Adaptive Feature Consolidation Network for Burst Super-Resolution

Nancy Mehta[1] Akshay Dudhane[2] Subrahmanyam Murala[1]
Syed Waqas Zamir[3] Salman Khan[2,4] Fahad Shahbaz Khan[2,5]
[1]CVPR Lab, Indian Institute of Technology Ropar [2]Mohamed bin Zayed University of AI
[3]Inception Institute of Artificial Intelligence [4]Australian National University [5]Linköping University

Abstract

Modern digital cameras generally count on image signal processing (ISP) pipelines for producing naturalistic RGB images. Nevertheless, in comparison to DSLR cameras, low-quality images are generally output from portable mobile devices due to their physical limitations. The synthesized low-quality images usually have multiple degradations - low-resolution owing to small camera sensors, mosaic patterns on account of camera filter array and sub-pixel shifts due to camera motion. Such degradation usually restrain the performance of single image super-resolution methodologies for retrieving high-resolution (HR) image from a single low-resolution (LR) image. Burst image super-resolution aims at restoring a photo-realistic HR image by capturing the abundant information from multiple LR images. Lately, the soaring popularity of burst photography has made multi-frame processing an attractive solution for overcoming the limitations of single image processing. In our work, we thus aim to propose a generic architecture, adaptive feature consolidation network (AFCNet) for multi-frame processing. To alleviate the challenge of effectively modelling the long-range dependency problem, that multi-frame approaches struggle to solve, we utilize encoder-decoder based transformer backbone which learns multi-scale local-global representations. We propose feature alignment module to align LR burst frame features. Further, the aligned features are fused and reconstructed by abridged pseudo-burst fusion module and adaptive group upsampling modules, respectively. Our proposed approach clearly outperforms the other existing state-of-the-art techniques on benchmark datasets. The experimental results illustrate the effectiveness and generality of our proposed framework in upgrading the visual quality of HR images.

1. Introduction

Super-resolution (SR) is a long standing research problem, intended to synthesize high-resolution (HR) image given low-resolution (LR) input. Depending on the number of LR inputs, super-resolution has been divided into two main categories - single image super-resolution (SISR) and multi-frame super-resolution (MFSR). SISR is the task of generating HR image using a single LR image. Numerous methods have been developed to solve the SISR problem [32,38]. However, the major hurdle lies in synthesizing high-frequency details in a single input image, consistent to the ground-truth HR image.

On the other hand, MFSR seeks the reconstruction of HR image by employing numerous degraded LR images of a scene. Critically, capturing LR images under the burst mode results in sub-pixel shifts [33] among the multiple LR burst images and thereby, generates different LR samplings of the underlying scene. However, the process of burst image acquisition brings its own issues. For example, during image capturing, any slight movement in scene objects and/or scene objects arises misalignment issues, thereby generating blurring and ghosting artifacts in the reconstructed image [34]. The existing MFSR approaches utilize pre-trained flow computation [5] or optical-flow [19] for aligning the multi-frame features. This explicit feature alignment causes the resulting errors in the flow estimation stage to be propagated to the image processing and warping stages, thereby negatively affecting the generated outputs.

To mitigate the aforementioned problems, we propose an Adaptive Feature Consolidation Network (AFCNet) for Multi-Frame Super-Resolution. The proposed AFCNet comprises of four steps: 1) Feature alignment, 2) Feature extraction, 3) Feature fusion and 4) Feature up-sampling. The features of RAW burst images are initially aligned through deformable convolution [39] followed by feature back-projection approach. This implicit feature alignment limits the error propagation inherent in cascaded explicit alignment approaches [5, 19]. Further, the aligned representations of each burst image are passed through a feature extractor [36] to extract multi-scale local-global representations. The feature fusion mechanism enables the inter-frame communication via abridged pseudo-burst generation such that each and every feature in the pseudo-burst encloses complimentary properties of all input burst images. Further-

more, we adopt an adaptive group up-sampling module [12] to select the reliable and desired information content from each burst image and thus obtain the high-quality HR result.

On account of above modules, our framework efficiently merges the image contents among multiple burst LR RAW frames in a coherent and effective way, generating HR RGB outputs with realistic textures and additional high-frequency details. Highlights of the proposed approach are outlined as follows:

1. We propose a simple but effective feature alignment module to align the burst image features with the base frame.

2. We utilise encoder-decoder based transformer backbone for feature extraction to enrich the aligned feature representations.

3. An efficient abridged pseudo-burst fusion module is utilized to aid inter-frame information exchange and feature consolidation.

4. Finally, adaptive group up-sampling is performed for progressive fusion and up-scaling of the burst features.

Comprehensive experiments have been performed on NTIRE-21 [3] and NTIRE-22 [4] synthetic as well as real-world benchmark datasets to validate the proposed AFCNet for burst SR. Our proposed methodology exemplifies favourable SR performance on real-world bursts, notably outperforming state-of-the-art (SOTA) techniques in a user study. Furthermore, we layout a detailed ablation study, for scrutinizing the influence of basic modules of the proposed AFCNet framework.

2. Related work

In this section, a detailed discussion of the existing approaches for multi-frame super-resolution, feature alignment, attention mechanism and upsampling techniques have been accomplished.

2.1. Multi-frame super-resolution

Compared with SISR, MFSR encounters new challenges while estimating the offsets among different images resulting from moving objects and camera movement. Tsai and Huang [30] were the first to put forward a frequency-domain based solution for MFSR problem. Since, frequency-domain resulted in visual artifacts while processing the images, Irani and Peleg [17] and Peleg *et al.* [26] proposed an iterative back-projection approach for sequentially estimating the HR image. Subsequent works [2, 13, 15] improved this approach with maximum a posteriori (MAP) model. Farsui *et al.* [14] proposed a hybrid method for performing demosaicking and super-resolution with MAP framework. Wronski *et al.* [35] proposed a MFSR algorithm that merges burst of raw images for supplanting the requirement of demosaicking in camera pipeline. Recently, few works resorted to incorporate deep learning for handling the MFSR problem. Deudon *et al.* [9] proposed HighRes-net, the first deep learning MFSR approach in satellite imagery, capable of learning all its subtasks in an end-to-end fashion. Molini *et al.* [25] designed a novel CNN-based algorithm for exploiting both temporal and spatial correlations to combine multiple images. Bhat *et al.* [6,7] addressed the problem of real-world MFSR from any handheld camera by attention-based fusion mechanism.

2.2. Feature Alignment

The major concern of multiple frames lies at solving the misalignment problem among multiple frames. Optical flow has been deployed in [8] for estimating the motion between frames. Working towards this realm, [5] made use of PWC-Net [28] as the flow estimator on account of its high accuracy and speed. Additional studies accomplishes implicit motion compensation by utilising deformable convolutions or dynamic filtering. Recently, [29, 31] used deformable convolution for aligning neighboring frame features with the current frame for the task of MFSR. Deformable convolution is quite effective while handling misalignment between inter-frames and addresses the problem arising out of explicit motion alignment approaches. In this direction, Dudhane *et al.* [12] proposed edge boosting feature alignment which enhances the initial feature representations using attention based feature processing module followed by deformable convolutions and edge boosting mechanism. In the proposed AFCNet, inspired from [12], we make use of deformable convolutions and back-projection mechanism for feature alignment. Unlike [12], we designed a lighter alignment module by eliminating the feature processing unit [12]. Instead, we achieve our target of feature consolidation in the subsequent stages.

2.3. Attention Mechanism

In existing literature, capturing long-range pixel dependencies for extracting global scene properties, has proved to be helpful for a wide range of image restoration tasks [37] (e.g., extreme low-light image enhancement [1] and image/video super-resolution [24]). Existing EDVR [31] and EBSR [23] make use of attention, non-local operation for fusing the information between multiple frames. Further, BIPNet [12] is composed of global context attention mechanism to refine the burst features. In recent times, self attention has shown its effectiveness in a variety of vision applications [18]. In particular, [20, 36] deployed multi-head attention for improving the final restoration results. In the same spirit, the proposed AFCNet makes use of encoder-

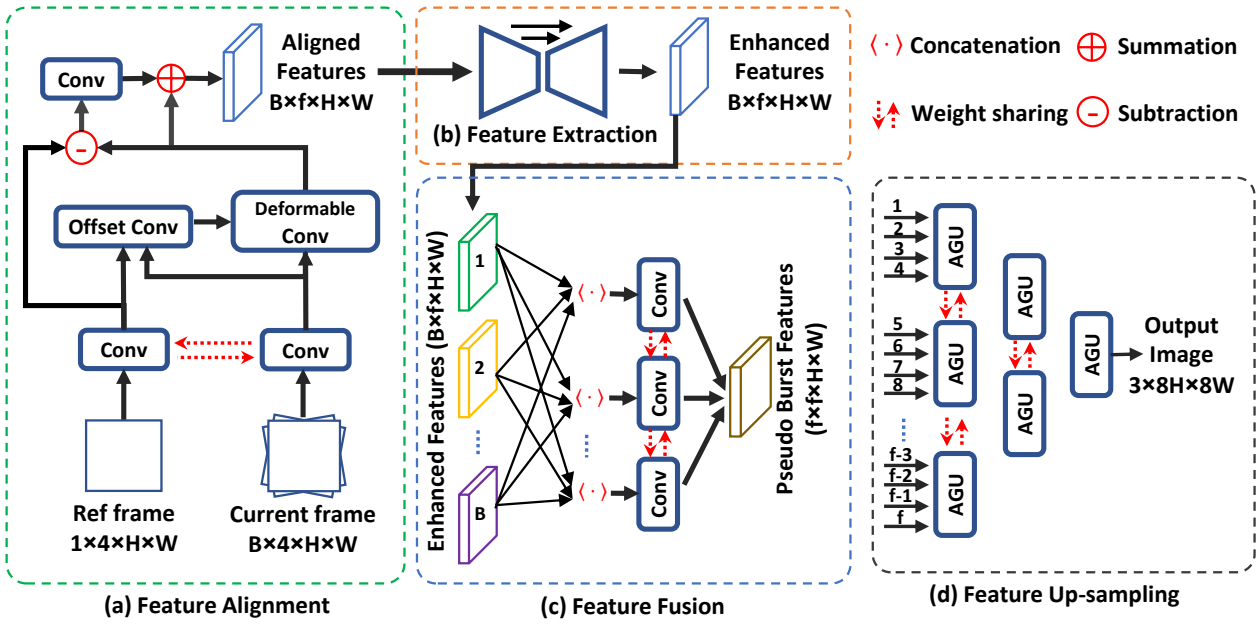

Figure 1. Overall pipeline of the proposed adaptive feature consolidation network (AFCNet) for burst SR. The proposed AFCNet processes input RAW burst image and generates a HR RGB image. It is divided into four parts: (a) Feature alignment module aligns the burst features with respect to the reference frame, (b) Feature extraction module extracts multi-scale local-global representations, (c) Feature fusion module enables the inter-frame communication and integrates the learned burst features to obtain the pseudo-bursts representation and (d) Feature up-sampling module performs adaptive and progressive weighted up-sampling on pseudo-bursts to produce HR RGB image.

decoder based transformer backbone [36] to capture the multi-scale local-global features and thus to improve the overall representation ability of the aligned features.

2.4. Image Upsampling

Image upsampling deals with resizing of input features, and is widely deployed in several image-related applications. The traditional interpolation techniques incorporate nearest neighbor, bicubic and bilinear interpolation. On account of the easy implementation of these methods, they are still quite a popular choice in various CNN-based SR models. Lately, learning-based upsampling methods are introduced into the SR field. Transposed Convolution [11] performs upsampling by a transformation opposite to normal convolution. Pixelshuffle [27] a learn-able upsampling layer, generates plurality of channels by first employing convolution and then reshaping them. In MFSR, adaptive group up-sampling [12] is proposed to handle the pseudo-burst features in groups and progressively perform feature up-sampling.

3. Proposed Method

On account of the rapid capture of images in a burst from a hand-held device, they inherit minute inter-frame offsets. This creates multiple aliased versions of the same scene, thus generating additional signal information for SR. Our proposed AFCNet processes multiple noisy, RAW, LR images to consider the merit of this shifted complementary information from multiple images and combines the information from individual LR images for generating HR RGB image as output. Our first challenge lies in alignment of the slight mismatches between multiple supporting frames and the reference frame. Following this, effective merging of the aligned features is equally important along with the reconstruction of HR image. In subsequent sub-sections, different modules of the proposed AFCNet are discussed.

3.1. Feature Alignment

The major hurdle in burst SR is the unknown inter-frame sub-pixel displacement. This displacement, stemming from camera motion and scene variations, results in misalignment among the frames [5]. Thus, to align the burst features with the reference frame, we utilized modulated deformable convolutions [39] as shown in Figure 1(a). Considering, $\{\boldsymbol{x}^b\}_{b\in[1:S]} \in \mathbb{R}^{S\times n\times H\times W}$, as an initial representation of burst having S images and n number of feature channels. Currently, each frame feature x^b is concatenated with the reference frame feature x^{b_r} and passed through convolution layer to get the offsets and modulated scalars required for the deformable convolution layer. With the obtained offsets and modulated scalars, burst features x^b are processed through modulated deformable convolutions which returns the aligned burst features $\bar{x}^b$.

Our alignment module consists of three deformable layers for improving the overall alignment capability to enhance the aligned burst features. Unlike [12], we processed and aligned the burst features without any pre-processing. We combine it with the feature extraction module where we compute the local-global feature representations. This reduces the extra overhead on feature alignment module and simplifies the overall architecture. Further, high-frequency residue is evaluated by calculating the difference between these aligned features and reference frame features followed by its addition to the aligned features [12] to enhance the high-frequency edge information.

3.2. Feature Extraction

For further strengthening the feature alignment and to rectify small misalignment errors, we utilize a encoder-decoder based transformer backbone (EDTB) [36] for capturing global context information among various frames. Unlike [12], which employ feature refinement module to capture long-range dependencies for modelling global scene properties prior to aligning the features, we leverage a EDTB, after the aligned features as depicted in Figure 1(b). EDTB processes the aligned features $\bar{x}^b$ and returns its enriched representation y^b. Following [36], we employ a 4-level encoder-decoder architecture with number of transformer blocks as [4, 6, 6, 8], attention heads in multi-head attention block are set to [1, 2, 4, 8], and number of channels are [64, 128, 256, 512], respectively.

3.3. Feature Fusion

For generating a merged feature embedding of the enriched aligned features, we designed an abridged pseudo-burst fusion (APBF) module inspired from [12]. It is a well proven fact that simple pooling operations like element-wise average or max pooling across the burst frames generates dissatisfying results [5]. The major reason tends to attribute towards the fact that fusion module requires adaptive merging on the basis of image content and noise levels. Furthermore, considering the benefits, and the indispensable role of inter-frame communication among the channels with multi-path network layout, for fusing the multi-frame features. We, thereby accomplish inter-frame connections through concatenation of the corresponding channel-wise burst feature maps and attain corresponding pseudo-bursts [12] as shown in Figure 1(c). Given the refined features set $y = \left\{\boldsymbol{y}_c^b\right\}_{c\in[1:n]}^{b\in[1:S]}$ of burst size S and n number of channels, the pseudo-burst is generated through,

$$\boldsymbol{P}^c = W^\rho\left(\left\langle \boldsymbol{y}_c^1,\ \boldsymbol{y}_c^2,\ \cdots,\ \boldsymbol{y}_c^S \right\rangle\right), \quad s.t. \quad c \in [1:n], \tag{1}$$

where, $\langle\cdot\rangle$ represents feature concatenation, $\boldsymbol{y}_c^1$ is the c^{th} feature map of 1^{st} aligned burst feature set $\boldsymbol{y}^1$, W^ρ denotes the convolution layer with f output channel, and $\boldsymbol{P} = \{\boldsymbol{P}^c\}_{c\in[1:n]}$ represents the pseudo-burst of size $n \times n \times H \times W$. We have set n = 64 for this module.

Currently, every feature map in the pseudo-bursts embrace complimentary information from all the actual burst frame features. Apart from simplifying the learning task, the inter-frame feature representation merges the required information through decoupling of the burst feature channels. In [12], the aligned burst features are used to obtain the pseudo-bursts followed by the multi-scale feature extractor (encoder-decoder sub-module). In the proposed AFCNet, we abridge this process and directly process the set of enriched features obtained from feature extraction stage (EDTB module) to obtain pseudo-bursts. The proposed abridged pseudo-burst fusion (APBF) scheme serves the dual benefits of, (1) merging the consolidated feature information, and (2) avoiding the computational overhead of processing pseudo-bursts through heavy multi-scale module which is happening in [12].

3.4. Feature Up-sampling

The final step for reconstructing HR image is up-sampling. In AFCNet, we utilized the adaptive group up-sampling (AGU) [12] to reconstruct the HR details shown in Figure 1(d). AGU takes the feature maps (P^c) produced by the abridged pseudo-burst fusion module as input and generates a super-resolved output via three-level progressive upsampling. In AGU, the pseudo-burst features are sequentially divided into groups of 4. Being mindful of the benefits of applying different fusion weights to texture-less and edge regions, we ought to predict the fusion weights through an attention mechanism. To do so, we initially obtain a dense attention map for each pseudo-burst and subsequently apply element-wise multiplication with the corresponding dense attention map. This adaptively rescaled feature response is further passed through transposed convolution layer to up-sample and thus reconstruct the final HR image.

Since, for burst SR we need to perform $\times 8$ up-sampling[1], we perform three levels, with each level performing up-sampling ($\times 2$). As we have 64 pseudo-bursts, for three levels of AGU, naturally it forms a group of 16, 4, 1 pseudo-bursts group.

4. Experiments

We evaluate the proposed AFCNet for both synthetic as well as real burst SR task. We follow the NTIRE-21 [3] and NTIRE-22 [4] competition guidelines to carry out network training and testing.

[1]The real task is to perform upsampling by $\times 4$, additional $\times 2$ is on account of mosaicked RAW LR frames.

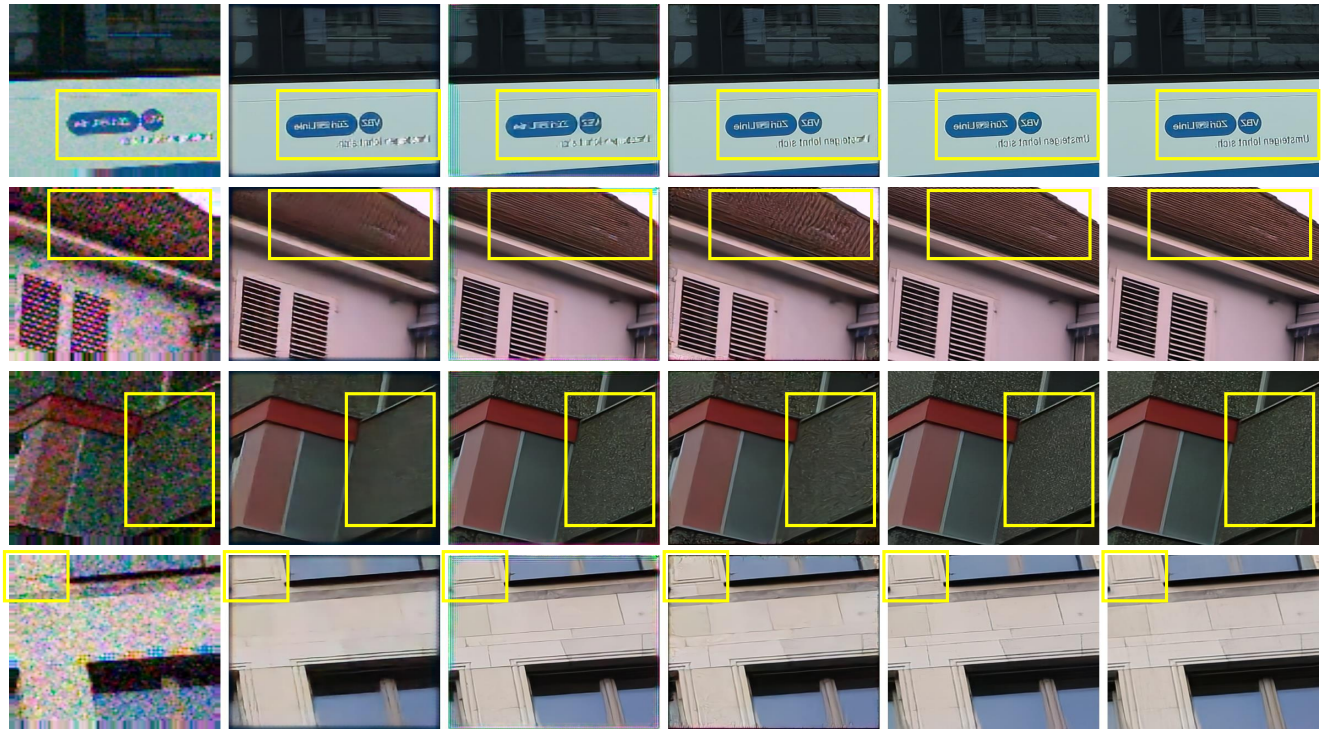

Figure 2. Comparisons for $\times 4$ burst super-resolution on SyntheticBurst dataset [3] (NTIRE-21 Track 1). Our AFCNet produces much cleaner and sharper results than other competing approaches (specifically the marked yellow box regions).

Figure 3. Comparisons on SyntheticBurst dataset [4] for $\times 4$ burst super-resolution (NTIRE-22 Track 1). First and second rows depicts results of base frame up-scaled using bilinear interpolation and the proposed AFCNet respectively.

4.1. Implementation details

Our AFCNet is a single end-to-end trainable network designed for burst SR and requires no pre-training of the proposed module. For overall network efficiency, all burst frames have been processed through shared AFCNet modules. AFCNet has been trained for 100 epochs on synthetic bursts generated by utilising 46,839 sRGB images from Zurich-RAW-to-RGB dataset [16]. We train AFCNet for burst SR task using L_1 loss only. While for real burst SR, we fine-tune our AFCNet with pre-trained weights on SyntheticBurst dataset using aligned L_1 loss [7]. The models are trained with Adam optimizer. Cosine annealing strategy [22] is deployed for steadily decreasing the learning rate from 10^{-4} to 10^{-6} during training. We augment our dataset using horizontal and vertical flips. It should be noted that unlike [23], we have not employed any kind of ensemble techniques to boost the evaluation metrics.

4.2. SyntheticBurst dataset (NTIRE-21 Track 1)

It consists of 300 RAW bursts for validation. Each burst contains 14 LR RAW images (each of size 48×48 pixels) that are synthetically generated from a single sRGB image [5]. Table 1 shows the quantitative evaluation on SyntheticBurst dataset [3]. Also, we have shown the visual comparison between the proposed and existing SOTA methods

Figure 4. Visual Comparisons for $\times 4$ burst super-resolution on Real BurstSR dataset [3] (NTIRE-21 Track 2). Our AFCNet generates crisper and sharper results than other competing techniques.

for $\times 4$ burst SR task in Figure 2. From Table 1 and Figure 2, it is clear that the proposed AFCNet outperforms other existing SOTA methods for $\times 4$ burst SR task.

4.3. Real BurstSR dataset (NTIRE-21 Track 2)

It consists of 5,405 and 882 patches for training and validation, respectively cropped from 200 real RAW bursts images. Each input crop has a size of 80×80 pixels. As shown in Table 1, the proposed AFCNet performs favorably well when compared to the other existing SOTA for $\times 4$ real burst SR task. Also, Figure 4 demonstrates that HR images produced by the AFCNet for $\times 4$ are sharper with vivid details as compared to the other existing SOTA.

4.4. SyntheticBurst dataset (NTIRE-22 Track 1)

It consists of 100 and 92 RAW bursts in validation and test set respectively. Each RAW burst contains 14 LR RAW images (each of size 256×256 pixels) synthetically synthesized from a single sRGB image [5]. Table 2 summarises the quantitative evaluation on validation and test dataset of the proposed AFCNet in comparison with the baseline approach on SyntheticBurst dataset [4]. While Figure 3 display the visual results produced by the proposed AFCNet for RAW bursts from validation set. Figure 3 shows the ability of the proposed AFCNet in producing HR images with enriched details. We have not fine-tuned the proposed AFCNet for this experiment and we directly tested the network trained on the training set as discussed in Section 4.1.

4.5. Ablation Study

In this section, we demonstrate the importance of each module in our proposed AFCNet. Commencing from our base network, we introduce different network models systematically, and exhibit its performance for burst SR application. Every network combination has been trained for 100 epochs on the training set discussed in 4.1. Table 3 marks all the ablation experiments conducted for $\times 4$ burst SR task on validation set of Zurich-RAW-to-RGB dataset [16]. For the baseline model, we employ Resblocks [21] as our feature extraction module, simple concatenation operation has been deployed as a fusion module, and we used transposed convolution for upsampling. The baseline network obtains 36.38 dB PSNR. After appending the proposed modules to the baseline, their seem to be a significant and consistent improvement in results. For example, inclusion of alignment module and back projection approach improves the PSNR by 2.54 and 0.58 dB respectively. While, feature extraction stage which is composed of EDTB [36] achieves

Methods	SyntheticBurst		(Real) BurstSR	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Single Image	36.17	0.91	46.29	0.982
HighRes-net [10]	37.45	0.92	46.64	0.980
DBSR [5]	40.76	0.96	48.05	0.984
LKR [19]	41.45	0.95	-	-
MFIR [7]	41.56	0.96	48.33	0.985
BIPNet [12]	41.93	0.96	48.49	0.985
AFCNet (Ours)	42.21	0.96	48.63	0.986

Table 1. Performance assessment on SyntheticBurst and real BurstSR validation datasets (NTIRE-21) [3] for $\times 4$ burst super-resolution.

Methods	Validation set		Test set	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Baseline [12]	42.24	0.97	-	-
AFCNet (Ours)	42.44	0.97	42.08	0.97

Table 2. Performance evaluation on validation and test set of SyntheticBurst dataset (NTIRE-22 Track 1) [4] for $\times 4$ burst super-resolution.

Table 3. Significance of AFCNet modules assessed on SyntheticBurst validation set [3] for $\times 4$ burst SR task.

Modules	A1	A2	A3	A4	A5	A6
Baseline	✓	✓	✓	✓	✓	✓
Alignment (§3.1)		✓	✓	✓	✓	✓
Back-projection (§3.1)			✓	✓	✓	✓
EDTB (§3.2)				✓	✓	✓
APBF (§3.3)					✓	✓
AGU (§3.4)						✓
PSNR	36.38	38.92	39.50	41.20	41.80	42.21

significant gain of 1.70 dB in the performance. Inclusion of APBF module contributes improvement of 0.60 dB whereas, adaptive group up-sampling block takes the gain to 42.21 dB. Overall, our AFCNet attains a captivating performance gain of 5.83 dB over the baseline.

5. Conclusion

In this paper, we propose an adaptive feature consolidation network (AFCNet) for burst super-resolution. The proposed AFCNet is end-to-end trainable with provision for implicit feature alignment mechanism as well as for inter-frame communication. Additionally, it utilizes adaptive group up-sampling technique to progressively up-scale the multi-frame features. With the help of experimental analysis, it is observed that the proposed sub-modules work jointly to reconstruct the high-resolution image with enriched details and thus, outperform other existing SOTA approaches for burst super-resolution task. Meticulously carried out ablation study show significant improvement in the network performance after inclusion of its sub-modules *viz.* feature alignment, feature extraction, fusion and up-sampling modules.

6. Acknowledgement

This work was supported by a grant from the Department of Science and Technology, Government of India, for the Technology Innovation Hub at the Indian Institute of Technology Ropar in the framework of National Mission on Interdisciplinary Cyber-Physical Systems (NM - ICPS).

References

[1] Aditya Arora, Muhammad Haris, Syed Waqas Zamir, Munawar Hayat, Fahad Shahbaz Khan, Ling Shao, and Ming-Hsuan Yang. Low light image enhancement via global and local context modeling. *arXiv:2101.00850*, 2021. 2

[2] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. Motion deblurring and super-resolution from an image sequence. In *European conference on computer vision*, pages 571–582. Springer, 1996. 2

[3] Goutam Bhat, Martin Danelljan, and Radu Timofte. Ntire 2021 challenge on burst super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 613–626, 2021. 2, 4, 5, 6, 7

[4] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2022 burst super-resolution challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2, 4, 5, 6

[5] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021. 1, 2, 3, 4, 5, 6

[6] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021. 2

[7] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2460–2470, 2021. 2, 5, 6

[8] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4778–4787, 2017. 2

[9] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*, 2020. 2

[10] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*, 2020. 6

[11] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 3

[12] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *CVPR*, 2022. 2, 3, 4, 5, 6

[13] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997. 2

[14] Sina Farsiu, Michael Elad, and Peyman Milanfar. Multiframe demosaicing and super-resolution from undersampled color images. In *Computational Imaging II*, volume 5299, pages 222–233. SPIE, 2004. 2

[15] Russell C Hardie, Kenneth J Barnard, John G Bognar, Ernest E Armstrong, and Edward A Watson. High-resolution

image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system. *Optical Engineering*, 37(1):247–260, 1998. 2

[16] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 536–537, 2020. 5, 6

[17] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991. 2

[18] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*. 2

[19] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *ICCV*, 2021. 1, 5, 6

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 2

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 6

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[23] Ziwei Luo, Lei Yu, Xuan Mo, Youwei Li, Lanpeng Jia, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Ebsr: Feature enhanced burst super-resolution with deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 471–478, 2021. 2, 5

[24] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, 2020. 2

[25] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deepsum: Deep neural network for super-resolution of unregistered multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3644–3656, 2019. 2

[26] Shmuel Peleg, Danny Keren, and Limor Schweitzer. Improving image resolution using subpixel motion. *Pattern recognition letters*, 5(3):223–226, 1987. 2

[27] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3

[28] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[29] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020. 2

[30] R Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984. 2

[31] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[32] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European Conference on Computer Vision*, pages 101–117. Springer, 2020. 1

[33] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019. 1

[34] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019. 1

[35] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019. 2

[36] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021. 1, 2, 3, 4, 6

[37] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 2

[38] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 1

[39] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 1, 3