EBSR: Feature Enhanced Burst Super-Resolution with Deformable Alignment

Ziwei Luo [1] Lei Yu [1] Xuan Mo [1] Youwei Li [1] Lanpeng Jia [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/EBSR
{luoziwei,yulei02,moxuan,liyouwei,jialanpeng,fhq,sunjian,liushuaicheng}@megvii.com

Abstract

We propose a novel architecture to handle the problem of multi-frame super-resolution (MFSR). The proposed framework is known as Enhanced Burst Super-Resolution (EBSR), which divides the MFSR problem into three parts: alignment, fusion, and reconstruction. We propose a Feature Enhanced Pyramid Cascading and Deformable convolution (FEPCD) module to align multiple low-resolution burst images in the feature level. And then the aligned features are fused by a Cross Non-Local Fusion (CNLF) module. Finally, the SR image is reconstructed by the Long Range Concatenation Network (LRCN). In addition, we build a cascading residual pathway structure (CR) to improve the performance. We conduct several experiments to analyze and demonstrate these modules. Our EBSR model won the champion in the real track and second place in the synthetic track in the NTIRE21 Burst Super-Resolution Challenge.

1. Introduction

Super-resolution (SR) is a widely studied problem [14, 10, 13, 28, 22, 1, 29], and the task of SR is generating high-resolution (HR) images reference given low-resolution (LR) images. According to the form of LR input, super-resolution can be divided into two categories: single image super-resolution (SISR) and multi-frame super-resolution (MFSR). Single image super-resolution is a task to generate high-resolution (HR) image with a single low-resolution image. Various methods focus on solving the problem of SISR [14, 10, 29, 22]. The main challenge is how to synthesize high-frequency details from an single LR input. The ill-posed problem makes it difficult to generate suitable high-frequencies details similar to ground-truth HR image.

On the other hand, the multi-frame super-resolution aims to reconstruct the original HR image using multiple LR images. The LR images are captured by a hand-held smartphone under the burst mode, where LR images contain shifts due to the camera motion. The shifts between the burst LR images, which called sub-pixel shifts [24], can provide different LR samplings of the underlying scene. Therefore, relative to SISR, MFSR approaches can obtain additional images signal information from the burst images obtained by natural hand tremors [24]. In general, MFSR approaches can exhibit better super-resolution performance relative to SISR.

Figure 1. The comparison between our method EBSR and other representative methods, EDSR [10], RCAN [29], and EDVR [19]

The NTIRE 2021 Burst Super-Resolution Challenge [2] uses a new dataset and has 2 tracks, namely Track 1: Synthetic and Track 2: Real-world. Given multiple noisy RAW images of a scene, the goal of the challenge is to predict a denoised higher-resolution RGB image by combining information from the multiple input frames. A burst sequence containing 14 images, where each image contains the RAW sensor data from a bayer filter (RGGB) mosaic, are provided as inputs.

Our Enhanced Burst Super-Resolution (EBSR) for multi-frame super-resolution is based on the convolutional neural network. Our proposed framework demonstrates superior performance in the NTIRE 2021 challenges as shown in Fig. 1. We compared different methods with our EBSR,

*Corresponding author.

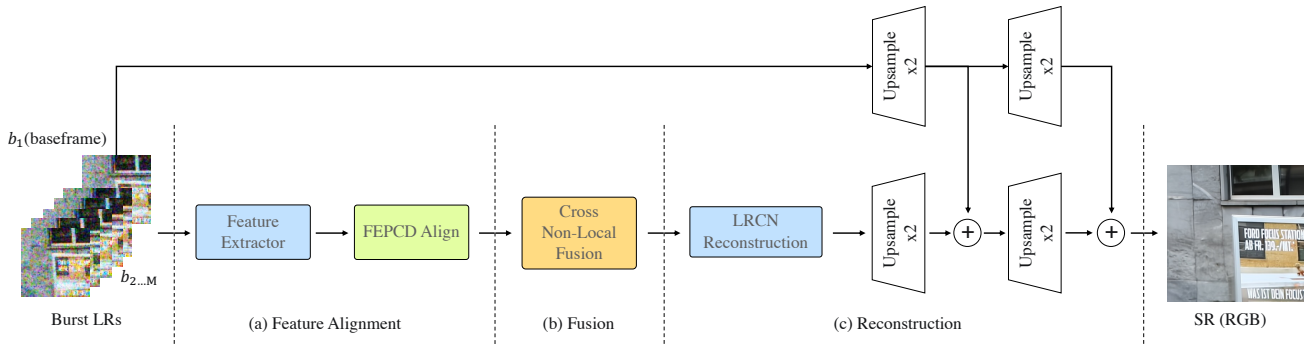

Figure 2. An overview of the proposed method Enhanced Burst Super-Resolution (EBSR). The EBSR can be divided into three parts: feature alignment, fusion and reconstruction. It contains four modules: Feature Enhanced Pyramid Cascading and Deformable convolution (FEPCD) module for alignment, Cross Non-Local Fusion (CNLF) module for fusion and Long-Range concatenation network (LRCN) for reconstruction. Further, we utilize the cascaded residual pathway structure (CR) to improve the performance.

including EDSR [12], RCAN [29] and EDVR [19]. Note that in the zoom area, our method can restore richer texture details than other methods. The main reason it achieves good results can be attributed as that we use effective modules to utilize the extra images signal information between the set of burst LR images. In addition, we design a reconstructed network with strong capability of learning rich features. In particular, our network directly operates on noisy RAW bursts captured from a hand-held camera, and the goal is to exploit the information from the multiple input images to generate a denoised, demosaicked, and super-resolved image as output. Compared with RGB images, Raw images have more original and rich signal information. Choosing RAW as the network input can provide richer information to the network in order to restore better quality super-resolution images.

We achieve this goal by our framework in three steps: 1) align, 2) fusion and 3) reconstruction. We extract the features of burst RAW images and align them by our Feature Enhanced Pyramid Cascading and Deformable convolution (FEPCD) module. We use multi-scale features extracted by Feature pyramid networks (FPN) [11] to enhance the feature representation ability of Pyramid Cascading and a Deformable convolution (PCD) [19]. This alignment step is to facilitate the fusion of different RAW image features in fusion module. We utilize a Cross Non-Local Fusion (CNLF) module as our fusion module. When we determine the reference image, each of the other image features are sent into CNLF with the reference image features. Then, the CNLF can computes the response at a position as a weighted sum of the features at all positions in the input feature maps.

The CNLF allows the features of other images to be weighted with reference to the features of the reference image, so as to better integrate the useful images signal information of other images for reconstruction module. The Long-Range concatenation network (LRCN), which proposed for reconstruction, utilizes long range features information by concatenating the feature at different levels. Therefore, LRCN have better feature representation capabilities for the reconstruction of super-resolved RGB images. Furthermore, we use the cascaded residual pathway structure (CR), which can improve the performance of the model for the noise suppression. Thanks to the above modules, our framework can combine the image contents between burst LR RAW images in a reasonable way, producing RGB prediction with natural textures and more high-frequency details that similar to the ground-truth HR images.

Contribution In this work, our main contributions are summarized as follows.

- We propose a Feature Enhanced PCD (FEPCD) module to improve the performance of features alignment.
- We propose a Cross Non-Local Fusion (CNLF) module for combining signal information from aligned features of multiple images.
- We propose a Long-Range concatenation network (LRCN) for better reconstruction of super-resolved images.
- We use a cascaded residual pathway (CR) structure to improve the performance for the noise suppression of our method.

2. Related Work

Single Image Super-Resolution. Single Image Super Resolution (SISR) is a long standing research topic due to

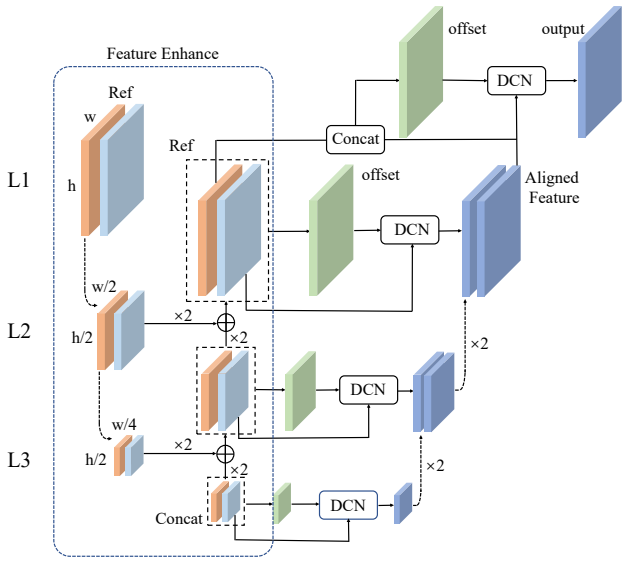

Figure 3. Feature Enhanced PCD convolution alignment.

its importance and ill-posed nature. Traditional learning-based methods adopts sparse coding [5, 15, 26] or local linear regression [17, 16, 25]. Deep learning (DL)-based method is first proposed by SRCNN [14] that employs a relatively shallow network and adopts the bicubic degradation for HR and LR pairs. Following which, various SISR approaches have been proposed, such as VDSR that adopts very deep network [14]; EDSR that modifies the ResNet for enhancement [10]; DnCNN that predicts high frequency details [28]; ESPCN that uses efficient sub-pixel CNN [13]; CDC that divides images into multiple regions [22], CARN that adopts cascading residual network [1]; and VGG loss [14], GAN loss [6] that improve the perceptual visual quality [9, 12, 21].

Alignment. How to solve the misaligned between multiple frames is always the focus of multi-frame super resolution. Optical flow is used in [4] to estimate the motion between frames. Another branch of studies achieve implicit motion compensation by dynamic filtering or deformable convolution. In EDVR [19], which won the champion of NTIRE19, proposed a block called PCD to solve this problem. In this work, we used the feature pyramid to enhance PCD and got good results.

Attention Mechanism. In many published works [18, 23, 29], attention has mentioned as a useful method to improve the final result. In EDVR [19], attention is used to fuse information between frames.And non-local Operation [20] is a commonly way to calculate the interrelationship between frames. Inspired by these works, we use non-local operation to compute the weight map between reference frame and the others, after that we use these maps to fuse the frames.

Upsampling. Image interpolation, a.k.a. image scaling, refers to resizing digital images and is widely used by image-related applica- tions. The traditional interpolation methods include nearestneighbor interpolation, bilinear and bicubic interpolation. Since these methods are interpretable and easy to implement, some of them are still widely used in CNN-based SR mode. In recent year, learning-based upsampling methods are introduced into SR field. Transposed convolution layer tries to perform transformation opposite a normal convolution. Pixelshuffle [7] is another end-to-end learnable upsampling layer, which generating a plurality of channels by convolution at first and then reshaping them. In this work, we proposed a cascaded residual pathway structure to improve the pixelshuffle.

3. Method

Given a RAW burst low-resolution sequence $\{b_i\}_{i=1}^{M}$ and upscaling factor γ, the goal of EBSR is to reconstruct a high-resolution image by taking advantage of the shifted complementary information from different images. Each image $b_i \in \mathbb{R}^{C\times H\times W}$ is obtained from the RAW sensor camera. In this work, we take the first image as base frame and align the rest neighboring images to it in the feature level by using a Feature Enhanced, Pyramid, Cascading and Deformable convolution (FEPCD) module. And we propose to use a Cross Non-Local Fusion (CNLF) module to fuse the aligned features. The details of FEPCD and CNLF are described in Sec. 3.1 and Sec. 3.2. The fused features are then passed to the long-range concatenation network (LRCN) to obtain a high-resolution RGB output. In addition, we take a two stage upsampling strategy to protect the network from raw noises. The overview of the proposed framework is shown in Figure 2.

3.1. Feature Alignment

One big challenge of burst SR is that the input consists of multiple noisy, disordered and shifted RAW images with unknown displacements which stem from both global camera motion and scene variations. To tackle this problem we adopt the modulated deformable modules [30] to perform feature level alignment between RAW images inspired by PCD [19]. However, PCD is inefficient in RAW images alignment since it didn't take the noises from input into account. Therefore, we propose the FEPCD module with feature pyramid extraction to eliminate the effect of noises.

The overall FEPCD module is shown in Figure 3. It is a double pyramid structure, in which the first pyramid is responsible for denoising and feature enhancement, and the second pyramid is responsible for features alignment

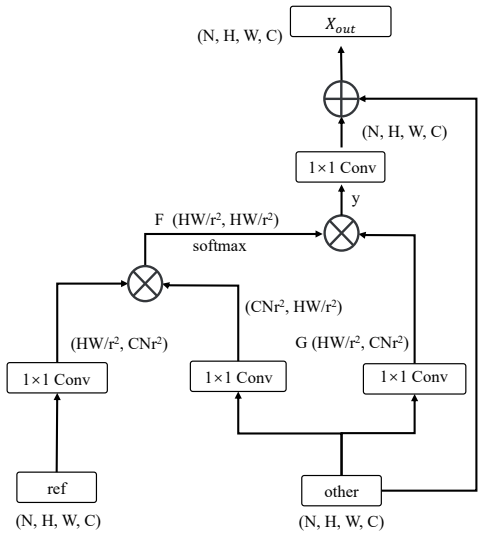

Figure 4. Cross Non-Local Fusion Module.

and refinement. To achieve a high performance alignment, we map each input image b_i to a deep feature representation F_i using a set of Wide Activation Residual Block (WARB) as introduced in [27]. And then the features are enhanced through a top-down, bottom-up pathway and implicitly aligned. Specifically, the top-down pathway produce pyramid levels of feature with different resolutions and channels, and laterally connects to corresponding bottom-up pathway to hallucinate cleaner and semantically stronger features. PCD module requires using same channels for different level of features. To start alignment, we simply attach a 1×1 convolution to obtain the final pyramid features. The noises are removed in the pyramid feature enhancement process. We take the first feature of input images as base frame and align other neighboring features to the base.

3.2. Fusion

The temporal relation between multiple frames plays a vital role in feature fusion, due to the blurry frames from camera perturbation and misalignment from preceding alignment module. To aggregate the aligned features dynamically, we propose a Cross Non-Local Fusion (CNLF) module by considering the non local relation between base frame and neighboring frames. The cross non-local operation is defined as follows:

$$\mathrm{y}_i = \frac{1}{C(x)} \sum_{\forall j} f(ref_i, other_j) g(other_j), \qquad (1)$$

where ref and $other$ are the base frame and other input frames, respectively. Function f produce the adaptive pixel-level weight vectors between two frames and function g produce a feature representation of the input frame. We consider the normalization factor $C(x) = \sum_{\forall j} f(ref_i, other_j)$. To avoid introducing too many additional parameters, the Gaussian function is used as the choice of f:

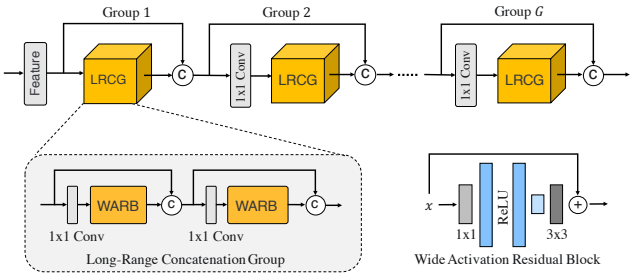

Figure 5. Long Range Concatenation Network.

$$f(ref_i, other_j) = e^{{ref_i}^T \cdot other_j} \qquad (2)$$

where ${ref_i}^T \cdot other_j$ is dot-product similarity, which is more convenient in deep learning platforms. As shown in Figure 4, we design the CNLF module to measure the similarity between every two pixels in multi-frames features. Different from the original non-local structure, we use ref and $other$ as the inputs of the network. Firstly, we change the dimension of the inputs ref and $other$ by 1×1 convolution, respectively. Then we implement the matrix multiplication for those two branches and the results are operating by the softmax function. By that means, the shape of feature map F becomes $HW/r^2 \times HW/r^2$ that is irrelevant of N, where r represents the reduction factor and N represents the number of batchsize. Finally, The feature map F and the result of g branch are matrix multiplied to obtain a correlation feature map y_i. The output of CNLF is defined as $Wy_i + other_i$, where W is implemented by 1×1 convolution, and $other_i$ denotes the residual learning. Generally, CNLF is an application of self-attention mechanism, the more similar the feature representations between two locations, the higher the correlation between them. According to this property, we enhance the regions in other frames which are similar to the reference frame. It should be noted that the matching between two features requires a large amount of computation and memory, so we performed these operations on the downsampled features.

3.3. Reconstruction

We reconstruct the RGB image by utilizing a Long-Range Concatenation Network (LRCN) as shown in Figure 5. The LRCN module is composed of G Long-Range Concatenation Groups (LRCG), and each LRCG contains B Residual Blocks with wide activation (WARB) which is inspired by WDSR [27]. Both the LRCG and WARB receives a multi-level feature that is obtained by concatenating

Method	Strategy	Track 1			Track 2		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
EDSR[10]	No Alignment	37.020	0.925	0.101	44.921	0.971	0.067
RRDB [21]		37.740	0.926	0.098	45.313	0.974	0.053
WDSR [27]		37.702	0.925	0.096	45.435	0.974	0.057
RCAN [29]		37.872	0.928	0.091	45.436	0.974	0.064
LRCN(Ours)		38.099	0.930	0.088	45.607	0.976	0.053
EDVR [19]	Deformable Alignment	42.268	0.969	0.035	46.168	0.977	0.047
EBSR(Ours)		43.350	0.973	0.029	46.586	0.979	0.041
EBSR*(Ours)		-	-	-	48.221	0.985	0.024

Table 1. The table shows a comparison between our methods and the other teams. The best one marks in red and the second best are in blue. '*' means the model is pretrained on synthetic dataset.

Name	Baseline	Feature Alignment			Fusion	Reconstruction
EDSR	√	√	√	√	√	√
WARB		√	√	√	√	√
PCD			√	√	√	√
FEPCD				√	√	√
CNLF					√	√
LRCN						√
Track 1	37.020	37.702	42.77	42.915	43.272	43.35
Track 2	44.921	45.435	46.098	46.247	46.408	46.586

Table 2. In this table, we directly use RAW images to produce RGB results and shows the PSNR of the both tracks. It also prove the benefits of adding different modules to the network. The baseline fusion model is a 1×1 Conv layer.

Team	Track 1			Track 2
	PSNR	SSIM	LPIPS	PSNR
raoumer	37.618	0.895	0.166	41.395
TakahiroMaeda	44.399	0.973	0.038	44.153
JohnDoe4598	44.762	0.969	0.034	x
chowy333	39.221	0.918	0.104	x
Noah TerminalVision	46.855	0.983	0.018	45.36
Ours	46.723	0.983	0.02	45.454

Table 3. The table shows a comparison between our methods and the other teams. The best one marks in red and the second best are in blue.

all previous features followed by a 1×1 Conv layer. Moreover, we introduce a progressively upsampling strategy that uses pixelshuffle [7] and a cascading residual pathway (CR) structure to reconstruct the final SR image. Different from EDVR, in which the base frame is upscaled $\times4$ with bilinear interpolation directly, our method learns two cascaded pixelshuffle($\times2$) layers and adds the outputs to the predicted image residual as shown in Figure 2(c). More specifically, after each upsampling, we add these two outputs to the corresponding $\times2$ and $\times4$ reconstructed high-resolution features. We observe that employing this two-stage progressively upsample technique, the noise in the upsampling process can be greatly reduced.

3.4. Loss Function

In the training of real and synthetic track models, we use L1 loss to evaluate model prediction errors. We use charbonnier loss for Fine-tune training, proposed by LapSRN [8], which is defined as:

$$Loss_{charbonnier} = \sum_{i=1}^{S} \sqrt{(E(I_{IN}^{i}) - I_{HR}^{i})^2 + \epsilon^2}, \quad (3)$$

where S is the number of training samples, and ϵ is 10^{-3}. E is our EBSR model. I_{IN} and I_{HR} are the input and HR image. This loss function not only improves the performance of our model, but also improve the convergence speed.

4. Experiment

4.1. Dataset and Implementation Details

Our method is evaluated on two datasets provided by the Burst Super-Resolution Challenge[1]: synthetic RAW burst dataset and real-world BurstSR dataset [3]. We flatten the RAW burst images from four channel (RGGB) to single channel so as to perform super-resolution with a scaling factor of $times4$. We notice that it is better to incorporate the task of demosaic within the network than to use the non-parametric post-processing demosaic outside the network. In other words, we learn a NN demosaic by ourselves. This is important for the performance improvements.

In the training phase, we use Adam optimizer and set exponential decay rates as 0.9 and 0.999. The initial learning rate is set to 4×10^{-4} and then reduced to half every 200 epochs. For each training data, we randomly crop 14 RAW burst image patches with the size of 64×64. The batch-size is set to 16. We implement the proposed EBSR using Pytorch framework with 4 NVIDIA 2080Ti GPUs.

Though the single EBSR model could achieve impressive results, we also observe that we can further improve the performance by employing a multi-model ensemble training strategy. We load three training completed models with frozen weights, and add few convolution layers to fusion their outputs.

[1]The challenge website is here: https://data.vision.ee.ethz.ch/cvl/ntire21/

Figure 6. Qualitative results of a comparison between our method and other classical methods in Track 1.

4.2. Ablation Studies

In this section, we mainly compared the effectiveness of each module. We chose the EDSR as the baseline which using 1×1 Conv layer as fusion module.

FEPCD As mentioned in Sec. 3.1, the FPCD is used to align images from different frames. We further exam the effectiveness of our FPCD block on EDSR by removing the Feature Ehance block. As show in Table 2, it can be proved that introducing multi-scale features improve the performance of PCD module.

CNLF The CNLF module can better tell the network how to fuse multiple frames of data. In order to prove its effectiveness, we replace this module with a simple 1×1 Conv layer as our baseline fusion module as shown in Table 2. PSNR scores in different tracks are all decreased, especially in track 1, the PSNR decreased 0.357 dB.

LRCN We introduced the structure of LRCN in Sec. 3.3. In order to prove its effectiveness, we removed the structure of LRCN, and the results are shown in Table 2. It can be seen that the LRCN module is of great help to the improvement of PSNR.

Pretrain Module In the competition, we found that fine-tuning the real data using the trained model of the synthetic dataset will get a higher PSNR, as shown in Table 1.

4.3. Comparisons with Existing Methods

In Table 3, we show the results compared to the other teams on the two tracks. In addition, we also selected some classical algorithms for qualitative and quantitative comparison on the two tracks. Note that we implement these single

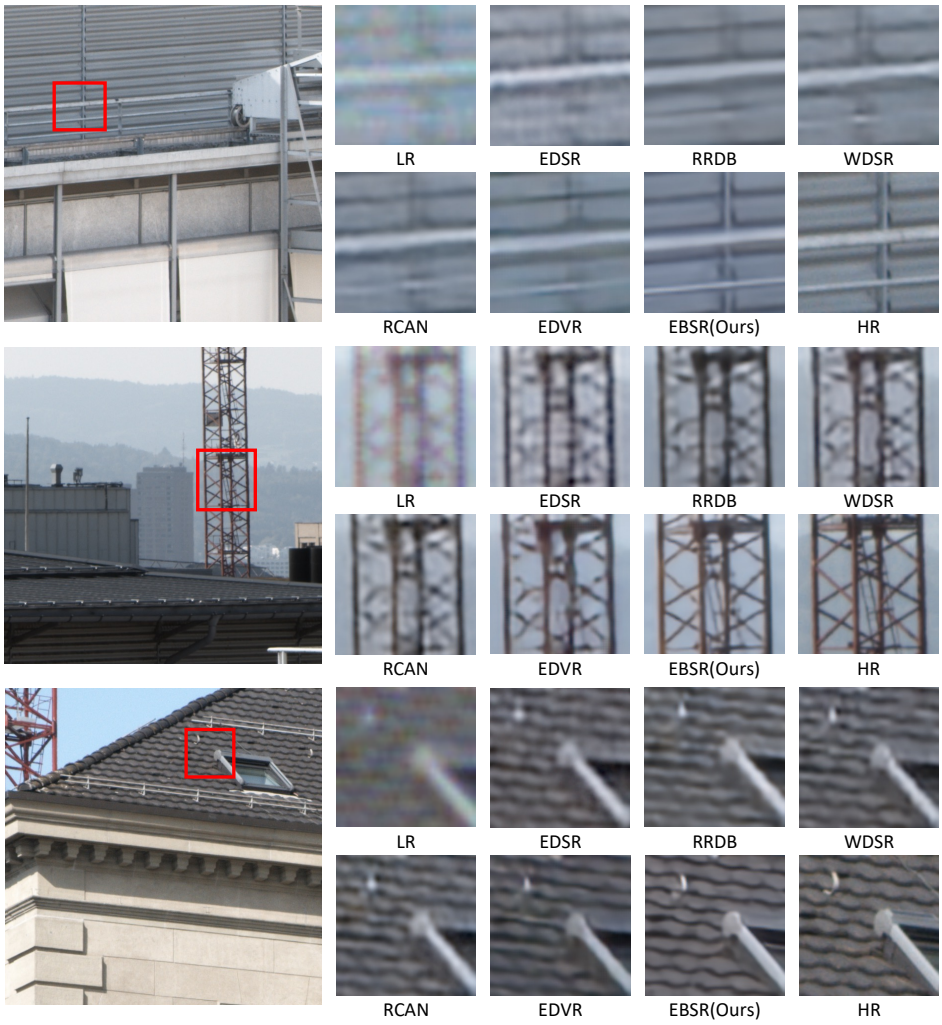

Figure 7. Qualitative results of a comparison between our method and other classical methods in Track 2.

image SR methods to deal with the RAW burst images by directly concatenating multiple images on the channel axis as their input without any alignment. We also modified the EDVR to use the first LR RAW image as reference. For a fair comparison, all models are trained from scratch except our final finetuned EBSR model in Track 2 (EBSR*). Among them, the results of track 1 are shown in Table 1 and Figure 6, and the results of track 2 are shown in Table 1 and Figure 7. As you can see from the last case in Figure 6, our method restores the details of the license plate very well, while other methods fail to do so.Similarly, in the last case in Figure 7, our method restores the edge of the tile better.

5. Conclusion

Quantitative and qualitative results prove that, EBSR can accomplish the burst image super-resolution task very well. Compared with the original alignment module (PCD), our FEPCD can greatly reduce the alignment failure caused by large motion between frames. The CNLF module has a dependable multi-frame images fusing performance, since we taking the similarity between feature representations in to the calculation of correlation. The long-range concatenation groups and progressively up-sampling module in LRCN, help the model to obtain clearer, high-fidelity super-resolution results. By combining and training these models, EBSR not only wins 1st and 2st places in NTIRE21 Challenges on real and synthetic burst image super-resolution track, but also demonstrates superior performance to most of the existing methods on burst image super-resolution.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No.61872067 and No.61720106004.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proc. ECCV*, pages 252–268, 2018. 1, 3

[2] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 challenge on burst super-resolution: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. *arXiv preprint arXiv:2101.10997*, 2021. 5

[4] 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 *Proc. CVPR*, July 2017. 3

[5] Dengxin Dai, Radu Timofte, and Luc Van Gool. Jointly optimized regressors for image super-resolution. In *Computer Graphics Forum*, volume 34, pages 95–104, 2015. 3

[6] 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. 3

[7] CK Huang and Hsiau-Hsian Nien. Multi chaotic systems based pixel shuffle for image encryption. *Optics communications*, 282(11):2123–2127, 2009. 3, 5

[8] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proc. CVPR*, pages 624–632, 2017. 5

[9] 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 *Proc. CVPR*, pages 4681–4690, 2017. 3

[10] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proc. CVPRW*, pages 136–144, 2017. 1, 3, 5

[11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. CVPR*, pages 2117–2125, 2017. 2

[12] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proc. ICCV*, pages 4491–4500, 2017. 2, 3

[13] 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 *Proc. CVPR*, pages 1874–1883, 2016. 1, 3

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 3

[15] Libin Sun and James Hays. Super-resolution from internet-scale scene matching. pages 1–12, 2012. 3

[16] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proc. ICCV*, pages 1920–1927, 2013. 3

[17] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Proc. ACCV*, pages 111–126, 2014. 3

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 3

[19] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proc. CVPRW*, pages 0–0, 2019. 1, 2, 3, 5

[20] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks, 2018. 3

[21] 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 *Proc. ECCVW*, pages 0–0. 3, 5

[22] 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 *Proc. ECCV*, pages 101–117, 2020. 1, 3

[23] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018. 3

[24] 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

[25] Chih-Yuan Yang and Ming-Hsuan Yang. Fast direct super-resolution by simple functions. In *Proc. ICCV*, December 2013. 3

[26] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *Proc. CVPR*, pages 1–8, 2008. 3

[27] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas Huang. Wide activation for efficient and accurate image super-resolution. *arXiv preprint arXiv:1808.08718*, 2018. 4, 5

[28] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. on Image Processing*, 26(7):3142–3155, 2017. 1, 3

[29] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proc. ECCV*, September 2018. 1, 2, 3, 5

[30] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proc. CVPR*, pages 9308–9316, 2019. 3