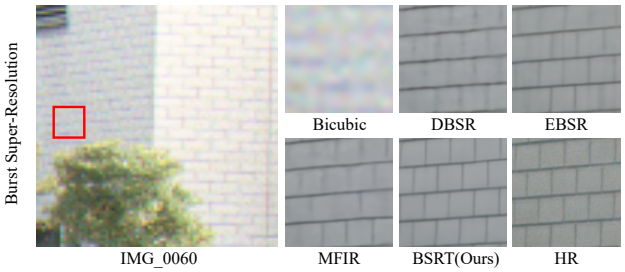
BSRT: Improving Burst Super-Resolution with Swin Transformer and Flow-Guided Deformable Alignment

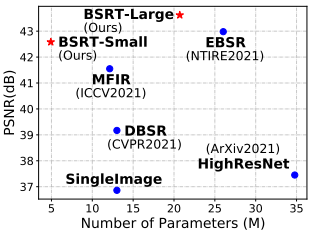
Ziwei Luo [1] Youwei Li [1] Shen Cheng [1] Lei Yu [1] Qi Wu [1] Zhihong Wen [1]
Haoqiang Fan [1] Jian Sun [1] Shuaicheng Liu [2,1*]
[1] Megvii Technology
[2] University of Electronic Science and Technology of China
https://github.com/Algolzw/BSRT

Abstract

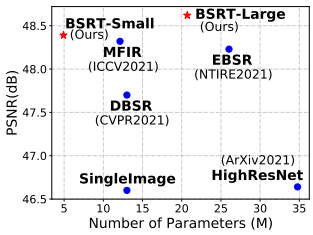
*This work addresses the Burst Super-Resolution (BurstSR) task using a new architecture, which requires restoring a high-quality image from a sequence of noisy, misaligned, and low-resolution RAW bursts. To overcome the challenges in BurstSR, we propose a **B**urst **S**uper-**R**esolution **T**ransformer (BSRT), which can significantly improve the capability of extracting inter-frame information and reconstruction. To achieve this goal, we propose a Pyramid Flow-Guided Deformable Convolution Network (Pyramid FG-DCN) and incorporate Swin Transformer Blocks and Groups as our main backbone. More specifically, we combine optical flows and deformable convolutions, hence our BSRT can handle misalignment and aggregate the potential texture information in multi-frames more efficiently. In addition, our Transformer-based structure can capture long-range dependency to further improve the performance. The evaluation on both synthetic and real-world tracks demonstrates that our approach achieves a new state-of-the-art in BurstSR task. Further, our BSRT wins the championship in the NTIRE2022 Burst Super-Resolution Challenge.*



(a) Visual comparison on BurstSR dataset.



(b) Synthetic dataset.



(c) Real-world dataset.

Figure 1. The comparison between our approach and other representative methods [4, 5, 37] on Synthetic dataset [23] and BurstSR dataset [4]. Our method achieves the best performance while being computationally efficient.

1. Introduction

Multi-frame super-resolution (MFSR) is a fundamental low-level vision problem [2, 4, 14, 54], which aims to restore a high-resolution (HR) image from a sequence of low-resolution (LR) images. Compared to single image super-resolution [15, 26, 33], MFSR approaches are able to aggregate sub-pixel information from multi-frames of the same scene, alleviating the ill-posed problem in super-resolution [29, 54]. But in recent years, the MFSR problem receives less attention than SISR. In this work, we tackle the practical problem of Burst Super-Resolution (BurstSR), in which the inputs are low-resolution RAW snapshots captured from real-world smartphone cameras [4]. These RAW bursts are usually noisy and misaligned, so in order to better extract information from multi-frames to recover high-quality images, we need a more efficient architecture to address these challenges.

The NTIRE2022 (New Trends in Image Restoration and Enhancement) contains the Burst Image Super-Resolution Challenge [3]. The challenge has 2 tracks, the first track is called Synthetic Track and the second track is Real-world Track. In the synthetic track, the input bursts are generated from RGB images using a synthetic data generation pipeline. Meanwhile, in the real-world track, the test set containing bursts captured from a handheld Samsung

*Corresponding author.

Galaxy S8 smartphone camera. The goal in both tracks is to reconstruct the original image as well as possible, and not to artificially generate a plausible, visually pleasing image [2]. This challenge promotes more research on BurstSR.

Some existing BurstSR methods solve this problem with the following steps: feature extraction, feature alignment, fusion and HR image reconstruction [2, 37]. To be more specific, firstly, CNN-based residual blocks are often used in feature extraction and reconstruction [2, 4, 37]. Secondly, both optical flow [40] and deformable convolution network (DCN) [12, 61] can be used to align features of multi-frames. Finally, attention mechanism [48] as well as non-local [49] techniques are widely-used in the fusion step to aggregate information from multiple aligned features. However, a general convolution is a local operator that is ineffective for long-range information interaction [32] and the individual flow/DCN-based alignment is not sufficient to deal with large, complex shifts between frames [10]. Foremost among these problems is that these rudimentary designs limit the efficacy of information aggregation and thus lead to poorer performance in rich details and occluded regions.

In this paper, we propose a **B**urst **S**uper-**R**esolution **T**ransformer (BSRT), which enhances the effectiveness of feature extraction, alignment, and reconstruction in the BurstSR task. The main components of BSRT are the Pyramid Flow-Guided Deformable Convolution Network (Pyramid FG-DCN) and the Transformer-based backbone. Specifically, as shown in Fig. 2, FG-DCN combines optical flow and DCN to predict coarse-to-fine distortion and offset, enabling the network to align images more effectively. Further, we apply a pyramid structure to improve the alignment on the top of the flow-guided DCN. On the other hand, the self-attention mechanism and Transformer have shown promising performance in most computer vision tasks [31, 32, 35]. Therefore, to better use the inter-frame information, we incorporate Swin Transformer blocks and groups in our architecture to capture both global and local contexts for long-range dependency modeling [32, 35].

Based on the aforementioned components, the proposed BSRT achieves an impressive performance and surpasses existing art methods in BurstSR by a large margin. Our approach recovers textures that are more similar to the ground-truth, with a more clear and plausible appearance, while being computationally efficient, as illustrated in Fig. 1. The main contributions are summarized as follows:

- We propose to use SpyNet [40] in BurstSR to obtain pyramid flows between multi-frames, which can guide the DCNs [12] to obtain multi-scale features with better alignment. This design can facilitate a more efficient aggregation of inter-frame information.

- We introduce the Transformer-based backbone into BurstSR task to capture global interactions between contexts, which can further improve the performance.

- Experiments on both synthetic and real-world tracks demonstrate that the proposed BSRT leads to a new state-of-the-art performance in the BurstSR problem. Further, our approach wins the championship in the Real-World track of the NTIRE2022 Burst Super-Resolution Challenge.

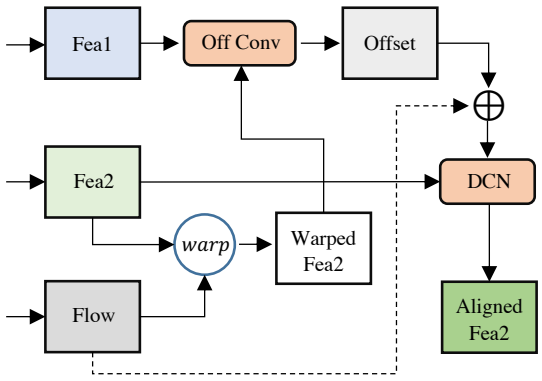


Figure 2. Details of the Flow-Guided Deformable Convolution Network (FG-DCN). There are three inputs: reference feature (*Fea1*), current feature (to be aligned, *Fea2*), and the pre-calculated flow between *Fea1* and *Fea2* from PyNet.

2. Related Work

Single Image Super-Resolution. Single Image Super-Resolution (SISR) is a long standing research topic due to its importance in computer vision. SRCNN [43] is the pioneering deep learning-based method that employs a three-layers-convolutions network and applied the bicubic degradation on HR images to construct HR and LR pairs. Since then, various approaches have been proposed to handle the SISR problem [13, 21, 26–28, 33, 36, 42, 45, 56, 58–60]. For example, VDSR [43] adopted a very deep network to improve performance and ESPCN [42] used an efficient sub-pixel strategy for upsampling. EDSR [33] further enhanced the network by modifying the residual blocks with a non-batchnorm design. Moreover, VGG loss [43], perceptual loss [25], and GAN loss [18] were also used to improve the perceptual visual quality [30, 41, 50]. However, these methods can hardly recover rich details for real-world complex images due to the ill-posed nature of SISR.

Multi-Frame Super-Resolution. To overcome the ill-posed problem in SISR, Multi-Frame Super-Resolution (MFSR) is proposed to aggregate pixels from multiple images of the same scene, which can provide complemen-

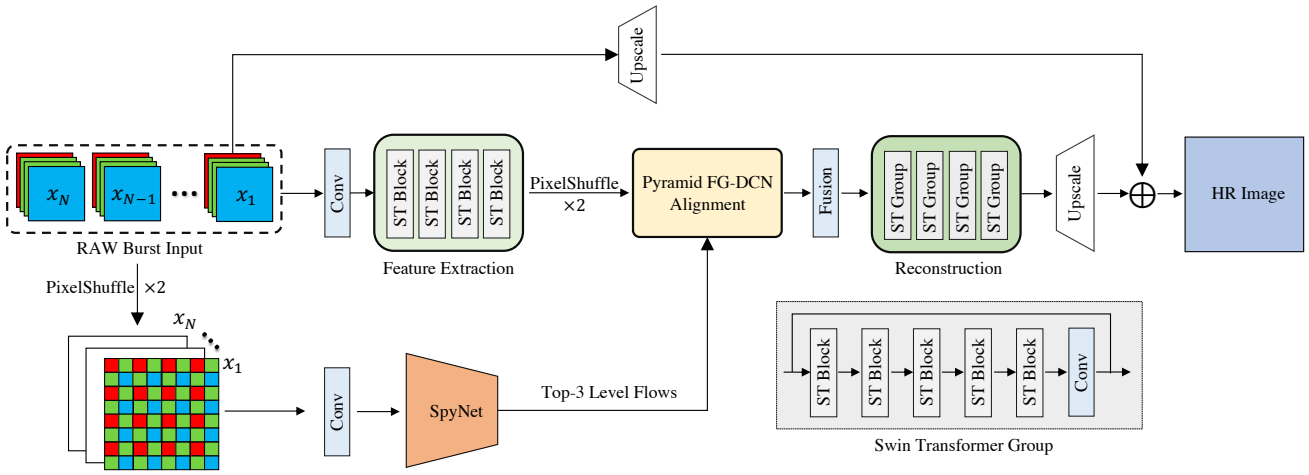


Figure 3. The network inputs a sequence of low-quality RAW images and outputs a high-quality RGB image. Firstly, all RAW inputs are upscaled to 1-channel 'RGGB' format by PixelShuffle and expanded to 3-channels through a 3×3 convolution. Then they are sent to the SpyNet [40] to obtain multi-scale optical flows between each frame and the reference frame. Meanwhile, we extract useful features from original RAW inputs and upscale them before alignment so that we can combine the pre-calculated flows with DCNs on multi-scale features. We fuse these aligned features by a 1×1 convolution and then restore the final HR image.

tary sub-pixel information for a better image reconstruction [19, 46, 47]. MFSR is also well-studied in the last three decades. Traditionally, Tsai and Huang [47] were the first that proposes to perform MFSR in the frequency domain. Peleg et al. [39] and Irani [24] proposed to iteratively minimize the reconstruction error between estimated HR image and the ground truth image. The subsequent works [1, 17, 20] extended it with a regularization term under the maximum a posteriori (MAP) framework.

On the other hand, deep learning based approaches have shown promising performance in processing MFSR problem. DeepSum [38] and HighResNet [14] were proposed for remote sensing applications. Bhat et al. [4] proposed a CNN-based encoder-decoder for RAW burst super-resolution and introduced an attention-based fusion into their network. They then further improved its performance in RepMFIR [5] with a deep reparameterization of the MAP framework. Meanwhile, Lecouat et al. [29] proposed an end-to-end approach for joint image alignment and super-resolution from raw burst inputs. Moreover, in the last NTIRE2021 Burst Super-Resolution Challenge, the winner method EBSR [37] presented a deformable convolution network (DCN) based alignment and non-local based fusion to enhance the performance.

Low-Level Vision Transformer. Attention-based network, i.e., Transformer, have shown great performance and gained much popularity in various high-level computer vision tasks [7, 9, 16, 34, 35, 53, 55]. Recently, Transformer has also been introduced for low-level vision and tends to learn global interactions to focus on enhancing details and important regions [8, 11, 31, 32, 52]. Chen et al. [11] were the first propose to use Transformer-based backbone IPT for various image restoration problems. Liang et al. [32] proposed an efficient structure, SwinIR, for image restoration based on the Swin Transformer [35]. Compared with IPT, SwinIR requires fewer parameters and training datasets and achieves a new art performance in single image super-resolution, JPEG compression artifact reduction and denoising.

3. Method

3.1. Overview of the Framework

The overview of the proposed BSRT framework is shown in Fig. 3. Let $I_{HR} \in \mathbb{R}^{3\times Hs\times Ws}$ be the ground truth HR image (RGB) and $\{x_i\}_{i=1}^{N}$ be the input bursts which are all 4-channels 'RGGB' RAW images (H, W is the image height and width, s is the scale factor, N is the number of bursts, $x_i \in \mathbb{R}^{4\times \frac{H}{2}\times \frac{W}{2}}$). For burst super-resolution task, each low-quality image is obtained by transforming the downsampling the HR image. The overall burst super-resolution problem can be formulated as

$$x_i = (T_i \circ I_{HR})_{\downarrow_s} + \eta_i \text{ for i} = 1, \ldots, \text{N}, \qquad (1)$$

where T_i is a transformation representing the scene motion, i.e., translation and rotation. $\circ$ is the warping operator and $\downarrow_s$ denotes bicubic downsampling. η_i represents some additive noise.

Our goal is to restore a high-quality image I_{SR} from a set of RAW bursts. Firstly, we flatten the inputs to single

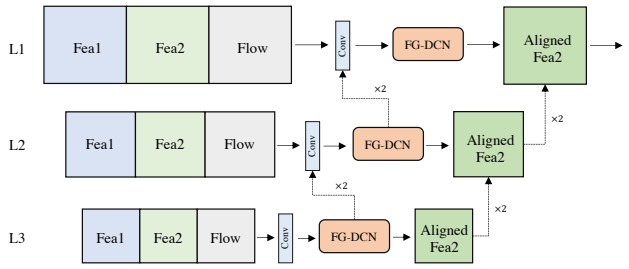


Figure 4. Detailed architecture of the proposed pyramid flow-guided deformable alignment module (Pyramid FG-DCN).

channel and convert them to 3-channels by a 3×3 Conv so that they be sent to the SpyNet to obtain three level optical flows which are calculated from each frame and the reference frame:

$$f_i^1, f_i^2, f_i^3 = L_{\text{SpyNet}}(L_{\text{Conv}}(x_i), L_{\text{Conv}}(x_{\text{ref}})), \quad (2)$$

where f_i^1, f_i^2, f_i^3 are the estimated pyramid flows on each level, L_{SpyNet} and L_{Conv} are the SpyNet and the convolution layer, repectively. Particularly, we use a pre-trained SpyNet and preserve the top-3 levels of flows to guide corresponding level's deformable convolution network (DCN) alignment. Meanwhile, the original 4-channels RAW inputs are sent to several Swin Transformer Blocks (ST Blocks) to extract informative features:

$$F_i = L_{\text{STB}}(x_i),\ F_i \in \mathbb{R}^{C\times\frac{H}{2}\times\frac{W}{2}} \quad (3)$$

where the L_{STB} denotes the ST Blocks and C is the feature channels. We then upscale these features using pixelshuffle [22] to match the sizes of the obtained flows and align them with the reference frame's feature via a pyramid flow-guided deformable alignment module, as shown in Fig. 2 and Fig. 4. After that, we fuse these features (1×1 Conv) to reconstruct the high-resolution image via several Swin Transformer Groups as:

$$I_{HR} = L_{\text{STG}}(L_{\text{Conv1}}(\{AF_i\}_{i=1}^N) \quad (4)$$

where $AF_i \in \mathbb{R}^{C\times H\times W}$ is the i-th aligned feature. L_{STG} and L_{Conv1} are the ST Groups and the 1×1 Conv fusion layer, respectively.

3.2. Pyramid Flow-Guided DCN Alignment

Inspired by BasicVSR++ [10], we combine the flow-based alignment and deformable alignment. Specifically, the pyramid optical flows $\{f_i^1, f_i^2, f_i^3\}_{i=1}^N$ estimated by the SpyNet can be regarded as a coarse alignment prior. Based on these flows, DCNs tend to learn more accurate and refined offsets for aligning features. The details of the Flow-Guided DCN (FG-DCN) are illustrated in Fig. 2. Given feature F_i and the corresponding flow f_i, we can get the coarsely warped feature $\hat{F}_i$ by

$$\hat{F}_i = \mathcal{W}(F_{\text{ref}}, f_i), \quad (5)$$

where $\mathcal{W}$ denotes the wrapping operator. Then we concatenate $\hat{F}_i$ with the reference feature to predict refined local offsets. Subsequently we add the fine offsets with flows as more accurate offsets:

$$\mathcal{O}_i = f_i \oplus L_{\text{offconv}}(\hat{F}_i, F_{\text{ref}}), \quad (6)$$

where $\oplus$ denotes the element-wise sum operator and L_{offconv} represents some convolution layers that predict the offsets. Based on these offsets, we warp the input feature to obtain the aligned feature AF_i through an original DCN alignment module as

$$AF_i = \mathcal{W}(F_i, \mathcal{O}_i). \quad (7)$$

Moreover, we design a 3-levels-pyramid structure to further improve the alignment as shown in Fig. 4. From level-3 to level-1 (L3-L1), the predicted offsets and aligned features are upsampled and subsequently concatenated with the next level's offsets and aligned features. By doing so, we can refine the output feature with multi-scale information and raise superior to noise reduction. In addition, we also add a feature enhancement network in front of the Pyramid FG-DCN model to alleviate the negative effect of noises as in EBSR [37].

3.3. Handling Features with Swin Transformer

To extract useful features and reconstruct high-quality images, we introduce the powerful Swin Transformer [32, 35] as our main backbone, as shown in Fig. 3. Compared to CNN-based structures, transformer is capable of capturing long-range dependencies to aggregate correlated high-frequency information. Inside of a ST Block, it consists of a standard multi-head self-attention (MSA) and a multi-layer perceptron (MLP). The layernorm is also added in front of the MSA and MLP as same as the original Transformer layer [48]. Let $X \in \mathbb{R}^{C\times H\times W}$ be the fused feature of multiple aligned features. The whole process of a ST Block can be formulated as

$$X = MSA(LN(X)) + X \quad (8)$$

$$X = MLP(LN(X)) + X. \quad (9)$$

The ST Group consists of several ST Blocks and a convolution layer (in the last). And the residual connection is also employed in this module.

Following the common practices in super-resolution, we use L1 loss between the restored image and the ground truth HR image as our objective function:

$$\mathcal{L} = ||SR(\{x_i\}_{i=1}^N; \theta) - I_{HR}|| \quad (10)$$

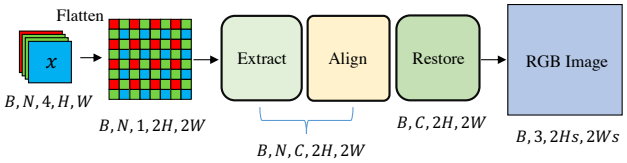
Method	#Parameters	Synthetic dataset			Real-world dataset		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SingleImage [4]	13.01M	36.86	0.919	0.113	46.60	0.979	0.039
HighResNet [14]	34.78M	37.45	0.924	0.106	46.64	0.980	0.038
DBSR [4]	13.01M	39.17	0.946	0.081	47.70	0.984	0.029
EBSR [37]	26.03M	42.98	0.972	0.031	48.23	0.985	0.024
MFIR [5]	12.13M	41.55	0.964	0.045	48.32	0.985	0.023
BSRT-Small(Ours)	4.92M	42.72	0.971	0.031	48.48	0.985	0.021
BSRT-Large(Ours)	20.71M	43.62	0.975	0.025	48.57	0.986	0.021

Table 1. The table shows a comparison between our methods and the other approaches. The best one marks in red and the second best are in blue. Note that the results of SingleImage and HighResNet are reported from [5], and all models for the real-world dataset are first pretrained on the synthetic dataset.

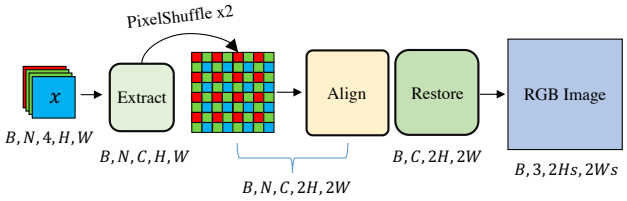
where 'SR' is the whole network, and θ denotes its learnable parameters.

3.4. Pipeline for RAW images

As shown in Fig. 5, we propose a new pipeline for processing misaligned RAW images. Note that EBSR [37] directly flatten the 4-channels RAW inputs (with size $H \times W$) to 1-channel 'RGGB' format (with a larger size $2H \times 2W$) before sending them to the network. Then EBSR performs feature extraction, alignment, fusion, and reconstruction all based on the size $2H \times 2W$. Such a strategy improves the performance but is computationally expensive. In practice, we have noticed that the performance improvement mainly comes from performing alignment and reconstruction on the large size feature maps. Address it, we modify the pipeline to that the feature extraction is applied on the low-resolution space, and scaled $2\times$ before alignment. Compared with EBSR, our approach is effective and computationally efficient, and thus can use a larger patch size and batch size to accelerate training.



(a) RAW processing pipeline of EBSR [37].



(b) New processing pipeline of our approach.

Figure 5. Illustration of the proposed pipeline for processing RAW bursts. In our method, the feature extraction is applied on the low-resolution space, and scaled $2\times$ before alignment, which is effective and computationally efficient.

4. Experiment

4.1. Dataset and Implementation Details

As previous works explored [4,5,37], our method is evaluated on both synthetic and real-world datasets provided by the NTIRE2022 Burst Super-Resolution Challenge [3]. The synthetic dataset [23] contains 46839 cropped RGB images (with sizes fix to 448×448) that are used to synthesize sets of low-quality RAW burst images, with randomly translated and rotated. The noises are also added in the RGB-to-RAW inverse camera pipeline [6]. The real-world dataset contains 5405 real-world RAW burst patches captured by a Samsung Galaxy S8 smartphone, with sizes of 160×160, and the HR images are captured from a DSLR camera. In addition, 300 synthetically generated images (size 96×96) and 882 real-world patches (size 160×160) are used for evaluation, with $4\times$ scaling factor.

4.2. Training and Testing

As a common practice, our model is first trained on the synthetic dataset, then finetuned on the real-world dataset for real-world track. All of the inputs are 4-channels 'RGGB' RAW images, and the outputs are 16-bit RGBs which can be converted to be visually pleasant by the provided post processing scripts. For synthetic training, we optimize the whole model using ℓ_1 loss as introduced in Sec. 3.3. For real-world data training, since the ground truth images are not pre-aligned with any inputs, we use $aligned$ ℓ_1 loss which firstly aligns the ground truth image with the super-resolved image by utilizing a pre-trained PWC-Net [44], and then calculates the ℓ_1 based on the well-aligned images as the same as [2,4]. Note that the proposed BSRT learns the demosaic process implicitly, so that our

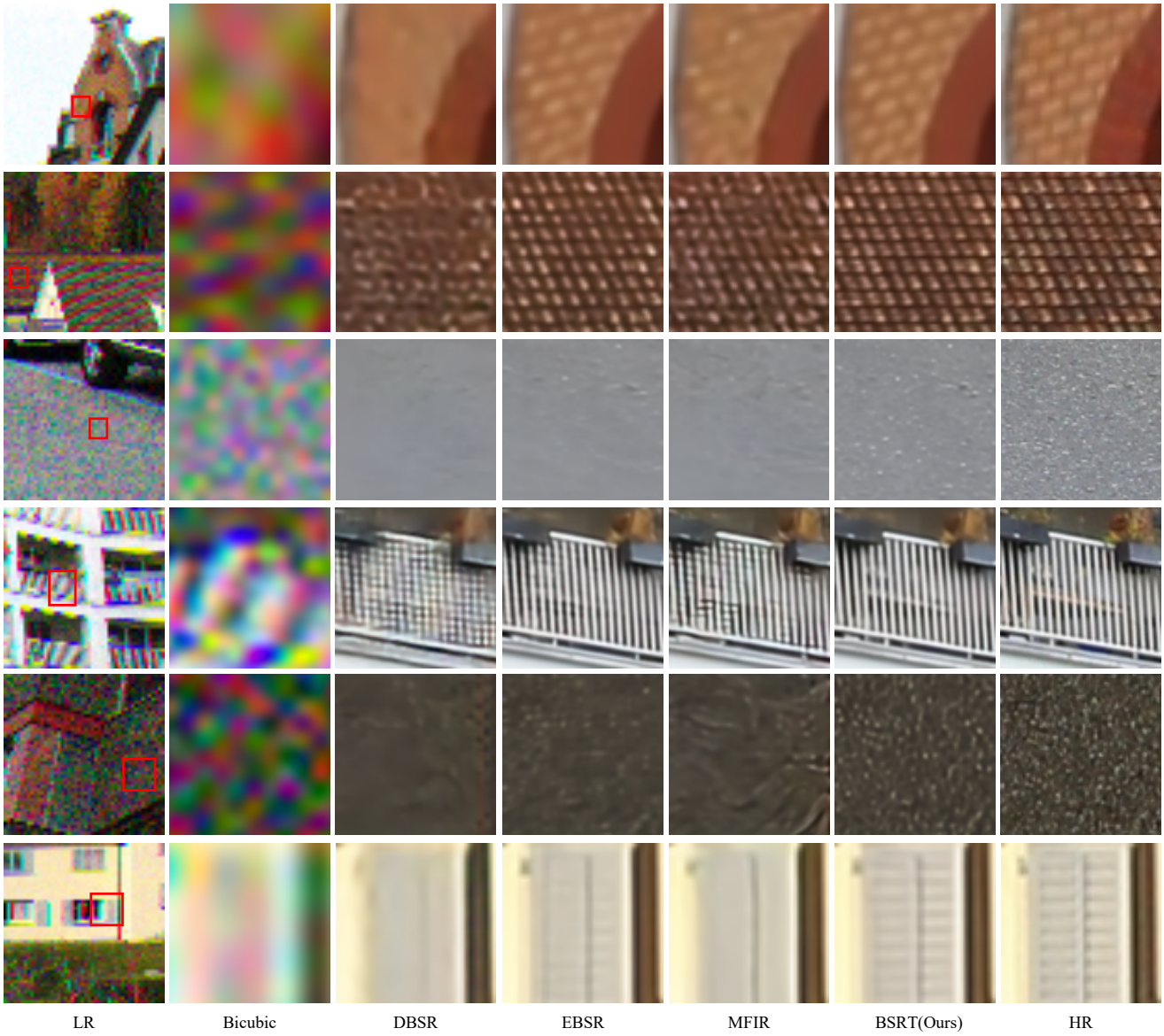


Figure 6. Comparison of our method with other state-of-the-art approaches on synthetic dataset.

network can be trained in an end-to-end manner. For both datasets, we use Adam optimizer and set exponential decay rates as 0.9 and 0.999. The initial learning rate is set to 8×10^{-5} and then reduced to half every 150 epoch. In each training batch, the HR images are cropped to 256×256, then we randomly synthesize 14 burst LR image patches based on the HR image. We implement the proposed BSRT with PyTorch framework and 8 NVIDIA 2080Ti GPUs, taking around 14 days.

In practice, we also find that a large patch size can further improve the performance. So it is better to finetune the trained model with a patch size of 384×384 for HR images. However, we can not train the model on such a large patch size directly due to the limited computing resource and memory, and we choose to freeze the model's weights and only finetune the alignment module and a portion of Conv layers.

4.3. Comparisons with Existing Methods

We compare our method with state-of-the-art BurstSR approaches including HighResNet [14], DBSR [4], EBSR [37] and MFIR [5]. DBSR is the first deep learning-based burst SR method, which uses optical flows to align frames and proposes an attention-based fusion strategy. The

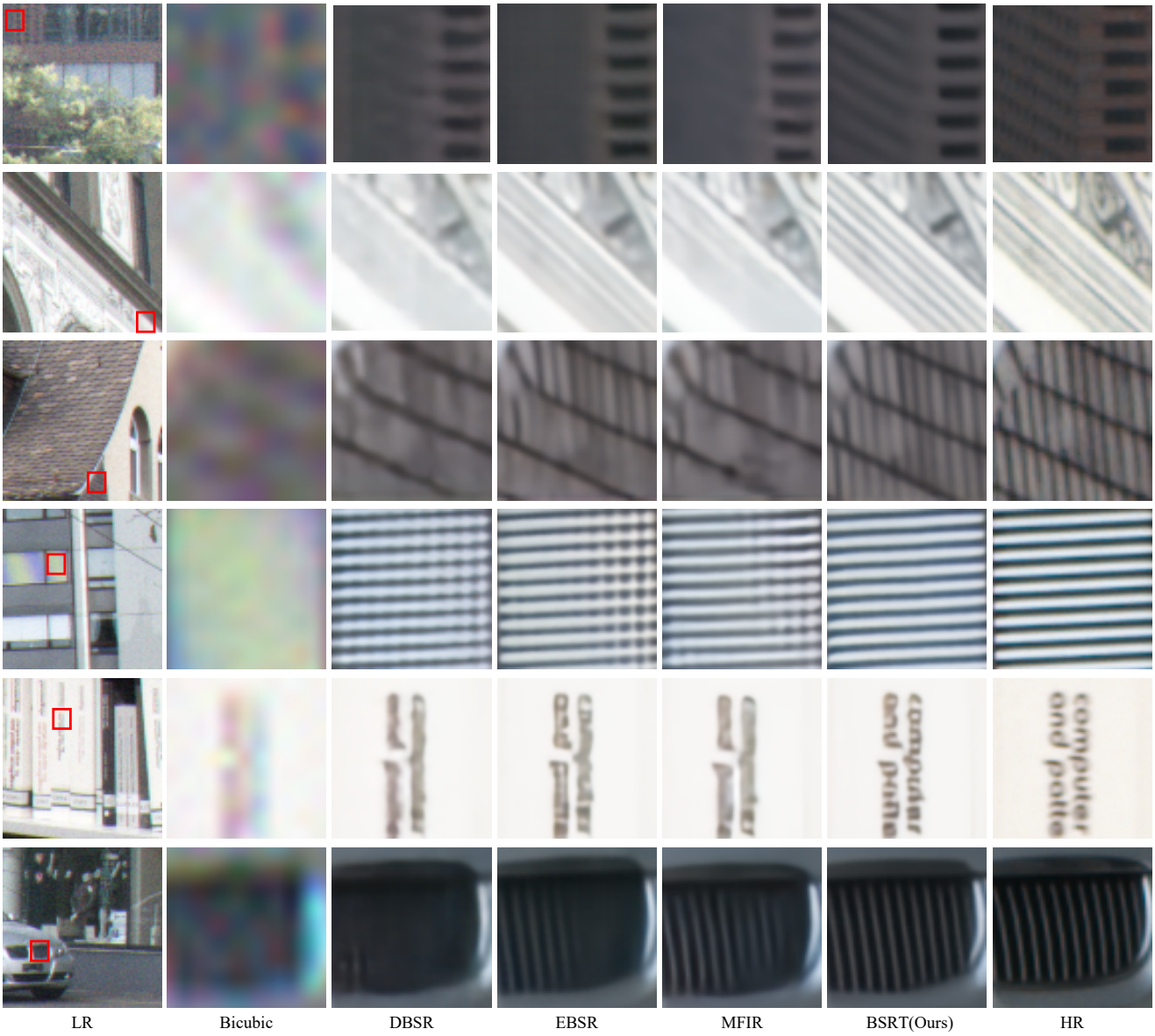


Figure 7. Comparison of our method with other state-of-the-art approaches on real-world dataset.

encoder and decoder networks are employed to extract features and reconstruct HR images. MFIR is the improved version of DBSR, which also incorporates flow estimations to align frames and restores the HR image with an advanced deep reparameterization formulation. EBSR is the winner method in BurstSR Challenge of NTIRE2021 [2], which is a CNN-based restoration network and only utilizes DCN in the alignment. In addition, we also provide a single image method that uses the same architecture as DBSR but with a single RAW image as input. For our approach, we provide two models that have a fewer and greater number of parameters: BSRT_Small and BSRT_Large. We use PSNR, SSIM [51] and LPIPS [57] as the evaluation metrics for a more convincing comparison.

The quantitative results on both datasets are shown in Table 1. As we can see, all multi-frame super-resolution methods perform better than single image method. MFIR [5] outperforms DBSR [4] by 2.3dB and 0.6dB on synthetic data and real-world data, respectively, in terms of PSNR. EBSR [37] achieves an impressive result on the synthetic dataset, but its performance dropped when finetuned on the real-world dataset. Our approach, the BSRT-Large, outperforms all other methods on both datasets by a big margin. And the efficient one, BSRT-Small, also achieves a good

(a)	(b)	(c)	(d)	PSNR↑	SSIM↑	LPIPS↓
✗	CNN	✗	CNN	42.98	0.972	0.031
√	CNN	✗	CNN	43.12	0.972	0.030
√	CNN	√	CNN	43.29	0.973	0.029
√	CNN	√	STG	43.39	0.973	0.027
√	STB	√	STG	43.62	0.975	0.025

Table 2. Ablation studies of the main components on synthetic dataset. (a) Use new pipeline; (b) Network structure in feature extraction; (c) Use Pyramid FG-DCN; (d) Network structure in reconstruction. STB and STG are Swin Transformer blocks and groups, repectively.

performance on the synthetic dataset and outperforms other methods on the real-world images, even if the number of parameters is less than 5M. The visual results on synthetic data and real-world data are shown in Fig. 6 and Fig. 7, respectively. It is obvious that the proposed method produces the best visually pleasant images on both datasets. The proposed BSRT is robust to noises and meanwhile preserves rich details. For example, as shown in the 3rd row and the 5th row of Fig. 6, our method produces clean results while the details are all preserved. In contrast, all other approaches failed to handle the noisy details. Moreover, it can be seen that our method can restore more information from real-world burst images. As illustrated in the second row and the last row of Fig. 7, only our method recovers the whole lines on the wall and in front of the car.

4.4. Ablation Study

In this section, we illustrate the effectiveness of the main components of the proposed BSRT, including the new RAW processing pipeline, Pyramid FG-DCN and Swin Transformer blocks and groups. Here, we chose the original EBSR [37] as the baseline, which uses normal pyramid DCN alignment and residual blocks, performing burst SR under the old RAW processing pipeline. The results are shown in Table 2, which show that the new processing pipeline improves the baseline's performance overall metrics. Based on the new proposed pipeline, the Pyramid FG-DCN alignment module can further improve the results. Moreover, Swin Transformer plays an important role in both feature extraction and HR image reconstruction. Especially, Swin Transformer blocks can extract more effective features compared to residual blocks, which improves the performance of the network. This enhancement can also demonstrate that long-range dependencies of the self-attention have positive effects on BurstSR task.

5. Result on NTIRE2022 BurstSR Challenge

Our method wins 1st place in the NTIRE2022 Burst Super-Resolution Challenge Real-World Track. The top-5 ranked teams are shown in Table 3. The evaluation is based on a user study on a test set containing 20 real-world burst sequences captured from a handheld Samsung Galaxy S8 smartphone camera. The results demonstrate that the super-resolved images produced by our method are more pleasant and plausible compared with other teams.

Team	MegSR*	HIT-IIL	S&C	Noah_TerminalVision	VDSL
Rank	1	2	3	4	5

Table 3. The top-5 ranked teams for Track 2 (Real-World Track). Our team is marked by '*'.

6. Conclusion

A more efficient approach, called BSRT, to BurstSR is proposed in this paper. The main components of the BSRT include the Pyramid Flow-based Deformable alignment module (Pyramid FG-DCN) and the Swin Transformer-based backbone. Compared with the previous methods, the proposed Pyramid FG-DCN can greatly improve the alignment performance and alleviate the effect of noises. Meanwhile, Swin Transformer blocks and groups in our backbone can make more effective use of global contextual information in multi-frames and further improve the performance through the self-attention mechanism. Our results on both synthetic and real-world datasets demonstrate that our method achieves a state-of-the-art performance and recovers more plausible and pleasing visual results. Furthermore, our proposed BSRT wins 1st place in real-world track of the NTIRE 2022 Burst Super-Resolution Challenge.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under grants No.61872067.

References

[1] 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. 3

[2] 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. 1, 2, 5, 7

[3] 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. 1, 5

[4] 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, 5, 6, 7

[5] 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. 1, 3, 5, 6, 7

[6] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 5

[7] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 3

[8] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 3

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[10] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *arXiv preprint arXiv:2104.13371*, 2021. 2, 4

[11] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 3

[12] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2

[13] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 2

[14] 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. 1, 3, 5, 6

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014. 1

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[17] 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. 3

[18] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 2

[19] Russell Hardie. A fast image super-resolution algorithm using an adaptive wiener filter. *IEEE Transactions on Image Processing*, 16(12):2953–2964, 2007. 3

[20] 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. 3

[21] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018. 2

[22] CK Huang and Hsiau-Hsian Nien. Multi chaotic systems based pixel shuffle for image encryption. *Optics communications*, 282(11):2123–2127, 2009. 4

[23] 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. 1, 5

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

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2

[26] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 1, 2

[27] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 2

[28] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 2

[29] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2370–2379, 2021. 1, 3

[30] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 2

[31] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 2, 3

[32] 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, 3, 4

[33] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 1, 2

[34] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020. 3

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3, 4

[36] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. *arXiv preprint arXiv:2202.07508*, 2022. 2

[37] 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. 1, 2, 3, 4, 5, 6, 7, 8

[38] 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. 3

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

[40] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 2, 3

[41] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, pages 4491–4500, 2017. 2

[42] 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 *CVPR*, pages 1874–1883, 2016. 2

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[44] 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. 5

[45] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017. 2

[46] Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on image processing*, 16(2):349–366, 2007. 3

[47] R Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984. 3

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[49] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks, 2018. 2

[50] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[52] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021. 3

[53] Christoph Wick, Jochen Zöllner, and Tobias Grüning. Transformer for handwritten text recognition using bidirectional post-decoding. In *International Conference on Document Analysis and Recognition*, pages 112–126. Springer, 2021. 3

[54] 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

[55] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 3

[56] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 2

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[58] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep

residual channel attention networks. In *ECCV*, pages 286–301, 2018. 2

[59] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2

[60] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2020. 2

[61] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 2