NTIRE 2022 Burst Super-Resolution Challenge

Goutam Bhat* Martin Danelljan* Radu Timofte* Yizhen Cao Yuntian Cao
Meiya Chen Xihao Chen Shen Cheng Akshay Dudhane Haoqiang Fan
Ruipeng Gang Jian Gao Yan Gu Jie Huang Liufeng Huang Youngsu Jo
Sukju Kang Salman Khan Fahad Shahbaz Khan Yuki Kondo Chenghua Li
Fangya Li Jinjing Li Youwei Li Zechao Li Chenming Liu Shuaicheng Liu
Zikun Liu Zhuoming Liu Ziwei Luo Zhengxiong Luo Nancy Mehta
Subrahmanyam Murala Yoonchan Nam Chihiro Nakatani Pavel Ostyakov
Jinshan Pan Ge Song Jian Sun Long Sun Jinhui Tang Norimichi Ukita
Zhihong Wen Qi Wu Xiaohe Wu Zeyu Xiao Zhiwei Xiong Rongjian Xu
Ruikang Xu Youliang Yan Jialin Yang Wentao Yang Zhongbao Yang
Fuma Yasue Mingde Yao Lei Yu Cong Zhang Syed Waqas Zamir
Jianxing Zhang Shuohao Zhang Zhilu Zhang Qian Zheng Gaofeng Zhou
Magauiya Zhussip Xueyi Zou Wangmeng Zuo

Abstract

Burst super-resolution has received increased attention in recent years due to its applications in mobile photography. By merging information from multiple shifted images of a scene, burst super-resolution aims to recover details which otherwise cannot be obtained using a simple input image. This paper reviews the NTIRE 2022 challenge on burst super-resolution. In the challenge, the participants were tasked with generating a clean RGB image with 4× higher resolution, given a RAW noisy burst as input. That is, the methods need to perform joint denoising, demosaicking, and super-resolution. The challenge consisted of 2 tracks. Track 1 employed synthetic data, where pixel-accurate high-resolution ground truths are available. Track 2 on the other hand used real-world bursts captured from a handheld camera, along with approximately aligned reference images captured using a DSLR. 14 teams participated in the final testing phase. The top performing methods establish a new state-of-the-art on the burst super-resolution task.

1. Introduction

Burst mode shooting has seen increased popularity in recent years. Instead of capturing only a single picture, burst mode captures multiple photos of the scene in quick succession. In addition to allowing the photographer to select the best picture among the multiple images, burst mode capture also provides the possibility of combining information from the multiple images to generate a single higher quality image. For instance, since the noise in the different images are approximately independent, the burst images can be merged together to perform denoising [35, 22]. Furthermore, if the burst images are captured using different exposure settings, they can be combined to perform HDR imaging.

A fundamental challenge when fusing information from the burst images is the spatial shifts between the images introduced due to natural hand motion. These shifts are a nuisance for tasks such as burst denoising, as it necessitates a separate alignment step. However, recent works [51] have shown that these shifts can in fact provide multiple aliased samplings of the underlying scene. By merging information from these multiple low-resolution samplings, a higher-resolution version of the scene can be recovered. This task, referred to as burst super-resolution, thus allows recovering high-frequency details from the scene, which otherwise cannot be obtained using a simple image. Consequently, burst super-resolution has many practical applications in mobile photography, where the resolution and quality of the captured images are limited by small sensor size.

The goal of the NTIRE 2022 Burst Super-Resolution

*Goutam Bhat, Martin Danelljan, and Radu Timofte are the NTIRE 2022 challenge organizers. The other authors participated in the challenge and are listed alphabetically.
Appendix A contains the participants' team names and affiliations.
NTIRE 2022 webpage:
https://data.vision.ee.ethz.ch/cvl/ntire22/

challenge is to promote further research in the burst super-resolution task, and to establish the current state-of-the-art. As part of the challenge, the participants were required to generate a clean high-resolution image give multiple noisy RAW images as input. Thus the task required performing joint denoising, demosaicking, and super-resolution which are fundamental steps in an ISP. The challenge contained two tracks, namely Track 1 and Track 2. In Track 1, the input bursts were generated synthetically using the pipeline introduced in [4]. Since an accurate pixel-wise ground truth is available in this case, the performance of various network architectures can be quantitatively evaluated using standard fidelity metrics like PSNR. To further evaluate the real-world performance of the participating methods, Track 2 employed real-world bursts from the BurstSR dataset [4] for evaluation. The methods were ranked using a human study in this track.

The NTIRE 2022 Burst Super-Resolution challenge saw participation from 14 different teams. The top ranked methods set a new state-of-the-art on the burst super-resolution task. This report briefly describes the solutions proposed by the participating methods, and reports their performance on the test sets of both Track 1 and Track 2.

2. NTIRE 2022 Challenge

The NTIRE 2022 Burst Super-Resolution challenge aims to stimulate research in the burst SR problem and benchmark the current approaches using a standard evaluation protocol. This challenge is one of the NTIRE 2022 associated challenges: spectral recovery [3], spectral demosaicing [2], perceptual image quality assessment [13], inpainting [39], night photography rendering [11], efficient super-resolution [20], learning the super-resolution space [28], super-resolution and quality enhancement of compressed video [53], high dynamic range [37], stereo super-resolution [45]. This is the second edition of the challenge. The participants were required to develop methods which takes a noisy RAW burst containing 14 images as input, and generates a clean RGB image with 4 times higher spatial resolution. As a starting point, the participants were provided a public toolkit (https://github.com/goutamgmb/NTIRE22_BURSTSR) which contained integration of standard burst SR datasets, and basic tools for training and evaluation. The challenge contained two separate tracks which are described in the next sections.

2.1. Track 1: Synthetic

In Track 1, the input burst are generated synthetically using a single sRGB image. This ensures that an accurate pixel aligned ground truth is easily available to both evaluate the methods, as well as for training. Consequently, the impact of different architectural choices on super-resolution performance can be readily analysed in this setting, using standard fidelity metrics such as PSNR.

In order to generate the synthetic bursts, we use the pipeline introduced in [4]. Here, an sRGB image is first converted to linear sensor space using an inverse camera pipeline [7]. Random translations (maximum 24 pixels) and rotations (maximum 1 degree) are then applied to this linear image to obtain a burst. A RAW low-resolution burst is then obtained by downsampling each of the images by a factor of 4 using bilinear interpolation and mosaicking them using bayer pattern. The RAW burst is then corrupted with independent read and shot noise to obtain the noisy input burst.

The participants were provided the scripts to generate synthetic bursts for training and evaluation via the public toolkit. The participants were free to use any image dataset except the validation split of the BurstSR dataset [4] to generate synthetic bursts for training. The official validation set for Track 1 was generated using DSLR images from the validation split of the BurstSR dataset [4]. The validation set consisted of 100 bursts, each containing 14 RAW images of resolution 256×256. The participants could evaluate their methods on the validation set during the development phase, using an evaluation server (https://codalab.lisn.upsaclay.fr/competitions/1750). The server also provided a live public leaderboard. The final test set containing 92 bursts of size 14 and resolution 256×256 was generated similarly using DSLR images from the test split of BurstSR [4]. The participants were only provided the input bursts asked to submit their predictions. These were then evaluated by the challenge organizers to obtain the final ranking.

2.2. Track 2: Real-World

Track 2 aims at evaluating the real-world performance of the burst super-resolution methods. For this purpose, we employ the BurstSR dataset introduced in [4] for our evaluation. The BurstSR dataset contains 200 RAW bursts, each containing 14 images, captured using a hand held Samsung Galaxy S8 camera. For each burst, the dataset also provides a higher resolution RGB image captured using a DSLR for reference. However, since this HR image is captured using a separate camera, there exists spatial mis-alignment as well as color space differences between the burst images and the DSLR reference. This makes training and evaluation of the model on the real-world bursts challenging. We refer to [4] for more details about the BurstSR dataset.

The test set for Track 2 consisted of 20 bursts of resolution 256×256, each containing 14 images. The test set was constructed using bursts from the test split of BurstSR dataset, as well as new bursts captured using the same camera which was used to collect BurstSR dataset, *i.e.* Samsung Galaxy S8 camera. The participants were allowed to utilize

the training split of the BurstSR dataset, as well as any other real-world or synthetic datasets to train and validate their models. Due to the difficulties associated with evaluating the models on real-world bursts, no evaluation server was provided in Track 2. Instead the participants were required to submit their predictions for the final test phase, which were then ranked via a human study.

3. Challenge Results

Here, we evaluate the participating methods on the test set of both the tracks. In total, 14 teams participated in the challenge, out of which 13 submitted results for Track 1, while 11 submitted results for Track 2. A brief description of the participating methods is provided in Section 4. The details of the participating teams are provided in Appendix A.

3.1. Evaluation Metrics

In this section, we briefly describe the evaluation metrics used to rank the methods in Track 1 and Track 2. The aim in burst super-resolution is to recover the original high frequency details by multi-frame fusion, instead of hallucinating realistic looking details. Thus we utilize fidelity based metrics to rank the methods.

Track 1: As discussed in Section 2.1, the bursts in Track 1 are generated using a synthetic pipeline. Hence an accurately aligned high resolution ground truth is readily available. Thus we utilize the fidelity-based metric Peak Signal-to-Noise Ratio (PSNR) to rank the methods. Additionally, we also report the Structural Similarity Index (SSIM) [50] and LPIPS [55] score for all the methods. Note that the metrics are computed in the linear sensor space.

Track 2: Unlike in Track 1, an accurately aligned ground truth image is not available for the real-world bursts in Track 2. As a result, it is not possible to directly use image metrics such as PSNR to evaluate the methods. The previous edition of the challenge [5] employed a spatial and color alignment strategy utilized in [4] in order to align the network predictions to the ground truth before computing metrics such as PSNR. However, such an evaluation strategy can introduce a bias towards methods which are trained using an identical loss. Thus, in order to encourage participants to investigate alternate training strategies, we did not employ this evaluation strategy. Instead the methods were ranked purely using a human study.

The human study was conducted by two doctoral students working on super-resolution. This was to ensure that the reconstruction ability of the methods was rewarded over generating visually pleasing but fake details, which can be the case when using inexperienced evaluators from *e.g.* Amazon Mechanical Turk (AMT). We first performed an initial round of study to select the top 5 methods. The human study participants were then asked to rank the predictions of these methods for each of the 20 bursts in the test set, based on their similarity w.r.t. a high resolution reference image. The mean of these rankings (MOR) was then used to obtain the final ranking of the methods.

	PSNR↑	SSIM↑	LPIPS↓
Noah_TerminalVision_SR	**46.50**	**0.986**	**0.017**
VIDAR_A	46.09	0.985	0.018
HIT-IIL	45.98	0.985	0.021
Ver	45.90	0.984	0.019
CUCteam	45.88	0.984	0.021
okfine	45.84	0.984	0.019
VIDAR_B	45.63	0.984	0.021
S&C	45.62	0.984	0.020
MegSR	45.38	0.984	0.025
VDSL	44.22	0.979	0.040
MultiTeam	44.19	0.979	0.024
TeamIITRPR	42.07	0.970	0.041
TTI_IIM_SR	37.89	0.929	0.101

Table 1. Comparison of the participating methods on the test set from Track 1, in terms of PSNR, SSIM, and LPIPS.

3.2. Track 1: Synthetic

Here, we report the results on the participating methods on the test set of Track 1. The mean PSNR, SSIM, and LPIPS scores over the 92 bursts in the test set are provided in Table 1. The best results in terms of all three metrics were obtained by team Noah_TerminalVision_SR, with a PSNR score of 46.50. The team employs an ensemble of 4 models, each of them based on NoahBurstSRNet, the winner of Track 1 in the previous edition of the challenge [5]. The second rank was obtained by team VIDAR_A, which uses a transformer based module, in addition to the PCD module [46] for alignment. Team HIT-IIL obtained the third best performance with a PSNR score of 45.98. Their approach is based on EBSR [30], with the key difference that a transformer based module is used for reconstruction. Team Ver, which also employs a variant of EBSR [30] obtained the fourth rank. The fifth rank was obtained by team CUCteam, which explicitly trains a network module to denoise the reference image, as part of the burst SR architecture.

A qualitative comparison between the participating methods is provided in Figure 1 and 2. We also visualize the first image in the burst (after demosaicking using OpenCV and bilinear upsampling) for reference. The submitted methods perform very well, generating cleaner images with better details compared to simple processing using OpenCV. Observe that the top ranking methods obtain results which are very close to the ground truth despite using noisy RAW images with 4 times lower resolution as input.

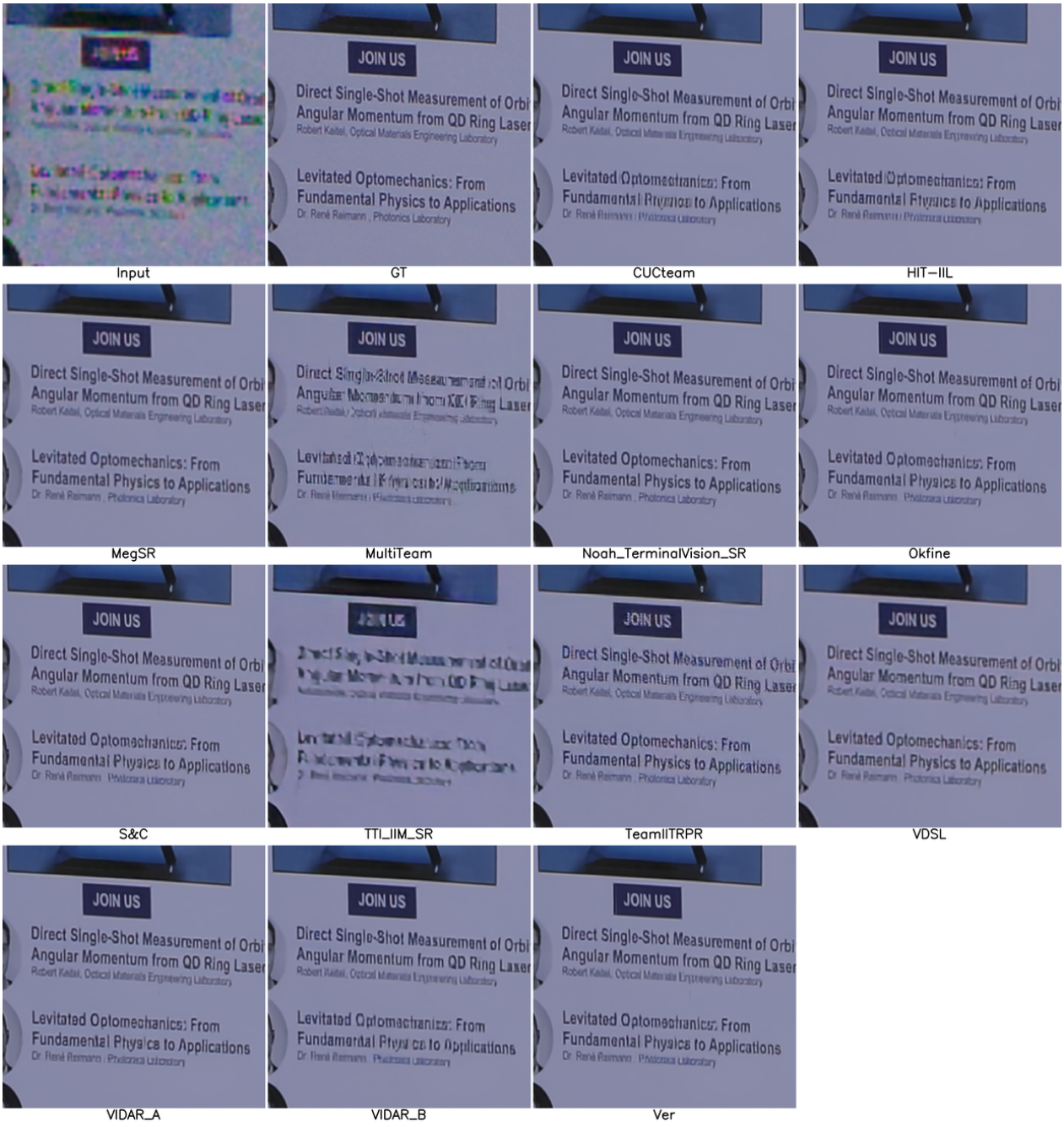

Figure 1. Qualitative comparison on test set of Track 1. Input denotes the reference image of the burst, after demosaicking using OpenCV and bilinear upsampling.

3.3. Track 2: Real-World

Here, we present the results for Track 2. In total, 11 teams submitted results to Track 2, namely: CUCteam, HIT-IIT, LD, MegSR, Multi-Team, Noah_TerminalVision_SR, S&C, TeamIITRPR, TTI_IIM_SR, VDSL, Ver. Out of these, HIT-IIT, MegSR, Noah_TerminalVision_SR, S&C, and VDSL were selected as the top 5 methods based on an initial human study. The mean rankings of these 5 methods over the bursts from the test set is provided in Table 2. MegSR obtained the best results with a mean ranking of 1.77. MegSR employs a flow-guided deformable convolution network for alignment, while also incorporating Swin Transformer blocks [27] for reconstruction. The second rank was obtained by team HIT-IIL, which uses a perceptual loss

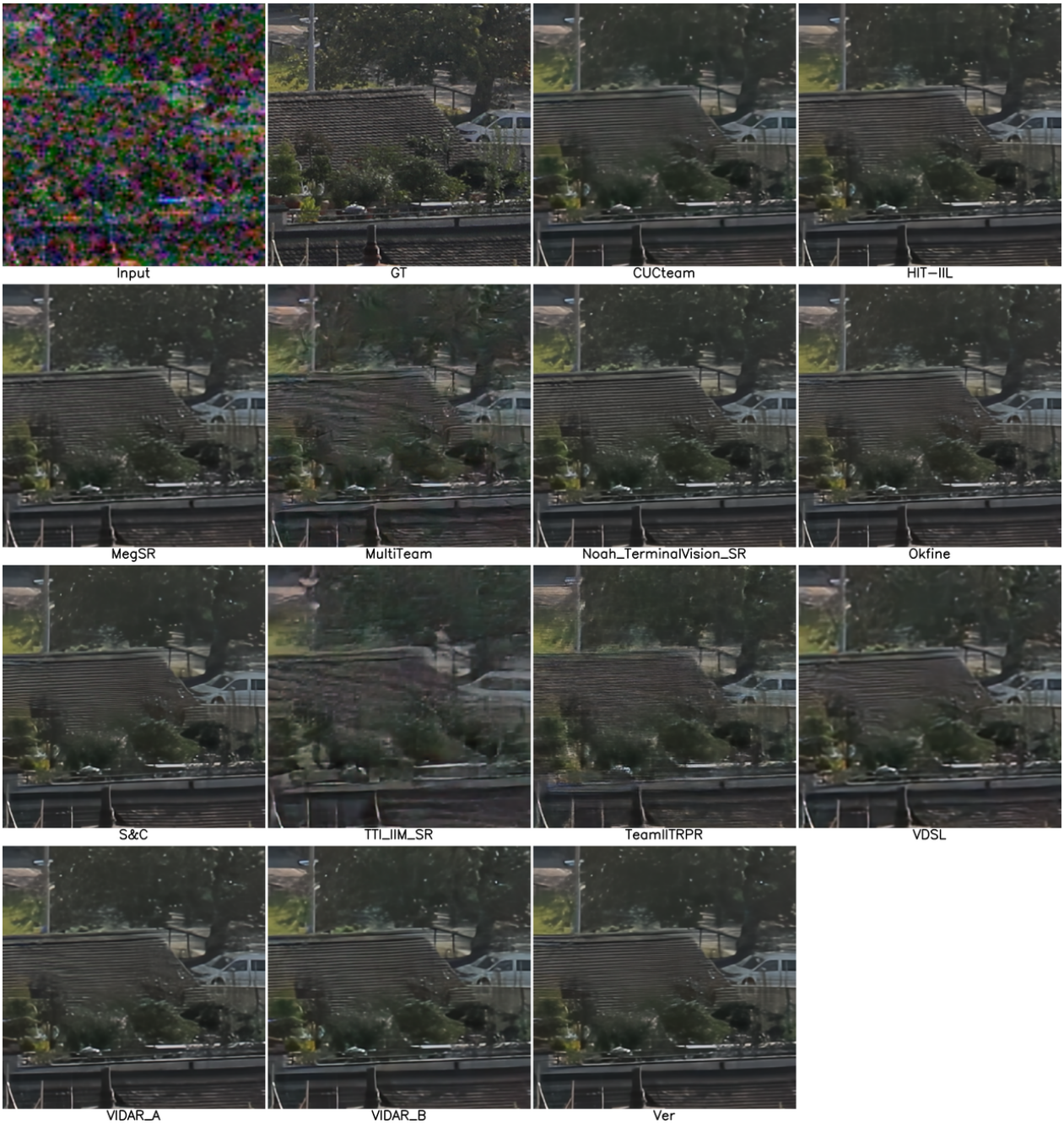

Figure 2. Qualitative comparison on test set of Track 1. Input denotes the reference image of the burst, after demosaicking using OpenCV and bilinear upsampling.

in addition to a fidelity-based loss for training the model. Team S&C who use EBSR [30] network architecture obtained the third rank. Noah_TerminalVision_SR which employs the NoahBurstSRNet [5] obtained the fourth place with a mean rank of 3.11. The fifth place was obtained by team VDSL.

A qualitative comparison between all the participating methods is provided in Figures 3, 4, 5 and 6. We additionally include a high-resolution reference captured using a DSLR, as well as the first image of the burst processed using LibRaw. We applied simple post-processing to the predictions to approximately match the color space to that of the ground truth. We observe that while CUCteam can recover high-frequency details, it also adds checkerboard artifacts. The predictions of TTI_IIM_SR contains artifacts. The images generated by LD, Ver, and MultiTeam are blurry. The

	MOR↓
MegSR	**1.77**
HIT-IIL	2.29
S&C	2.53
Noah_TerminalVision_SR	3.11
VDSL	3.50

Table 2. Results of human study on the test set from Track 2. The methods are ranked based on the mean ranking (MOR) of the method in the human study.

top ranking method in the human study, MegSR produces impressive results, recovering high-frequency details while also performing denoising. HIT-IIL produces sharper results in general due to the use of perceptual loss in the training.

4. Challenge Methods and Teams

Here, we briefly describe the methods proposed by the participating teams.

4.1. CUCteam

CUCteam propose a framework called RAW Burst Super-Resolution with Enhanced Multi-Residual Denoising Net (RBSR), as shown in Figure 7. RBSR solves the burst SR problem by two steps, a reconstruction step and a Multi-Residual step. The method employs a denoising architecture in burst SR task, in order to make information compensation for subsequent super resolution modules.

Network Architecture: In the reconstruction step, the network extracts shallow features of all 14 LR burst images. It aligns other neighboring features to the reference feature, using a Feature Alignment module based on FEPCD [30]. The aligned features are fused by a Temporal Fusion module. Then, Multi-Group Spatial Reconstruction (MGSR) module reconstructs the SR images. The output feature adopts the method of adaptive residual learning, which includes conv layers with different filter sizes to extract multi-scale features. In the Multi-Residual step, the reference frame is passed through a multi-residual framework, which includes a denoising residual enhancement flow and a RAW residual enhancement flow. Denoising residual enhancement flow consists of a Sep-Unet module and a 2x upsample module. After processing, it outputs denoised clean features to supplement the network with more noise-free information. The RAW residual enhancement flow consists of a 4x upsample module, which provides the model with RAW information lost by the network.

Training: For track 1, the method employs synthetic bursts generated using the data generation pipeline employed in [4]. Before read and shot noise are added to image to obtain the noisy RAW burst, the noise free RAW image as ground truth for training the Sep-Unet module. The network is trained with a combined loss which is reconstruction loss and denoising loss which are all set to $L1$ loss. Track 2 employs BurstSR dataset [4] which contains RAW bursts captured from a handheld Samsung Galaxy S8 smartphone camera.

Inference details: In Track 1, the method uses Test Time Augmentation(TTA) [41] data enhancement strategy during testing time to improve the results. In Track 2, an edge-enhanced filter is applied to enhance details and sharpen edges of the RGB image output by the RBSR network.

4.2. HIT-IIL

Network Architecture: The team proposes a transformer model for burst image super-resolution named TBSR, as shown in Fig. 8. TBSR borrows the alignment and fusion modules from EBSR [30], and takes the transformer module as the reconstruction module. The reconstruction module includes m transformer groups, and each transformer group includes n transformer blocks. The basic block proposed in Restormer [54] is employed as the transformer block. The block implicitly captures long-range pixel interactions by applying self-attention across channels. Thus, the computational complexity of the blocks grows linearly with the spatial resolution, while that of the transformer-based methods that apply self-attention across the spatial dimension grows quadratically. The efficient blocks make TBSR applicable to large images. During the experiment, they use $m = n = 8$. The total number of parameters for TBSR model is $\sim$ 24 M.

Training: For track 1, the team utilizes ℓ_1 loss to train TBSR end to end. The training burst data is synthesized from sRGB images in the Zurich RAW to RGB [15] training set. The model pre-trained in track 1 is used as the initialization model for track 2. Then the model is trained with a combination of ℓ_1 loss with alignment [57], VGG-based perceptual loss [42] and sliced Wasserstein (SW) loss on BurstSR [4] training set.

Inference: During testing, the team uses a self-ensemble strategy [24] for better quantitative performance.

4.3. LD

The team employs the Deep-Rep [6] model. Please refer to [6] for more details.

4.4. MegSR

MegSR proposes a **B**urst **S**uper-**R**esolution **T**ransformer (BSRT) which improves the capability in alignment and reconstruction processes by proposing a Pyramid Flow-Guided Deformable Convolution Network (**Pyramid FG-DCN**) and incorporating Swin Transformer Blocks [27] as the main backbone.

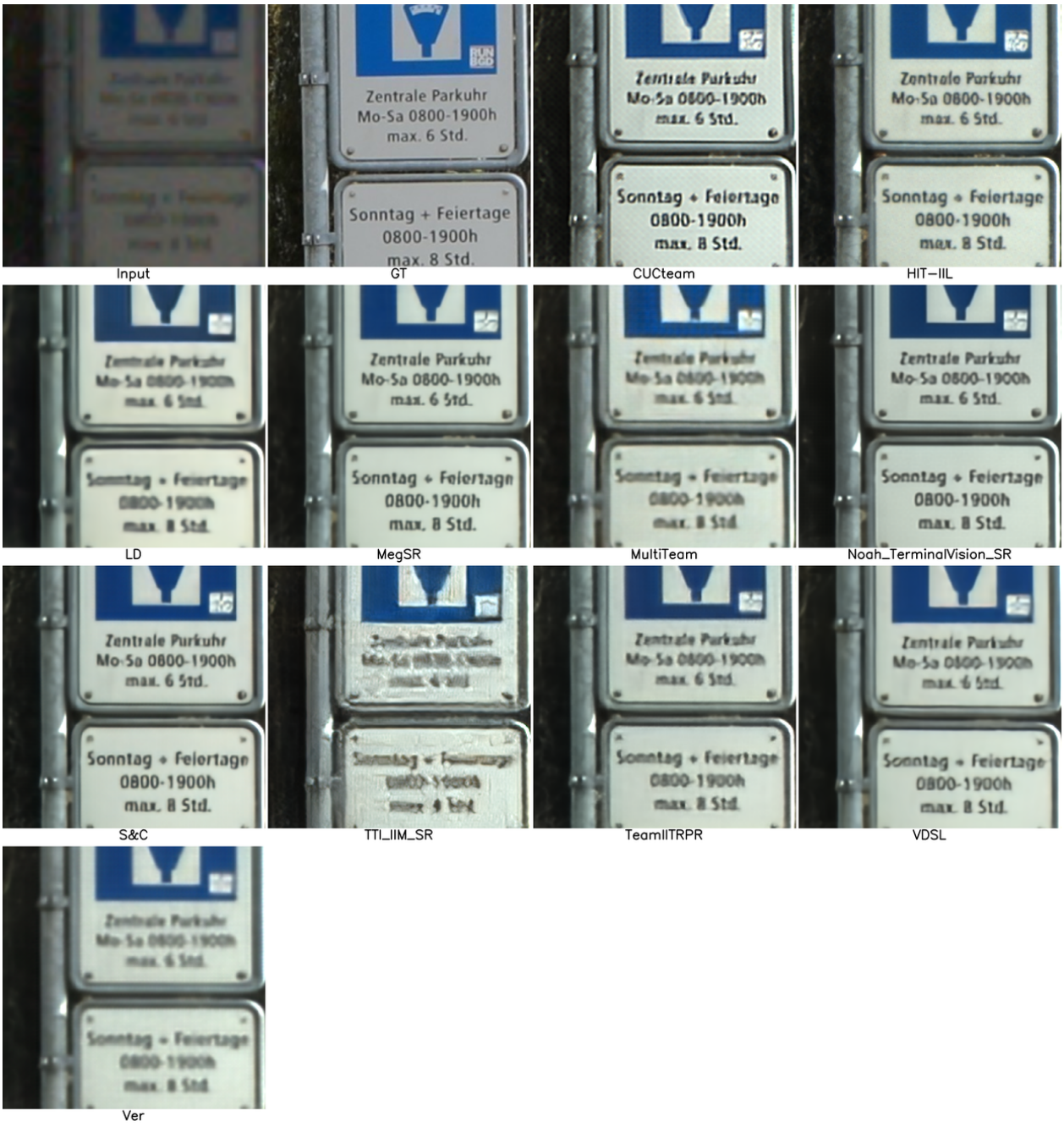

Figure 3. Qualitative comparison on test set of Track 2. Input denotes first image of the burst processed using LibRaw to obtain an RGB image. Note that the predictions of the methods have been post-processed to approximately match the color space of the ground truth.

Network Architecture: The overview of the proposed BSRT framework is shown in Figure 9. Inspired by BasicVSR++ [9], BSRT combines the flow-based alignment and deformable alignment. Specifically, the optical flow estimated by the SpyNet [38] can be regarded as a coarse alignment prior. Based on these flows, DCNs are used to learn more accurate and refined offsets for aligning features. The input images $\{x_i\}_{i=1}^{N}$ are 4-channel 'RGGB' RAW sequence. These are firstly flattened to single channel and passed through SpyNet [38] to obtain multi-level optical flows which are calculated from each frame and the reference frame. Particularly, BSRT uses pre-trained SpyNet weights and preserve the top-3 levels of flows to guide corresponding level's deformable convolution network (DCN) alignment. Meanwhile, the original 4-channel RAW inputs are send to several Swin Transformer Blocks (ST Blocks)

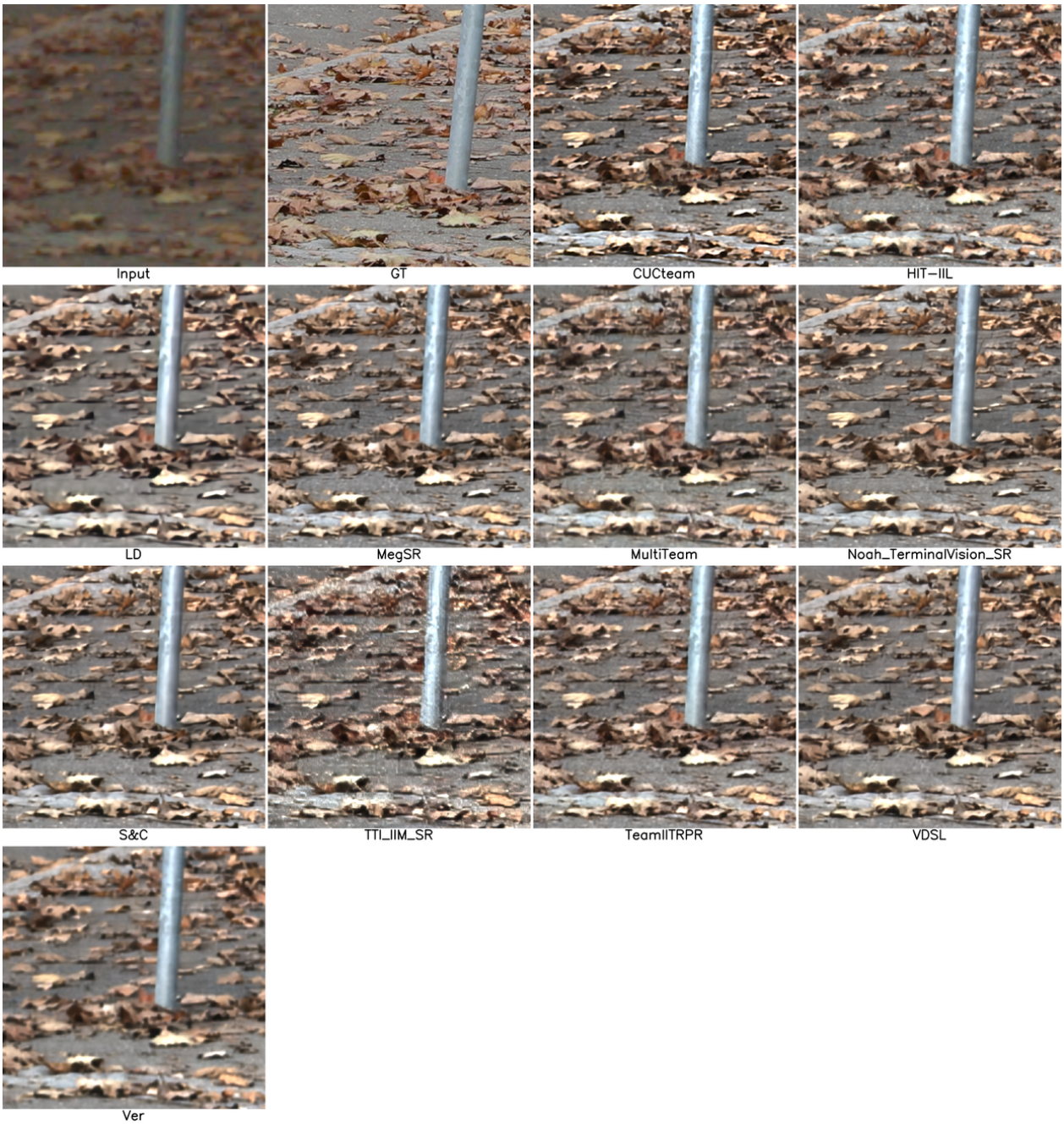

Figure 4. Qualitative comparison on test set of Track 2. Input denotes first image of the burst processed using LibRaw to obtain an RGB image. Note that the predictions of the methods have been post-processed to approximately match the color space of the ground truth.

to extract informative features. These features are then upscaled 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. The details of the Flow-Guided DCN (FG-DCN) level is illustrated in Figure 10. After that, these features are fused (1×1 Conv) and passed via several Swin Transformer Groups to reconstruct the high-resolution image. Please refer to [29] for more details.

Training and inference: The model is trained on the Synthetic dataset for Track 1, and then finetuned on the real world BurstSR dataset [4] for Track 2. Following [4, 5], the team employed AlignedL1 loss for real dataset (Track 2) with a pre-trained PWC-Net [43]. In testing, since this Challenge is on RAW domain, MegSR adopted the Test

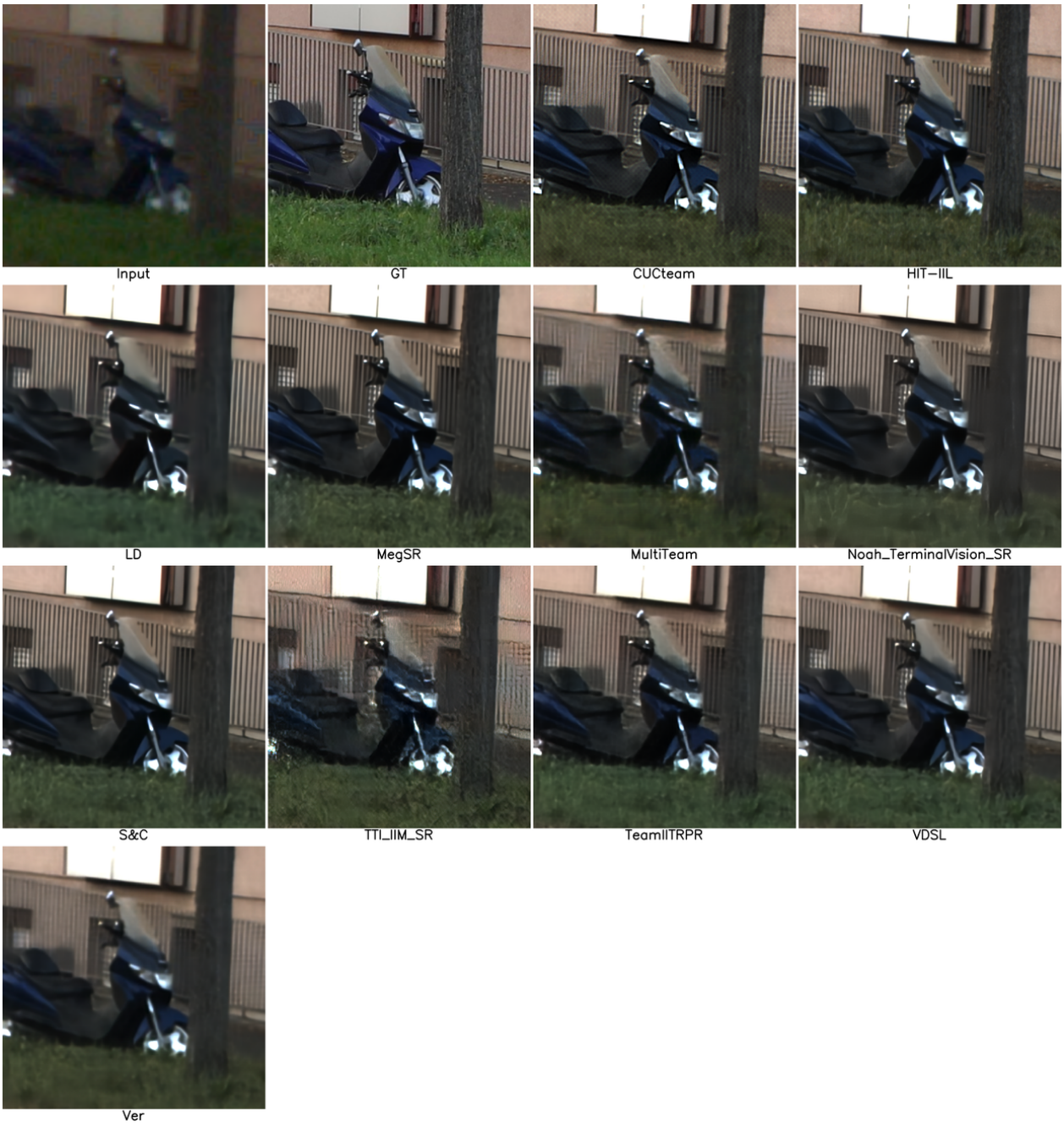

Figure 5. Qualitative comparison on test set of Track 2. Input denotes first image of the burst processed using LibRaw to obtain an RGB image. Note that the predictions of the methods have been post-processed to approximately match the color space of the ground truth.

Time Augmentation (TTA) strategy proposed by [25] for Synthetic data (Track 1).

4.5. MultiTeam

Network architecture: An overview of the network is provided in Fig. 11. The SR network contains six modules: Denoise Module, Feature Extractor, Align Module, Fusion Module and Upsampler. The Denoise Module is based on U-net [40]. In the Feature Extractor, five Wide Activation Residual Blocks are applied. The FEPCD module[31] is empolyed in Align Module. The Fusion Module is similar to BurstSR[4]. The Upsampler is realized by combined U-shape with residual block. The Discriminator Network consists of ResNet18[14] and three fully-connect layers.

Figure 6. Qualitative comparison on test set of Track 2. Input denotes first image of the burst processed using LibRaw to obtain an RGB image. Note that the predictions of the methods have been post-processed to approximately match the color space of the ground truth.

Training: In the Track 1, the synthetic dataset is used to train the model using L2 Loss. Then, Track 2 Real, MultiTeam combines the alignedL2 loss of real dataset [4] with the discriminator loss of GAN[36] to finetune the model trained by synthetic dataset.

4.6. Noah_TerminalVision_SR

The team employs previously proposed state-of-the-art NoahBurstSRNet [5] that consists of 4 modules: encoder, alignment module, weight prediction fusion, and reconstruction blocks (see Fig. 12). As an alignment module, NoahBurstSRNet uses Pyramid, Cascading, and Deformable (PCD) alignment module [46] that consists of de-

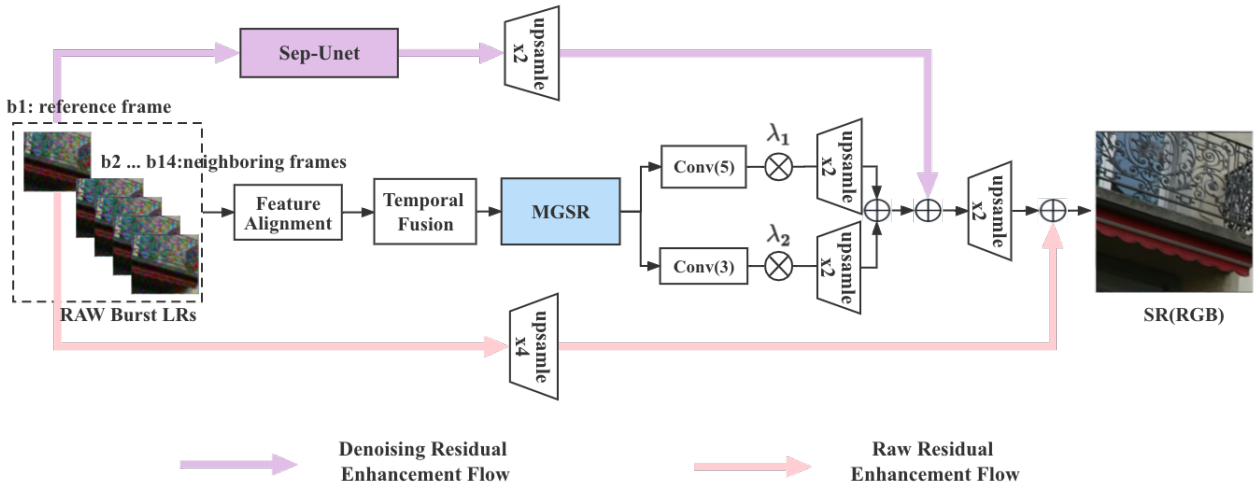

Figure 7. An overview of RAW Burst Super-Resolution with Enhanced Multi-Residule Denoising Net (RBSR) proposed by team CUCteam. The net contains two steps, a reconstruction step and a Multi-Residule step. The reconstruction step contains a Feature Alignment module, a Temporal Fusion module, and a Multi-group Spatial Resconstruction (MGSR) module; The Multi-Residule step contains a denoising residual enhancement flow and a RAW residual enhancement flow. Each lambda (λ) is a trainable scalar parameter.

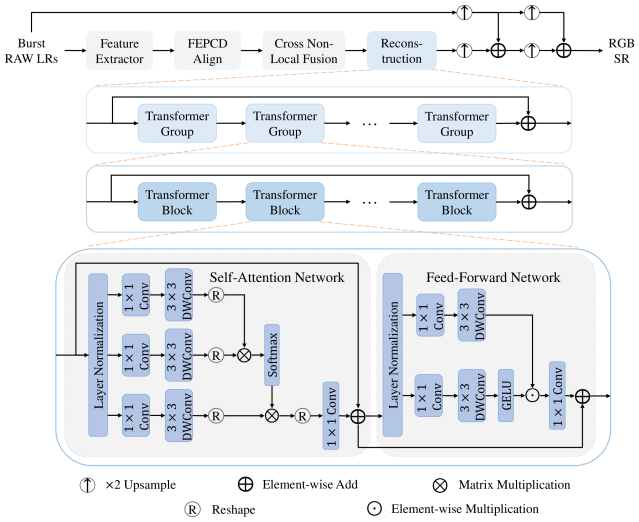

Figure 8. Overview of the network architecture for TBSR proposed by team HIT-IIL

formable convolution (DConv) layers [58]. According to [8], DConv suffers from unstable training, since the offset overflow may cause severe performance degradation and is not implemented for smartphone devices. Thus, the team demonstrates that the PCD module can be replaced by a conventional CNN-based module with negligible loss of performance. Particularly, they replace all DConv layers from PCD with small ResUNet with depth 2.

Training and Inference: For track 1, besides previous year's state-of-the-art model NoahBurstSRNet, the team trained other 3 models with the same procedure: (a) NBSR-PCD, which is a NoahBurstSRNet with less number of RFDB blocks [23], (b) NBSR-ResUNet A and B are two NBSR models with ResUNet instead of DConv. The weighted ensemble of 4 models with the self-ensemble technique achieves the best performance. The inference time of the ensemble solution is 16.85 sec. per 256×256 burst sequence. For track2, the team employed a single NoahBurstSRNet model, which was trained as described in [5].

4.7. okfine

The team proposes an efficient model with deformable alignment and adaptive feature fusion for burst SR (AFF-BSR), which can be divided into four parts: feature extraction, alignment, fusion and reconstruction.

Network architecture: The architecture of the proposed model is shown in Fig 13. Several RCABs[56] are first stacked to extract features from raw burst images. The features are then aligned by a pyramid alignment module (PAM). PAM directly processes features on three different scales, and then concatenates all the aligned features into a channel attention layer for finer alignment. The aligned features are then passed to a Conv3d-based residual block for feature fusion. In this step, the input features are progressively reduced on temporal dimension using Conv3d convolutions for adaptive feature fusion. Finally, the SFT Layer[48] is introduced to embed feature priors for better reconstruction. The network takes features before the upsampling layer of the pretained EBSR [30] as a prior, and EBSR is fixed during training. With the embedded features, a progressive upsample module with PixelShuffle operation

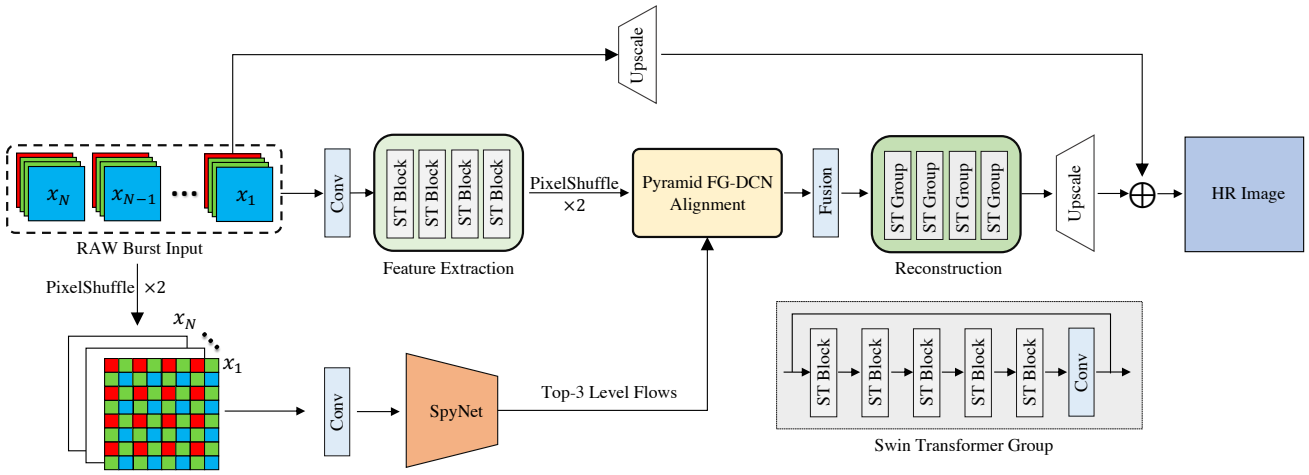

Figure 9. An overview of the BSRT framework proposed by team MegSR.

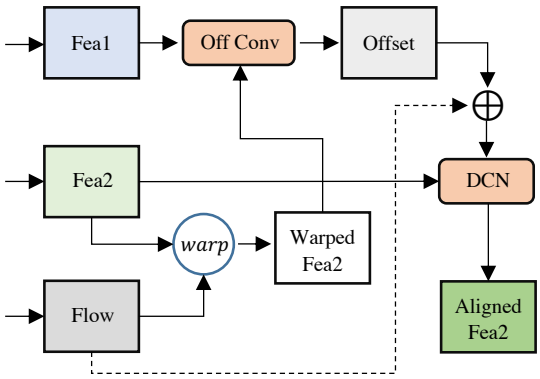

Figure 10. Flow-Guided Deformable Convolution Network (FG-DCN) employed by team MegSR.

is applied to generate the final SR image.

Training and inference: The team randomly crops 16 patches of size 64×64 from the raw burst images as input for each training mini-batch. Data augmentation is performed on the training set, such as random rotations and horizontal flips. The proposed model is trained by minimizing L1 loss function with Adam optimizer and L2 loss is also used for fine-tuning. For the final submission, the team uses self-ensemble technique[31] to boost the performance.

4.8. S&C

The team uses the same network architecture as EBSR [30]. The pre-trained EBSR model is fine-tuned according to spatial resolution of the validation dataset used in both track 1 and track 2 to reduce domain gap.

4.9. TeamIITRPR

The team proposes adaptive feature consolidation network (AFCNet) for burst super-resolution. The proposed architecture combines and processes the information from individual low-resolution (LR) images for generating high-resolution (HR) output. AFCNet works in four steps: (a) Feature alignment, (b) Feature extraction, (c) Feature fusion and (d) Feature up-sampling. Burst features are initially aligned using a deformable convolution based feature alignment module. Further, high-frequency residue is evaluated by taking the difference between these aligned features and reference frame features followed by its addition to the aligned features [10]. The aligned features are processed through multi-head multi-level transformer backbone [54] which overall enhances the aligned features. These consolidated burst features are further passed through a fusion module similar to [10] to enable inter-frame communication by generating pseudo bursts. The final high-resolution image is reconstructed through a progressive adaptive group up-sampling (AGU) module [10]. For more details, please refer to [34].

Training: The shared parameters across several stages are jointly learned by minimizing the L_1 loss with respect to network parameters. For Track 1, the network is trained on synthetic bursts generated by utilising 46,839 sRGB images from Zurich-RAW-to-RGB dataset [15] for 100 epochs. For Track 2, network trained on synthetic data is fine-tuned for 25 epochs on BurstSR train set with alignedL1 loss [4].

4.10. TTI_IIM_SR

The team extends Single Image Super-Resolution (SISR) network architectures to the Burst Image Super-Resolution (BISR) task. This is achieved by changing the number of channels in the first convolution layer so that a fixed num-

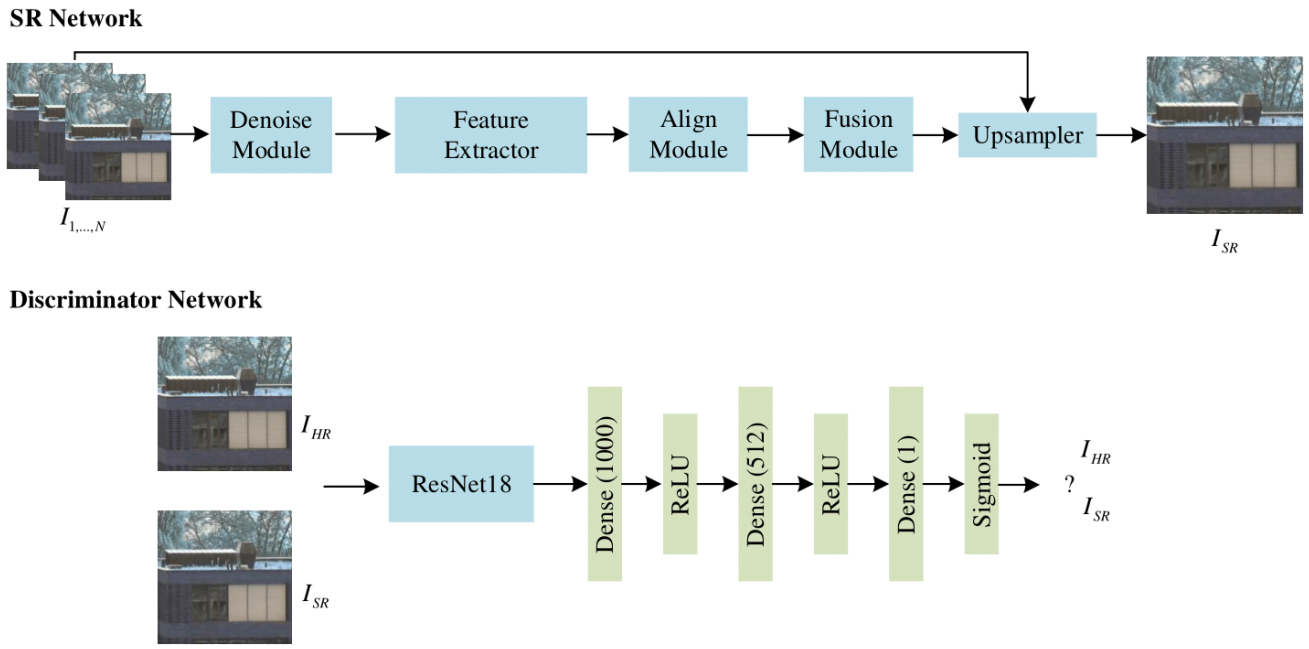

Figure 11. Overview of network architecture employed by MultiTeam.

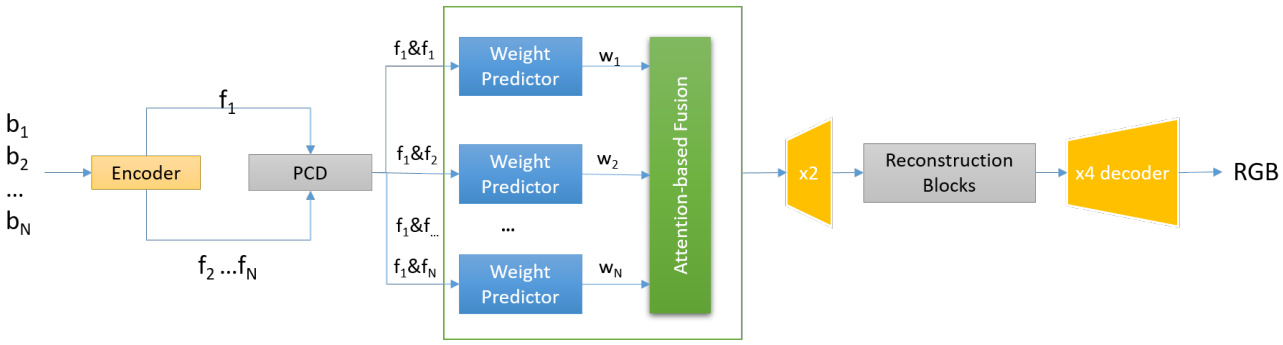

Figure 12. The architecture overview of the NoahBurstSRNet employed by the Noah_TerminalVision_SR team.

ber of multiple images could be fed into the network. The overview of the model is shown in Fig. 15. SwinIR [21] is used as the base model in track 1, while Real-ESRGAN [47] is used in track 2.

Training: The team uses a SISR model trained on several datasets (i.e., DIV2K [1], Flickr2K [44], OST [49], WED [32], FFHQ [17], Manga109 [33], and SCUT-CTW1500 [26]) as the base model. The first layer of the model is modified to receive 14 RAW images as input. For track 1, the modified network is finetuned on the synthetic burst dataset using L1 loss. In track 2, the loss function used in Real-ESRGAN [47] is used to finetune on the BurstSR dataset [4]. Specifically, the loss function is a weighted linear sum of adversarial loss L_{adv} [12, 19] which restores high-frequency components, reconstruction loss L_{recon} which restores global structure, and perceptual loss L_{percep} [16] which enhances the perceptual quality.

4.11. VDSL

The method presented by team VDSL is inspired by DBSR [4] and DeepRep[6], improving the performance in two ways. First, the original alignment module in DBSR is replaced with Enhanced Residual Deformable Alignment module (ERDA) which enables more accurate alignment. Second, after the fusion of the aligned features, the MAP step [6] is implemented between the upsampling stages of X2 and X4. Furthermore, in the reconstruction stage, the MAP step enables the reconstruction error to minimized by learning the latent space.

Network Architecture: An overview of the network architecture is provided in Figure 16. In order to obtain accurate alignment, the network adds the Residual DCN alignment along with the optical flow based PWCNet[43] employed in DBSR[4]. The flow warped feature and the base feature are first aligned to produce offsets for DCN. The resid-

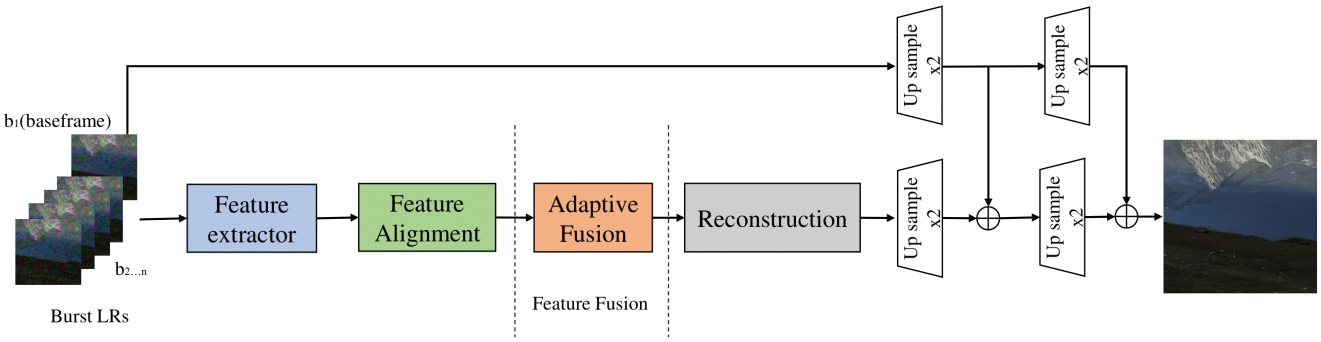

Figure 13. Overview of the network architecture of the AFFBSR proposed by team okfine.

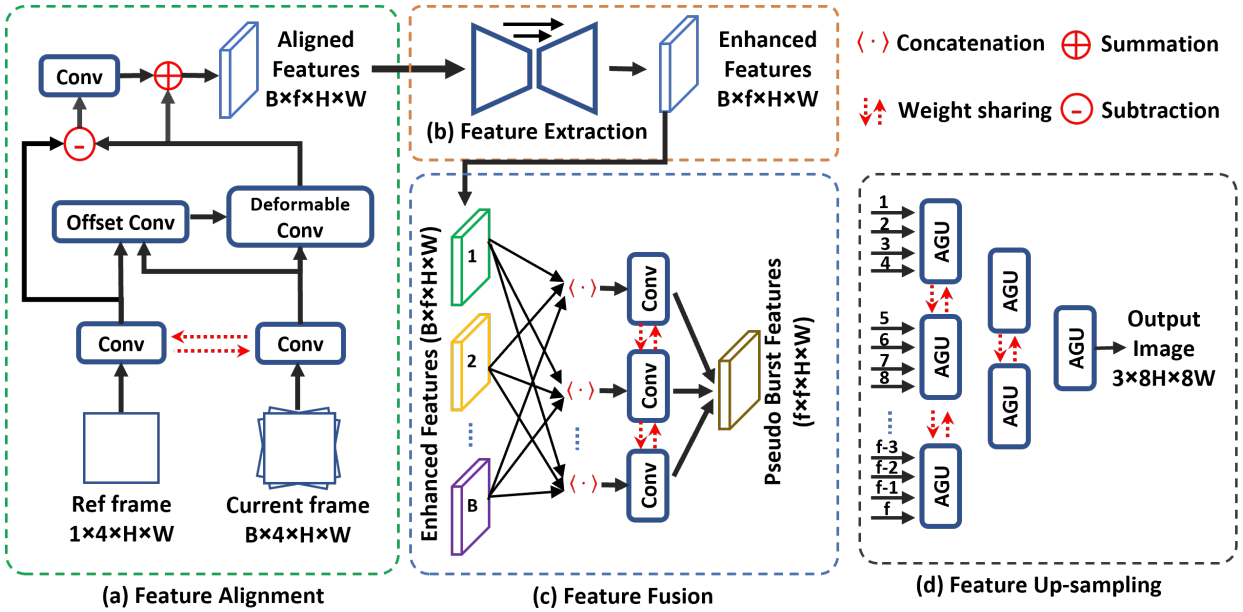

Figure 14. Overall pipeline of the adaptive feature consolidation network for burst super-resolution network (AFCNet) proposed by TeamI-ITRPR.

ual feature between the base feature and the DCN alignment feature is passed through a conv layer and added to DCN alignment feature to obtain final aligned feature [10] (see Figure 17). The aligned features are merged using the attention-based fusion from DBSR[4]. Motivated by the DeepRep[6] model, VDSL performs reparameterized maximum a posteriori estimation in the deep feature space. The MAP step operates to minimize the reconstruction error between the fused feature and a simulated feature map. The optimized latent space representation is up-sampled using

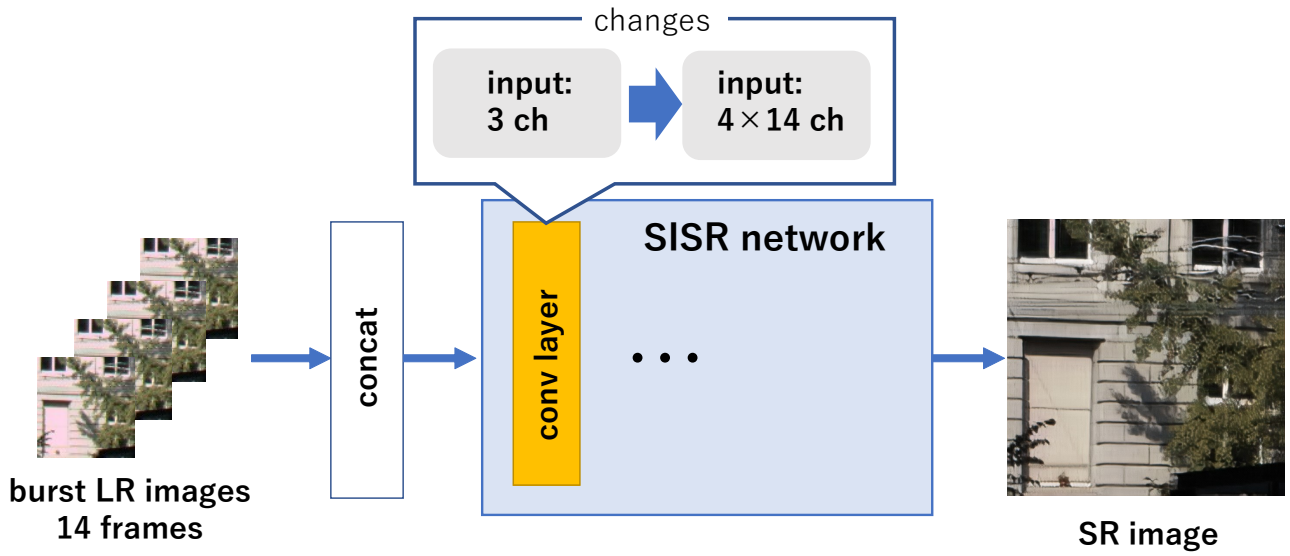

Figure 15. The overview of the approach used by team TTI_IIM_SR.

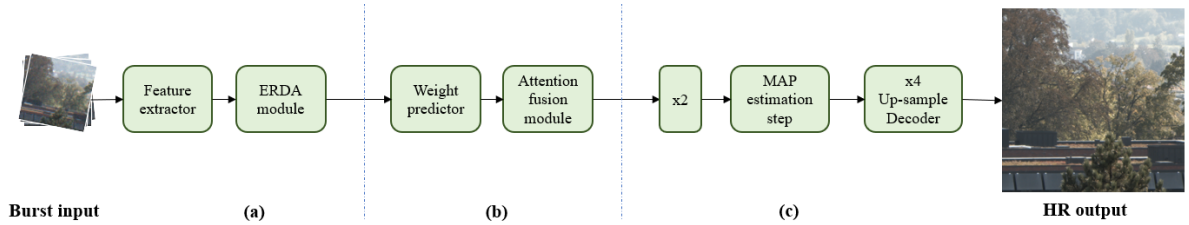

Figure 16. Overview of the network architecture employed by team VDSL.

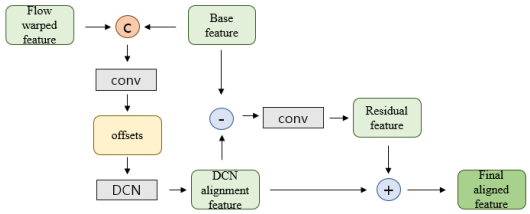

Figure 17. Enhanced Residual Deformable Alignment module employed by team VDSL

the pixel shuffle to obtain the output image.

Training: The model in track 1 is trained using an ADAM optimizer and minimizing L1 loss for 600 epochs on synthetic bursts generated using Zurich RAW to RGB dataset. This model is fine-tuned for 50 epochs on BurstSR dataset[4] using aligned L1 loss for track 2.

4.12. Ver

The team uses EBSR [31] as the backbone of the proposed method. The specific modifications are: (1) Learn high frequency information to improve texture.(2) Better learning rate strategies. (3) Additional high quality datasets. The team explores the use of multi-scale non-local fusion. Next, in order to improve the details after super-resolution, wavelet calculation is added to the loss to constrain high-frequency information more strongly. Residual channel attention is also employed in the model backbone to improve the performance.

4.13. VIDAR_A

The team proposes a Multi-fusion Network for burst super-resolution, as shown in Figure 19. Motivated by [31], they solve the burst SR problem in four steps: alignment, fusion, reconstruction, and refinement. Previously, information from multiple burst images are fused through concatenation [31] or Weight Predictor (WP) [4]. Instead, team VIDAR_A proposes to fuse information in both spatial and frequency domains.

Network architecture: Following [31], the network extracts features with a feature enhance pyramid network and aligns the features from the burst images using the well-known Pyramid, Cascading and Deformable (PCD) module [46]. To conquer the drastic dimension reduction in the fusion module, the team proposes a Multi-Fusion Module. As shown in Figure 19, the aligned features are fused in

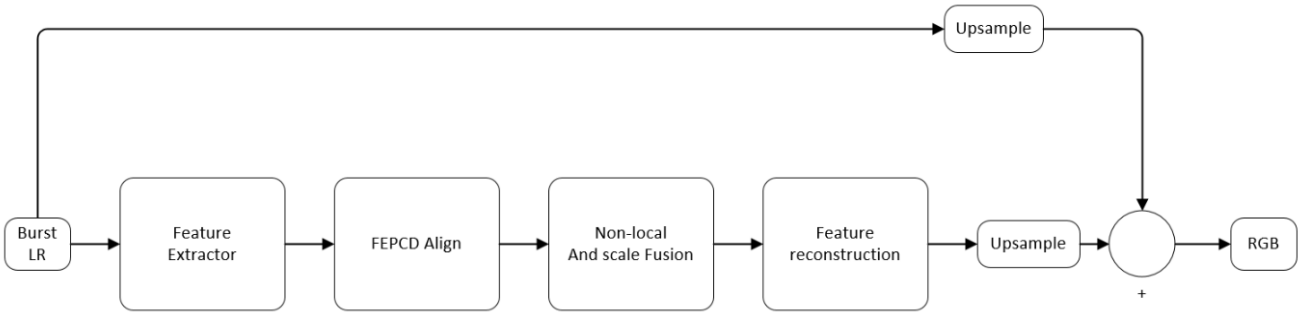

Figure 18. An overview of the architecture employed by team Ver.

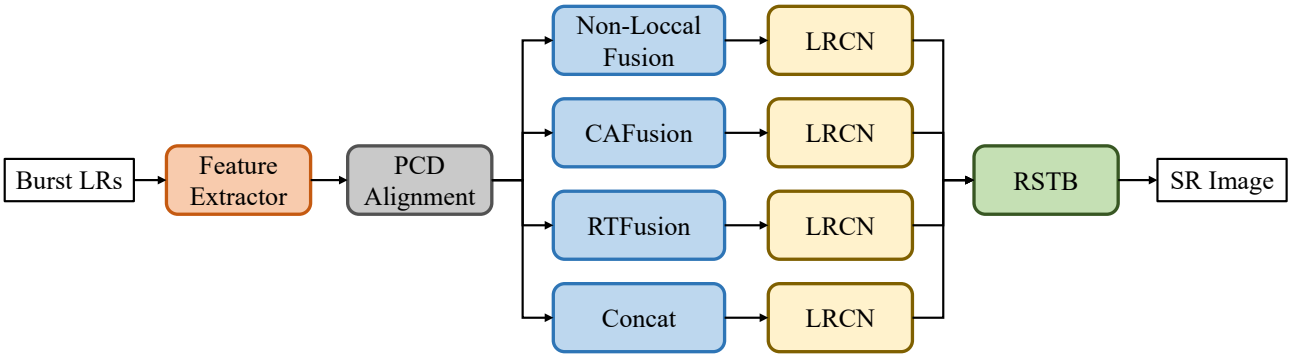

Figure 19. Overview of the multi-fusion network architecture employed by team VIDAR_A

different ways, including Channel Attention Fusion (CAFusion), Residual Frequency domain Fusion (RFFusion), Non-Local Fusion, and a normal concatenation fusion. The Long Range Concatenation Network (LRCN) [31] is then used to reconstruct an intermediate SR image and multiple Residual Swin Transformer Blocks (RSTB) [21] are used to refine the SR image.

Training: The network is first pre-trained on the synthetic bursts generated from Zurich raw to RGB dataset [15] and then finetuned on the BurstSR Dataset [4], using L_1 loss.

4.14. VIDAR_B

The team proposes multi-scale transformer for burst image super-resolution (TBSR), which divides this task into two key parts: alignment and fusion (see Figure 20). For alignment, TBSR uses a novel transformer-based module, named as MS-TR module, which captures the correspondence of input features leveraging the multi-scale attention heads (see Figure 21). Additionally, it also uses the PCD alignment module [46] to align the input features. The aligned features of the two modules are concatenated and fed to fusion module. For fusion, TBSR uses the pyramid temporal-spatial attention module [52] to fuse the aligned features. The fused features are then passed through a reconstruction module, which is a cascade of residual channel-wise attention blocks. The upsampling operation is performed at the end of the network to increase the spatial size. Finally, the high-resolution frame is obtained by adding the predicted image residual to a direct upsampled image.

Training and inference: The network is trained using L_1 loss function. An ensemble strategy with multiple checkpoints and multiple models is used to enhance the reconstructed results. Four other models in addition to TBSR are used to obtain the prediction. For model 1, the PSA module is replaced with "CrossNonLocal Fusion" [30]. For model 2, the feature extractor and reconstruction module are replaced with coupling layers [18]. For model 3, the PSA module, feature extractor and reconstruction module are replaced as in model 1 and 2. For model 4, the original EBSR [30] is employed. These four models together with TBSR generates five results which are fused to achieve a better performance.

5. Conclusion

This paper reviews the NTIRE 2022 Burst Super-Resolution challenge. In the challenge, the participants were tasked with performing joint denoising, demosaicking, and super-resolution using multiple input images. That is, given a burst containing multiple noisy RAW images, the task is to combine these images to generate a clean, high-

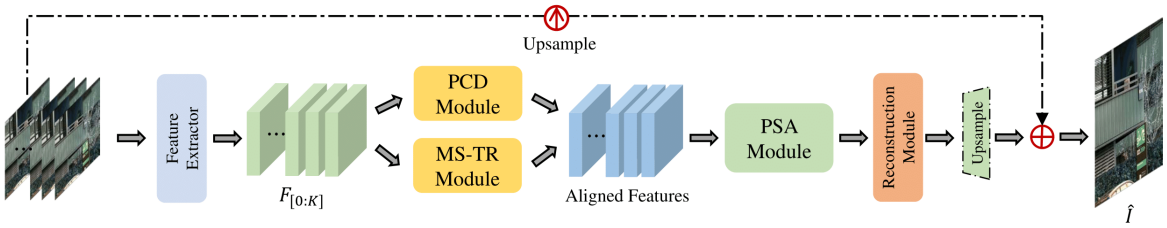

Figure 20. Overview of the network architecture employed by team VIDAR_B.

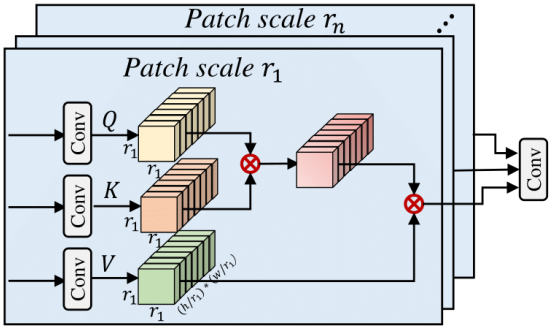

Figure 21. The structure of Multi-Scale Head Transformer Module employed by team VIDAR_B.

resolution RGB image as output. The challenge was held in two tracks, one employing synthetically generated data, while the other used real-world bursts for evaluation. 14 teams participated in the challenge, employing diverse network architectures and training strategies to tackle the burst super-resolution task.

Acknowledgments

We thank the NTIRE 2022 sponsors: Huawei, Reality Labs, Bending Spoons, MediaTek, OPPO, Oddity, Voyage81, ETH Zürich (Computer Vision Lab) and University of Würzburg (CAIDAS).

Appendix

A. Teams and Affiliations

NTIRE2022 Organizers

Members:
Goutam Bhat[1] (goutam.bhat@vision.ee.ethz.ch)
Martin Danelljan[1] (martin.danelljan@vision.ee.ethz.ch)
Radu Timofte[1,2] (radu.timofte@uni-wuerzburg.de)
Affiliations:
[1] Computer Vision Lab, ETH Zürich, Switzerland
[2] Julius Maximilian University of Würzburg, Germany

CUCteam

Title: Raw Burst Super-Resolution with Enhanced Multi-Residual Denoising Net

Team Leader:
Qian Zheng (zhengqian@cuc.edu.cn)
Members:
Qian Zheng, Communication University of China
Yuntian Cao, Communication University of China
Ruipeng Gang, UHDTV Research and Application Laboratory
Chenghua Li, Chinese Academy of Sciences' Institute of Automation
Jinjing Li, Communication University of China
Fangya Li, Communication University of China
Yizhen Cao, Communication University of China
Chenming Liu, UHDTV Research and Application Laboratory

HIT-IIL

Title: Transformer for Burst Image Super-Resolution

Team Leader:
Zhilu Zhang (cszlzhang@outlook.com)
Members:
Zhilu Zhang, Harbin Institute of Technology
Rongjian Xu, Harbin Institute of Technology
Shuohao Zhang, Harbin Institute of Technology
Xiaohe Wu, Harbin Institute of technology
Wangmeng Zuo, Harbin Institute of Technology

LD

Title: Real-world Burst Super-Resolution with Deep Reparametrization

Team Leader:
Wentao Yang (wente_young@foxmail.com)
Members:
Wentao Yang, School of Electronic and Information Engineering, South China University of Technology
Liufeng Huang, School of Electronic and Information Engineering, South China University of Technology
Zhuoming Liu, School of Electronic and Information Engineering, South China University of Technology

MegSR

Title: BSRT: Improving Burst Super-Resolution with Swin Transformer and Flow-Guided Deformable Alignment
Team Leader:
Ziwei Luo (luoziwei@megvii.com)
Members:
Ziwei Luo, Megvii Technology
Shen Cheng, Megvii Technology
Lei Yu, Megvii Technology
Zhihong Wen, Megvii Technology
Qi Wu, Megvii Technology
Youwei Li, Megvii Technology
Haoqiang Fan, Megvii Technology
Jian Sun, Megvii Technology
Shuaicheng Liu, Megvii Technology, University of Electronic Science and Technology of China (UESTC)

MultiTeam

Team Leader:
Meiya Chen (2560612536@qq.com)
Members:
Meiya Chen, Xiaomi
Cong Zhang, Xiaomi
Gaofeng Zhou, Xiaomi

Noah_TerminalVision_SR

Title: NoahBurstSRNet for Raw Burst Image Super-Resolution
Team Leader:
Magauiya Zhussip (magauiya.zhussip1@huawei.com)
Members:
Magauiya Zhussip, Noah‘s Ark Lab, Huawei
Pavel Ostyakov, Noah‘s Ark Lab, Huawei
Xueyi Zou, Noah‘s Ark Lab, Huawei
Youliang Yan, Noah‘s Ark Lab, Huawei

okfine

Title: AFFBSR: Adaptive Feature Fusion for Burst Image SR
Team Leader:
Zhongbao Yang (121106010692@njust.edu.cn)
Members:
Zhongbao Yang, Nanjing University of Science and Technology
Long Sun, Nanjing University of Science and Technology
Jinhui Tang, Nanjing University of Science and Technology
Zechao Li, Nanjing University of Science and Technology
Jinshan Pan, Nanjing University of Science and Technology

S&C

Title: Fine-Tuned EBSR Model
Team Leader:
Jianxing Zhang (jx2018.zhang@samsung.com)
Members:
Jianxing Zhang, SRC-B
Zhengxiong Luo, CASIA
Zikun Liu, SRC-B
Jian Gao, SRC-B

TeamIITRPR

Title: Adaptive Feature Consolidation Network for Burst Super-Resolution
Team Leader:
Subrahmanyam Murala (subbumurala@iitrpr.ac.in)
Members:
Nancy Mehta, Indian Institute of Technology Ropar (IIT Ropar)
Akshay Dudhane, Mohamed bin Zayed University of AI (MBZUAI)
Subrahmanyam Murala, Indian Institute of Technology Ropar (IIT Ropar)
Syed Waqas Zamir, Inception Institute of Artificial Intelligence (IIAI)
Salman Khan, MBZUAI, Australian National University (ANU)
Fahad Shahbaz Khan, MBZUAI, Linkoping University

TTI_IIM_SR

Title: Expansion of Single Image Super-Resolution Networks to Multi-input
Team Leader:
Yuki Kondo (sd18037@toyota-ti.ac.jp)
Members:
Yuki Kondo, Toyota Technological Institute (TTI)
Fuma Yasue, Toyota Technological Institute (TTI)
Chihiro Nakatani, Toyota Technological Institute (TTI)
Norimichi Ukita, Toyota Technological Institute (TTI)

VDSL

Title: Burst Super Resolution using Enhanced Residual Deformable Alignment and MAP estimation
Team Leader:
Yoonchan Nam (dbscks0825@gmail.com)
Members:
Yoonchan Nam, Sogang University

Youngsu Jo, Sogang University
Sukju Kang, Sogang University

Ver

Title: WBSR:Burst Super-resolution used Wavelet Calculation
Team Leader:
Yan Gu (18583993892@163.com)
Members:
Yan Gu, UESTC
Ge Song, USTC
Jialin Yang, WHU

VIDAR_A

Title: Multi-Fusion Network for Burst Super-Resolution
Team Leader:
Zhiwei Xiong (zwxiong@ustc.edu.cn)
Members:
Zhiwei Xiong, University of Science and Technology of China
Xihao Chen, University of Science and Technology of China
Zeyu Xiao, University of Science and Technology of China
Mingde Yao, University of Science and Technology of China
Ruikang Xu, University of Science and Technology of China
Jie Huang, University of Science and Technology of China

VIDAR_B

Title: Multi-Scale Transformer Alignment for Burst Super-Resolution
Team Leader:
Zhiwei Xiong (zwxiong@ustc.edu.cn)
Members:
Zhiwei Xiong, USTC
Mingde Yao, USTC
Ruikang Xu, USTC
Xihao Chen, USTC
Zeyu Xiao, USTC
Jie Huang, USTC

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 13

[2] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, et al. NTIRE 2022 spectral demosaicing challenge and dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[3] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, et al. NTIRE 2022 spectral recovery challenge and dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[4] Goutam Bhat, Martin Danelljan, L. Gool, and R. Timofte. Deep burst super-resolution. In *CVPR*, 2021. 2, 3, 6, 8, 9, 10, 12, 13, 14, 15, 16

[5] 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. 3, 5, 8, 10, 11

[6] 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. 6, 13, 14

[7] T. Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and J. Barron. Unprocessing images for learned raw denoising. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11028–11037, 2019. 2

[8] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. *arXiv preprint arXiv:2009.07265*, 4(3):4, 2020. 11

[9] 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. 7

[10] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. *arXiv preprint arXiv:2110.03680*, 2021. 12, 14

[11] Egor Ershov, Alex Savchik, Denis Shepelev, Nikola Banic, Michael S Brown, Radu Timofte, et al. NTIRE 2022 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 13

[13] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy Ren, Radu Timofte, et al. NTIRE 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 9

[15] A. Ignatov, L. Gool, and R. Timofte. Replacing mobile camera isp with a single deep learning model. *2020 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2275–2285, 2020. 6, 12, 16

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, 2016. 13

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 13

[18] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 16

[19] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 13

[20] Yawei Li, Kai Zhang, Radu Timofte, Luc Van Gool, et al. NTIRE 2022 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[21] 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. 13, 16

[22] O. Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, J. Barron, Dillon Sharlet, Ryan Geiss, S. W. Hasinoff, Y. Pritch, and M. Levoy. Handheld mobile photography in very low light. *ACM Transactions on Graphics (TOG)*, 38:1 – 16, 2019. 1

[23] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. *arXiv preprint arXiv:2009.11551*, 2020. 11

[24] Jiaming Liu, Chi-Hao Wu, Yuzhi Wang, Qin Xu, Yuqian Zhou, Haibin Huang, Chuan Wang, Shaofan Cai, Yifan Ding, Haoqiang Fan, et al. Learning raw image denoising with bayer pattern unification and bayer preserving augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6

[25] Jiaming Liu, Chi-Hao Wu, Yuzhi Wang, Qin Xu, Yuqian Zhou, Haibin Huang, Chuan Wang, Shaofan Cai, Yifan Ding, Haoqiang Fan, et al. Learning raw image denoising with bayer pattern unification and bayer preserving augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 9

[26] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, and Sheng Zhang. Detecting curve text in the wild: New dataset and new solution. *CoRR*, 2017. 13

[27] 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. 4, 6

[28] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2022 challenge on learning the super-resolution space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[29] Ziwei Luo, Youwei Li, Shen Cheng, Lei Yu, Qi Wu, Wen Zhihong, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. BSRT: Improving burst super-resolution with swin transformer and flow-guided deformable alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 8

[30] 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. 3, 5, 6, 11, 12, 16

[31] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 9, 12, 15, 16

[32] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016. 13

[33] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 13

[34] Nancy Mehta, Akshay Dudhane, Subrahmanyam Murala, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Adaptive feature consolidation network for burst super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 12

[35] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018. 1

[36] Yudai Nagano and Yohei Kikuta. Srgan for super-resolving low-resolution food images. In *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*, pages 33–37, 2018. 10

[37] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Richard Shaw, Ales Leonardis, Radu Timofte, et al. NTIRE 2022 challenge on high dynamic range imaging: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[38] 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. 7

[39] Andres Romero, Angela Castillo, Jose M Abril-Nova, Radu Timofte, et al. NTIRE 2022 image inpainting challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 9

[41] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. When and why test-time augmentation works. *arXiv e-prints*, pages arXiv–2011, 2020. 6

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6

[43] Deqing Sun, X. Yang, Ming-Yu Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 8, 13

[44] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 13

[45] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. NTIRE 2022 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[46] 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. 3, 10, 15, 16

[47] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021. 13

[48] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 11

[49] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 13

[50] Zhou Wang, A. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. 3

[51] B. Wronski, Ignacio Garcia-Dorado, M. Ernst, D. Kelly, Michael Krainin, Chia-Kai Liang, M. Levoy, and P. Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38:1 – 18, 2019. 1

[52] Ruikang Xu, Zeyu Xiao, Jie Huang, Yueyi Zhang, and Zhiwei Xiong. Edpn: Enhanced deep pyramid network for blurry image restoration. In *CVPRW*, 2021. 16

[53] Ren Yang, Radu Timofte, et al. NTIRE 2022 challenge on super-resolution and quality enhancement of compressed video: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[54] 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. 6, 12

[55] Richard Zhang, Phillip Isola, Alexei A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 3

[56] 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. 11

[57] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *ICCV*, pages 4348–4358, 2021. 6

[58] 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. 11