# NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study

Eirikur Agustsson
CVL, ETH Zurich, Switzerland
aeirikur@vision.ee.ethz.ch

Radu Timofte
CVL, ETH Zurich & Merantix GmbH
radu.timofte@vision.ee.ethz.ch

## Abstract

*This paper introduces a novel large dataset for example-based single image super-resolution and studies the state-of-the-art as emerged from the NTIRE 2017 challenge. The challenge is the first challenge of its kind, with 6 competitions, hundreds of participants and tens of proposed solutions. Our newly collected DIVerse 2K resolution image dataset (DIV2K) was employed by the challenge. In our study we compare the solutions from the challenge to a set of representative methods from the literature and evaluate them using diverse measures on our proposed DIV2K dataset. Moreover, we conduct a number of experiments and draw conclusions on several topics of interest. We conclude that the NTIRE 2017 challenge pushes the state-of-the-art in single-image super-resolution, reaching the best results to date on the popular Set5, Set14, B100, Urban100 datasets and on our newly proposed DIV2K.*

## 1. Introduction

Example-based single image super-resolution (SR) aims at full restoration of rich details (high frequencies) in images based on prior examples under the form of low resolution (LR) and corresponding high resolution (HR) images. The loss of details / contents can be due to various degrading factors such as blur, decimation, noise or hardware limitations (*e.g*. camera sensors). SR is an ill-posed problem because for each LR image patch the number of corresponding HR image patches can be very large.

Single image super-resolution as well as image restoration research literature spans over decades [36, 20, 4, 13, 16, 3, 15, 14, 6, 32, 54, 30, 17, 23, 12, 47, 48, 10, 21]. Nonetheless, the recent years showed tremendous progress as shown in Fig. 1. The performance of the top methods have continuously improved [54, 48, 21, 26] as the field has reached maturity.

There is a continuous need for standardized SR benchmarks to allow for comparison of different proposed methods under the same conditions. Most of the recent SR works adopted a couple of datasets like the 91 train images

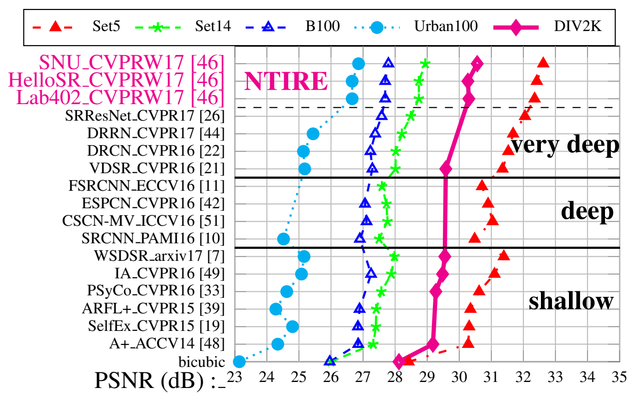



Figure 1. Representative methods from the recent years and their average PSNR performance on five datasets for scale $\times 4$.

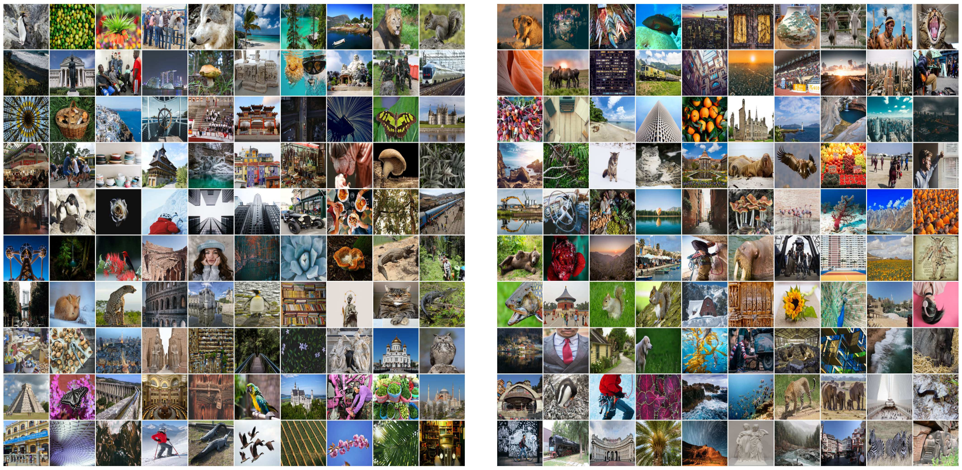
proposed by Yang *et al*. [54] and the validation datasets Set5 [5], Set14 [56], B100 [31, 48] brought together by Timofte *et al*. [47, 48], or the more recent Urban100 [19]. The bicubic downscaling (imresize from Matlab) is the most used degradation operator to simulate the HR to LR transformation.

In this work, we propose a novel DIV2K dataset with DIVerse 2K resolution high quality images collected from Internet. It has 1000 images with considerable higher resolution than the popular datasets mentioned before. Moreover, we organized the first example-based single image super-resolution online challenge which used the DIV2K dataset. The NTIRE 2017 SR challenge [1] [46] employs two types of degradations: the standard bicubic and the unknown downscaling operators *aka* downscaling operators known only through train data of LR and corresponding HR images.

Another contribution of this paper is a study of our newly proposed DIV2K in relation with the achieved performance by the winners of the NTIRE 2017 SR Challenge and representative methods from recent years. We report results using a selection of image quality measures and investigate correlations and limits in SR benchmarking.

The remainder of the paper is structured as follows. Section 2 introduces the DIV2K dataset. Section 3 reviews the NTIRE 2017 SR Challenge and its settings. Section 4 intro-

[1] http://www.vision.ee.ethz.ch/ntire17/



**DIV2K 100 validation images** **DIV2K 100 test images**

Figure 2. Visualization of proposed DIV2K validation and test images. DIV2K contains also 800 train images.

duces the image quality assessment (IQA) measures, Section 5 - the datasets, and Section 6 - the methods from our study. Section 7 discusses the experiments and interprets the achieved results, while Section 8 concludes the paper.

## 2. Proposed DIV2K dataset

We propose the DIV2K dataset [2], a novel DIVerse 2K resolution image dataset for benchmarking example-based single image super-resolution (see Fig. 2. DIV2K is intended to complement the existing SR datasets (see Fig. 5) and to further increase the (content) diversity.

**Source:** We manually crawled 1000 color RGB images from Internet paying special attention to the image quality, to the diversity of sources (sites and cameras), to the image contents and to the copyrights. DIV2K is meant for research purposes.

**Resolution and quality:** All the 1000 images are 2K resolution, that is they have 2K pixels on at least one of the axes (vertical or horizontal). All the images were processed using the same tools. For simplicity, since the most common magnification factors in the recent SR literature are of $\times 2$, $\times 3$ and $\times 4$ we cropped the images to multiple of 12 pixels on both axes. Most of the crawled images were originally above 20M pixels. The images are of high quality both aesthetically and in the terms of small amounts of noise and other corruptions (like blur and color shifts).

**Diversity:** We collected our images from dozens of sites. A preference was made for sites with freely shared high quality photography (such as `https://www.pexels.com/` ). Note that we did not use images from Flickr, Instagram, or other legally binding or copyright restricted images. We only seldom used keywords to assure the diversity for our dataset. DIV2K covers a large diversity of contents, ranging from people, handmade objects and environments (cities, villages), to flora and fauna, and natural sceneries including underwater and dim light conditions.

**Partitions:** After collecting the DIV2K 1000 images we computed image entropy, bit per pixel (bpp) PNG compression rates and CORNIA scores (see Section 7.6) and applied bicubic downscaling $\times 3$ and then upscaling $\times 3$ with bicubic interpolation (imresize Matlab function), ANR [47] and A+ [48] methods and default settings. We randomly generated partitions of 800 train, 100 validation and 100 test images until we achieved a good balance firstly in visual contents and then on the average entropy, average bpp, average number of pixels per image (ppi), average CORNIA quality scores and also in the relative differences between the average PSNR scores of bicubic, ANR and A+ methods. Table 1 summarizes the main characteristics of DIV2K validation and test partitions in comparison with the most popular SR datasets. Fig. 2 visualizes the 100 images for validation and the 100 images for testing of the DIV2K dataset.

## 3. NTIRE 2017 SR Challenge

The NTIRE 2017 challenge on example-based single image super-resolution [46] was the first of its kind and had as objectives: to gauge the state-of-the-art in SR, to facilitate comparison of different solutions on a novel large dataset - DIV2K, and to propose more challenging SR settings.

**Tracks and competitions** The challenge had two tracks: *Track 1 for bicubic downscaling ('classic')* and *Track 2 for Unknown downscaling*. For Track 1 the degradation is the popular bicubic downscaling (Matlab imresize function) and facilitates easy deployment of the recent solutions that assumed this degradation. Track 2 is more challenging as it uses a combination of blur and decimation 'unknown' under explicit form to the challenge participants, but known through exemplars of LR and corresponding HR images. Each track corresponds to 3 competitions for the

[2] `https://data.vision.ee.ethz.ch/cvl/DIV2K/`

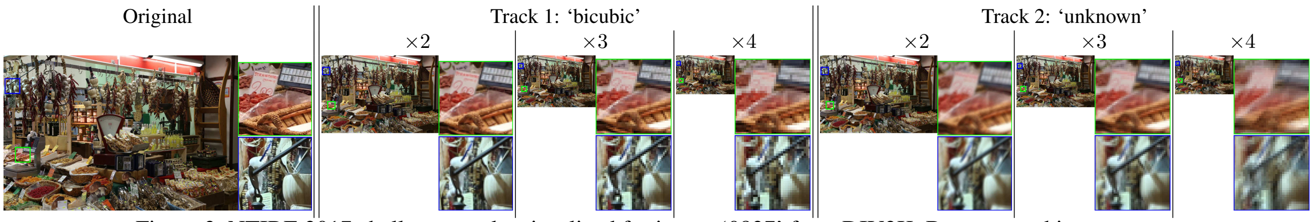


Figure 3. NTIRE 2017 challenge tracks visualized for image '0837' from DIV2K. Best zoomed in on screen.

usual 3 downscaling factors ($\times 2, \times 3, \times 4$). A visualization of the tracks and competitions is shown in Fig. 3. The hosting platform for the competitions is CodaLab [3]. For each competition the LR and HR train images (from the DIV2K train set) were provided for learning models during the *development (training) phase*. The following *validation phase* gave the opportunity to the participants to test their solutions on the LR validation images (from DIV2K validation set) and compare their scores through an online validation server and associated leaderboard. The *final evaluation (test) phase* provided the LR test images (from DIV2K testing set) and invited the submission of the HR results before the challenge deadline. PSNR and SSIM (see Section 4) are the challenge main quantitative quality assessment measures for the image restoration results. A $6 + scalefactor$ pixels image boundary is ignored in the evaluation.

**Challenge results** Each competition had on average 100 registered participants and 20 teams submitted results, code/executables and factsheets for the final test phase. All these competing solutions and the achieved results on the DIV2K test data are described in the NTIRE 2017 SR challenge report [46]. All the proposed challenge solutions, except WSDSR [7], employ end-to-end deep learning of convolutional neural networks (CNN) [25] and use GPU for both training and testing. They propose a diversity of ideas and design details and generally build upon and go beyond the very recent proposed SR works [10, 49, 21, 26]. In Fig. 4 we plot the average PSNR vs. runtime results of the challenge solutions in comparison with several other representative methods and in Table 2 we show results for a selection of them. The top challenge solutions are consistent across all 6 competitions, showing that the solutions proposed for Track 1 with bicubic downscaling generalize well to Track 2 unknown downscaling if sufficient training examples are provided. The PSNR and the SSIM scores correlate well. The scores on Track 2 are generally worse than on Track 1 for the same methods/solutions and reflects the increased difficulty of the unknown downscaling setup.

## 4. Image Quality Assessment (IQA)

There is a large interest in the automatic assessment of the image quality and numerous measures have been proposed [41, 57, 34, 29]. According to the presence and the use of a ground truth reference image there are two main categories: full reference measures and no-reference measures.

When a ground truth image $G$ with $N$ pixels is available the quality of a corresponding (degraded or restored) image $I$ can be defined as the pixel-level fidelity to the ground truth. Representatives are **Mean Square Error (MSE)** defined by $MSE = \frac{1}{N}\sum_{i=1}^{N}(G_i - I_i)^2$ where $G_i$ (or $I_i$) is the $i$-th pixels of $G$ (or $I$), and **Peak Signal-to-Noise Ratio (PSNR)**, $PSNR = 10\log_{10}\left(\frac{MAX_G^2}{MSE}\right)$ where $MAX_G$ is the maximum possible pixel value of the image, here 255. However, small shifts in the content of $I$ leads to (very) poor MSE and PSNR scores even when the contents are identical. Therefore, another group of measures amounts for such structural similarity. If MSE and PSNR measure absolute errors the **Structural Similarity index (SSIM)** [50] [4] is a perception-based model that considers image degradation as perceived change in structural information and **Information Fidelity Criterion (IFC)** [40] [5] assesses the image quality based on natural scene statistics.

From the no-reference measures we chose **Codebook Representation for No-Reference Image Assessment (CORNIA)** [55] [6], a model learned to map images to average human quality assessments. CORNIA is a perceptual measure which does not use a reference image.

All the above selected measures were not intended for measuring the quality of a super-resolved image. However, they tend to generalize well for different kinds of image distortions. In particular, IFC was shown to have a strong correlation with the human assessed perceptual quality for image super-resolution results [53]. While working without a reference image, CORNIA proved superior to many full reference measures in assessing perceptual quality. Very recent CORNIA was shown to achieve high correlation to the human perception also for the image super-resolution task on a large dataset with and without retraining its model [29]. We use the CORNIA with the original model and the default settings.

The SR task aims at recovering the original contents (de-

[3] `https://competitions.codalab.org/`

[4] `https://ece.uwaterloo.ca/~z70wang/research/ssim/`

[5] `http://live.ece.utexas.edu/research/quality/ifcvec_release.zip`

[6] `http://www.umiacs.umd.edu/~pengye/research/CORNIA_release_v0.zip`

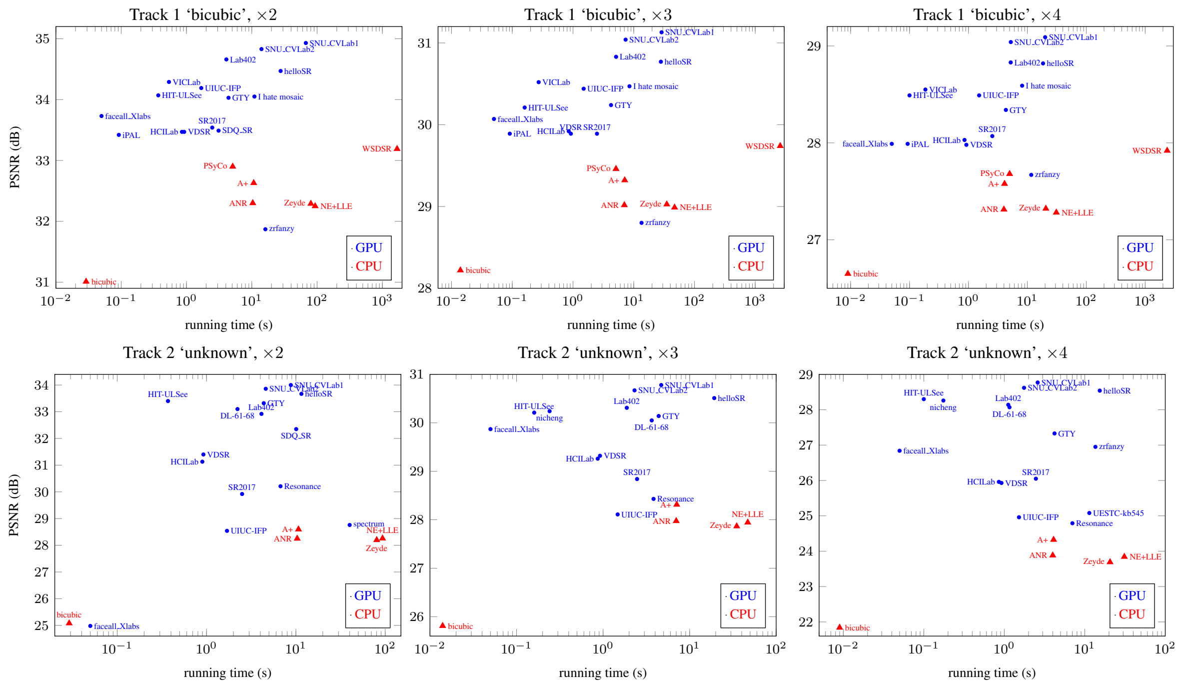



Figure 4. Runtime vs. PSNR results on NTIRE 2017 Challenge tracks, DIV2K test data.

| Dataset | images | CORNIA | ppi | bpp PNG | entropy |
|---|---|---|---|---|---|
| Set5 | 5 | 21.50 (±20.00) | 113491 | 12.45 (±1.77) | 7.37 (±0.51) |
| Set14 | 14 | 21.56 (±12.52) | 230202 | 12.57 (±3.89) | 7.03 (±0.89) |
| B100 | 100 | 18.46 (±12.74) | 154401 | 14.12 (±2.76) | 7.30 (±0.44) |
| Urban100 | 100 | 5.80 (±19.26) | 774313 | 14.01 (±2.76) | 7.57 (±0.36) |
| DIV2K validation | 100 | 15.86 (±15.38) | 2835028 | 12.69 (±2.69) | 7.33 (±0.83) |
| DIV2K test | 100 | 17.47 (±14.55) | 2757182 | 12.64 (±2.33) | 7.50 (±0.47) |
| DIV2K test ⇓ 2 bic. | 100 | 17.93 (± 13.09) | 689295 | 13.41 (± 2.38) | 7.49 (± 0.47) |
| DIV2K test ⇓ 4 bic. | 100 | 18.26 (± 13.08) | 172323 | 14.40 (± 2.45) | 7.48 (± 0.47) |
| DIV2K test ⇓ 8 bic. | 100 | 23.20 (± 14.09) | 43166 | 15.37 (± 2.44) | 7.47 (± 0.46) |
| DIV2K test ⇓ 16 bic. | 100 | 31.34 (± 15.76) | 10849 | 16.49 (± 2.32) | 7.45 (± 0.45) |
| DIV2K test ⇓ 2 crop | 100 | 16.93 (± 14.55) | 689295 | 13.79 (± 2.45) | 7.52 (± 0.35) |
| DIV2K test ⇓ 4 crop | 100 | 22.59 (± 15.75) | 172323 | 14.09 (± 2.67) | 7.37 (± 0.45) |
| DIV2K test ⇓ 8 crop | 100 | 30.85 (± 18.66) | 43166 | 14.12 (± 3.06) | 7.14 (± 0.62) |
| DIV2K test ⇓ 16 crop | 100 | 41.32 (± 19.76) | 10849 | 14.20 (± 3.49) | 6.88 (± 0.82) |

Table 1. Main characteristics of the SR datasets. We report average and standard deviation.

tails / high frequencies), therefore an ideal IQA measure should use and reflect a fidelity to the ground truth when available. In practice, however, often the ground truth is not available and, therefore, plausible and perceptually qualitative super-resolved images are desirable as long as the information from the LR image is preserved. For perceptual image super-resolution studies we refer to [53, 26, 29].

Most of the recent SR literature deploys and validates models on either the luminance component of the image (Y channel from YCbCr color space) or on the full RGB image. The texture is captured by the luminance component while the chroma is less important since the human eye is less sensitive to the changes from the chroma components. Typically the models working on Y channel reconstruct the color super-resolved image by processing the Y channel and simply upscaling the chroma Cb and Cr channels through interpolation (such as bicubic) to then convert the result to RGB color space.

## 5. Datasets

In this study we use the most common datasets from SR literature (shown in Fig. 5). We mention also LIVE1 [41, 53], L20 [49, 27], ImageNet [38], Kodak [7] or SuperTex136 [8] that are less popular for single-image SR.

**Train91** was proposed by Yang *et al.* [54] for training. It has 91 RGB images with mainly small sized flower images.

**Set5** was used in [5] and adopted under the name 'Set5' in [47]. It contains five popular images: one medium size image ('baby', 512 × 512) and four smaller ones ('bird', 'butterfly','head', 'women').

**Set14** was proposed by Zeyde *et al.* [56]. It contains 14 commonly used images in the image processing literature. The images in Set14 are larger and more diverse than those in Set5.

**B100** represents the set of 100 testing images from the Berkeley Segmentation Dataset [31] as adopted in [48]. It covers a large variety of real-life scenes.

**Urban100** was introduced by Huang *et al.* [19]. It consists from 100 clean from urban environments with repetitive patterns and high self-similarity.

**DIV2K** is our proposed dataset as introduced in Section 2 and is used for the NTIRE 2017 SR Challenge.

In Table 1 we summarize main characteristics of the SR datasets. According to the perceptual image quality assessed by CORNIA all the datasets have good image quality, Urban100 and DIV2K being at the top. The average image size (pixels per image or ppi) varies from 113491

[7] http://r0k.us/graphics/kodak/

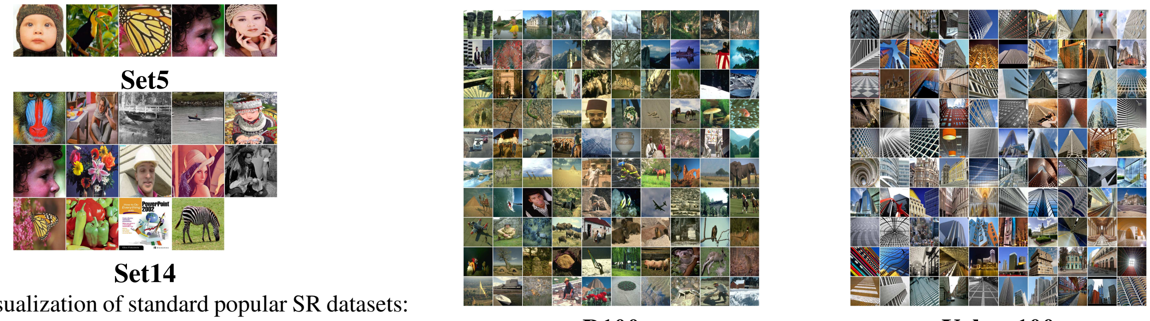



Figure 5. Visualization of standard popular SR datasets: Set5, Set14, B100, and Urban100.

pixels for Set5 to 2.8million pixels for DIV2K. DIV2K images are about 4 times larger than those from Urban100. In terms of entropy computed over the grayscale images the datasets are comparable, but Set14 has the lowest entropy. Also, the datasets are comparable in terms of bits per pixel (bpp) required by PNG for lossless compressing the images. Both bpp and image entropy are indicators of the amount of information present in the image per image pixel. Since CORNIA score, bpp, and entropy are comparable for all the datasets, the differences are made by the number and the resolution of the images. Thus, as intended, DIV2K has the best diversity in semantic contents (similar with B100) and the highest resolution ($\sim 4\times$ more than Urban100).

## 6. Methods

In this study we use the top methods from NTIRE 2017 SR Challenge [46], as well as several representative methods from the recent literature.

### 6.1. NTIRE 2017 SR challenge methods [46]

**SNU_CVLab1** of Lim *et al.* [28] is the winner of the NTIRE 2017 challenge. It builds onto SRResNet architecture [26]. The building block removes the batch normalization layers from the residual block (ResBlock) in [18] and a residual scaling layer (constant multiplication with 0.1) is added after the 2nd convolutional layer of the block. The method has 36 ResBlocks end-to-end trained on the DIV2K train data, and for the Track 2 of the challenge used also crawled Flickr images to generate additional train data besides the DIV2K train data.

**SNU_CVLab2** is a compactly designed approach meant for HR estimation at multiple scales simultaneously. Most of the implementation and architecture design is shared with the SNU_CVLab1 single-scale solution. SNU_CVLab2 trades the brute performance of SNU_CVLab1 for increased efficiency at train and runtime.

**HelloSR** is a winner of NTIRE 2017 challenge based on a stacked residual-refined network design.

**Lab402** proposed by Bae *et al.* [2] is the third winner of NTIRE 2017 challenge. Lab402 solution consists from a 41 layers Haar wavelet residual network.

**WSDSR** is a very recent self-similarity based method based on BM3D and a newly proposed Wiener filter. It was proposed by Cruz *et al.* [7].

### 6.2. Other representative methods

We use a selection of methods from [48, 49, 33], the seminal CNN methods [9, 21, 26], and additionally a self-similarity based method [19].

**Bicubic interpolation** is probably the most employed technique for interpolation of images in practice and often a basic component in more involved SR methods (such as A+ [48] and VDSR [21]). Each pixel in the upscaled image is a bicubic interpolation over a support LR patch of pixels.

**Yang** of Yang *et al.* [54] employs sparse coding and sparse dictionaries for compact modeling of the LR-HR train examples and sharp HR reconstruction.

**Zeyde** method of Zeyde *et al.* [56] builds upon Yang and efficiently learns sparse dictionaries using K-SVD [1] and Orthogonal Matching Pursuit for sparse solutions.

**ANR** (Anchored Neighborhood Regression) of Timofte *et al.* [47] relaxes the sparse decomposition from Yang and Zeyde to a ridge regression solved offline and stored per each dictionary anchor for large speed benefits.

**A+** or Adjusted ANR of Timofte *et al.* [48] improves over ANR by learning regressors from all the training patches in the local neighborhood of the anchor.

**IA** is the Improved A+ method [49] which uses a couple of proposed techniques such as: data augmentation, enhanced prediction, cascading, hierarchical search with larger dictionaries, and context reasoning.

**SRCNN** of Dong *et al.* [9, 10] directly learns to map patches from LR to HR images with a deep CNN model [25].

**VDSR** is a VGG16 architecture [43] based CNN model proposed by Kim *et al.* [21]. In comparison with SRCNN it goes ‘very deep’ with the convolutional layers and significantly boosts the achieved performance.

**PSyCo** proposed by Perez *et al.* [33] builds upon A+ framework and efficiently reduces the manifold span to better use the train data and improve the model capacity.

**SRResNet** is a ResNet architecture [18] based CNN model proposed by Ledig *et al*. [26] which goes deeper than VDSR for better performance.

**SelfEx** is a self-similarity based method introduced by Huang *et al*. [19] optimizing over the LR image.

# 7. Experimental results

## 7.1. Scaling factors to benchmark

Most recent SR works validate their methods for 3 downscaling factors. The Pearson correlation is above 0.97 ($\rho > 0.97$) for the PSNR scores on the DIV2K test data of any two competitions (scaling factors) of Track 1 bicubic of the NTIRE 2017 SR Challenge [46]. The SSIM scores reported for $\times 2$ and $\times 3$ have $\rho = 0.99$, while between $\times 2$ and $\times 4$ $\rho = 0.88$. It is clear that validating on all these scaling factors for the same degradation operator (bicubic downscaling) is redundant and perhaps the efforts should be placed elsewhere, on better IQA measures and more challenging SR setups. The PSNR differences between the results of different challenge methods are larger for the smallest scale factor $\times 2$, while for SSIM the differences get larger for higher scales. The higher scales are the more challenging ones and the perceptual differences are also potentially larger between different methods. If we analyze the reported results on the Set5, Set14, B100, Urban100 we come to the same conclusion that for low scaling factors ($\times 2$) both PSNR and SSIM scores of the recent methods are rather large and difficult to assess by the human perception. Therefore we recommend to work with $\times 4$ and to push further to $\times 8$ and above for extreme SR validation and benchmarking. Already a couple of SR works [53, 26] report on $\times 8$ settings.

In the unknown downscaling setting (Track 2 challenge) the degradation operators are different for each scaling setup and both the achieved performances of the SR methods and their correlations are lower. Noteworthy is that most of the top methods from Track 1 generalize well and deliver consistent performances also on Track 2 while keeping their relative ranking.

## 7.2. Ensembles for enhanced prediction

The top ranked entries SNU_CVLab1, SNU_CVLab2, HelloSR, UIUC-IFP and 'I hate mosaic' in the NTIRE 2017 SR Challenge use ensembles of HR predictions. They are flipping and/or rotating by 90° the LR input image then process them to achieve HR corresponding results and align these results back for the final HR average result. This enhanced prediction is one of the seven ways to improve SR described in [49]. In our experiments the top methods without the enhanced prediction achieve 0.1 to 0.25dB lower PSNR results on DIV2K test data. Since the unknown downscaling depends on the image orientation, SNU_CVLab1, SNU_CVLab2, and HelloSR propose to train different models eventually with different losses and to average their predictions. In this case the improvements are only marginal on the DIV2K dataset.

| Method | computed on Y from YCbCr | | | | computed on RGB | | | |
|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | IFC | CORNIA | PSNR | SSIM | IFC | CORNIA |
| SNU_CVLab1 | 30.56 | 0.853 | 3.39 | 26.4 | 29.09 | 0.837 | 3.41 | 26.2 |
| SNU_CVLab2 | 30.51 | 0.852 | 3.38 | 27.1 | 29.04 | 0.836 | 3.40 | 27.1 |
| Lab402 | 30.30 | 0.846 | 3.28 | 28.9 | 28.83 | 0.830 | 3.30 | 28.8 |
| HelloSR | 30.27 | 0.846 | 3.26 | 28.2 | 28.80 | 0.830 | 3.28 | 28.1 |
| VDSR[21] | 29.58 | 0.828 | 2.91 | 35.2 | 27.98 | 0.808 | 2.93 | 35.3 |
| A+[48] | 29.15 | 0.815 | 3.05 | 46.3 | 27.58 | 0.793 | 2.86 | 46.5 |
| Bicubic | 28.12 | 0.782 | 2.39 | 65.5 | 26.64 | 0.760 | 2.27 | 65.3 |

Table 2. Quantitative results on the test set of DIV2K with $\times 4$. More results are available at `https://www.vision.ee.ethz.ch/ntire17/SR_challenge/`

## 7.3. IQA scores on luma (Y) vs. on RGB

Most of the recent SR literature report performance scores on the luma (intensity) component of the image, typically Y from YCbCr color space. The human visual system is much more sensitive to the image texture as represented by the luma component than to the chroma components. Moreover, the luma component captures most of the image high frequencies / rich details while the chroma generally does not. The consequence is that bicubic downscaling of an image with a typical 2 or 4 factor leads to a dramatic loss of high frequencies (rich details) while the low frequencies are usually preserved and thus the luma component amount for most of the lost information.

NTIRE 2017 SR Challenge works on RGB images and uses the RGB color space for evaluation. We computed also the performance results on the luma component Y from the YCbCr color space after converting the RGB images. Some results are in Table 2. There is a very strong correlation (Pearson correlation $\rho > 0.96$) between the PSNR results computed on Y luma and those on RGB for the NTIRE 2017 SR Challenge methods on each DIV2K validation and testing datasets. However, the correlation is weaker for SSIM which is closer correlated to the human perception. We conclude that if for reconstruction fidelity (measured by MSE or PSNR) reporting on Y is comparable with reporting on RGB, whenever the perceptual quality is targeted (measured by full-reference SSIM, IFC or no-reference CORNIA) for color images the RGB space could be better suited.

## 7.4. IQA measures for SR

As previously mentioned our selection of IQA measures is motivated by reconstruction fidelity (absolute errors) and prior studies on the most robust automatic measures for perceptual image quality [34, 55, 53, 29]. From the considered IQA measures CORNIA is the best at perceptual IQA being followed by IFC, SSIM, PSNR, MSE, in this order, as shown in [53, 29] for the SR task. Some results are in Table 2.

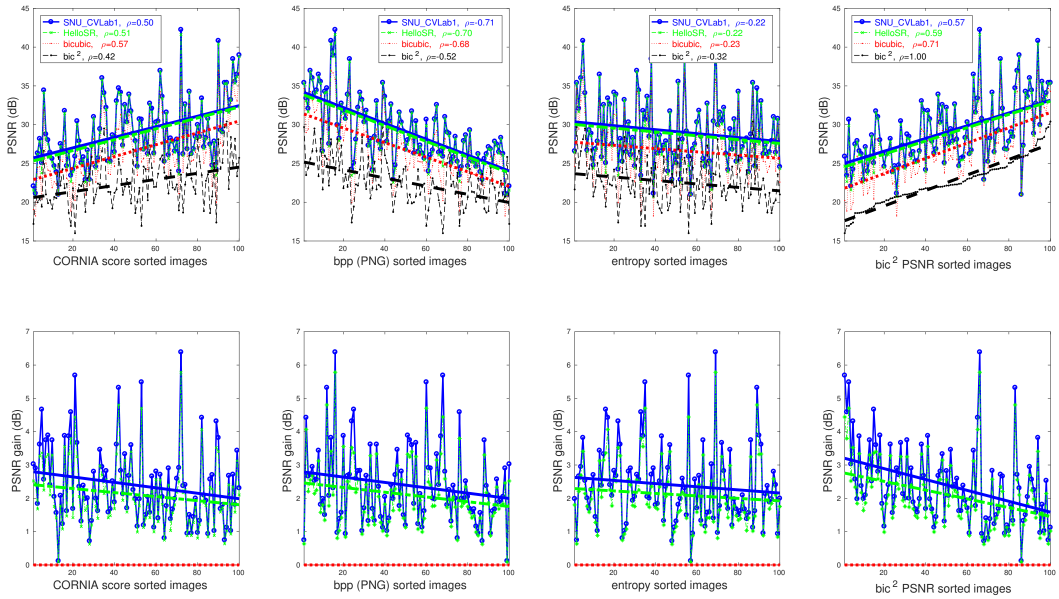



Figure 6. PSNR performance on DIV2K (bicubic downscaling $\times 4$) can be predicted by the image quality (CORNIA score), PNG compression rate (bpp), image entropy, or the PSNR of the bicubic downscaled and upscaled LR image (bic$^2$). For each predictor we sort the images and compute the Pearson correlation ($\rho$) with the PSNR results of the methods: SNU_CVLab1, HelloSR, bicubic and bic$^2$. PSNR gains over bicubic are reported on the bottom row.


Figure 7. Bicubic downscaling versus centered cropping as seen for a central image patch of fixed size.

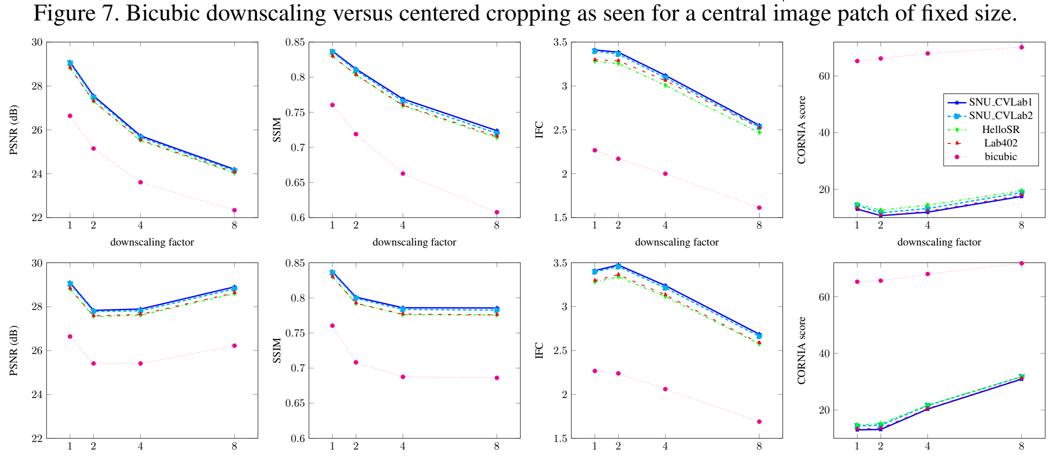



Figure 8. Reducing the ppi of the DIV2K test images by bicubic downscaling or by centered cropping has little influence on the relative performance of the winners of NTIRE 2017 SR Challenge.

### 7.5. Runtime and performance

Fig. 4 shows the trade-off between runtime and performance for SR methods. The runtimes of the non-CNN models are reported on CPU for Matlab implementations. RAISR [37] is reportedly the fastest from the anchored family of methods (ANR, A+, IA, PSyCo) with CPU runtimes comparable to the fastest GPU methods from our study, but its PSNR performance is below A+. At the other extreme

WSDSR [7] is the slowest and the best performing among the non-CNN methods on bicubic downscaling settings. The best performing methods are also among the slowest ones. Fig. 1 reports a timeline performance ($\times 4$, bicubic downscaling, PSNR on Y channel) on five datasets for representative top methods. NTIRE 2017 challenge methods largely improve over prior art.

### 7.6. Predictors for SR performance

How much a SR method can restore / improve a LR image? We propose and study a couple of predictors for SR performance in the absence of the ground truth.

**CORNIA** is a top no-reference perceptual IQA.

**bpp** or bit per pixel is the standard measure for image compression. If we retain all the details in the image (lossless compression) then bpp is a strong indicator on the quantity of image information (takes into account also pixel contexts) and thus relevant to the SR task. We use the bpp achieved by the PNG lossless compression.

**Image entropy** is a standard a measure of the image information. It does not take into consideration the spacial correlations from the 2D image but treats each pixel as a separate observation. To correct this a couple of image information-spacial entropy measures were proposed [24, 35]. However, entropy is usually a good predictor for the image information and the lossless compression rates achievable by current methods. Another good image entropy can be computed based on the differences between adjacent pixels.[8]

**bic**$^2$ is our proposed simple predictor based on the PSNR achieved by the bicubic interpolation (Matlab imresize function) of the bicubic downscaling of the LR image by the intended SR magnification factor. If the bicubic interpolation performs well on the further downscaled LR image then we expect the same to happen for the LR image. Note that this should happen also for other SR methods, as the downscaled LR image is highly correlated in content with the LR image. Liang *et al*. [27] apply the same procedure for very deep CNN models to predict their performance on LR images.

**Best performance predictors** In Fig. 6 we report the PSNR and PSNR gain over bicubic (computed on RGB color space) for the top SNU_CVLab1 and HelloSR methods for the 100 DIV2K test images sorted according to the above listed predictors, as well as the linear regression of the PSNR scores and the Pearson correlation ($\rho$) between the achieved PSNR scores and the performance predictors. For reference we add bicubic interpolation and bic$^2$ scores. Entropy is the poorest predictor ($\rho = -0.22$), somehow expected since it discards the relations between the (neighboring) pixels. CORNIA scores correlate well ($\rho = 0.5$), bic$^2$ correlates better ($\rho = 0.59$), while bpp (PNG) has the strongest correlation ($\rho = -0.70$) with the achieved PSNR performance by the SR methods. Noteworthy is that bic$^2$ is the best predictor for the performance of the bicubic interpolation. A more thorough study on predictors for the performance of SR methods is necessary, we conjecture that such predictors could be used for fusion [52, 45] or SR method selection to super-resolve LR images [27].

Another observation is due: if the best predictor (bpp) without knowledge of the ground truth HR image gets $\rho = 0.70$ with the SR methods, the Pearson correlation of the PSNR achieved results by bicubic interpolation and SNU_CVLab1 reaches $\rho = 0.962$, while the SNU_CVLab1 and HelloSR are heavily correlated in performance with $\rho = 0.999$. Therefore, the known quantitative results of a SR method are reliable predictors on the performance gains of other SR methods.

### 7.7. Image resolution vs. performance

How critical for a SR evaluation dataset is the image resolution in terms of ppi? To answer this we conduct a couple of experiments by reducing the HR images of the DIV2K test and then evaluating the best NTIRE 2017 challenge methods. In Fig. 7 we depict the two ways we use to reduce the ppi of the images: (i) by bicubic downscaling with a factor and thus preserving mainly much of the low frequencies and (ii) by centered cropping of a subimage corresponding to a downscaling factor. In Table 1 we show the main characteristics of the DIV2K test datasets derived for the 2, 4, 8, and 16 downscaling factors. In Fig. 8 we report the IQA scores of the selected methods for $\times 4$ magnification versus the factors used to reduce the images of the DIV2K test dataset. Surprisingly, there is little to no effect of the test image resolution (ppi) on the relative performance achieved by the methods, the ranking is preserved and this regardless the IQA measures used. We note also the clear drop in performance when using bicubic downscaling and the relatively smaller drop when using the centered cropping strategy (or uniform sampling). As our experiments shows, given sufficiently large number of diverse images in the dataset, the image resolution is, in comparison, less important for benchmarking SR methods. We could easily use DIV2K $\Downarrow 8$ and still achieve meaningful benchmarking results.

## 8. Conclusions

In this paper we introduced DIV2K, a new dataset for super-resolution benchmarking, studied the winning solutions from the NTIRE 2017 SR challenge in comparison with representative methods from the recent literature, and investigated topics of interest such as predictors for super-resolution performance, image resolution and quality, image quality assessment measures, magnification factors for benchmarking and challenges.

[8] http://www.astro.cornell.edu/research/projects/compression/entropy.html

## Acknowledgements

We thank the NTIRE 2017 sponsors: NVIDIA Corp., SenseTime Group Ltd., Twitter Inc., Google Inc., and ETH Zurich.

## References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11), November 2006.

[2] W. Bae, J. Yoo, and J. C. Ye. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. *arXiv preprint arXiv:1611.06345 (submitted to NTIRE 2017)*, 2016.

[3] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

[4] M. R. Banham and A. K. Katsaggelos. Digital image restoration. *IEEE signal processing magazine*, 14(2):24–41, 1997.

[5] M. Bevilacqua, A. Roumy, C. Guillemot, and M. line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 135.1–135.10. BMVA Press, 2012.

[6] H. Chang, D.-Y. Yeung, and Y. Xiong. Super-resolution through neighbor embedding. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2004.

[7] C. Cruz, R. Mehta, V. Katkovnik, and K. Egiazarian. Single image super-resolution based on wiener filter in similarity domain. *arXiv preprint arXiv:1704.04126*, 2017.

[8] D. Dai, R. Timofte, and L. Van Gool. Jointly optimized regressors for image super-resolution. In *Computer Graphics Forum*, volume 34, pages 95–104, 2015.

[9] C. Dong, C. C. Loy, K. He, and X. Tang. *Learning a Deep Convolutional Network for Image Super-Resolution*, pages 184–199. Springer International Publishing, Cham, 2014.

[10] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, Feb 2016.

[11] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*, pages 391–407. Springer, 2016.

[12] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

[13] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997.

[14] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004.

[15] M. A. Figueiredo and R. D. Nowak. An em algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.

[16] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.

[17] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 349–356. IEEE, 2009.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[20] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991.

[21] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[22] J. Kim, J. Kwon Lee, and K. Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1645, 2016.

[23] 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*, 32(6):1127–1133, 2010.

[24] K. G. Larkin. Reflections on shannon information: In search of a natural information-entropy for images. *CoRR*, abs/1609.01117, 2016.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[26] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.

[27] Y. Liang, R. Timofte, J. Wang, Y. Gong, and N. Zheng. Single image super resolution-when model adaptation matters. *arXiv preprint arXiv:1703.10889*, 2017.

[28] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[29] 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*, 158:1 – 16, 2017.

[30] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on image processing*, 17(1):53–69, 2008.

[31] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological

statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 416–423. IEEE, 2001.

[32] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

[33] E. Perez-Pellitero, J. Salvador, J. Ruiz-Hidalgo, and B. Rosenhahn. Psyco: Manifold span reduction for super resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57 – 77, 2015.

[35] Q. Razlighi and N. Kehtarnavaz. A comparison study of image spatial entropy. In *IS&T/SPIE Electronic Imaging*, pages 72571X–72571X. International Society for Optics and Photonics, 2009.

[36] W. H. Richardson. Bayesian-based iterative method of image restoration. *JOSA*, 62(1):55–59, 1972.

[37] Y. Romano, J. Isidoro, and P. Milanfar. Raisr: Rapid and accurate image super resolution. *IEEE Transactions on Computational Imaging*, 3(1):110–125, March 2017.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[39] S. Schulter, C. Leistner, and H. Bischof. Fast and accurate image upscaling with super-resolution forests. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[40] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128, 2005.

[41] 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*, 15(11):3440–3451, 2006.

[42] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[44] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2015.

[45] R. Timofte. Anchored fusion for image restoration. In *23rd International Conference on Pattern Recognition (ICPR)*, 2016.

[46] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[47] R. Timofte, V. De Smet, and L. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[48] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In D. Cremers, I. Reid, H. Saito, and M.-H. Yang, editors, *Computer Vision – ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV*, pages 111–126, Cham, 2014. Springer International Publishing.

[49] R. Timofte, R. Rothe, and L. Van Gool. Seven ways to improve example-based single image super resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[50] 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*, 13(4):600–612, April 2004.

[51] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[52] J. Wu, R. Timofte, and L. Van Gool. Generic 3d convolutional fusion for image restoration. In *ACCV Workshops*, 2016.

[53] C.-Y. Yang, C. Ma, and M.-H. Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision*, pages 372–386. Springer, 2014.

[54] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[55] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1098–1105. IEEE, 2012.

[56] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24 - 30, 2010, Revised Selected Papers*, pages 711–730, 2012.

[57] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.