BSRAW: Improving Blind RAW Image Super-Resolution

Marcos V. Conde, Florin Vasluianu, Radu Timofte
Computer Vision Lab, CAIDAS & IFI, University of Würzburg
{marcos.conde, radu.timofte}@uni-wuerzburg.de
https://github.com/mv-lab/AISP

Abstract

In smartphones and compact cameras, the Image Signal Processor (ISP) transforms the RAW sensor image into a human-readable sRGB image. Most popular super-resolution methods depart from a sRGB image and upscale it further, improving its quality. However, modeling the degradations in the sRGB domain is complicated because of the non-linear ISP transformations. Despite this known issue, only a few methods work directly with RAW images and tackle real-world sensor degradations.

We tackle blind image super-resolution in the RAW domain. We design a realistic degradation pipeline tailored specifically for training models with raw sensor data. Our approach considers sensor noise, defocus, exposure, and other common issues. Our BSRAW models trained with our pipeline can upscale real-scene RAW images and improve their quality. As part of this effort, we also present a new DSLM dataset and benchmark for this task.

1. Introduction

The specialized on-board Image Signal Processor (ISP) transforms RAW sensor readings into more refined and perceptually meaningful RGB representations [20,45,52,68]. Photographers often opt for RAW over RGB processing to yield perceptually superior images, attributing this preference to two key benefits of RAW data: (i) RAW information carries a broader data range, generally ranging from 12-14 bits, compared to the 8-bit RGB output from the ISP. (ii) RAW data linearly correlates with scene radiance, which simplifies the correction of degradations such as noise. Conversely, ISP operations are non-linear and undergo an irrevocable loss of information across various stages [6,27], complicating subsequent image restoration [6,54]. For these considerations, RAW imagery is a more advantageous starting point than RGB for a multitude of low-level vision applications like image denoising [1,6,36,38,59,62], color balance [21], deblurring [30], exposure adjustment [11,19,69], and image super-resolution [42,54,60,68].

Nevertheless, only a small fraction of low-vision research directly engages with RAW data, mainly due to the greater abundance and accessibility of general-purpose sRGB images. Consequently, the most popular deep learning architectures for image restoration predominantly operate on RGB images [13,51,61,62]. The majority of cutting-edge Single Image Super-Resolution (SISR) techniques [4,15,26,31,63] rely on deep convolutional networks or Transformer [47] architectures, and operate in the RGB color space. Yet, these approaches have notable drawbacks. Primarily, they are often trained on artificially generated data, leading to poor generalization on real-world scenes. Secondly, modeling accurately the degradations in RGBs is challenging due to the non-linear transformations and information loss incurred during ISP processing (RAW to RGB).

The classical single-image super-resolution model [17,54,66] is formulated as:

\[ y = (x \otimes k) \downarrow s + n \]  

(1)

It assumes the observed - or captured- low resolution (LR) image \( y \) is obtained from an underlying high resolution (HR) image \( x \) by applying a degradation kernel (or PSF) \( k \) [16], followed by a downsampling operation \( \downarrow s \) with scale factor \( s \) (e.g. Bicubic [46,57]), and the addition of the noise \( n \) [18]. The more realistic these factors are modelled and applied, the better the restoration and SISR models perform and generalize in real-world scenarios [54,66].

This introduces the need for more complex degradation pipelines, to reduce the generalization gap observed in the real world domain. Recently, Zhang et al. [66] designed a practical degradation pipeline for the RGB domain SISR, achieving promising results in real scene super-resolution.

RAWSR from Xu et al. [54,55] is one of the first works that apply a simple degradation pipeline (Eq. 1) for synthesizing low-resolution RAW images and train SISR models that successfully generalize on real scenes. Their approach is illustrated in Figure 1c, ultimately they aim at enhancing RGB images by leveraging both the sensor RAW data and the RGB produced by the ISP. For this reason, their studies are performed on the resultant RGBs, and their
In this work we study **RAW SISR** -see Figure 1d-, using exclusively RAW sensor data, thus, the signal depends only on the scene and the imaging system (*e.g.* light sources, user-camera interaction). Moreover, the degradations such as noise [1, 6] are not altered by the ISP stages [14, 27]. We build a realistic RAW degradation pipeline that extends and improves previous methods from Zhang *et al.* [66] and Xu *et al.* [54, 55]. Our baseline model, BSRAW, is an efficient blind RAW SISR method, trained using our degradation pipeline, and real data from DSLR and DSLM cameras. The *main contributions* of this paper are:

1. A controllable degradation pipeline to train deep neural models using RAW images and considering realistic downsampling, noise and blur.


3. We study BSRAW and other methods for this task, and we provide a new benchmark for RAW SISR.

### 2. Related Work

#### Image Signal Processing

Digital camera systems incorporate a specialized in-camera Image Signal Processor (ISP) [14, 20, 27, 45, 52], which converts the captured RAW sensor data into images that are both less deteriorated and visually appealing to human perception.

In the recent years many deep learning approaches have been proposed to learn the ISP transformation [12, 34]. Some popular methods in this direction are CycleISP [62], CameraNET [32], and PyNET [25].

In this work, we use ISP pipelines [27] with known parameters and operations, however, our proposed method can help to improve learned ISPs for real scene processing.

#### Degradation Models

Creating authentic degradation models for synthesizing data is a critical endeavor in the field of low-level vision. Standard degradation models frequently adopt a sequential approach involving blur, downsampling, and noise injection (as described in Eq. 1). Since most degradation models are far from the real-world complexity, some works aimed at solving such problem by designing complex degradation pipelines [7, 35, 54, 66].

Zhang *et al.* [64, 66] devised a pragmatic degradation methodology specifically tailored for training sRGB image restoration models, achieving encouraging outcomes in Single Image Super-Resolution (SISR). Extensions of this approach to video denoising have also been reported [10].

Xu *et al.* [54, 55] implement the degradation equation (Eq. 1) to synthesize low-resolution RAW images and then...
train models that exhibit strong generalization capabilities in real-world scenarios. In their work, they use heteroscedastic Gaussian noise, varying disk kernels for defocus blur, and minimal motion blur.

Little research exists on elaborating more intricate degradation phenomena, apart from noise, in the RAW image space [1, 18, 30]. In response, we introduce improvements such as enhanced noise profiles, anisotropic defocusing, motion blur variation, exposure inconsistencies, and compression artifacts. Details of our enhanced degradation model are elaborated in Section 3.

2.1. Image Super-Resolution and Restoration

In the recent years blind image restoration and super-resolution using deep learning has become a popular task [13, 15, 31, 49, 65, 66, 71], this is because blind methods do not require prior knowledge about the degradation factors or the camera sensor, therefore they are general purpose models i.e. practically, such methods can process any degraded RGB image. SwinIR [31] is an example of general purpose SISR and restoration method.

The vast majority of blind restoration and SISR methods operate in the RGB color space. We also focus on the blind super-resolution problem, yet, in the RAW domain.

There are far fewer works that solve complex inverse problems in the RAW domain. Most methods focus on RAW image denoising, since estimating and removing the noise on linear raw data is well-studied [1, 6, 12, 38]. SIDD [1] and DND [40] are the most popular RAW denoising benchmarks. These benchmarks show that state-of-the-art methods can achieve high reconstruction (> 50dB PSNR) given a noisy RAW image, indicating that new challenging data should be used. Liang et al. studied RAW image deblurring on a simple setup using images from one DSLR camera [30]. We also find works focused on exposure correction for low-light image enhancement [11, 12, 19, 23]. Xu et al. introduced RAWSR [54, 55] for super-resolution and enhancement using both RAW and RGB images as inputs.

We aim to study the RAW SISR problem assuming (i) unknown real-world degradations, and (ii) cross-device generalization i.e. our method can be applied to images from different sensors, as the most popular RGB restoration models [13, 31, 61].

3. Controllable RAW Degradation Pipeline

We improve previous degradation models to synthesize realistic degraded RAW images [54] for training SR models. We consider the most common limiting factors related to the acquisition and the processing of high quality RAW images: (i) noise and low resolution — related to the size of the sensor [27]. (ii) exposure. (iii) motion and defocus blur – related to stabilization issues. Note that previous methods [54, 55] did not consider multiple noise profiles, or complex blur and PSFs. We provide samples from our dataset in Figure 2.

3.1. Noise

Noise (n) is an omnipresent element across all stages of image capture and processing [1, 18]. Its occurrence is influenced by a multitude of variables, such as exposure duration and lighting conditions. In the domain of photography, noise removal —or denoising— commonly takes place in the RAW stage, where its linear characteristics have been thoroughly investigated [1, 6, 18, 19, 37]. This makes it more manageable to address prior to the non-linear transformations introduced by the ISP [27]. Most methods use Homoscedastic Gaussian noise [65–67] in their degradation models. Xu et al. [54] use sharper Heteroscedastic Gaussian distribution [1, 40].

We adopt a more practical shot-read noise [6, 18, 70]. In Equation 2, we can observe the intensity y as a sample of a Gaussian distribution having as mean the input signal x and the variance depending on the λs (shot) and λr (read) parameters [6]. This is derived from a Poisson-Gaussian noise model [50]:

\[ y \sim \mathcal{N}(\mu = x, \sigma^2 = \lambda_r + \lambda_s x) \]  

Utilizing Zhang et al. methodology [70], we derive noise profiles specific to our imaging devices, in addition to adopting established profiles from SIDD [1, 2] and DND [40] for smartphone and DSLR sensors, respectively. This diverges from earlier works that relied on a singular noise distribution [6, 54]. Our approach incorporates varied real-world noise profiles, which manifest stochastically. We also emulate the interaction effects (e.g., attenuation, amplification) between introduced noise and other image degradations (e.g., blur).

3.2. Blur

Blur (B_k) is also one of the most common degradations appearing in image capturing e.g., camera shake in hand-held photography, motion blur, and defocus blur [3, 22, 58]. Capturing aligned blurry-clean pairs in real scenarios is extremely difficult, for this reason, the most popular deblurring datasets are synthetic. Most approaches adopt a uniform blur by convolving the image with iso/anisotropic Gaussian kernels [66]. Xu et al. [54, 55] implemented defocus blur as a disk kernel, and a modest motion blur [44].

We create our blur degradation by convolving the image with a diverse pool of kernels: classical isotropic and anisotropic gaussian blur kernels [29, 66], real estimated motion blur kernels [39, 43, 56], and real estimated PSFs (point-spread-function) from [53, 63, 72].
This task is fundamental since the current trend in reducing size and weight observed for the most popular DSLR or mirrorless cameras, and for the smartphone cameras, makes it challenging to implement sensor stabilization systems handling extreme levels of sensor instability during point-and-shoot acquisition.

3.3. Exposure

Exposure levels (EE) can introduce undesirable artifacts in photographic images. Both underexposure and overexposure can result in signal loss or increased noise in the imagery [11, 12, 19]. While our inclusion of this degradation isn’t intended to address HDR issues [19], it does allow the model to acquire exposure correction skills. Additionally, the interplay between simulated exposure levels and noise augments the complexity of the degradation environment. We adopt a low exposure model inspired by the work of Punnappurath et al. [41], employing a linear ISO scaling technique to generate the resultant darkened image. It is worth noting that our real-world testing dataset comprises images with exposure-related challenges, along with their respective, carefully-selected ground truth.

3.4. Downsampling

One of the most important differences between cameras is the size of the sensor [14, 25, 27]. Photographers can use telephoto lenses to obtain HR images, but the resolution of the scene is limited by the size of the sensor. Thus, it is more common to capture a wide-range scene as a LR image using a wide-angle lens, and then apply SISR [54, 55]. Considering this, it is desirable to be able to upscale RAW images [54, 55]. However, this is not a trivial task -as for RGB images- since most traditional methods do not preserve the Bayer pattern (RGGB).

Since obtaining real LR-HR pairs is extremely expensive and time-consuming, we follow [54, 66] to synthesize LR RAW images from the assumed HR real captures (e.g. 60MP RAWs). We downsample the high-quality RAW images so that each pixel could have its ground truth red, green and blue values. We denote this operation as ↓s.

3.5. Complete RAW Degradation Model

The degradation pipeline covers most of the factors affecting the quality of RAW images. One important aspect is the noise sampling from multiple noise profiles, as well as the PSF sampling (and even application of two PSFs). The order of the degradations is defined by the optics theory [17] i.e. first we apply the kernel (k) to the RAW signal, then we simulate a small sensor with downsampling and noise injection -that can be influenced by the exposure- and later we apply the other degradations including additional motion blur and downsampling.

We define four different degradation levels in our experiments: 1) classical image restoration [17, 31]. Only blur
and noise are considered. II) SISR degradation model (see Eq. 1) III) SISR considering additional exposure degradations IV) Finally we apply our complete realistic degradation pipeline, including multiple noise profiles, the application of multiple kernels, exposure, and downsampling. We provide more details about the pipeline in the appendix.

### 4. Experiments

We conduct several experiments to prove the benefits of our approach for RAW image restoration and SISR. We propose BSRAW as an efficient baseline method for this task. Due to the pages limit, we provide additional implementation details, dataset information, and results in the supplementary material.

#### 4.1. BSRAW Dataset

Our goal is to solve blind super-resolution for RAW images, and apart from a realistic degradation prior (Sec. 3), an image prior also contributes to the success of the deep model [66]. Our sources of real RAW data and our train/test split are summarized in Table 1. The quality of the DSLR RAW images is substantially better than in smartphones due to the larger optics and sensor, this implies that the RAW images show less noise and minimal optical aberrations.

Following previous work [52, 54, 55], we use images from the Adobe MIT5K dataset [8], which includes two DSLR cameras. The DSLR images are manually filtered to ensure diversity and natural properties (i.e. remove extremely dark or overexposed images), we also remove the blurry images (i.e. we only consider all-in-focus images).

The pre-processing is as follows: (i) we normalize all RAW images depending on their black level and bit-depth. (ii) we convert (“pack”) the images into the well-known RGGG Bayer pattern (4-channels), which allows to apply the transformations and degradations without damaging the original color pattern information [33]. (iii) Following previous work on RAW processing and learned ISPs [24, 54], we train using image patches of dimension $248 \times 248 \times 4$.

### Synthetic Dataset

We apply our complete degradation pipeline (see Sec. 3.5) to the training images to generate aligned degraded-clean pairs. The synthetic test dataset is generated by applying our degradation pipeline, at different levels, to the corresponding test images – see Table 1.

#### Real-World Captures

We believe the models trained using our pipeline have notable generalization capabilities (also concluded in [54, 66]), therefore, we collect a novel dataset with challenging real scenes. As indicated in Table 1, we capture images using two mirrorless full-frame cameras: Sony α7R4 (60MP) and Nikon Z7 (45.7MP). In Figure 3 we show samples from our data capturing setup.

This is a paired dataset consisting of variations over 100 real-world indoor and outdoor scenes. All the devices used for data capturing allow for manual operation of the camera ISP parameters and optics, and are fixed on a tripod. For each “clean” reference image, we capture several variations, where each variation represents a degradation with a different degree of intensity. The scene variants represented there are degraded in terms of defocusing blur, exposure time, high ISO noise, and motion blur. Since interaction with the camera device was needed during the acquisition, we perform a post-processing alignment (similar to [5]) to further improve the quality and usability of the data.

---

**Table 1. Dataset split.** We use diverse DSLR images from the MIT5K as previous works [8, 54] (Canon EOS 5D and Nikon D700). The sensors highlighted in gray correspond to our new dataset of real-world captures. The train/test split ensures no overlap and proper evaluation.

<table>
<thead>
<tr>
<th>Device</th>
<th>Sensor</th>
<th># Train</th>
<th># Test</th>
<th>Res.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSLRs [8, 54]</td>
<td>Diverse</td>
<td>1000</td>
<td>150</td>
<td>10MP</td>
</tr>
<tr>
<td>Sony α7R4</td>
<td>IMX551</td>
<td>90</td>
<td>10</td>
<td>2MP</td>
</tr>
<tr>
<td>Nikon Z7</td>
<td>IMX309BQJ</td>
<td>90</td>
<td>10</td>
<td>2MP</td>
</tr>
</tbody>
</table>
4.2. Experimental Results

We propose BSRAW as an efficient baseline, able to process 4K resolution images in standard GPUs (120 ms for a 4K image on a Nvidia 3090Ti 24GB GPU), without patching or tiling strategies. The architecture is illustrated in Figure 4. BSRAW combines NAFNet [13] baseline blocks with dynamic convolutions and simplified stereo channel attention in the decoder part of the model. In comparison to NAFNet [13] -already quite efficient- our method has $10 \times$ less parameters and $\approx 20\%$ less MACs. We upscale the images using an adapted pixel-shuffle head.

**SISR results** We compare with previous method RAWSR [54] for SISR. Our realistic degradation pipeline helps our efficient model BSRAW to generalize better on complex degradations. We upscale the images using BSRAW, and later process them using a basic ISP [27] e.g. Dcraw as [54] to obtain the resultant RGB. Note that the training dataset derived from MIT5K [8] from Xu et al. RAWSR [54] is not publicly available. Therefore we train using our own curated DSLR dataset (see Table 1): this, together with BSRAW being trained using our degradation pipeline, allow us to achieve notably better results. We also compare with TENet [42] (based on RDN [71]) pipeline for “Denoising (DN) → SR” for RAWs.

---

### Table 2. Comparison for RAW SISR ×2

<table>
<thead>
<tr>
<th>Methods</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepISP</td>
<td>21.71</td>
<td>0.7323</td>
<td>RAW</td>
</tr>
<tr>
<td>RDN [71]</td>
<td>29.93</td>
<td>0.7804</td>
<td>RGB</td>
</tr>
<tr>
<td>TENet [42]</td>
<td>30.20</td>
<td>0.800</td>
<td>RAW</td>
</tr>
<tr>
<td>BSRGAN [66]</td>
<td>30.01</td>
<td>0.794</td>
<td>RGB</td>
</tr>
<tr>
<td>RAWSR [54]</td>
<td>30.79</td>
<td>0.804</td>
<td>RAW</td>
</tr>
<tr>
<td>BSRAW II (Ours)</td>
<td>31.02</td>
<td>0.812</td>
<td>RAW</td>
</tr>
<tr>
<td>BSRAW IV (Ours)</td>
<td>31.40</td>
<td>0.843</td>
<td>RAW</td>
</tr>
<tr>
<td>BSRAW † (Ours)</td>
<td>31.86</td>
<td>0.860</td>
<td>RAW</td>
</tr>
</tbody>
</table>

We present the results in Table 2. We distinguish three variants of our model, the ablation study is as follows:

(i) BSRAW II is trained using a simple SISR degradation pipeline, similar to [54] -our level II-. (ii) BSRAW IV is trained using our complete degradation pipeline. We can appreciate how this helps to improve results notably. Finally, (iii) BSRAW † extends the previous variant, and we include training data from our dataset. This leads to the best results in terms of fidelity metrics.

In Table 3 we show the results from our model on the synthetic dataset at each degradation level.

### Effectiveness on Real Images

As we show in Figures 6 and Table 2 for RAW restoration and SISR, our model can generalize and provide reasonable results on real scenes. Figures 5 and 6 qualitative results show that BSRAW is able to produce noise-free sharp images with proper textures and color distributions, even using RAW images from unseen sensors such as Sony Quad sensors or SIDD phones [1].

Super-Resolution, RAW or RGB? We also compare the following two strategies for image processing:

(I) SwinIR [31] and BSRGAN [66], general-purpose models trained for real-world blind SISR on RGBs. Their input is a degraded RGB image obtained after applying an ISP to the original degraded RAW image. We use images from the SIDD dataset [1] with noise and low exposure.

(II) Our approach. BSRAW takes directly the degraded RAW image, applies restoration and upsampling, then we apply the ISP with known parameters and operations.

The ISP [27] is fixed for a fair comparison. The images are real-world captures from “unknown” sensors to the model.
In Figure 6 we compare these two possible variants of super-resolution. The quantitative results form Table 2, and qualitative results (Figure 6) show the advantages and potential of our approach. We provide further analysis and samples in the supplementary material.

4.3. Implementation Details

The method was implemented in PyTorch. We use the basic $L_1$ loss between the ground-truth and the restored image. We set the minibatch size to 32. We use Adam [28] optimizer with standard hyper-parameters. The initial learning rate is set to $1e^{-4}$, we use a linear LR decay scheduler, until a minimum of $1e^{-5}$. We perform experiments using multiple NVIDIA RTX 3090Ti and 4090Ti.

The degradation pipeline can be integrated into the dataloader as an augmentation. All the degradations are applied on 4-channel RAW images (RGGB).

Note that the training dataset derived from MIT5K [8] from Xu et al. RAWSR [54] is not publicly available, and therefore cannot be used for comparison. We cannot perform a qualitative comparison because RAWSR weights and results are not publicly available i.e. the material at ¹ is obsoleted and links do not work.

In the RAW domain, we can compare the HR RAW ($1460\times2192\times4$) and the LR RAW ($365\times548\times4$) — which correspond to RGBs of resolution ($2920\times4384\times3$) and ($730\times1096\times3$) respectively. Upscaling the LR RAW using Bicubic interpolation implies a reconstruction of 27.13dB (and 0.81 SSIM) w.r.t the HR RAW.

4.4. Limitations

Downsampling RAW images is not trivial, this operation can cause color artifacts besides blurriness, and irreversible structure loss. Also, the MIT5K [8] images were capture with “old” cameras, for instance, 12MP DSLRs, in comparison with our 60MP DSLM cameras. We need further analysis using modern cameras, specially smartphones.

5. Conclusion

In this paper we discuss a controllable degradation pipeline to synthesize realistic degraded RAW images for training deep blind super-resolution models. We have curated a dataset with different DSLR cameras. As part of this effort, we provide a new dataset of diverse real scenes captured using DSLM cameras. Our experiments demonstrate that models trained with our degradation pipeline can restore real-world degraded RAWs. This represents a powerful alternative in image signal processing, and can be of benefit to other low-level downstream tasks.

¹https://github.com/xuxy09/RAWSR

Future Work We plan to extend this analysis to smartphone images, where the ISP is more challenging, and the possible degradations in the RAW images might be more complex due to the limited optics i.e. more noise and blur.

Acknowledgments This work was partly supported by the The Alexander von Humboldt Foundation (AvH).
<table>
<thead>
<tr>
<th>Degradation</th>
<th>Super-Resolution ×2</th>
<th></th>
<th>Super-Resolution ×4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BSRAW-light</td>
<td>BSRAW</td>
<td>BSRAW-light</td>
<td>BSRAW</td>
</tr>
<tr>
<td></td>
<td>PSNR↑</td>
<td>SSIM↑</td>
<td>PSNR↑</td>
<td>SSIM↑</td>
</tr>
<tr>
<td>I</td>
<td>42.57</td>
<td>0.949</td>
<td>43.30</td>
<td>0.952</td>
</tr>
<tr>
<td>II</td>
<td>43.26</td>
<td>0.961</td>
<td>43.98</td>
<td>0.965</td>
</tr>
<tr>
<td>III</td>
<td>43.53</td>
<td>0.962</td>
<td>43.96</td>
<td>0.966</td>
</tr>
<tr>
<td>IV</td>
<td>43.78</td>
<td>0.964</td>
<td>44.25</td>
<td>0.971</td>
</tr>
</tbody>
</table>

Table 3. Ablation study for BSRAW on different degradation levels using our synthetic MIT5K testset. The metrics are calculated in the RGB domain, thus, the ISP setting was fixed for all the compared configurations, to avoid possible biases, we use a simple ISP with the canonical steps (demosaicing, white balance, color correction, gamma and tone mapping correction.) [27]. BSRAW-light is a variant with less encoder blocks and channels, described in the supplementary material.

Figure 6. From left to right: sRGB rendered using the input RAW and a simple ISP [1], our restored RAW (we apply the ISP after restoring and upsampling), BSRGAN [66] and Real-ESRGAN [9, 48] restored images. The approaches are illustrated in Figure 1. We believe both blind single image super-resolution (SISR) approaches might be complementary and represent future research.
References

[27] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In ECCV, pages 429–444, 2016. 1, 2, 3, 4, 6, 8


