NTIRE 2021 Challenge on Burst Super-Resolution: Methods and Results

Goutam Bhat*   Martin Danelljan*   Radu Timofte*   Kazutoshi Akita   Wooyeong Cho
Haoqiang Fan   Lanpeng Jia   Daeshik Kim   Bruno Lecouat   Youwei Li
Shuaicheng Liu   Ziluan Liu   Ziwei Luo   Takahiro Maeda   Julien Mairal
Christian Micheloni   Xuan Mo   Takeru Oba   Pavel Ostyakov   Jean Ponce
Sanghyeok Son   Jian Sun   Norimichi Ukita   Rao Muhammad Umer   Youliang Yan
Lei Yu   Magauiya Zhussip   Xueyi Zou

Abstract

This paper reviews the NTIRE2021 challenge on burst super-resolution. Given a RAW noisy burst as input, the task in the challenge was to generate a clean RGB image with 4 times higher resolution. The challenge contained two tracks; Track 1 evaluating on synthetically generated data, and Track 2 using real-world bursts from mobile camera. In the final testing phase, 6 teams submitted results using a diverse set of solutions. The top-performing methods set a new state-of-the-art for the burst super-resolution task.

1. Introduction

Super-resolution (SR) is a fundamental computer vision problem with numerous applications in *e.g.* mobile photography, remote sensing, medical imaging. Given a single or multiple images of a scene, SR aims to generate a higher-resolution output by adding missing high-frequency details. In recent years, the SR community has mainly focused on the single-image super-resolution (SISR) task [8, 9, 24, 28, 22, 64, 49, 21, 27, 58, 53, 34]. Thanks to the development of specialized network architectures [9, 24, 28, 22, 64, 49] and training strategies [21, 27, 58, 53], the SISR methods have achieved impressive SR performance. Despite these advances, the SISR approaches are fundamentally limited by the available information (single frame), and thus only rely on learned image priors to recover the missing details.

In contrast, multi-frame super-resolution (MFSR) approaches combine information from multiple low-resolution (LR) images to generate a HR output. If the in-put images are captured using a non-stationary camera and thus contain sub-pixel shifts w.r.t. each other, they provide multiple LR samplings of the same underlying scene. By effectively fusing this information, the MFSR methods can reconstruct high-frequency details which otherwise cannot be recovered using a single input image. This makes MFSR especially tempting for the popular mobile burst photography applications. Since the burst images contain sub-pixel shifts due to natural hand tremors [55], MFSR can be employed to improve the image resolution which is otherwise restricted by hardware constraints.

Despite the aforementioned advantages, the MFSR problem has received limited attention in recent years, compared to the SISR task. The recent work [4] by Bhat *et al*. aims to address this issue by introducing a synthetic, as well as a real-world dataset for burst super-resolution, in addition to a MFSR network architecture. The NTIRE 2021 Challenge on Burst Super-Resolution aims to further stimulate research in the burst super-resolution task. The challenge consists of two tracks. In Track 1, the methods are evaluated on the synthetic burst dataset introduced in [4], and ranked using standard fidelity score PSNR. Track 2 evaluates the real world performance on the BurstSR dataset introduced in [4]. The BurstSR dataset consists of bursts captured using a hand held camera, along with a corresponding HR image captured using a DSLR. The methods are ranked using a combination of fidelity score as well as a human study.

In total, 6 teams participated in the NTIRE 2021 Challenge on Burst Super-Resolution. The participating teams employed a variety of fusion approaches, alignment modules, and reconstruction networks. 4 of the 6 participating teams outperformed DBSR [4] on Track 1, setting a new state-of-the-art on the burst super-resolution task.

2. NTIRE 2021 Challenge

The goal of the NTIRE 2021 Challenge on Burst Super-Resolution is to encourage further research in the burst SR

---

task and provide a common benchmark for evaluating different methods. This challenge is one of the NTIRE 2021 associated challenges: nonhomogeneous dehazing [2], defocus deblurring using dual-pixel [1], depth guided image relighting [10], image deblurring [39], multi-modal aerial view imagery classification [30], learning the super-resolution space [33], quality enhancement of heavily compressed videos [56], video super-resolution [47], perceptual image quality assessment [12], burst super-resolution, and high dynamic range [42]. The burst super-resolution challenge contained two tracks. In both tracks, the methods are provided a noisy RAW burst containing 14 images. The task is to perform joint denoising, demosaicking, and super-resolution to generate a clean RGB image with 4 times higher resolution. The participants were provided a public toolkit (https://github.com/goutamgmb/NTIRE21_BURSTSR) containing tools for training and evaluation for both tracks. Next, we describe the two challenge tracks in more detail.

2.1. Track 1: Synthetic

Track 1 employs synthetic bursts generated using the data generation pipeline employed in [4]. Given an sRGB image, an inverse camera pipeline introduced in [5] is employed to convert the sRGB image to linear sensor space. Next, a synthetic burst is generated by applying random translations from the range $[-24, 24]$ pixels, and random rotations from the range $[-1, 1]$ degrees. Each image in the burst is then downsampled by a factor of 4 using bilinear interpolation and then mosaicked using Bayer pattern. Finally, independent read and shot noise is added to each image to obtain the noisy RAW burst. Due to the use of synthetically generated data, an accurately aligned ground truth image is readily available in Track 1. This allows evaluating the impact of different architectural choices and loss functions on the SR performance, computed in terms of pixel-wise image quality metrics such as PSNR.

The public toolkit provided to the participants contained data generation scripts which could be used to generate synthetic bursts for training. We used the pre-generated synthetic burst dataset introduced in [4] as the validation set for Track 1. The dataset consists of 300 bursts, which have been generated using sRGB images from the Zurich RAW to RGB [19] test set. Each burst contains 14 RAW images of resolution 96×96. The participants could evaluate their methods on the validation set using an evaluation server during the development phase of the challenge. A public leaderboard (https://competitions.codalab.org/competitions/28078#results) was also made available. The dataset for the final test phase, consisting of 500 bursts, was synthetically generated using the DSLR images from the BurstSR [4] test set. Similar to the validation set, each burst in the test set contains 14

96×96 RAW images. The participants were only provided the RAW LR bursts, and asked to submit the predictions of their methods.

2.2. Track 2: Real-World

In this track, we employ the BurstSR dataset introduced in [4] for evaluating the methods. The BurstSR dataset consists of 200 RAW bursts captured from a hand held mobile camera. A corresponding higher-resolution image captured using a DSLR is also provided for each burst to serve as the ground truth. Compared to the synthetic dataset used in Track 1, the BurstSR dataset allows evaluating the performance of the methods on real-world degradation and noise. However, as the input burst and the HR ground truth are captured using different cameras, there are spatial mis-alignment and color differences between the two. This poses additional challenges on the training of the methods as well as evaluation.

The BurstSR dataset is split into train, validation, and test splits consisting of 160, 20, and 20 bursts, respectively. We extracted 160×160 crops from the bursts to obtain our training, validation, and test sets consisting of 5405, 882, and 639 crops, respectively. The participants were allowed to use the provided training set, in addition to external synthetic data, for training their methods. During the development phase, the participants were also provided the validation set, along with the ground truth, for evaluating different design choices. Unlike in Track 1, there was no evaluation server in Track 2. For the final test phase, the participants were provided only the LR bursts from the test set, and asked to submit the network predictions.

3. Challenge Results

In this section, we report the final results on the test sets of both Track 1 and Track 2. During the final test phase, the participants were asked to submit their predictions on the provided test data. In Track 1, there were 6 teams which submitted their methods, while 5 different teams submitted methods in Track 2. All the submitted methods are briefly described in Section 4, while the members and affiliations for each team are listed in Appendix A.

3.1. Evaluation Metrics

The aim in MFSR is to reconstruct the original HR image by fusing information from multiple LR observations. Thus, we employ fidelity based image metrics to evaluate the prediction quality for different methods. Since there are spatial and color mis-alignments between the input bursts and HR ground truth in the BurstSR dataset employed for Track 2, we additionally conducted a human study to evaluate the top ranking methods.

Track 1: Due to the use of synthetically generated dataset for evaluation, an accurately aligned ground truth is avail-

able in Track 1. This enables the use of pixel-wise image quality metrics for evaluating the performance of different methods. We use the fidelity-based Peak Signal-to-Noise Ratio (PSNR) score to rank the methods. Additionally, we also report the Structural Similarity Index (SSIM) [54] as well as the learned perceptual score LPIPS [62] for all the methods. The emphasis of the challenge is on learning to recover the HR signal, rather than learning any postprocessing steps. Thus, all metrics are computed in the linear sensor space, before applying white-balancing, gamma correction, contrast enhancement *etc*.

Track 2: We follow the same evaluation procedure employed in [4] in order to handle the spatial and color misalignments between the input bursts and HR ground truth. The network prediction is first spatially aligned to the ground truth, using pixel-wise optical flow estimated using PWC-Net [48]. A linear color mapping between the input burst and the ground truth, modeled as a 3x3 color correction matrix, is then estimated and used to transform the spatially aligned network prediction to the same color space as the ground truth. The spatially aligned and color corrected prediction is then compared with the ground truth to compute standard image quality metrics. The evaluation script was included in the public toolkit released to the participants. We refer to [4] for more details about the evaluation procedure.

We also conducted a human study on Amazon Mechanical Turk (AMT) to evaluate the top performing methods on Track 2. We manually selected 100 bursts from the test set with diverse texture. Next, we extract 3 random 80×80 crops from the HR predictions of the methods for each of these test images. The crops are then resized to 320×320 using nearest neighbor interpolation. The participants in the user study were shown the full ground truth image, as well as the network prediction crops, and then asked to rank the predictions based on visual quality. We obtained 5 independent rankings for each crop. The mean ranking (MOR) over the 300 crops, as well as the percentage of times a method was ranked first (%Top) are used as evaluation scores to rank the methods.

3.2. Baselines

We compare the participating methods with two additional baselines.

SingleImage: We evaluate a baseline single frame SR method. Our single image baseline passes the first image in the burst through a series of residual blocks [15] without batch normalization [20]. The extracted feature map is upsampled using the sub-pixel convolution layer [46], and passed through additional residual blocks to obtain the HR RGB image.

DBSR [4]: We also evaluate the DBSR burst super-resolution model introduced in [4]. DBSR employs the

	PSNR↑	SSIM↑	LPIPS↓
Noah_TerminalVision_SR	**46.85**	**0.983**	**0.018**
MegSR	46.72	**0.983**	0.020
Inria	44.76	0.969	0.034
TTI	44.40	0.973	0.038
BREIL	39.22	0.918	0.104
MLP_BSR	37.62	0.895	0.166
SingleImage	39.28	0.921	0.103
DBSR [4]	42.58	0.960	0.055

Table 1. Challenge results on the synthetic test set from Track 1, in terms of PSNR, SSIM, and LPIPS. The top section contains results for the participating methods, while baseline approaches are included in the bottom section.

PWCNet [48] optical flow network align the input images. The aligned images are merged using an attention-based fusion approach.

3.3. Track 1: Synthetic

In this section, we present results of the submitted methods on Track 1. The mean PSNR, SSIM, and LPIPS scores over the 500 bursts from the synthetic test set are provided in Table 1. The team Noah_TerminalVision_SR obtains the best results in terms of all 3 metrics with a PSNR score of 46.85. Noah_TerminalVision_SR employs the PCD module [51] to align the input images, which are then merged using the attention-based fusion approach proposed in [4]. MegSR achieves the second best performance with a PSNR of 46.72. MegSR uses a modified version of PCD, denoted FEPCD module for alignment. Fusion is performed using a cross non-local fusion module based on Non-Local network [52]. Team Inria, employing an optimization based approach, obtains a PSNR score of 44.76. The team utilizes the forward image formation model and jointly optimizes the HR estimate as well as the motion vectors. Note that four of the participating teams, namely Noah_TerminalVision_SR, MegSR, Inria, and TTI outperformed the DBSR method [4], thus setting a new state-of-the-art on the burst SR problem.

A qualitative comparison between the participating methods is provided in Figure 1. The two top performing methods Noah_TerminalVision_SR and MegSR obtain impressive results which are very close to the ground truth. The multi-frame approaches Noah_TerminalVision_SR, MegSR, Inria, TTI, and DBSR better recover the high-frequency details compared to the single image baseline, thanks to the use of additional information from multiple frames.

3.4. Track 2: Real-World

Here, we present the results on the real-world BurstSR test set from Track 2. The mean PSNR, SSIM, and LPIPS
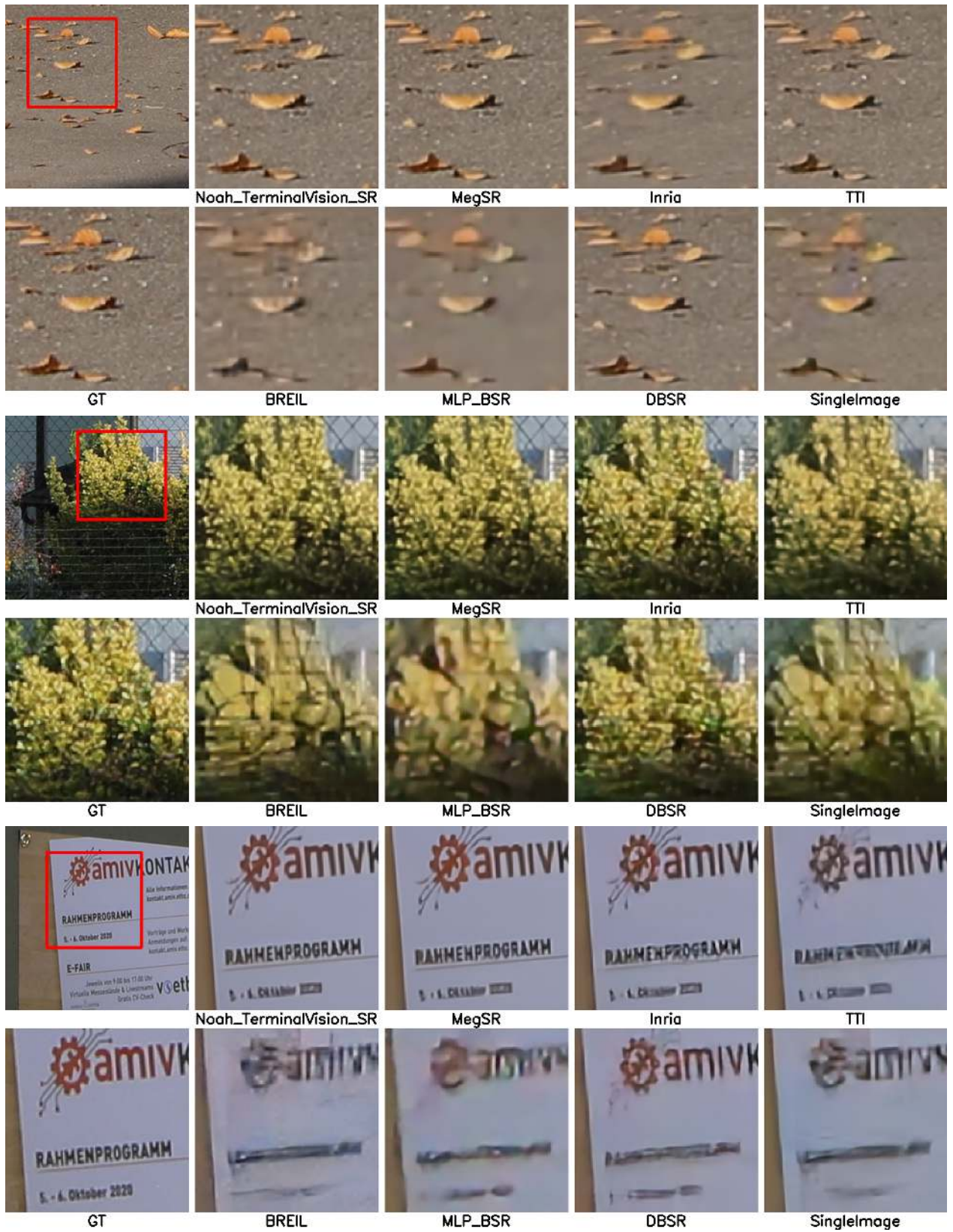
| Noah_TerminalVision_SR | MegSR | Inria | TTI |

| GT | BREIL | MLP_BSR | DBSR | SingleImage |

| Noah_TerminalVision_SR | MegSR | Inria | TTI |

| GT | BREIL | MLP_BSR | DBSR | SingleImage |

| Noah_TerminalVision_SR | MegSR | Inria | TTI |

| GT | BREIL | MLP_BSR | DBSR | SingleImage |

Figure 1. Qualitative comparison on Track 1 test set (4x super-resolution).

	PSNR↑	SSIM↑	LPIPS↓	MOR↓	%Top↑	Avg. Rank↓
MegSR	**45.45**	**0.979**	0.032	3.09	20.9	**1.5**
Noah_TerminalVision_SR_B	45.26	0.978	**0.026**	**2.96**	**29.7**	2.0
Noah_TerminalVision_SR_A	45.36	**0.979**	0.035	3.41	13.1	2.5
TTI	44.16	0.974	0.040	3.87	14.0	5.0
MLP_BSR	41.40	0.952	0.101	-	-	-
BREIL	29.93	0.797	0.141	-	-	-
DBSR [4]	45.17	0.978	0.037	3.57	11.4	4.0
SingleImage	44.02	0.972	0.051	4.10	10.9	6.0

Table 2. Challenge results on BurstSR test set from Track 2. PSNR, SSIM, and LPIPS scores are computed after spatial and color alignment of the network prediction to the ground truth. MOR denotes the mean ranking of the method from a human study, while %Top corresponds to the percentage of time a method was ranked as the best in the human study. The last column reports the average of a methods ranking in terms of PSNR and MOR.

scores over the test set are provided in Table 2. Note that all the metrics are computed after spatial and color alignment of the network prediction to the ground truth, as described in Section 3.1. MegSR achieves the best results in terms of PSNR and SSIM, with scores of 45.45dB and 0.979, respectively. Noah_TerminalVision_SR_A, which is trained using L_1 loss with alignment similar to MegSR, obtains the second best PSNR score of 45.36dB. In contrast, Noah_TerminalVision_SR_B which is trained using perceptual loss achieves the best LPIPS score of 0.026.

We also report results of the human study in Table 2, in terms of Mean Opinion Ranking (MOR) and %Top metrics. Noah_TerminalVision_SR_B obtains the best MOR score of 2.96, while the second best results are obtained by MegSR. A qualitative comparison between the methods is provided in Figure 2. As in Track 1, MegSR, Noah_TerminalVision_SR, TTI, and DBSR generate more detailed images compared to the single image baseline. The results of Noah_TerminalVision_SR_B are in general more sharper compared to that of MegSR. This can be attributed to the use of perceptual loss during training. However the predictions of Noah_TerminalVision_SR_B contain slight high-frequency artefacts, as seen in third example in Figure 2.

4. Challenge Methods and Teams

In this section, we provide a brief description of the participating methods. Training and inference times for different methods are summarized in Table 3.

4.1. Noah_TerminalVision_SR

The team proposes NoahBurstSRNet for the task of Joint Demosaicking, Denoising and Super Resolution (JDDSR) of smart-phone burst Raw images. NoahBurstSRNet is inspired by two recent works: EDVR [51] and DBSR [4]. The network architecture is shown in Figure 3.

Burst Raw images are passed to an encoder to extract features. Those features are then aligned to the base feature (the feature of the first Raw image) by a Pyramid, Cascading and Deformable (PCD) [51] module. With the aligned features, a Weight Predictor (WP) is used to predict the fusion weights between the base frame features and the other frame features. Based on the predicted fusion weights, an Attention-based Fusion (ABF) module is used to fuse the aligned futures. The WP and ABF modules are very similar to those proposed in [4], except that the optical flow is not used as an input of WP. The spatial resolution of the fused features is first increased by a factor of two using a Pixelshuffle layer [46]. The upsampled feature map is then processed by a sequence of Residual Feature Distillation Blocks (RFDBs) [31] in order to reconstruct the high-frequency details. Finally, the spatial resolution of the feature map is further increased by a factor of 4 and the final RGB image is predicted using pixelshuffle and convolution layers.

For track 1, the NoahBurstSRNet was trained in a fully-supervised manner on synthetic bursts generated using sRGB images from the Zurich RAW to RGB training set. The network was trained using the L1 loss. For the final submission, the team used a self-ensemble technique proposed in [32] which can augment the input data while preserving the bayer pattern.

For track2, the team employed a two-stage training strategy. For the first stage, since the training data is weakly-paired (*i.e.* they are not pixel-wise aligned), the model was trained using the spatial and color alignment strategy proposed in [4]. This model is termed as Noah_TerminalVision_SR_A. Next, the ground truth linear RGB images were spatially aligned and color matched to the SR predictions of the model trained in stage 1 using an in-house tool. The resulting aligned ground truth images were then used to fine-tune the model in stage 2 to obtain Noah_TerminalVision_SR_B. The fine-tuning was performed using the SSIM and LPIPS [62] loss as these losses can tolerate small misalignments in the training data.

4.2. MegSR

The team MegSR propose Feature Enhanced Burst Super-Resolution with Deformable Alignment (EBSR) framework, as shown in Fig. 4. EBSR solves the burst SR problem in three steps: align, fusion and reconstruction. First, it extracts high-level features of the LR burst images. The features are then aligned by a Feature Enhanced Pyramid, Cascading and Deformable convolution (FEPCD) module. Next, 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, EBSR builds a progressively scaled residual pathway structure to further improve the performance. Please refer to [35] for a detailed description.

Feature Alignment: EBSR extracts high level features using Wide activation Residual Block (WARB) introduced in [57]. The features are aligned by a Feature Enhance PCD
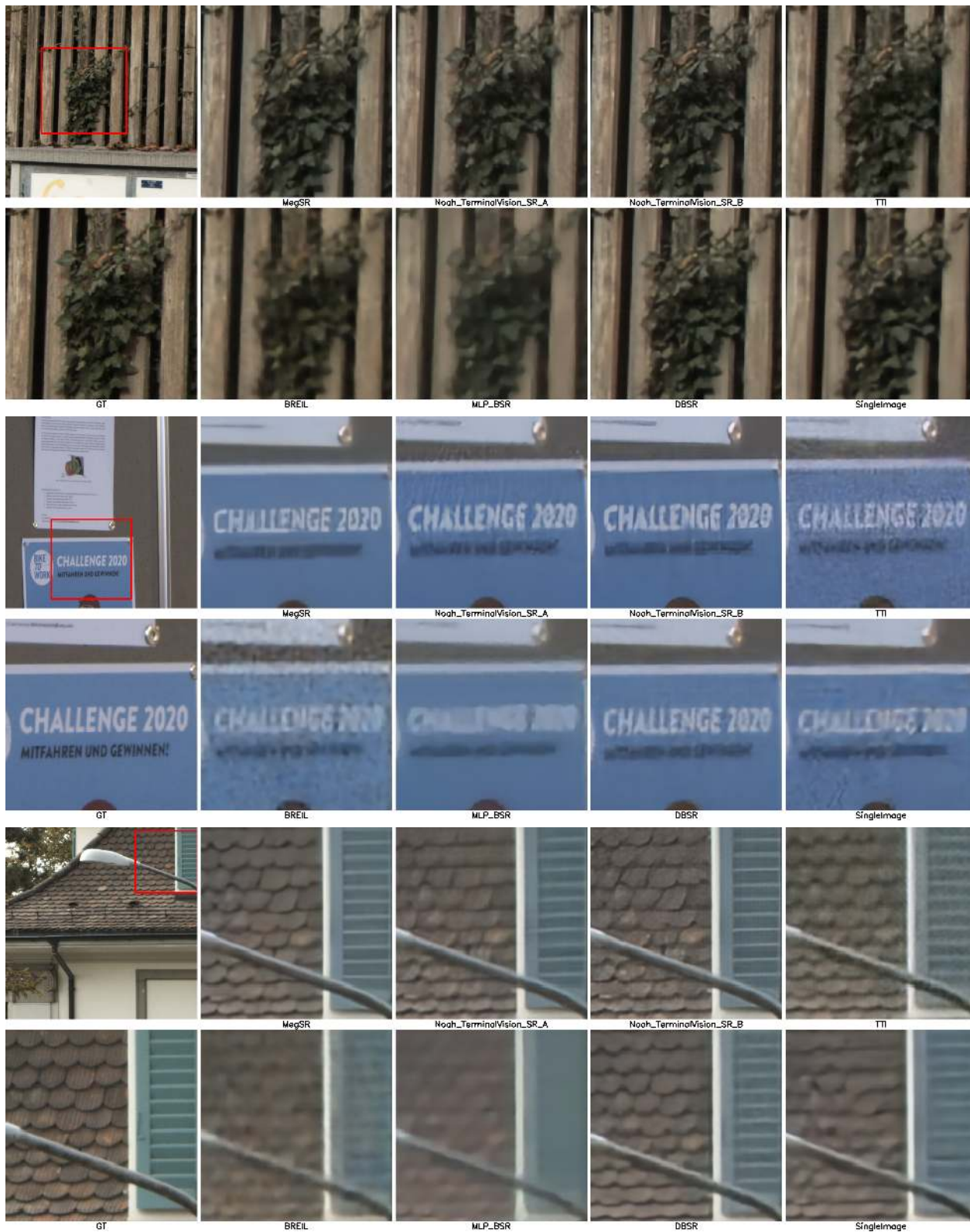
Figure 2. Qualitative comparison on Track 2 test set (4x super-resolution).

Team Name	Codalab Username	Track 1		Track 2	
		Train Time (days)	Runtime (sec)	Train Time (days)	Runtime (sec)
Noah_TerminalVision_SR	Noah_TerminalVision	5	1.60	2	0.30
MegSR	megviiLuo	7	0.15	3	0.18
Inria	JohnDoe4598	2.5	0.65	-	-
TTI	TakahiroMaeda	18	2.14	16	5.50
BREIL	chowy333	2.5	0.01	0.1	0.01
MLP_BSR	raoumer	10	0.33	-	0.88

Table 3. Information about the participating teams. We report the training time and inference time per burst provided by the teams.
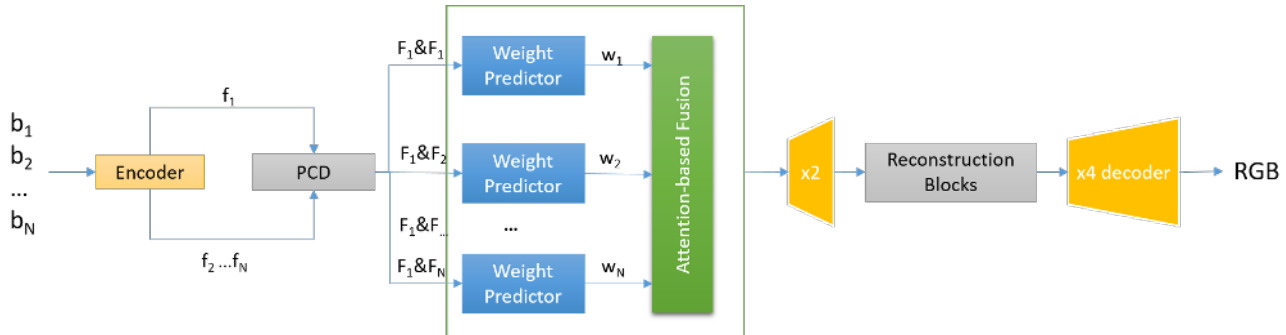

Figure 3. Overview of the NoahBurstSRNet network architecture employed by team Noah_TerminalVision_SR

(FEPCD) module with multi-scale features. FEPCD is an extension of the PCD module [51] where an initial feature pyramid is used to first denoise and enhance the image features The feature of the first frame from the LR image is chosen as the reference, and the features from other LR images are aligned to this reference.

Fusion: In order to properly fuse the features from different frames, EBSR introduces Cross Non-Local (CNL) module which is based on Non-Local network [52]. The CNL fusion module measures the similarity between every two pixels from reference frame and other frame feature maps. The more similar the feature representations between two locations, higher the correlation between them. According to this property, the valid regions of other frames are fused into the reference frame.

Reconstruction: The final SR output is reconstructed by the Long Range Concatenation Network (LRCN). It consists of two parts: backbone module and upsample module. The backbone module is composed of G Long-Range Concatenation Groups (LRCG), and each LRCG contains N Long-Range Residual Blocks with wide activation which is inspired by WDSR [57]. Moreover, EBSR introduces a progressive upsample module with pixelShuffle [17] and a residual pathway structure to reconstruct the final SR image. The reference frame goes through a pixelshuffle layer, and is added to the outputs of pixelshuffle layers of reconstruction module.

Training: For Track 1, the network is trained using the charbonnier loss proposed by LapSRN [25]. The trained model is then fine-tuned on the BurstSR train set to obtain the model for Track 2. When training on BurstSR dataset, the network predictions are first aligned to the ground truth as described in [4]. In order to further improve the performance, the team employs a multi-model ensemble training strategy. Three trained EBSR models are loaded with frozen weights, and a few additional convolution layers are trained to fuse their outputs.

Testing: Team MegSR uses the Test Time Augmentation (TTA) strategy which can be seen as a self-ensemble approach. Specifically, given the original input burst, each image in the burst is transposed and the images in the burst are shuffled. The two augmented bursts, along with the original burst, are then passed through the network. The resulting outputs are averaged after reversing the augmentation effect to obtain the final prediction.

4.3. Inria

The method presented in this section corresponds to the paper [26], where all details can be found. The brief description next, including figures, is borrowed from this paper. Starting from a low resolution burst of raw images, team Inria estimates a coarse block-parametric displacement field using a robust multiscale Lucas-Kanade algorithm [3]. Next, an estimate of the high resolution image is obtained and the motion parameters are subsequently refined alternatively. This iterative process involves a data-driven prior—here, a convolutional neural network—which helps removing artefacts. Model parameters are learned end-to-end by backpropagating on real HR/synthetic LR ex-
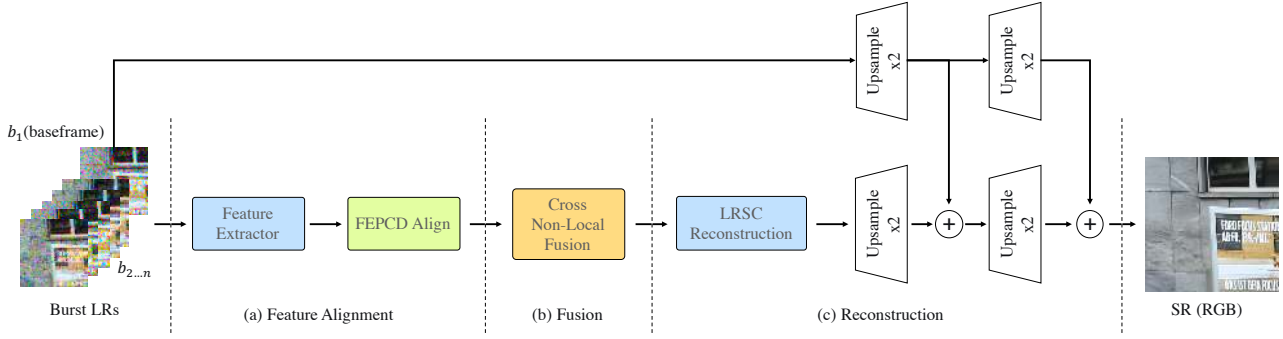
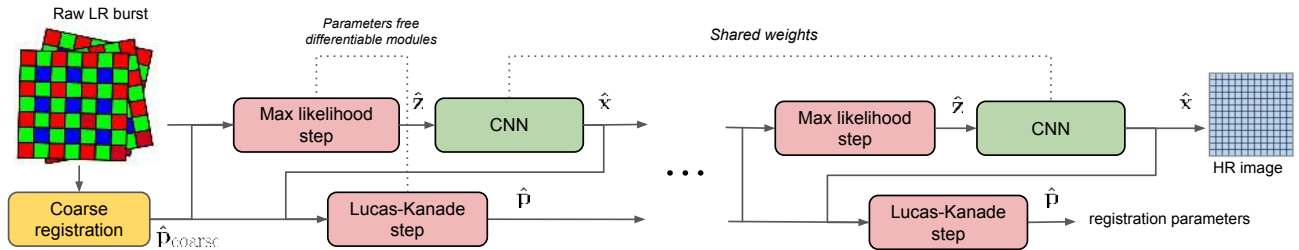Figure 4. Overview of the network architecture employed by team MegSR.

Figure 5. Overview of the network architecture employed by team Inria

amples. An overview of the network architecture is provided in Figure 5. The HR estimate as well as the refined motion parameters are obtained using the image formation model described next.

Inverse problem and optimization: The burst images are obtained through the following forward model:

$$\mathbf{y}_k = DBW_{\mathbf{p}_k} \mathbf{x} + \varepsilon_k \text{ for } k = 1, \ldots, K, \quad (1)$$

where ε_k is additive noise. Here, both the HR image $\mathbf{x}$ and the frames $\mathbf{y}_k$ of the burst are flattened into vector form. The operator $W_{\mathbf{p}_k}$, parameterized by $\mathbf{p}_k$, warps $\mathbf{x}$ to compensate for misalignments between $\mathbf{x}$ and $\mathbf{y}_k$ caused by camera or scene motion between frames and resamples the warped image to align its pixel grid with that of $\mathbf{y}_k$. Finally, the corresponding HR image is blurred to account for integration over space, and it is finally downsampled in both the spatial and spectral domains by the operator D. The spectral part corresponds to mosaicking operation of selecting one of the three RGB values to assemble the raw image. The model (1) can be rewritten as $\mathbf{y} = U_{\mathbf{p}}\mathbf{x} + \varepsilon$, where

$$U_{\mathbf{p}} = \begin{bmatrix} DBW_{\mathbf{p}_1} \\ \vdots \\ DBW_{\mathbf{p}_K} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_K \end{bmatrix}, \mathbf{p} = \begin{bmatrix} p_1 \\ \vdots \\ p_K \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_K \end{bmatrix}. \quad (2)$$

Given the image formation model of Eq. (1), recovering the HR image $\mathbf{x}$ from the K LR frames $\mathbf{y}_k$ in the burst can

be formulated as finding the values of $\mathbf{x}$ and $\mathbf{p}$ that minimize

$$\frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x}), \quad (3)$$

where ϕ_θ is a parameterized regularizer, which will be chosen as a convolutional neural network, and λ is a parameter balancing the data-fidelity and regularization terms. The objective (3) is minimized using the quadratic penalty method [40, Sec. 17.1] often called half-quadratic splitting (or HQS) [11]. Here, the original objective is replaced by

$$E_\mu(\mathbf{x}, \mathbf{z}, \mathbf{p}) = \frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\mathbf{z}\|^2 + \frac{\mu}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x}), \quad (4)$$

where $\mathbf{z}$ is an auxiliary variable, and μ is a parameter increasing at each iteration, such that, as $\mu \to +\infty$, the minimization of (4) with respect to $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{p}$ becomes equivalent to that of (3) with respect to $\mathbf{x}$ and $\mathbf{p}$ alone. The sequence of weights $(\mu^t)_{t\geq 0}$ are learned end-to-end.

Estimating the HR image x: The estimate $\mathbf{x}$ is updated as

$$\mathbf{x}^t \leftarrow \text{argmin}_{\mathbf{x}} \frac{\mu_{t-1}}{2}\|\mathbf{z}^t - \mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x}),$$

which amounts to computing the proximal operator of the prior ϕ_θ. Following the "plug-and-play" approach [6, 44, 50], the proximal operator is replaced by a parametric function $f_\theta(\mathbf{z}_t)$ (here, a CNN).

Initialization by coarse alignment: Each LR frame is aligned to an arbitrary one from the burst (*e.g.*, the first one)

8

by using the robust Lucas-Kanade forward additive algorithm [3, 45] which is known to be robust to noise. The raw images are first converted to grayscale format by using bilinear interpolation. Although sub-optimal, such a procedure is sufficient for obtaining coarse motion parameters.

Learned data prior: Good image priors are essential for solving ill-posed inverse problems. Instead of using a classical one, such as total variation (TV), team Inria learns an implicit prior parameterized by a convolutional neural network f_θ in a data-driven manner. The ResUNet architecture proposed in [61] is employed for this purpose.

Training: The model is trained using synthetic bursts generating using sRGB images from the training split of the Zurich raw to RGB dataset [19], by minimizing the ℓ_1 loss.

4.4. TTI

The team TTI adopt the Recurrent Back Projection Network (RBPN) [14] for super-resolving burst frames. RBPN constructs the projection module inspired by DBPN [13, 14] to iteratively refine an input frame using temporally-nearby frames, along with corresponding flow vectors. An overview of the approach is provided in Figure 6. In addition to the basic functions in RBPN, TTI applies deformable convolutions to align features to the input frame. A flow refinement module is also proposed to minimize the flow estimation error caused by noisy frames. This module is trained explicitly using the ground-truth flow in track 1 but trained implicitly solely on SR reconstruction loss in track 2. The residual blocks in the projection module are initialized using Fixup Initialization [60] to stabilize the training process of the normalization-free network.

Flow Estimation and Refinement: For initial flow estimation between a reference frame I_0 and frame I_t, TTI utilizes PWCNet [48] with fixed pretrained weights. In order to further refine the predicted flow, a UNet-like flow refinement module is employed. The estimated flow is concatenated with the reference frame I_0 and frame I_t, and fed into the refinement module to get the residual flow, which is then added to the initial flow estimate.

Alignment Extractor: In the original RBPN, a single convolution layer is used for feature extraction from each input frame, neighboring frame and the corresponding optical flow triplet. In contrast, team TTI first concatenates features from reference frame I_0 and frame I_t, along with the corresponding optical flow $F_{0,t}$ and passes this through a convolution layer to obtain offsets. These offsets are then used to align the frame I_t features to the base frame using a deformable convolution layer.

Training: For Track 1, the model is trained on synthetic burst data generated using sRGB image from [19], using the L_1 loss. Additionally, L_1 loss between the ground-truth flow $\hat{F}$ and refined flow F is also used to train the flow refinement module. In Track 2, the model was trained on

BurstSR training set, using the spatial and color alignment strategy employed in [4], by minimizing the L_2 loss. Since ground-truth flow $\hat{F}$ is not available in this case, the flow refinement module was trained using only the final SR reconstruction loss, without any direct supervision.

Inference: To achieve a fast training process and memory efficiency, the model is trained using bursts containing 8 images. During inference, original burst containing 14 frames is divided into 7 subsets, each including the reference (first) frame. The 7 SR predictions are averaged to output the final SR image. No model-ensemble and self-ensemble are used.

4.5. MLP_BSR

The team MLP_BSR propose a deep iterative Burst SR learning method (BSRICNN) that solves the Burst SR task in an iterative manner. To solve the Raw Burst Super-Resolution task, they rely on the forward observation model,

$$\mathbf{y}_i = \mathbf{MHS}_i(\tilde{\mathbf{x}}) + \eta_i, \quad i = 1, \dots, B \qquad (5)$$

where, $\mathbf{y}_i$ is an observed LR burst containing B images, $\mathbf{M}$ is a mosaicking operator that corresponds to the CFA (Color Filter Array) of a camera (usually Bayer), $\mathbf{H}$ is a down-sampling operator (*i.e.* bilinear, bicubic, etc.) that resizes an HR image $\tilde{\mathbf{x}}$ by a scaling factor r, $\mathbf{S}_i$ is an affine transformation of the coordinate system of the image $\tilde{\mathbf{x}}$ (*i.e.* translation and rotation), and η_i is an additive heteroscedastic Gaussian noise related to shot and read noise. Due to the ill-posed nature of inverse problem, the recovery of $\mathbf{x}$ from $\mathbf{y}_i$ mostly relies on variational approaches for combining the observation and prior knowledge, and the solution is obtained by minimizing the following objective function as,

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2\sigma^2 B} \sum_{i=1}^{B} \|\mathbf{y}_i - \mathbf{MHS}_i(\mathbf{x})\|_2^2 + \lambda \mathcal{R}(\mathbf{x}), \quad (6)$$

where, the first term is a data fidelity term that measures the proximity of the solution to the observations, the second term (*i.e.* $\mathcal{R}(\mathbf{x})$) is the regularization term that is associated with image priors, and λ is the trade-off parameter that governs the compromise between the data fidelity and the regularizer term. The objective (6) is minimized using Majorization-Minimization (MM) framework [18] which has been previously employed in image restoration tasks [38, 23].

The proposed Burst SR scheme is shown in Figure 7. BSRICNN is unrolled into K stages, where each stage computes the refined estimate of the solution. $\mathbf{y}_i$ is an input Raw LR burst, $\mathbf{x}^0$ is an initial estimate, and $\mathbf{x}^K$ is a final estimated SR image. In Figure 7, the Encoder-Resnet-Decoder (ERD) architecture employed is similar to the one used in [38]. Each network stage performs the efficient
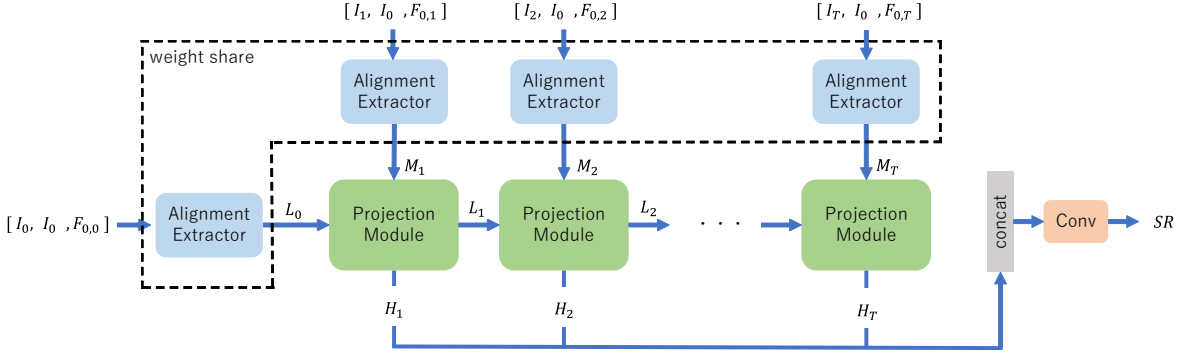
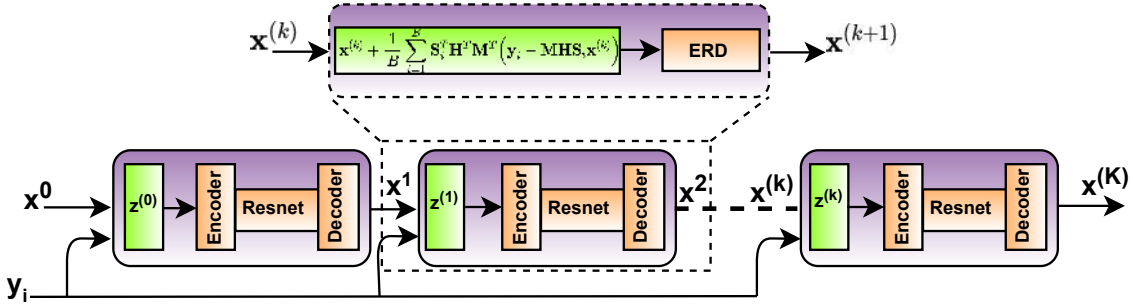Figure 6. Overview of the network architecture employed by team TTI.

Figure 7. The overview the architecture employed by team MLP_BSR.

Proximal Gradient Descent [41] updates to solve the optimization problem. The shared parameters across stages are learned jointly by minimizing the L_1 loss w.r.t. to all network parameters. For Track 1, the network is trained using synthetic bursts generated using 46,839 sRGB images from the Zurich RAW to RGB dataset [19]. The same model was then also employed for Track 2, without any finetuning on the BurstSR dataset.

4.6. BREIL

The proposed model takes multiple LR RAW burst images with noise $\{b_i\}_{i=1}^T$ and predicts denoised single HR RGB image I_{HR}^i. The kernel prediction network (KPN) [37, 36] which has been recently employed for burst image processing is utilized for aligning the images from the burst. The overall model shown in Figure 8 consists of two parts, i) a KPN-based module that aligns the burst images and ii) a reconstruction network that increases the resolution of images while fusing the output of the alignment module. Each of these modules are briefly described next. Please refer to [7] for more details.

Weighted Multi-Kernel Prediction: The team BREIL propose a weighted multi-kernel prediction network (WMKPN) for aligning the images from the input burst. WMKPN is an extension of MPKN approach introduced in [36]. MKPN predicts multiple kernels of different sizes, which are then used to align and fuse the burst images.

All the kernels are weighted equally in the MKPN framework. In contrast, WMKPN predicts additional weights for each kernel to adaptively weight the impact different kernels. Furthermore, the predicted kernels are only used to align the input images, while the fusion is performed by a seperate network. WMKPN employs a modified U-net architecture, resembling [37, 36, 59]. In the encoder part, the spatial sizes of the feature maps are reduced by the average pooling layer while using the convolution layer and ReLU activation function. On the other side, the decoder increases the spatial sizes of the feature maps by a bilinear upsampling layer. WMKPN also exploits the convolution layer and attention module, proposed in [59], which is composed of a series of the channel attention (CA) [63] and the spatial attention (SA) [16]. The encoded feature maps are concatenated to the decoder side that have the same spatial sizes like U-net architecture [43]. This modified U-net consists of two branches, the kernel prediction branch and the kernel weight branch which predict kernels and weights, respectively. Separable kernels are used for memory efficiency. To provide discriminative mechanism, the kernel weights are normalized using softmax operation applied across the different kernels. The weighted kernels are then convolved with the input burst images in order to align them.

SR Reconstruction Network: The aligned input images are concatenated and passed through a reconstruction network to obtain the SR image. The reconstruction network
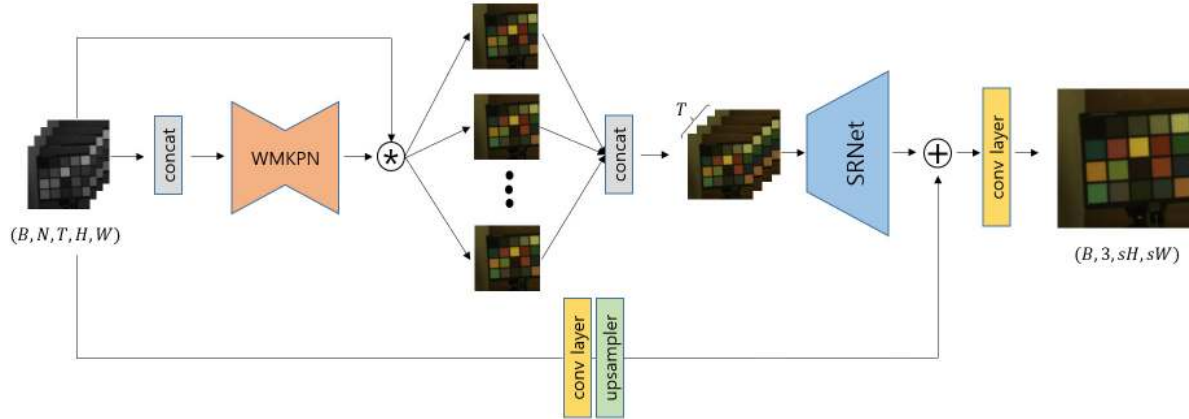
10

Figure 8. Overview of the network architecture employed by team BREIL.

leverages the residual blocks with short connections which are exploited in EDSR approach [29]. The SR network extracts deep features by passing the input through ten residual blocks. The output is then upsampled using three sub-pixel convolution layers [46] to obtain the final RGB image.

Training: The network is trained using a combination of L_1 and SSIM losses.

5. Conclusion

This paper describes the NTIRE2021 challenge on burst super-resolution. Given multiple images of a scene captured in quick succession, the burst super-resolution task aims to generate a super-resolved output by merging information from the input frames. The challenge tackled the problem of RAW burst super-resolution, where the goal is to predict a 4x super-resolved RGB image, given a RAW noisy burst as input. The challenge contained two tracks, namely Track 1 and Track 2. In Track 1, a synthetically generated burst dataset was used for evaluation, while Track 2 focused on real-world SR using bursts captured from a hand held camera. 6 teams submitted results in the final testing phase. The participating methods, described in this report, employed a diverse set of approach for the burst SR problem obtaining promising results.

Appendix

A. Teams and Affiliations

NTIRE2021 Organizers

Members:
Goutam Bhat (goutam.bhat@vision.ee.ethz.ch)
Martin Danelljan (martin.danelljan@vision.ee.ethz.ch)
Radu Timofte (radu.timofte@vision.ee.ethz.ch)
Affiliation: Computer Vision Lab, ETH Zurich

Noah_TerminalVision_SR

Title: Real-World Joint Demosaicking, Denoising and Super Resolution of Smart-phone Burst Raw Images
Team Leader:
Xueyi Zou (zouxueyi@huawei.com)
Members:
Xueyi, Zou, Noah's Ark Lab, Huawei
Magauiya, Zhussip, Noah's Ark Lab, Huawei
Pavel, Ostyakov, Noah's Ark Lab, Huawei
Ziluan, Liu, CBG AITA, Huawei
Youliang, Yan, Noah's Ark Lab, Huawei

MegSR

Title: EBSR: Feature Enhanced Burst Super-Resolution with Deformable Alignment
Team Leader:
Lanpeng Jia (jialanpeng@megvii.com)
Members:
Ziwei Luo, Megvii Technology
Lei Yu, Megvii Technology
Xuan Mo, Megvii Technology
Youwei Li, Megvii Technology
Lanpeng Jia, Megvii Technology
Haoqiang Fan, Megvii Technology
Jian Sun, Megvii Technology
Shuaicheng Liu, Megvii Technology

Inria

Title: End-to-End Super-Resolution from Raw Image Bursts

Team Leader:

Bruno Lecouat (bruno.lecouat@inria.fr)

Members:

Bruno Lecouat, Inria

Jean Ponce, Inria

Julien Mairal, Inria

TTI

Title: Recurrent Back Projection Network for Burst Super-Resolution

Team Leader:

Takahiro Maeda (sd19445@toyota-ti.ac.jp)

Members:

Takahiro Maeda, Toyota Technological Institute

Kazutoshi Akita, Toyota Technological Institute

Takeru Oba, Toyota Technological Institute

Norimichi Ukita, Toyota Technological Institute

MLP_BSR

Title: Deep Iterative Convolutional Neural Network for Raw Burst Super-Resolution

Team Leader:

Rao Muhammad Umer (engr.raoumer943@gmail.com)

Members:

Rao Muhammad Umer, University of Udine, Italy

Christian Micheloni, University of Udine, Italy

BREIL

Title: Weighted Multi-Kernel Prediction Network for Burst Image Super-resolution

Team Leader:

Wooyeong Cho (chowy333@kaist.ac.kr)

Members:

Wooyeong, Cho, KAIST (Korea Advanced Institute of Science and Technology)

Sanghyeok, Son, KAIST (Korea Advanced Institute of Science and Technology)

Daeshik, Kim, KAIST (Korea Advanced Institute of Science and Technology)

References

[1] Abdullah Abuolaim, Radu Timofte, Michael S Brown, et al. NTIRE 2021 challenge for defocus deblurring using dual-pixel images: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[2] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2021 nonhomogeneous dehazing challenge report. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[3] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004. 7, 9

[4] Goutam Bhat, Martin Danelljan, L. Gool, and R. Timofte. Deep burst super-resolution. In *CVPR*, 2021. 1, 2, 3, 5, 7, 9

[5] 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

[6] Stanley H Chan, Xiran Wang, and Omar A Elgendy. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016. 8

[7] Wooyeong Cho, Sanghyeok Son, and Daeshik Kim. Weighted multi-kernel prediction network for burst image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 10

[8] C. Dong, Chen Change Loy, Kaiming He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 1

[9] C. Dong, Chen Change Loy, Kaiming He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:295–307, 2016. 1

[10] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. NTIRE 2021 depth guided image relighting challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[11] Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE transactions on Image Processing*, 4(7):932–946, 1995. 8

[12] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Yu Qiao, Shuhang Gu, Radu Timofte, et al. NTIRE 2021 challenge on perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[13] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1664–1673. IEEE Computer Society, 2018. 9

[14] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3897–3906. Computer Vision Foundation / IEEE, 2019. 9

[15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[16] Yanting Hu, Jie Li, Yuanfei Huang, and Xinbo Gao. Channel-wise and spatial feature modulation network for

single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3911–3927, 2019. 10

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

[18] David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, pages 30–37, 2004. 9

[19] 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. 2, 9, 10

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *ICML*, 2015. 3

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 1

[22] Jiwon Kim, J. Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016. 1

[23] Filippos Kokkinos and Stamatios Lefkimmiatis. Iterative joint image demosaicking and denoising using a residual denoising network. *IEEE Transactions on Image Processing*, 28:4177–4188, 2019. 9

[24] Wei-Sheng Lai, Jia-Bin Huang, N. Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5835–5843, 2017. 1

[25] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 7

[26] Bruno Lecouat, Jean Ponce, and Julien Mairal. Aliasing is your ally: End-to-end super-resolution from raw image bursts. *ArXiv*, abs/2104.06191, 2021. 7

[27] C. Ledig, L. Theis, Ferenc Huszár, J. Caballero, Andrew Aitken, Alykhan Tejani, J. Totz, Zehan Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 1

[28] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 1

[29] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 11

[30] Jerrick Liu, Oliver Nina, Radu Timofte, et al. NTIRE 2021 multi-modal aerial view object classification challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[31] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. *arXiv preprint arXiv:2009.11551*, 2020. 5

[32] 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. 5

[33] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 learning the super-resolution space challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[34] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. SRFlow: Learning the super-resolution space with normalizing flow. In *ECCV*, 2020. 1

[35] 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. 5

[36] Talmaj Marinč, Vignesh Srinivasan, Serhan Gül, Cornelius Hellge, and Wojciech Samek. Multi-kernel prediction networks for denoising of burst images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2404–2408. IEEE, 2019. 10

[37] 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. 10

[38] Rao Muhammad Umer, Gian Luca Foresti, and Christian Micheloni. Deep iterative residual convolutional network for single image super-resolution. In *Proceedings of the International Conference of Pattern Recognition (ICPR)*, January 2021. 9

[39] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on image deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[40] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering, 2006. Second edition. 8

[41] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends in Optimization*, pages 127–239, 2014. 10

[42] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Aleš Leonardis, Radu Timofte, et al. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image com-

puting and computer-assisted intervention, pages 234–241. Springer, 2015. 10

[44] Ernest K Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *ICML*, 2019. 8

[45] Javier Sanchez. The inverse compositional algorithm for parametric registration. *Image Processing On Line*, 6:212–232, 2016. 9

[46] W. Shi, J. Caballero, Ferenc Huszár, J. Totz, A. Aitken, R. Bishop, D. Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 3, 5, 11

[47] Sanghyun Son, Suyoung Lee, Seungjun Nah, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on video super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[48] 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. 3, 9

[49] Ying Tai, Jian Yang, and X. Liu. Image super-resolution via deep recursive residual network. In *CVPR*, 2017. 1

[50] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013. 8

[51] 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, 5, 7

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 7

[53] Xintao Wang, K. Yu, Shixiang Wu, Jinjin Gu, Yi-Hao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and X. Tang. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018. 1

[54] 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

[55] 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

[56] Ren Yang, Radu Timofte, et al. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[57] 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. 5, 7

[58] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, 2016. 1

[59] Bin Zhang, Shenyao Jin, Yili Xia, Yongming Huang, and Zixiang Xiong. Attention mechanism enhanced kernel prediction networks for denoising of burst images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2083–2087. IEEE, 2020. 10

[60] Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 9

[61] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *CVPR*, 2020. 9

[62] 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, 5

[63] 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. 10

[64] Yulun Zhang, Yapeng Tian, Y. Kong, B. Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 1