Towards Real-World Burst Image Super-Resolution: Benchmark and Method

Pengxu Wei¹ Yujing Sun² Xingbei Guo¹ Chang Liu³ Guanbin Li¹ Jie Chen² Xiangyang Ji³ Liang Lin¹

¹Sun Yat-sen University
²School of Electronic and Computer Engineering, Peking University, Shenzhen, China
³Tsinghua University

{weipx3, liguanbin}@mail.sysu.edu.cn, yujingsun1999@gmail.com, guoxb7@mail2.sysu.edu.cn,
{liuchang2022,xyji}@tsinghua.edu.cn, chenj@pcl.ac.cn, linliang@ieee.org

Abstract

Despite substantial advances, single-image super-resolution (SISR) is always in a dilemma to reconstruct high-quality images with limited information from one input image, especially in realistic scenarios. In this paper, we establish a large-scale real-world burst super-resolution dataset, i.e., RealBSR, to explore the faithful reconstruction of image details from multiple frames. Furthermore, we introduce a Federated Burst Affinity network (FBAnet) to investigate non-trivial pixel-wise displacements among images under real-world image degradation. Specifically, rather than using pixel-wise alignment, our FBAnet employs a simple homography alignment from a structural geometry aspect and a Federated Affinity Fusion (FAF) strategy to aggregate the complementary information among frames. Those fused informative representations are fed to a Transformer-based module of burst representation decoding. Besides, we have conducted extensive experiments on two versions of our datasets, i.e., RealBSR-RAW and RealBSR-RGB. Experimental results demonstrate that our FBAnet outperforms existing state-of-the-art burst SR methods and also achieves visually-pleasant SR image predictions with model details. Our dataset, codes, and models are publicly available at https://github.com/yjsunnn/FBANet.

1. Introduction

As a fundamental research topic, Super-Resolution (SR) attracts long-standing substantial interest, which targets high-resolution (HR) image reconstruction from a single or a sequence of low-resolution (LR) observations. In recent years, we have witnessed the prosperity of Single Image Super-Resolution (SISR), e.g., SRCNN [7], EDSR [19], SRGAN [16], RDN [33] and ESRGAN [28]. Nevertheless, SISR intrinsically suffers from a limited capacity of restoring details from only one LR image, typically yielding oversmooth LR predictions, especially for large-scale factors. With real detailed sub-pixel displacement information, Multi-Frame Super-Resolution (MFSR) [31, 1, 2, 21, 20] provides a promising potential to reconstruct the high-quality image from multiple LR counterparts, which is valuable for many sensitive realistic applications, e.g., medical imaging, and remote satellite sensing.

After the pioneering work [25] of Tsai and Huang in 1984, the research on MFSR has not achieved as tremendous progress as SISR. Typically, they are overwhelmed by two challenges: 1) the difficulty of fusing multiple LR inputs, which especially is aggravated for real-world data; 2) the limitation of artificially-synthesized data, accounting for a poor generalization for real-world scenarios; To address those challenges, a recent work [1] has made seminal contributions to the first real-world burst SR dataset benchmark, BurstSR, and a novel architecture, DBSR. Subsequently, MFIR proposes a deep reparametrization to reformulate the classical MAP objective in a deep feature space [2]. BIPNet [8] introduces a set of pseudo-burst features for information exchange among multiple burst frames. BSRT [20] employs a

Figure 1: SR predictions with different numbers of burst image inputs in our RealBSR dataset, where more burst inputs facilitate more accurate reconstruction of image details.
pyramid flow-guided deformable convolution network in a Transformer architecture.

Despite great progress achieved, two aspects still need to be revisited. 1) Method: Align-fusion-reconstruction paradigm-based methods usually fuse multiple burst images according to their similarity to a reference image, following their alignment via the optical flow or deformable convolution. However, this fusion strategy largely relies on the reference image and is limited to exploring more information among burst images. 2) Dataset: BurstSR captures multiple LR images with a smartphone in burst mode and a corresponding HR image with a DSLR camera. Thus, several unexpected issues are nontrivial: a) data misalignment (even distortion) among burst LRs and their HR counterparts; b) cross-device gap between LRs and HR captured by different cameras; and c) unfair model evaluation on warped SR predictions by introducing GT HR. Moreover, BurstSR can be cast as a coupled task of burst image SR and enhancement.

To address these issues, we propose the Federated Burst Affinity Network (FBAnet), and make an attempt to build a new real-world burst image SR dataset, named RealBSR. Our RealBSR dataset is captured in quick succession a sequence of LR images and one HR image under a continuous shooting mode with the optical zoom strategy, like RealSR [3]. It provides a real-world benchmark for image detail reconstruction of real-world burst SR applications, avoiding the color style change in terms of the original LR data, especially, burst RAW inputs have no ISP process and often, for faithful high-resolution image predictions, especially for sensitive applications, e.g., medical imaging.

Our FBAnet employs a simple-yet-effective alignment algorithm via a homography matrix from a structural and global aspect. Then, a Federated Affinity Fusion (FAF) module is introduced to aggregate inter- and intra-frame information through affinity difference maps, aiming to not only focus on pixels consistent with the reference frame, following the over-cost collection of real-world LR-HR image pairs, those methods turn to map synthetic LR images to their HR counterparts, which is constantly criticized for poor model generalization in practical scenarios. To facilitate the exploration of real-world image SR, great efforts have been made on building functional benchmarks, e.g., SRRaw [32], RealSR [3], and DRealSR [30], following the optical zoom manner to capture paired LR-HR images. Meanwhile, LP-KPN [3] has been proposed to employ a Laplacian-based network for non-uniform kernel estimation. Encountering heterogeneous image degradation, CDC [30] proposes a gradient-weighted loss to adapt to diverse challenges in reconstructing different regions.

2. Related Work

2.1. Single Image Super-Resolution

SRCNN [7] pioneers CNN to image SR, inspiring numerous follow-ups. Fueled by the evolving of deep neural networks [13, 10, 14, 26], a series of seminal SISR methods have been built to achieve significant advances, e.g., VDSR [15], EDSR [21], SRResNet [17], ESRGAN [28], DRN [11], SwinIR [18], etc. Nevertheless, considering the over-cost collection of real-world LR-HR image pairs, those methods turn to map synthetic LR images to their HR counterparts, which is constantly criticized for poor model generalization in practical scenarios. To facilitate the exploration of real-world image SR, great efforts have been made on building functional benchmarks, e.g., SRRaw [32], RealSR [3], and DRealSR [30], following the optical zoom manner to capture paired LR-HR images. Meanwhile, LP-KPN [3] has been proposed to employ a Laplacian-based network for non-uniform kernel estimation. Encountering heterogeneous image degradation, CDC [30] proposes a gradient-weighted loss to adapt to diverse challenges in reconstructing different regions.

2.2. Multi-Frame Super-Resolution

With great potential to remedy the intrinsic ill-posed SISR problem, MFSR pursues absorbing authentic sub-pixel details contained in the image sequences towards real-world applications. In the early times of MFSR, Tsai and Huang [25] contribute the first fair solution. Afterward, taking advantage of deep learning, TDAN [24] introduces deformable convolutions to mitigate the misalignment problem between neighboring frames and the reference frames. Similarly, EDVR [27] and EBSR [21] build a pyramid structure facilitating the motion compensation during the alignment procedure. MFIR [2] presents a deep reparametrization algorithm that transforms Maximum A Posteriori (MAP) formulation to the latent space for better reconstruction. BIPNet [8] introduces a pseudo-burst feature fusion method to allow flexible information exchange among frames. In addition, BSRT [20] builds the reconstruction module based on Swin Transformer, which further improves the performance.

For real-world burst image SR, Bhat et al. [1] establish a dataset consisting of LR burst images captured from a smartphone and HR counterparts from a DSLR camera and introduce an encoder-decoder-based model to deal with unknown pixel-wise displacement with optical flow estimation and merge aligned frames with an attention mechanism.

• We have conducted extensive experiments on RAW and RGB versions of RealBSR to benchmark existing state-of-the-art methods. Empirically, the efficacy of our FBAnet has been justified with superior SR performances from quantitative and qualitative aspects.
3. RealBSR: A new Benchmark

BurstSR is the only existing dataset for real-world burst image super-resolution and enhancement, which has three typical issues. 1) Data Misalignment. The distortion between LRs and their HR counterparts is distinct. It possibly results in a severe misalignment between paired LR and HR images. Such serious mismatches yield counterproductive super-resolution results with few details reconstructed from burst LR images. 2) Cross-Device Distribution. Since LR sequences and HR counterparts are captured by a smartphone and a DSLR, respectively, the difference of imaging devices would inevitably lead to a cross-device gap between them. Therefore, it has to cast this task on the BurstSR dataset as a combination of burst image super-resolution and enhancement. 3) Evaluation Deficiency. The evaluation routine for BurstSR is that a generated final SR image is warped with the reference of its ground-truth HR and then this warped SR image is used to compute the evaluation metrics with the same ground-truth HR. This is rather problematic and even not fair to truly evaluate the model performance with the aid of GTs. Besides, the calculated metric values (e.g., PSNR) cannot well reflect the visual quality, which means pursuing a higher PSNR on the BurstSR dataset is not positively related to better reconstruction quality. This evaluation strategy greatly attributes to data misalignment and cross-device distribution, inviting a great challenge for evaluation.

In this work, we build a real-world burst super-resolution dataset, named RealBSR. It consists of 579 groups (RAW version) and 639 groups (RGB version) for the scale factor 4. Each group has 14 burst LR images and a GT HR image.

3.1. Collection and Processing

We use the optical zoom strategy for data collection, similar to RealSR [4] and DRealSR [30]. With a Sony DSLR camera (Alpha 7R), we capture a sequence of 14 LR images by pressing the camera shutter and optically zoom the camera to shoot an HR image. Those images are collected in various scenes, e.g., buildings (museum, church, office building, tower, etc.), posters, plants/trees, sculptures, and ships. Our indoor and outdoor images have 21 and 618 groups, respectively. For each group of burst data, since LR and HR images have different fields of view, we employ SIFT to crop LR sequences under the reference of collected HR counterpart.

Considering the distortion of RAW images is not addressed by the camera, their center regions are cropped into our RAW version dataset, named RealBSR-RAW. Besides, the RGB version of RealBSR, termed RealBSR-RGB, is also provided. Since the collected RGB images are processed by...
we derive five second-order statistic features from all the wise displacements to facilitate realistic detail reconstruction.

4.1. Overview

In comparison with SISR, MFSR pursues favorable pixel-wise displacements to facilitate realistic detail reconstruction.

4.2. Homography Alignment

Captured in a quick succession, the pixel-wise displacements among burst images mainly stem from camera motion and scene variations, which are usually regarded to be complementary for reconstructing more details. Before fusing them, we align those images firstly, avoiding the information confusing or discrepancy and leading to blurry super-resolution predictions even with unpleasant artifacts.

We employ a simple homography matrix $H_i$ for the alignment from a global and structural manner. Specifically, a $3 \times 3$ homography matrix $H_i$ between $i$-th frame $x_i$ and the base frame $x_1$ indicates the transformation with respect to Correlation Coefficient Maximization (ECC) criterion.

Since it is not easy to exactly figure out the displacement association among different burst LR images, how to fuse burst images remains intractable. What’s worse, stemming from physical camera shake during imaging, it occurs more unexpected and non-uniform pixel shifts. To address this issue, we propose a federated burst affinity network to move towards real-world burst SR by effectively integrating informative sub-pixel details contained in multiple frames.

Our FBAnet follows a conventional alignment-to-fusion paradigm, Fig. 4. Formally, given an LR image sequence $\{x_i\}_{i=1}^N$ of $N$ burst observations as the input, our model will yield a high-resolution image prediction $\hat{y}_i$ for a scale factor $s$, where their ground-truth (GT) HR counterpart is denoted as $y_i$. With the randomness of pixel-wise shift among different burst images, the fused features would not be perfect enough to directly support the reconstruction of image details. Without loss of generality, the first frame $x_1$ is regarded as the reference frame to align the other images in the sequence by their homography matrix $H$. Then, FBAnet employs a federated affinity fusion strategy to aggregate multiple frames and utilizes two hourglass Transformer blocks to take over the fused features for the final decoding phase of high-resolution image prediction.

4.2. Homography Alignment

Captured in a quick succession, the pixel-wise displacements among burst images mainly stem from camera motion and scene variations, which are usually regarded to be complementary for reconstructing more details. Before fusing them, we align those images firstly, avoiding the information confusing or discrepancy and leading to blurry super-resolution predictions even with unpleasant artifacts.

We employ a simple homography matrix $H_i$ for the alignment from a global and structural manner. Specifically, a $3 \times 3$ homography matrix $H_i$ between $i$-th frame $x_i$ and the base frame $x_1$ indicates the transformation with respect to Correlation Coefficient Maximization (ECC) criterion [9].
Each frame is warped with the homography matrix by taking the base frame as reference for alignment.

### 4.3. Federated Affinity Fusion

To take full advantage of potential information, the aligned images pass through the fusion module. As fused outputs should not only be consistent with the base frame but also incorporate additional signals from other frames, we propose a Federated Affinity Fusion (FAF) module to aggregate inter-frame and intra-frame information, Fig. 4. Our FAF determines how informative the final result could be by assigning pixel-wise fusion weights on each frame, which serves the core of the whole burst paradigm. It is noteworthy that one-order affinity maps, i.e., the differences between two affinity maps, are leveraged to determine the weights, rather than the affinity or attention maps. Specifically, we extract deep features \( F_i \) from aligned images with two convolutional layers. The affinity map \( A \) between two frames is the dot product of their features, i.e., \( A_{i,j} = F_i \cdot F_j \).

1) **Vanilla Affinity Fusion (VAF):** Following wisdom that higher similarity or affinity indicates more important pixels for fusion, it is intuitive that those affinity maps are utilized to weight each frame, which is common in existing works, e.g., TSA in EDVR [27]. The fused feature map of VAF can be formulated as \( M = \sum_{i=1}^{N} A_{1,i} \odot F_i \), where \( \odot \) is the element-wise product. As intuitively illustrated in Fig. 5a, VAF focuses on pixels from other frames consistent with the reference ones in the base frame. Consequently, information similar to the base frame would be kept (e.g., Pixel-A in Fig. 5a), while important details only appear in other frames are ignored (e.g., Pixel-B in Fig. 5b).

2) **FAF:** Despite the higher affinity of \( x_i \) (\( \forall i \neq 1 \)) indicating the higher similarity to the base frame \( x_1 \), there also underlies two adverse effects, especially for real-world burst images: (a) Their easy reconstruction regions (e.g., the flat) would also have large affinity values for \( x_i \), \( \forall i \neq 1 \), which would unexpectedly drive the model to pay more attention to those regions, resulting in over-fitting. (b) As for the imperfect alignment and pixel shift, even the key pixels in detail-rich regions may not have large affinity values to be highlighted for fusion.

To address these issues, our FAF additionally considers the affinity difference maps to distinguish specific differences between one frame from other frames. Consequently, our FAF would pay attention to those complementary details do not appear in the base frame, e.g., Pixel-B in Fig. 5a. The affinity difference map of the \( i \)-th frame can be expressed as

\[
D_i = A_{1,i}; \quad D_i = d(A_{1,i}, A_{1,1}), \text{ when } i \neq 1, \tag{1}
\]

where \( d(\cdot) \) is the difference function.

The final fused feature can be defined as,

\[
M = \sum_{i=1}^{N} D_i \odot F_i
\]

\[
= A_{1,1} \odot F_1 + \sum_{i=2}^{N} d(A_{1,i}, A_{1,1}) \odot F_i \tag{2}
\]

**Analysis:** As Eq. (2) indicates, the fused feature map consists of two components, i.e., (i) attentive features of the base frame based on its self-affinity, (ii) frame-specific features that are relatively independent of the base frame providing more complement from other frames. Given \( D_i \) computed in the Euclidean space, Eq. (2) can be derived as

\[
M = A_{1,1} \odot F_1 + \sum_{i=2}^{N} (F_i - F_1) \odot F_i \odot F_i \tag{3}
\]

The second terms in Eq. (2) & (3) can be regarded as the combination of difference maps \( (F_i - F_1) \) and correlation maps \( F_i \odot F_i \). The former would alleviate the issue that VAF encourages the fusion of redundant information similar to
the base frame too much, which is too easy for reconstruction, e.g., the flat. The latter would alleviate the adverse effects derived from the misalignment due to large motions. Thus, FAF rectifies the fusion information to alleviate the adverse effects resulting from the large misalignment and the overfitting to the easy reconstruction of regions with high affinity, as illustrated in Fig. 5b.

3) FAF*: Following the similar federating spirit, we can further extend this design of FAF. That is, the affinity maps and their different maps can take more complex federated interactions of frames into consideration, rather than only taking the base frame as a reference. Specifically, for t-th frame, its affinity difference map can be compared to any other frame. Thus, $D_t = d(A_{1,i}, A_{1,m}), i, m \neq 1$ and the fused features is computed similar to Eq. (2), $M = \sum_{k=1}^{N} (A_{k,k} \circ F_k) + \sum_{i=1, j \neq k}^{N} d(A_{i,i}, A_{k,k}) \circ F_i$. (4)

4.4. Burst Representation Decoding

To aggregate global information for finer high-frequency detail reconstruction, we utilize the self-attention mechanism to model long-range pixel relations. Specifically, our FBAnet adopts a burst representation decoding module to explicitly model inter-dependencies among channels. This module has two cascaded blocks, shown in Fig. 4. A block has an encoder and a decoder, both of which cascade three Locally-enhanced Window (LeWin) Transformer blocks [29]. Each block has a LayerNorm, multi-head self-attention, a LayerNorm, and a Locally-Enhanced Feed-Forward (LeFF) layer [29]. The module is followed by pixelshuffle [22] for producing the final HR predictions.

Our training objective includes a Mean Absolute Error (MAE) loss for SR image reconstruction. In addition, on the RAW-version dataset, to mitigate the negative effects brought by a slight misalignment of the RAW-version dataset, we also introduce the CoBi loss [32] to ease the training and enhance the visual quality of final results. While on the RGB-version dataset, we adopt the Gradient Weighted (GW) loss [30] for high-frequency detail reconstruction.

5. Experiments

5.1. Experimental Settings

Datasets. We conduct experiments on the two versions (RAW and RGB) of the proposed RealBSR benchmark at scale factor 4, real-world BurstSR [1] and a synthetic burst SR dataset, SyntheticBurst [1, 8], with fair comparisons.

Implementation Details. We align frames in a burst sequence using OpenCV to estimate homography matrices, before training. Input images are augmented using flip and rotation in the training stage. The AdamW optimizer is employed and the initial learning rate is set to be 1e-4. Besides, we adopt the cosine annealing schedule to set the learning rate of each parameter group.

<table>
<thead>
<tr>
<th>Task</th>
<th>Method</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
<th>LPIPS ↓</th>
<th>PSNR-Linear ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burst-SR</td>
<td>DBSR [1]</td>
<td>20.906</td>
<td>0.635</td>
<td>0.134</td>
<td>30.484</td>
</tr>
<tr>
<td></td>
<td>MFIR [2]</td>
<td>21.562</td>
<td>0.638</td>
<td>0.131</td>
<td>30.799</td>
</tr>
<tr>
<td></td>
<td>BSRT [20]</td>
<td>22.579</td>
<td>0.662</td>
<td>0.111</td>
<td>30.826</td>
</tr>
<tr>
<td></td>
<td>BIPNet [8]</td>
<td>22.896</td>
<td>0.641</td>
<td>0.144</td>
<td>31.311</td>
</tr>
<tr>
<td></td>
<td>FBAnet (ours)</td>
<td><strong>23.423</strong></td>
<td><strong>0.677</strong></td>
<td><strong>0.125</strong></td>
<td><strong>32.256</strong></td>
</tr>
</tbody>
</table>

Table 1: Performance comparisons on our RealBSR-RAW.

<table>
<thead>
<tr>
<th>Task</th>
<th>Method</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
<th>LPIPS ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>SISR</td>
<td>LP-KPN [4]</td>
<td>29.268</td>
<td>0.863</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>CDC [30]</td>
<td>30.014</td>
<td>0.880</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>SwinIR [18]</td>
<td>29.924</td>
<td>0.876</td>
<td>0.139</td>
</tr>
<tr>
<td>MFISR</td>
<td>DBSR [1]</td>
<td>30.715</td>
<td>0.899</td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td>MFIR [2]</td>
<td>30.895</td>
<td>0.899</td>
<td><strong>0.098</strong></td>
</tr>
<tr>
<td></td>
<td>BSRT [20]</td>
<td>30.782</td>
<td><strong>0.900</strong></td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td>BIPNet [8]</td>
<td>30.665</td>
<td>0.892</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>FBAnet (ours)</td>
<td><strong>31.012</strong></td>
<td>0.898</td>
<td>0.102</td>
</tr>
</tbody>
</table>

Table 2: Performance comparisons on our RealBSR-RGB.

<table>
<thead>
<tr>
<th>Method</th>
<th>SyntheticBurst [1]</th>
<th>PSNR</th>
<th>SSIM</th>
<th>LPIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bicubic</td>
<td>36.17</td>
<td>0.909</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>DBSR [1]</td>
<td>40.76</td>
<td>0.960</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>MFIR [2]</td>
<td>41.56</td>
<td>0.960</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>FBAnet (ours)</td>
<td><strong>42.23</strong></td>
<td><strong>0.970</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: Evaluation on synthetic and real BurstSR data. '*' indicates that the LPIPS results are not provided in [1] and thus are omitted.

Evaluation Metric. On RealBSR-RAW, we adopt four evaluation metrics, i.e., PSNR, SSIM, LPIPS, and PSNR-Linear [1]. The first three metrics are computed in the RGB space, and the last one is in the linear sensor space. On RealBSR-RGB, it follows the evaluation routine in the RGB image space and thus three metrics (PSNR, SSIM, LPIPS) are adopted. On BurstSR, the predicted SR images have to be warped by taking GT HRs as a reference before computing metrics [1], while without post-processing on our RealBSR.

5.2. Comparison with the State-of-the-art Methods

We compare our model with four state-of-the-art burst SR methods, i.e., DBSR [1], MFIR [2], BSRT [20], and BIPNet [8], and three video SR methods, including EDVR [27], BasicVSR [5] and BasicVSR++ [6].

Comparison with burst SR methods: On RealBSR-RAW, Tab. 1, our FBAnet achieves the best results on PSNR, SSIM, and PSNR-Linear metrics. Notably, FBAnet improves the performance by $\sim 0.5$ dB in terms of PSNR and $\sim 0.945$ dB in terms of PSNR-Linear. Similar to the performance on RealBSR-RAW, PSNR of our FBAnet on RealBSR-RGB is also superior to other burst SR methods,
validating the effectiveness of our method. Fig. 6 visualizes SR results of competing methods and ours in both RealBSR-RGB and RealBSR-RAW datasets. On RealBSR-RGB, it is clear that the state-of-the-art burst SR methods are prone to generate realistic but blurry textures, e.g., the building in Fig. 6. On RealBSR-RAW, the SR predictions of DBSR, MFIR and BIPNet have differences in image color, compared with that of our FBAnet. Moreover, on SyntheticBurst and real-world BurstSR, all the evaluation methods are trained from scratch in Tab. 3 and performance gains are also achieved by our FBAnet over existing methods.

**Comparison with video SR methods:** To further evaluate the results of video SR methods in the real-world burst SR task, we also introduce three state-of-the-art video SR methods (i.e., EDVR, BasicVSR and BasicVSR++) for comparison. Since the video SR algorithms are always based on RGB dataset, we train all these methods from scratch on RealBSR-RGB dataset. In Tab. 2, our FBAnet outperforms the state-of-the-art video SR algorithms by ~0.5dB gains (vs. BasicVSR++) at least and 1.489dB gains (vs. EDVR) at most. In Fig. 6a, it is clearly observed that visualization results of EDVR, BasicVSR and BasicVSR++ produce blurry details of the building, while our proposed FBAnet reconstructs realistic and sharp textures.

**SISR vs. Burst SR:** To verify the benefits brought by burst SR data, we provide comparisons under the real-world SISR task. The compared methods are two representative real-world SISR methods (i.e., LP-KPN [4] and CDC [30]) and a Transformer-based SISR method (i.e., SwinIR [18]). Those SISR methods only take the base frame of burst sequences as input. Compared to MFSR methods, SISR methods are characterized by generating relatively sharp and clean outputs, which could be observed from Fig. 6a, while suffering from the absence of informative details.

### 5.3. Evaluation and Analysis

**Alignment:** We have ablatively investigated the homography alignment module and also compared it with other different alignment methods including flow-based alignment [1, 2] and deformable-based alignment [27, 24]. As shown in Tab. 4, compared with optical flow alignment [1] and deformable convolutional alignment [27, 20], our approach outperforms them with performance improvements by 0.155dB.
After Alignment Before Alignment

Figure 7: Three examples on pixel shift among a sequence of 14 burst images before and after our homography alignment. Each color indicates one frame in a sequence.

Figure 8: Visualization of affinity maps and affinity difference maps.

and 0.230dB in PSNR, respectively. This demonstrates that our method is effective to align the pixel shifts in real-world burst frames, even though it is simple.

To verify our solution, we analyze motion patterns among a sequence of burst images. Taking the base frame as reference, pixel shifts of each image in a sequence are image-dependent and global-structural, Fig. 7. Namely, they present a relatively consistent displacement of pixels. This evidences that it is reasonable and effective to align images via homography matrix in the real-world burst super-resolution task.

This is different from many existing burst algorithms that usually adopt pixel-wise alignment methods (e.g., optical flow and deformable convolutions), which rarely consider the image-wise structural motion pattern prior of the frame.

Federated Affinity Fusion: In Tab. 4, we have evaluated the proposed federated affinity fusion module. In comparison with VAF using only affinity maps, our FAF introduces affinity difference maps and achieves the performance gains by 0.288dB in PSNR. And our FAF* further improves the performance by 0.185dB gains in PSNR. This indicates that our federated affinity fusion provides complementary information to the subsequent module.

To further analyze FAF, Fig. 8 provides the visualization of affinity maps and affinity difference maps. As discussed in Sec. 4.3, the affinity values in the flat region with few details would be rather large. Since VAF takes the affinity of one frame to base frame as fusion weight, this encourages the model to pay more attention to those easy reconstruction regions. Instead, FAF uses the affinity difference maps to lower their weights to alleviate this negative effect.

Besides, for the difference map of Frame2 in Fig. 8, it could be seen that the highlighted attention is different from that of Frame1 and Frame3, indicating that Frame2 also provides additional details to the fusion process. This can be further validated through the presented residual between HR predictions of FAF and VAF, which demonstrates that our FAF achieves better detail reconstruction than VAF, as highlighted in the prediction difference image.

Burst Representation Decoding. In Tab. 4, we compare our decoding module to that of BSRT with a Transformer design, under a similar architecture with the same alignment and FAF modules. Our decoding has achieved gains by 0.122dB.

The Number of Burst Image Inputs. We investigate the impact of different numbers of burst images in a sequence and compare it with a single-frame baseline. And all the training processes are based on our proposed architecture.

<table>
<thead>
<tr>
<th>Burst Inputs Number</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
<th>LPIPS ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.139</td>
<td>0.879</td>
<td>0.132</td>
</tr>
<tr>
<td>2</td>
<td>30.616</td>
<td>0.891</td>
<td>0.113</td>
</tr>
<tr>
<td>4</td>
<td>30.818</td>
<td>0.894</td>
<td>0.108</td>
</tr>
<tr>
<td>8</td>
<td>30.945</td>
<td>0.898</td>
<td>0.101</td>
</tr>
<tr>
<td>10</td>
<td>30.980</td>
<td>0.899</td>
<td>0.098</td>
</tr>
<tr>
<td>14</td>
<td>31.012</td>
<td>0.898</td>
<td>0.102</td>
</tr>
</tbody>
</table>

Table 5: Evaluation on the number of burst inputs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Burst Inputs Data</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
<th>LPIPS ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBSR [1]</td>
<td>(base frame)×14 14 burst images</td>
<td>29.389</td>
<td>0.867</td>
<td>0.150</td>
</tr>
<tr>
<td>MFIR [2]</td>
<td>(base frame)×14 14 burst images</td>
<td>29.325</td>
<td>0.865</td>
<td>0.151</td>
</tr>
<tr>
<td>BIPNet [8]</td>
<td>(base frame)×14 14 burst images</td>
<td>30.001</td>
<td>0.878</td>
<td>0.136</td>
</tr>
<tr>
<td>BSRT [20]</td>
<td>(base frame)×14 14 burst images</td>
<td>29.501</td>
<td>0.869</td>
<td>0.151</td>
</tr>
<tr>
<td>FBAnet (Ours)</td>
<td>(base frame)×14 14 burst images</td>
<td>30.086</td>
<td>0.868</td>
<td>0.152</td>
</tr>
<tr>
<td>FBAnet (Ours)</td>
<td>14 burst images</td>
<td>30.866</td>
<td>0.897</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>14 burst images</td>
<td>31.012</td>
<td>0.898</td>
<td>0.102</td>
</tr>
</tbody>
</table>

Table 6: Evaluating the burst inputs’ complementary content.
Tab. 5 reveals that there has been a giant gap in the performance between the single-image baseline and multi-frame restoration results. Specifically, with the burst size increasing from 2 to 14, the performance also experiences a marked rise from 30.616dB to 31.197dB, which tends to be relatively saturated with the input number being close to 1.

**The Complementary Content of Burst Image Inputs.** To verify the influence of contents in burst frames, we compare models trained on 14 shifted frames with models trained on 14 identical images (i.e. the base frame and its 13 copies), the results of which are reported in Tab. 6. For the five models (i.e. DBSR, MFIR, BIPNet, BSRT, and Ours) adopted, the performance gains among (base frame)×14 and 14 burst images range from 0.664dB to 1.194dB, which proves the necessity and effectiveness of complementary information provided by sub-pixel information among shifted frames.

### 6. Conclusions, Limitations, and Future Work

We release a real-world burst image super-resolution dataset, named RealBSR, which is expected to facilitate exploring the reconstruction of more image details from multiple frames for realistic applications, and a Federated Burst Affinity network (FBAnet), targeting addressing the fusing issue of burst images. Specifically, our FBAnet employs simple homography alignment from a structural geometry aspect, evidenced by the relatively consistent pixel shift for a sequence of burst images. Then, a Federated Affinity Fusion (FAF) strategy is proposed to aggregate the complementary information among frames. Extensive experiments on RealBSR-RAW and RealBSR-RGB datasets with improved performance have justified the superiority of our FBAnet.

**Limitations and future work:** Our FBAnet employs a simple homography alignment. But it is not easy to extend to the video SR task with large motions, which will be addressed in our future work. Since noise is inevitable, addressing real-world burst super-resolution and denoising at the same time is more practical. We will be devoted to this real-world benchmark and the solutions in future work.

### Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant No. U21A20470, and National Key R&D Program of China under Grant No. 2021ZD0111601. Thanks a lot for the valuable help from Xiaoxiao Sun.

### References


In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017. 2


