, M. K. Park, and M. G. Kang, "Super-Resolution Image Reconstruction: A Technical Overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[69] M. Bätz, J. Koloda, A. Eichenseer, and A. Kaup, "Multi-Image Super-Resolution Using a Locally Adaptive Denoising-Based Refinement," in *Proc. IEEE Workshop on Multimedia Signal Processing (MMSP)*, 2016, pp. 1–6.

[70] M. Bätz, A. Eichenseer, J. Seiler, M. Jonscher, and A. Kaup, "Hybrid Super-Resolution Combining Example-based Single-Image and Interpolation-based Multi-Image Reconstruction Approaches," in *Proc. International Conference on Image Processing (ICIP)*, 2015, pp. 58–62.

[71] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[72] X. Zeng and L. Yang, "A Robust Multiframe Super-Resolution Algorithm based on Half-Quadratic Estimation with Modified BTV Regularization," *Digital Signal Processing*, vol. 23, no. 1, pp. 98–109, 2013.
[73] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3441–3452, 2006.
[74] D. P. Capel and A. Zisserman, "Computer Vision Applied to Super Resolution," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 75–86, 2003.
[75] Y. Liu, J. Wang, S. Cho, A. Finkelstein, and S. Rusinkiewicz, "A No-Reference Metric for Evaluating the Quality of Motion Deblurring," *ACM Transactions on Graphics*, vol. 32, no. 6, pp. 175–1, 2013.
[76] A. Punnappurath, V. Rengarajan, and A. N. Rajagopalan, "Rolling shutter super-resolution," in *Proc. International Conference on Computer Vision (ICCV)*, vol. 2015 Inter, 2015, pp. 558–566.
[77] R. Köhler, M. Hirsch, B. Mohler, B. Schölkopf, and S. Harmeling, "Recording and playback of camera shake: benchmarking blind deconvolution with a real-world database," in *Proc. European Conference on Computer Vision (ECCV)*, 2012, pp. 27–40.