Simple Base Frame Guided Residual Network for RAW Burst Image Super-Resolution

Anderson Nogueira Cotrim[1] [a], Gerson Barbosa[2,3] [b], Cid Adinam Nogueira Santos[2] [c]
and Helio Pedrini[1] [d]

[1] *Institute of Computing, University of Campinas, Campinas, SP, 13083-852, Brazil*
[2] *Eldorado Research Institute, Campinas, SP, 13083-898, Brazil*
[3] *São Paulo State University, Guaratinguetá, SP, 12516-410, Brazil*

Keywords: Super-Resolution, Deep Learning, RAW Image, Multi-Frame, Burst.

Abstract: Burst super-resolution or multi-frame super-resolution (MFSR) has gained significant attention in recent years, particularly in the context of mobile photography. With modern handheld devices consistently increasing their processing power and the ability to capture multiple images even faster, the development of robust MFSR algorithms has become increasingly feasible. Furthermore, in contrast to extensively studied single-image super-resolution (SISR), burst super-resolution mitigates the ill-posed nature of reconstructing high-resolution images from low-resolution ones by merging information from multiple shifted frames. This research introduces a novel and effective deep learning approach, SBFBurst, designed to tackle this challenging problem. Our network takes multiple noisy RAW images as input and generates a denoised, super-resolved RGB image as output. We demonstrate that significant enhancements can be achieved in this problem by incorporating base frame-guided mechanisms through operations such as feature map concatenation and skip connections. Additionally, we highlight the significance of employing mosaicked convolution to enhance alignment, thus enhancing the overall network performance in super-resolution tasks. These relatively simple improvements underscore the competitiveness of our proposed method when compared to other state-of-the-art approaches.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

In recent times, smartphones have become the preferred choice for capturing image devices, surpassing the popularity of traditional digital cameras (Lafenetre et al., 2023). While smartphone cameras have made significant strides in enhancing image quality and resolution, they still face limitations when compared to professional DSLR cameras. A primary limitation concerns the small sensor size, which demands a fine balance between noise levels and resolution. While increasing resolution using smaller pixel sensors is possible, it comes at the cost of compromising image quality, particularly in low-light conditions, where noise becomes more prominent. Furthermore, optical zoom capabilities in smartphone cameras remain significantly inferior to those in professional cameras, primarily due to physical constraints. Large lenses, which are essential for high-quality optical zoom, would be impractical in portable devices.

Given these constraints inherent to portable cameras, which are progressively more demanding to address only through sensor enhancements, coupled with the ongoing evolution of artificial intelligence and onboard processing capabilities, researchers have dedicated their efforts to developing deep learning algorithms aimed at improving image quality. These techniques encompass various domains, including super-resolution, noise reduction, and high dynamic range (HDR).

Super-resolution (SR) is an extremely challenging and relevant task that consists of the generation of a high-resolution (HR) image from one or several low-resolution (LR) observations. This advancement could potentially reduce even more the gap in image quality between small devices and professional cameras, enabling improved zoom capabilities without requiring larger lenses.

In recent years, the SR community has primar-

[a] https://orcid.org/0009-0006-8115-589X
[b] https://orcid.org/0000-0002-1147-2519
[c] https://orcid.org/0000-0002-9278-5356
[d] https://orcid.org/0000-0003-0125-630X

Cotrim, A., Barbosa, G., Santos, C. and Pedrini, H.
Simple Base Frame Guided Residual Network for RAW Burst Image Super-Resolution.
DOI: 10.5220/0012348300003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 77-87
ISBN: 978-989-758-679-8; ISSN: 2184-4321

Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

ily focused on the single-image super-resolution task (SISR), where an HR image is estimated from a single LR input. This is a very ill-posed problem, and methods strive to hallucinate high-frequency information, which is limited from previously learned image information. Conversely, the multi-frame super-resolution (MFSR) approach has recently gained significant interest (Figure 1). In MFSR, the objective is to reconstruct the original HR image using multiple LR images. These images inherently feature spatial shifts induced by natural hand movements, and recent research (Wronski et al., 2019) has convincingly shown that these shifts can produce multiple aliased representations of the underlying scene. This phenomenon allows for the aggregation of subpixel information from several images of the same scene, thereby mitigating the ill-posed nature of single-image super-resolution.

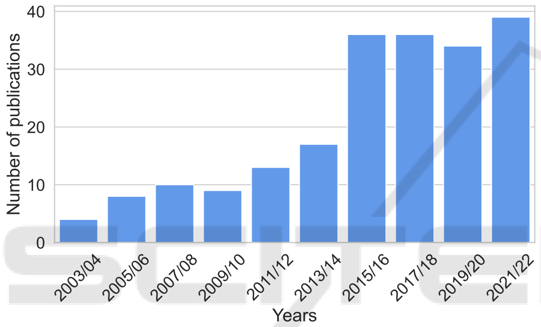


Figure 1: Evolution of the number of publications biyearly in major conferences or journals since 2003 dedicated to the topic of multi-frame or burst super-resolution, according to Web of Science.

This work develops a simple but effective deep-learning method, denoted as Simple Base Frame Burst (SBFBurst), to address the burst image super-resolution problem. This is particularly significant given the context previously outlined. Additionally, our emphasis is placed on RAW images, as they contain more information than processed ISP (Image Signal Processing) RGB images, potentially leading to improved accuracy in high-resolution image reconstruction.

The main contributions of this work are as follows:

- Proposition of a simple but effective deep convolutional architecture that benefits from base frame guidance and deformable convolution (Dai et al., 2017) for solving burst image super-resolution.
- Experimental evaluation of the effectiveness of incorporating the base frame guidance into the convolutional network flow design, whether through concatenation or skip connections.
- Use of SpyNet (Ranjan and Black, 2017) on a mosaicked convolutional feature map (Cilia et al., 2023) to obtain optical flows between frames, which can guide the deformable convolutions (Dai et al., 2017; Chan et al., 2022) to obtain features with better alignment.
- Development of mixed gradient loss (Lu and Chen, 2019) in order to guide our network to a better edge reconstruction.
- Contrary to some methods, the proposed method can deal with an arbitrary number of frames in an invariant permutation way without losing accuracy.
- Our experiments, conducted on both synthetic and real-world datasets, indicate that our approach not only surpasses state-of-the-art methods in both quantitative and qualitative measures but also exhibits efficient inference capabilities (Figure 2).

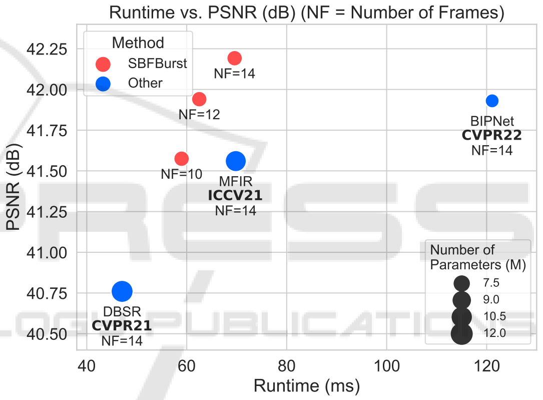


Figure 2: Comparison of performance and runtime on SyntheticBurst dataset (Bhat et al., 2021a). Our method outperforms others, even when utilizing a smaller number of frames.

The text is structured as follows. In Section 2, we provide a concise overview of relevant works in the field of super-resolution literature, with a particular focus on multi-frame super-resolution. Following that, in Section 3, we introduce our proposed method, providing an in-depth explanation of the architectural design and loss function. Moving forward to Section 4, we describe the adopted experimental setup, present the results we have obtained, and engage in a thorough discussion comparing our findings with other relevant approaches from the literature. Additionally, we include a brief ablation study to underscore the effectiveness of our architectural choices. Finally, in Section 5, we conclude our discussion with some final remarks and outline some directions for future work.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

2 RELATED WORK

Single Image Super-Resolution (SISR). SISR has been a longstanding field of research in computer vision that focuses on reconstructing a high-resolution image from its degraded low-resolution version. Significant strides have been made in this field thanks to deep convolutional learning-based approaches. Since the first work, SRCNN (Dong et al., 2015), numerous other learning-based methods have emerged in an effort to address this problem.

Some studies have focused on refining architectural designs (Kim et al., 2016a; Shi et al., 2016a; Zhang et al., 2018c; Haris et al., 2018; Lim et al., 2017), including recursive learning (Kim et al., 2016b; Ahn et al., 2018), progressive reconstruction (Wang et al., 2015; Lai et al., 2017), attention mechanisms (Zhang et al., 2018b; Dai et al., 2019; Zamir et al., 2020), generative adversarial networks (Ledig et al., 2017; Wang et al., 2019), vision transformers (Lu et al., 2022) and different loss functions (Johnson et al., 2016; Lugmayr et al., 2020). However, even with all these improvements, it is still really hard to recover rich details for real-world images because of the extremely ill-posed nature of this problem.

Multi-Frame Super-Resolution (MFSR). In order to address the challenging nature of ill-posed problems in SISR, the concept of MFSR has been introduced. MFSR involves the fusion of pixel data from multiple images of the same scene, each with some spatial displacement, thereby providing supplementary sub-pixel information to enhance the quality of image reconstruction (Tsai and Huang, 1984; Hardie, 2008).

The journey of MFSR began with (Tsai and Huang, 1984), who introduced a pioneering frequency domain approach, assuming that the translations between input images are known. HighRes-Net (Deudon et al., 2020), on the other hand, was explicitly tailored for satellite imagery. It implicitly aligns each frame with a reference frame and employs recursive fusion techniques to enhance image quality.

DBSR (Bhat et al., 2021a) took a step further by introducing a weighted-based fusion mechanism, which predicted element-wise weights between the base frame and the other frames, allowing for more precise fusion. They have also introduced a real-world dataset named BurstSR, which has played a crucial role in motivating and encouraging new research endeavors in this field. The evolution continued with MFIR (Bhat et al., 2021b), which extended this fusion mechanism by diving into deep feature space to handle both SR and denoising. LKR (Lecouat et al., 2021) advances proposing an end-to-end approach for joint image alignment and super-resolution from raw burst inputs.

Recent developments in the field, such as BIP-Net (Dudhane et al., 2022), have recognized the importance of comprehensive fusion. BIPNet introduces a pseudo-burst fusion strategy by fusing temporal features channel-by-channel, resulting in a more robust approach. However, it is worth noting that BIPNet requires fixing the input frame number, which can be a limitation in specific scenarios.

Another work (Cilia et al., 2023) introduces a novel convolutional block, namely mosaicked convolution feature extractor (MCFE), in order to improve feature extraction directly from raw mosaicked sensor data. The key concept behind MCFE is to extract high-level features while preserving the Bayer color arrangement. Through a series of experiments, the authors demonstrated the effectiveness of this block and presented competitive results.

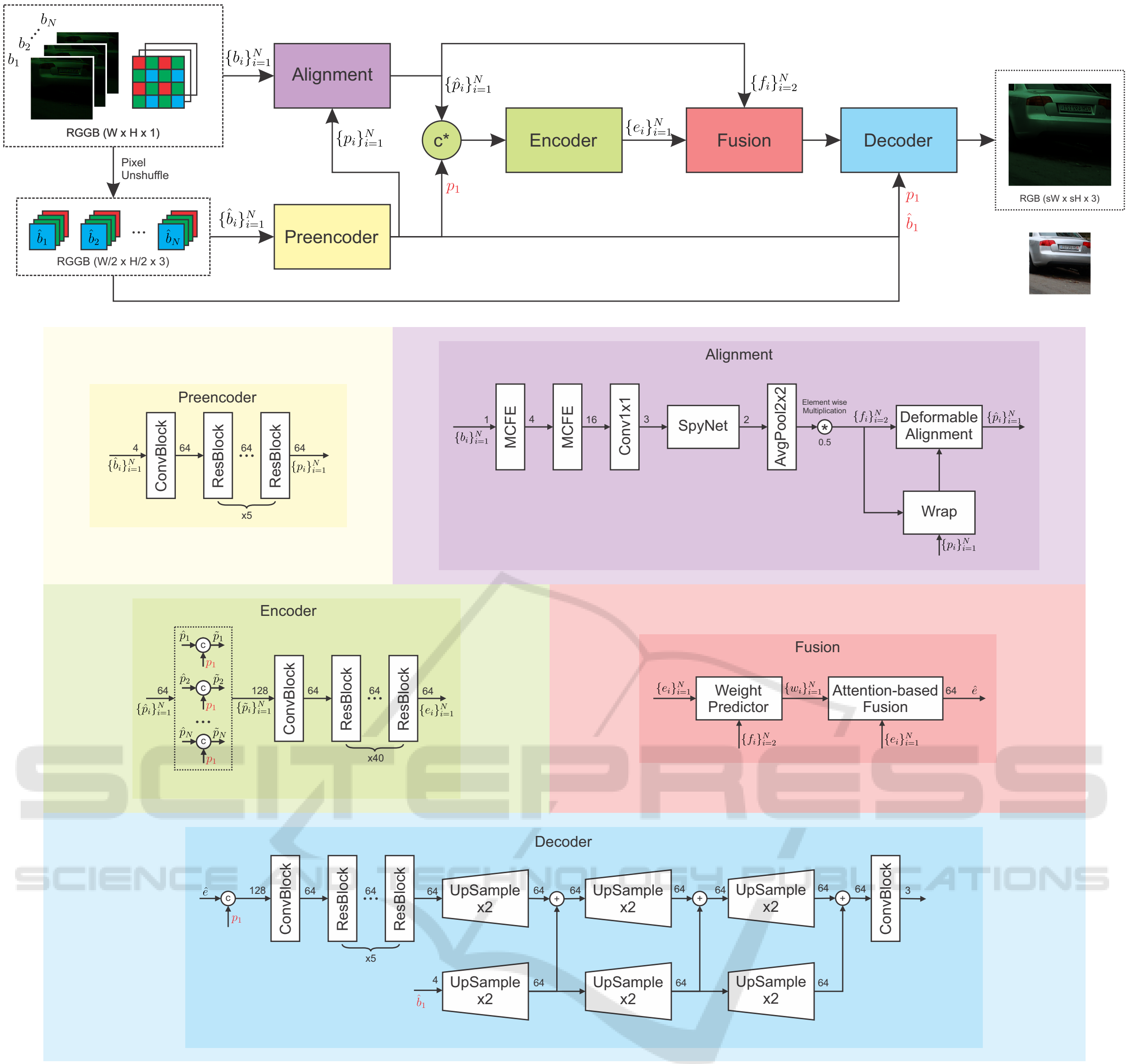
In addition to fusion strategies, some recent works (Dudhane et al., 2023; Mehta et al., 2023) have explored the use of inter-frame attention-based mechanisms to enhance feature interaction. These mechanisms enable cross-attention on the channel or spatial dimension, leading to better information exchange among frames. While these approaches hold great promise, it is essential to acknowledge that they can be computationally intensive.

3 PROPOSED METHOD

In this section, we present an in-depth explanation of our burst super-resolution network, SBFBurst. It takes multiple low-resolution RAW images captured quickly in a burst as input and performs denoising, demosaicking, and super-resolution simultaneously, resulting in a high-quality RGB output. Since burst images might have slight misalignments due to quick capture, they provide extra information for super-resolution. By effectively combining all the burst data, our network can better reconstruct the scene, resulting in a higher-quality output than single-frame methods.

Figure 3 contains an overview of the proposed architecture. Our network takes a series of RAW images captured in a burst, denoted as $\{b_i\}_{i=1}^{N}$, where N can be of any size. Each image, represented as $b_i \in \mathbb{R}^{W \times H}$, contains the RAW sensor data from the camera.

Before proceeding further, we independently extract feature representations from each unshuffled burst $\hat{b}_i$, resulting in $\{p_i\}_{i=1}^{N}$.



SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 3: The overall architecture of the proposed SBFBurst for raw image burst super-resolution, which follows a comprehensive pipeline to transform a RAW burst of degraded images into a clean, high-quality RGB image in five main components: the preencoder, alignment module, encoder, fusion, and decoder. The preencoder relies on a series of residual blocks for feature extraction from each burst. Subsequently, deformable convolutional alignment is employed to align these bursts, aided by SpyNet (Ranjan and Black, 2017), and MCFE (Cilia et al., 2023). Following alignment, each burst is concatenated with the reference frame, generating a deeper embedding by the encoder. All embeddings are then subjected to a fusion process (Bhat et al., 2021a) that reduces to a single one. Finally, the decoder block plays a crucial role by upsampling the fused embedding, which is concatenated with the base frame feature. Additionally, it maintains a skip connection with an upsampled version of the base frame.

Subsequently, we perform alignment and warping of each feature map p_i to align them with the reference frame, denoted as b_1. This alignment is achieved using offsets estimated through mosaicked convolution (Cilia et al., 2023), SpyNet (Ranjan and Black, 2017), and deformable convolutions (Dai et al., 2017).

Additionally, we employ a secondary deeper encoder to create deep feature representations, guided by the base frame p_1, leading to $\{e_i\}_{i=1}^N$.

In the fusion process, we utilize an attention-based module built upon the approach presented in (Bhat et al., 2021a). This module predicts fusion weights at the element level, enabling the network to dynamically select the most relevant information from each image within the burst. This process produces a unified feature map denoted as $\hat{e}$.

Finally, the merged feature map $\hat{e}$, along with the base frame representation p_1, is fed into the decoder module. This module gradually upsamples the input feature map, enhanced by a skip connection derived from the pixel shuffle operation applied to the base frame $\hat{b}_1$. This process generates the ultimate RGB image denoted as $y \in \mathbb{R}^{sW \times sH \times 3}$, where s is the super-resolution factor.

The following subsections detail each module of our proposed method and the loss function.

3.1 Preencoder

We begin by arranging the raw Bayer pattern into 2×2 blocks along the channel dimension, yielding a 4-channel image at half the initial resolution ($\hat{b}_i \in \mathbb{R}^{\frac{W}{2} \times \frac{H}{2} \times 4}$). This lower-resolution image ($\hat{b}_i$) is subsequently passed through the encoder, which consists of an initial convolutional layer followed by a sequence of residual blocks, yielding 64 feature maps for each burst. This process could be represented in Equation 1.

$$p_i = \text{Preencoder}(\hat{b}_i) \tag{1}$$

3.2 Alignment

The absence of precise information regarding pixel-level displacements among the images is one of the main problems faced in burst super-resolution. Several elements, such as scene changes and overall camera motions, could have been responsible for these displacements. It becomes essential to align the information to effectively fusion multiple frames. Our approach tackles this challenge using the SpyNet (Ranjan and Black, 2017) for optical flow estimations and deformable convolutions (Dai et al., 2017) for pixel alignment refinements.

We decided to build an alignment module that relies on b_i instead of $\hat{b}_i$. This could be preferable because b_i keeps spatial information, which may be important for a better optical flow estimation. Firstly, we employ two mosaicked convolutional layers on each flattened RAW image burst (b_i) to avoid disrupting the Bayer color arrangement, generating 16 maps. These maps are then downsized to 3 by 1x1 convolutional layers to obtain a suitable input for the pre-trained SpyNet designed for RGB images.

SpyNet is able to calculate calculating dense pixel-wise optical flow vectors denoted as $\hat{f}_i \in \mathbb{R}^{W \times H \times 2}$ between each burst image b_i. However, as our network relies on wrapping the feature maps $p_i \in \mathbb{R}^{\frac{W}{2} \times \frac{H}{2} \times 2}$, $\hat{f}_i \in \mathbb{R}^{W \times H \times 2}$ should be reduced by half. To achieve the desired optical flow output, we apply average pooling followed by element-wise multiplication with a factor of 0.5 on $\hat{f}_i$, as the displacements were reduced by half on the unshuffle operation of burst b_i. This results in the final pixel-wise flow information $f_i \in \mathbb{R}^{\frac{W}{2} \times \frac{H}{2} \times 2}$, which not only captures global camera motion but also accommodates any object motion within the scene, concerning to the base image $\hat{b}_1$.

The estimated flow vectors, f_i, are subsequently employed to warp the preencoded feature maps p_i to align with the base frame. These aligned maps are further refined using a deformable convolutional network (DCN) inspired by BasicVSR++ (Chan et al., 2022), which outputs the final aligned maps denoted by $\hat{p}_i$.

This entire process could expressed in Equations 2 and 3.

$$f_i = \text{FlowEst}(b_i, b_1) \tag{2}$$

$$\hat{p}_i = \text{DCN}(\text{Wrap}(p_i, f_i), f_i, p_1) \tag{3}$$

3.3 Encoder

The encoder process takes as input the aligned feature map ($\hat{p}_i$) concatenated with reference feature (p_1). These concatenated maps subsequently passed the initial convolutional layer followed by a sequence of residual blocks, yielding 64 feature maps for each burst, expressed in Equation 4.

$$e_i = \text{Encoder}(\hat{p}_i, p_1) \tag{4}$$

3.4 Fusion

The fusion module's purpose is to combine information from individual burst images ($\{e_i\}_{i=1}^{N}$) to create a unified feature representation called $\hat{e}$. In this study, we adopt the fusion module introduced by (Bhat et al., 2021a), an attention-based method that adaptively merges information based on factors such as image content and noise levels.

The fusion process relies on aligned embeddings e_i and flow vectors f_i to estimate attention weights for each embedding. It is worth noting that, unlike the approach outlined in (Bhat et al., 2021a), we chose to employ the embeddings e_i in their raw form without projecting them into a lower dimension, as initially suggested by the authors.

Subsequently, this merged feature map ($\hat{e}$) serves as input for a decoder module responsible for producing the final output. The dependences of $\hat{e}$ follows Equation 5,

$$\hat{e} = \text{Fusion}(\{e_i, f_i\}_{i=1}^{N}) \tag{5}$$

SCIENCE AND TECHNOLOGY PUBLICATIONS

3.5 Decoder

The decoder module is responsible for generating the high-resolution RGB output image from the fused feature map $\hat{e}$. Initially, we concatenate the base feature map p_1 with fusion embedding $\hat{e}$, then pass it through a residual network.

To achieve the desired resolution of $sH \times sW$, we employ sub-pixel convolution (Shi et al., 2016b) for upsampling. This operation is also applied to the base frame image ($\hat{b}_1$), and the data is merged using element-wise summation at each appropriate resolution. Lastly, we apply a final convolution operation to obtain the ultimate RGB image output ($\hat{Y}$). This whole process could be expressed in Equation 6.

$$\hat{Y} = \text{Decoder}(\hat{e}, p_1, \hat{b}_1) \tag{6}$$

3.6 Loss Function

In our training process for both real and synthetic track models, we employ mixed gradient loss (Lu and Chen, 2019) with l_1 to assess the model's prediction errors, defined in Equation 7.

$$\text{MixGE}(Y, \hat{Y}) = l_1(Y, \hat{Y}) + \lambda(l_1(G(Y), G(\hat{Y}))) \tag{7}$$

where l_1 stands for the mean absolute error function and G represents the gradient function obtained from the Sobel operator. Additionally, λ is a weighting factor applied to the gradient differences, effectively penalizing high-frequency reconstructions such as edges aligned with the objective of the super-resolution task.

4 EXPERIMENTS

This section presents the experimental setting, results, discussion, and an ablation study achieved through the proposed burst image super-resolution method.

4.1 Experimental Settings

In this subsection, we describe the datasets used in the experiments, training settings and evaluation metrics.

4.1.1 Datasets

As in previous works (Bhat et al., 2021a; Bhat et al., 2021b; Dudhane et al., 2022), our approach is subjected to comprehensive evaluation using both synthetic and real-world datasets, as provided by the authors of (Bhat et al., 2021a). The synthetic dataset comprises 46,839 RGB images sourced from the Zurich RAW to RGB Dataset (Ignatov et al., 2020) by Canon 5D Mark IV DSLR Images. This dataset serves for generating sets of low-quality RAW burst images through random translations, rotations, and the introduction of additional noise in the RGB-to-RAW inverse camera pipeline (Brooks et al., 2018). This process yields synthetic low-resolution and noisy burst images.

The real-world dataset, known as BurstSR, includes 5,405 RAW burst patches captured in real-world conditions using a Samsung Galaxy S8 smartphone, each with dimensions of 160×160 pixels. The corresponding high-resolution images are obtained from a Canon DSLR camera.

For our evaluation, we used a validation set consisting of 300 synthetically generated images (sized at 96×96 pixels) and 882 real-world patches (sized at 160×160 pixels).

4.1.2 Training Settings

In our training and testing procedures, we adhere to established conventions. Initially, our model undergoes training using the synthetic dataset and subsequently undergoes fine-tuning with the real-world dataset.

Throughout all experiments, we maintain a fixed scale factor s set to 4. During both synthetic and real training phases, we employ the MixGE loss function with parameter λ set to 0.01, as detailed in Section 3.6, to optimize the entire model. When training with real-world data, as the ground truth images are not initially aligned with the input data, we incorporate an aligned MixGE loss that accounts for both spatial and color disparities between the low-resolution (LR) input bursts and high-resolution (HR) ground truth. This alignment process involves initially matching the ground truth image with the super-resolved image using a pre-trained PWC-Net (Sun et al., 2018). This loss function enables the model to learn the generation of HR images that closely align with the ground truth, considering both spatial and color information.

For both datasets training, we employ the AdamW optimizer with exponential decay rates of 0.9 and 0.999. During synthetic training, we conduct 500 epochs with a batch size of 16, starting with an initial learning rate of 2×10^{-4} and halving it at epochs 80, 120, 280, 350, 410, and 460. To facilitate better convergence, we freeze the weights of the pre-trained SpyNet for the first 130 epochs. In each training batch, HR images are cropped to dimensions of 384×384 pixels, and 14 burst LR image patches (96×96) are randomly synthesized based on the HR image.

For real-world training, we fine-tune the model, which was originally pre-trained on synthetic data, for an additional 80 epochs. This fine-tuning process employs a batch size of 8, beginning with an initial learning rate of 5×10^{-5} and then reducing it by half every 15 epochs.

Our implementation of the proposed method is realized using the PyTorch framework and takes advantage of an NVIDIA Tesla A100 GPU, with completion times of approximately 3 days for synthetic and 1 day for real-world scenarios.

4.1.3 Evaluation Metrics

We adhere to the established evaluation protocols and datasets utilized in prior studies (Bhat et al., 2021a; Bhat et al., 2021b; Dudhane et al., 2022) to assess our approach.

Our evaluation metrics encompass the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) (Wang et al., 2004), and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018a). While these metrics can be directly employed for synthetic datasets, we employ aligned PSNR, SSIM and LPIPS for assessing our model on real-world data, as has been the practice in previous works (Bhat et al., 2021a; Bhat et al., 2021b; Dudhane et al., 2022).

4.2 Results and Discussion

We compare SBFBurst with state-of-the-art burst super-resolution approaches including HighResNet (Deudon et al., 2020), DBSR (Bhat et al., 2021a), LKR (Lecouat et al., 2021), MFIR (Bhat et al., 2021b) and BipNet (Dudhane et al., 2022).

Table 1 presents the quantitative results for both testing datasets. When comparing SBFBurst to BipNet, our approach surpasses 0.26dB and 0.38dB in terms of PSNR on synthetic and real data, respectively, without substantially increasing the number of parameters. It is noteworthy that these improvements are nearly equal on both synthetic and real datasets, indicating the potential for our approach to generalize effectively when fine-tuned on diverse datasets.

When dealing with a variable number of input frames, our method also exhibits superior performance compared to the previous ones, as visually demonstrated in Figure 4. Notably, our SBFBurst achieves comparable results using only 4 input frames, matching the performance of DBSR (Bhat et al., 2021a), which requires 14 frames. Moreover, even with just 10 frames as input, our SBFBurst achieves performance on par with MFIR (Bhat et al., 2021b).

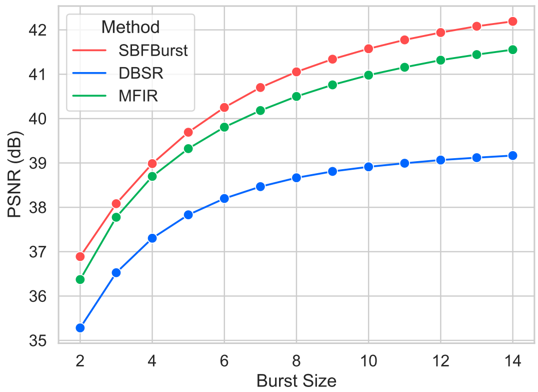


Figure 4: Comparison in terms of PSNR with other approaches when facing variable input frame numbers on Synthetic Dataset (Bhat et al., 2021a). It is worth noting that BipNet encounters a limitation in handling input of variable sizes during prediction.

For a visual representation of our results using both synthetic and real-world data, please refer to Figures 5 and 6. The visual representations in Figures 5 and 6 illustrate that our SBFBurst method excels in recovering highly detailed textures with better color while exhibiting fewer artifacts compared to alternative approaches.

For instance, in the 1st row of Figure 5, our method excels in producing a clean and denoised signaling cone, preserving all the crucial details. Similarly, in the 4th row of Figure 5, our method shows superior edge reconstruction for Venetian blinds, underscoring its capability to effectively handle high-frequency information. In contrast, all other approaches failed to handle the noisy details.

Furthermore, our method demonstrates its proficiency in restoring additional information from real-world burst images, as we can see in Figure 6. In the third row, SBFBurst practically eliminates ghosting artifacts on letters, a challenge often unaddressed by other methods. Moreover, there is an enhancement in color continuity, as evidenced by the letter “r” when compared to the MFIR method.

4.3 Ablation Study

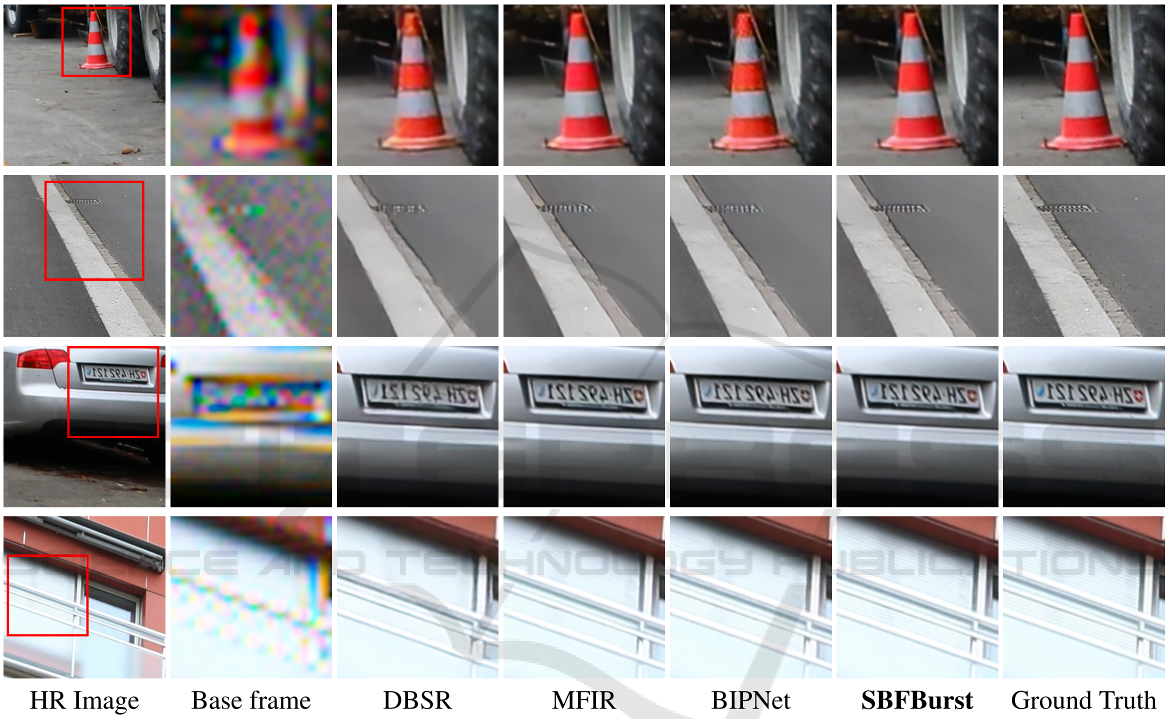
In our ablation study, we rigorously evaluate the effectiveness of our alignment module and the guidance provided by the network’s base frame. Table 2 presents a comprehensive summary of our results as we progressively incorporate the proposed mechanisms, indicated by checkmark symbols (✓).

Our baseline experiments begin with the architecture excluding base frame guidance, denoted as p_1 on the encoder, p_1 and $\hat{b}_1$ on the decoder, and the align-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 1: Comparison between SBFBurst and the other approaches. The best one marks in red and the second best are in blue. All results are reported for a 4× super-resolution task.

Methods	Authors	#Parameters	Synthetic			Real-World		
			PSNR (dB) ↑	SSIM ↑	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
HighResNet	(Deudon et al., 2020)	34.78M	37.45	0.924	0.106	46.64	0.980	0.038
DBSR	(Bhat et al., 2021a)	13.01M	39.17	0.946	0.081	47.70	0.984	0.029
LKR	(Lecouat et al., 2021)	-	41.45	0.950	-	-	-	-
MFIR	(Bhat et al., 2021b)	12.13M	41.55	0.964	0.045	48.32	0.985	0.023
BipNet	(Dudhane et al., 2022)	6.66M	41.93	0.967	0.035	48.49	0.985	0.026
SBFBurst		7.64M	42.19	0.968	0.036	48.87	0.987	0.022



SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 5: Qualitative results of a comparison between our and other approaches in Synthetic Dataset (Bhat et al., 2021a).

ment module applied directly to the unshuffled image, following prior work (Bhat et al., 2021a).

The introduction of the **Base Frame Decoder** entails the inclusion of p_1 and $\hat{b}_1$ in the decoder, while the **Base Frame Encoder** involves the addition of p_1 to the decoder. Lastly, the **MCFE Alignment** represents the incorporation of mosaicked convolutional features on a flattened image, which serves as a strategy to preserve spatial information during the alignment process.

As illustrated in Table 2, it becomes evident that all the relatively straightforward enhancements contribute incrementally to our baseline method. These improvements play a crucial role in driving our baseline method to surpass all the state-of-the-art approaches, as evidenced in Table 1.

Table 2: Ablation experiments to assess the impact of SBFBurst's contributions, we evaluate PSNR performance on the SyntheticBurst dataset for a 4× super-resolution task. Our findings indicate that simple design decisions, such as base frame guidance and preserving spatial information on raw alignment can lead to significant improvements.

Improvements	AS1	AS2	AS3	AS Final
Baseline	✓	✓	✓	✓
Base Frame Decoder		✓	✓	✓
Base Frame Encoder			✓	✓
MCFE Alignment				✓
PSNR (dB)	41.34	41.58	42.12	**42.19**

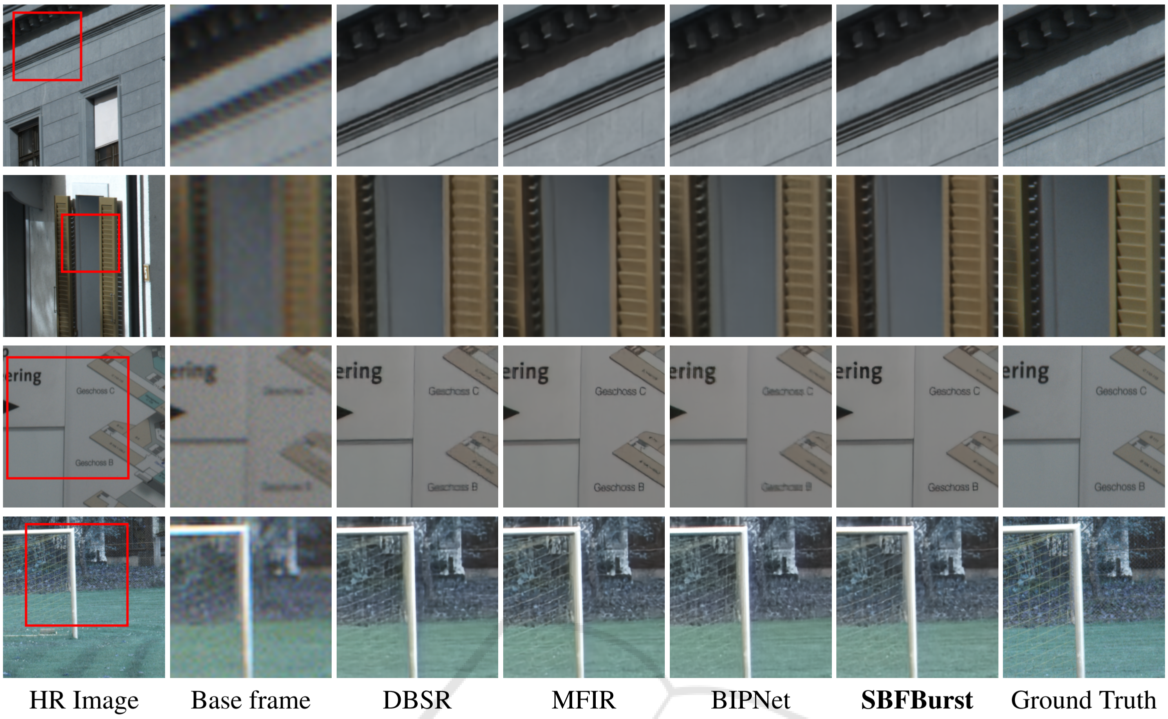


Figure 6: Qualitative results of a comparison between our and other approaches in Real World Dataset BurstSR (Bhat et al., 2021a).

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

5 CONCLUSIONS

This research presents a significant advancement in addressing the burst image super-resolution problem, presenting a straightforward yet highly effective deep-learning method. By focusing on RAW images and leveraging base frame guidance, deformable convolution, SpyNet, and mixed gradient loss, this work has made several key contributions to the field. Importantly, the proposed method demonstrates remarkable performance versatility by accommodating an arbitrary number of frames without sacrificing accuracy.

Through comprehensive experiments on synthetic and real-world datasets, this approach not only outperforms existing state-of-the-art methods in both quantitative and qualitative assessments but also demonstrates its efficiency in terms of inference speed. Therefore, this research opens up new possibilities for improving the quality of high-resolution image reconstruction, with broad implications for various applications in image processing, especially on portable cameras.

While the contributions of this work are indeed promising, there remain several directions that could be explored to further enhance burst image super-resolution techniques. Firstly, addressing the challenge of scenes featuring fast-moving objects, which represents a significant hurdle in this field. Moreover, it would be desirable to explore ways to reduce computational complexity to make it more suitable for real-time applications on resource-constrained devices. Furthermore, developing new evaluation methods tailored to real-world scenarios is crucial for a better assessment of the methods, as aligned versions of traditional metrics such as PSNR, SSIM, or LPIPS might be biased for methods that employ aligned loss on training. Overall, these future efforts aim to push the boundaries of burst image super-resolution and unlock even greater potential for its application in various domains.

ACKNOWLEDGEMENTS

The authors would like to thank Eldorado Research Institute, Brazilian National Council for Scientific and Technological Development (CNPq #304836/2022-2) for their support.

REFERENCES

Ahn, N., Kang, B., and Sohn, K.-A. (2018). Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network. In *15th European Conference on Computer Vision*, page 256–272, Munich, Germany. Springer-Verlag.

Bhat, G., Danelljan, M., Van Gool, L., and Timofte, R. (2021a). Deep Burst Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218.

Bhat, G., Danelljan, M., Yu, F., Gool, L. V., and Timofte, R. (2021b). Deep Reparametrization of Multi-Frame Super-Resolution and Denoising. In *IEEE/CVF International Conference on Computer Vision*, pages 2440–2450, Los Alamitos, CA, USA. IEEE Computer Society.

Brooks, T., Mildenhall, B., Xue, T., Chen, J., Sharlet, D., and Barron, J. T. (2018). Unprocessing Images for Learned Raw Denoising. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11028–11037.

Chan, K. K., Zhou, S., Xu, X., and Loy, C. (2022). BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 5962–5971, Los Alamitos, CA, USA. IEEE Computer Society.

Cilia, M., Valsesia, D., Fracastoro, G., and Magli, E. (2023). Multi-Level Fusion for Burst Super-Resolution with Deep Permutation-Invariant Conditioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1–5.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017). SimDeformable Convolutional Networks. In *IEEE International Conference on Computer Vision*.

Dai, T., Cai, J., Zhang, Y., Xia, S.-T., and Zhang, L. (2019). Second-Order Attention Network for Single Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 11057–11066.

Deudon, M., Kalaitzis, A., Goytom, I., Arefin, M. R., Lin, Z., Sankaran, K., Michalski, V., Kahou, S. E., Cornebise, J., and Bengio, Y. (2020). HighRes-Net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery.

Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image Super-Resolution Using Deep Convolutional Networks.

Dudhane, A., Zamir, S., Khan, S., Khan, F., and Yang, M. (2022). Burst Image Restoration and Enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 5749–5758, Los Alamitos, CA, USA. IEEE Computer Society.

Dudhane, A., Zamir, S., Khan, S., Khan, F., and Yang, M. (2023). Burstormer: Burst Image Restoration and Enhancement Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 5703–5712, Los Alamitos, CA, USA. IEEE Computer Society.

Hardie, R. (2008). A Fast Image Super-Resolution Algorithm Using an Adaptive Wiener Filter. *IEEE Transactions on Image Processing*, 16:2953–64.

Haris, M., Shakhnarovich, G., and Ukita, N. (2018). Deep Back-Projection Networks for Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1664–1673, Los Alamitos, CA, USA. IEEE Computer Society.

Ignatov, A., Gool, L. V., and Timofte, R. (2020). Replacing Mobile Camera ISP with a Single Deep Learning Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, volume 1, pages 2275–2285, Los Alamitos, CA, USA. IEEE Computer Society.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *European Conference on Computer Vision*, pages 694–711, Cham. Springer International Publishing.

Kim, J., Lee, J. K., and Lee, K. M. (2016a). Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1646–1654.

Kim, J., Lee, J. K., and Lee, K. M. (2016b). Deeply-Recursive Convolutional Network for Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1637–1645.

Lafenetre, J., Facciolo, G., and Eboli, T. (2023). Implementing Handheld Burst Super-Resolution. *Image Processing On Line*, 13:227–257.

Lai, W.-S., Huang, J.-B., Ahuja, N., and Yang, M.-H. (2017). Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 5835–5843.

Lecouat, B., Ponce, J., and Mairal, J. (2021). Lucas-Kanade Reloaded: End-to-End Super-Resolution from Raw Image Bursts. In *IEEE/CVF International Conference on Computer Vision*, volume 1, pages 2350–2359, Los Alamitos, CA, USA. IEEE Computer Society.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 105–114, Los Alamitos, CA, USA. IEEE Computer Society.

Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, volume 1, pages 1132–1140.

Lu, Z. and Chen, Y. (2019). Single Image Super Resolution based on a Modified U-Net with Mixed Gradient Loss. *CoRR*, abs/1911.09428.

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., and Zeng, T. (2022). Transformer for Single Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, volume 1, pages 456–465, Los Alamitos, CA, USA. IEEE Computer Society.

Lugmayr, A., Danelljan, M., Van Gool, L., and Timofte, R. (2020). SRFlow: Learning the Super-Resolution

Space with Normalizing Flow. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *European Conference on Computer Vision*, pages 715–732, Cham. Springer International Publishing.

Mehta, N., Dudhane, A., Murala, S., Zamir, S., Khan, S., and Khan, F. (2023). Gated Multi-Resolution Transfer Network for Burst Restoration and Enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 22201–22210, Los Alamitos, CA, USA. IEEE Computer Society.

Ranjan, A. and Black, M. J. (2017). Optical Flow Estimation Using a Spatial Pyramid Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 2720–2729.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016a). Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1874–1883, Los Alamitos, CA, USA. IEEE Computer Society.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016b). Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1874–1883, Los Alamitos, CA, USA. IEEE Computer Society.

Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. (2018). PWC-Net: CNNs for Optical Flow using Pyramid, Warping, and Cost Volume. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943.

Tsai, R. Y. and Huang, T. S. (1984). Multiframe Image Restoration and Registration. *Multiframe image restoration and registration*, 1:317–339.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Loy, C. C. (2019). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Leal-Taixé, L. and Roth, S., editors, *European Conference on Computer Vision*, pages 63–79, Cham. Springer International Publishing.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Wang, Z., Liu, D., Yang, J., Han, W., and Huang, T. (2015). Deep Networks for Image Super-Resolution with Sparse Prior. In *IEEE International Conference on Computer Vision*, volume 1, pages 370–378.

Wronski, B., Garcia-Dorado, I., Ernst, M., Kelly, D., Krainin, M., Liang, C.-K., Levoy, M., and Milanfar, P. (2019). Handheld Multi-Frame Super-Resolution. *ACM Transactions on Graphics*, 38(4).

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. (2020). Learning Enriched Features for Real Image Restoration and Enhancement. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *European Conference on Computer Vision*, pages 492–511, Cham. Springer International Publishing.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018a). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 586–595, Los Alamitos, CA, USA. IEEE Computer Society.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. (2018b). Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *European Conference on Computer Vision*, pages 294–310, Cham. Springer International Publishing.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. (2018c). Residual Dense Network for Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS