

Burst Image Super-Resolution with Base Frame Selection

Sanghyun Kim* Min Jung Lee* Woohyeok Kim Deunsol Jung Jaesung Rim
Sunghyun Cho Minsu Cho

Pohang University of Science and Technology (POSTECH), South Korea

{sanghyun.kim, minjlee, woohyeok, deunsol.jung, jsrim123, s.cho, mscho}@postech.ac.kr

Abstract

Burst image super-resolution has been a topic of active research in recent years due to its ability to obtain a high resolution image using complementary information between multiple frames in the burst. In this work, we explore using burst shots with non-uniform exposures to confront real-world practical scenarios by introducing a new benchmark dataset, dubbed Non-uniformly Exposed Burst Image (NEBI), that includes the burst frames at varying exposure times to obtain a broader range of irradiance and motion characteristics within a scene. As burst shots with non-uniform exposures exhibit varying levels of degradation, fusing information of the burst shots into the first frame as a base frame may not result in optimal image quality. To address this limitation, we propose a Frame Selection Network (FSN) for non-uniform scenarios. This network seamlessly integrates into existing super-resolution methods in a plug-and-play manner with low computational cost. The comparative analysis reveals the effectiveness of the non-uniform setting for the practical scenario and our FSN on synthetic-/real- NEBI datasets.

1. Introduction

Burst image super-resolution (BISR) is a task to reconstruct a high-resolution (HR) image with vivid details, utilizing a rapid succession of low-resolution (LR) frames captured by handheld devices [2, 19, 27, 43]. It spans a wide range of computer vision problems such as medical imaging [15, 21, 29], satellite imaging [7, 9, 38], object detection [18, 30, 33] and image generation [32, 34, 37].

In contrast to conventional single-image super-resolution (SISR) techniques [10, 44, 45, 47] where the inherent challenges of ill-posed problems constrain their performance, BISR demonstrates the ability to achieve high-quality images in terms of signal-to-noise ratio (SNR) through the utilization of multiple images for restoration



Figure 1. Output quality decreases when a BISR model uses burst shots taken at sub-optimal exposure times compared to optimal time (c). On the other hand, as shown in (f), the model reconstructs a high-resolution (HR) image as if it were captured using optimally exposed burst shots by harnessing information from non-uniformly exposed burst shots.

and enhancement. Despite the advantages conferred by utilizing multiple input images for super-resolution, the successive frames are usually acquired with uniform exposure times [2, 19]. This may induce inferior-quality outputs of the enhancement network utilizing the burst shots captured at the suboptimal exposure time due to inadequate camera noise and motion blur present in those shots. As shown in Fig. 1, the output quality significantly declines when utilizing burst shots acquired far from the optimal exposure time in the uniformly exposed setting. This approach becomes impractical in real-world scenarios where ascertaining the optimal exposure time is challenging. Meanwhile, we observe that the model utilizing burst shots acquired with non-uniform exposures can reconstruct an HR image as if it uses the optimal exposure time.

Motivated by this observation, we explore using burst shots with non-uniform exposures for super-resolution in real-world scenarios, introducing a novel benchmark named

* These authors contributed equally to this work.

Non-uniformly Exposed Burst Image (NEBI). Unlike existing burst image super-resolution benchmarks [2, 19], our approach emphasizes non-uniform burst imaging, capturing images at varying exposure times. This enables the acquisition of images with a broader range of irradiance and motion characteristics within a scene for practical scenarios.

In contrast to burst shots with a uniform exposure, where all frames exhibit similar levels of degradation (e.g., blur, noise), burst shots with non-uniform exposures present varying degrees of degradation. The degradation of the base frame, which merges information from subsequent frames, may interfere with the feature alignment and fusion processes between the base frame and the other frames, resulting in a noticeable decline in the output image quality. However, existing methods [2, 3, 11] assume that the first frame of the input bursts serves as the base frame without considering the importance of the base frame on the image restoration. As shown in Fig. 2 (a), the quality of the output is decreased when the first frame is always used as the base frame. To address this limitation, we also introduce a Frame Selection Network (FSN) that automatically discerns the most suitable base frame as shown in Fig. 2 (b). Our FSN effectively eliminates the presumption that the first frame is invariably the optimal choice, improving the overall image quality. Furthermore, our FSN seamlessly integrates into prevailing burst enhancement networks, offering plug-and-play compatibility. Our proposed method achieves a noteworthy enhancement in quality compared to existing stand-alone super-resolution methods, as demonstrated by superior performance in perceptual metrics on synthetic and real NEBI datasets.

The contributions of our paper are threefold:

- We employ non-uniformly exposed burst shots for practical scenarios where optimal exposure times are unknown in the super-resolution task.
- We introduce the Non-uniformly Exposed Burst Image (NEBI) benchmark, featuring varied exposure times.
- We propose a novel Frame Selection Network (FSN) that selects the most suitable base frame for BISR in a plug-and-play manner.

2. Related Work

Burst image enhancement. Burst denoising [8, 14, 28] and burst deblurring [1, 42] have recently been studied with the development of handheld cameras. In addition to burst denoising, burst images have been utilized for low-light denoising to enhance images captured in very dark conditions [6, 19, 22, 27]. For the burst image super-resolution, a seminal classical paper [41] proposes a restoration model using multiple frames with known translation between frames at the frequency domain. Many following approaches [12, 13, 17, 23, 39] have extended it to the spatial domain to solve burst image super-resolution. Cur-

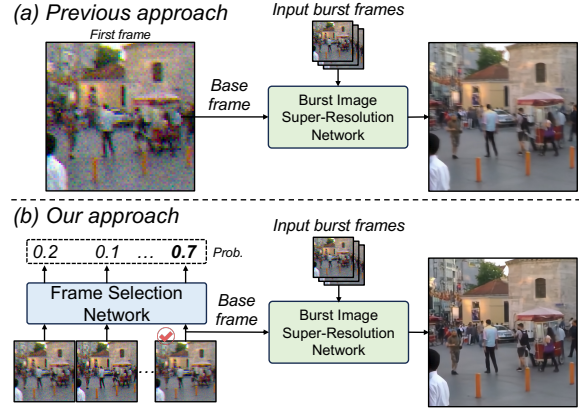


Figure 2. (a) Previous methods [2, 3, 11] fix the base frame to the first frame, which often suffers from degradation (e.g., noise, blur), resulting in poor image quality. (b) Our approach dynamically selects the base frame using a frame selection network, enhancing the quality of the final output image. This adaptive strategy helps the model generate a cleaner final output image.

rent learning-based methods [9, 24, 25] formulate the burst-frame super-resolution joint learning to align burst frames. DBSR [2] proposes a burst image enhancement benchmark and sets the baseline model to predict a DSLR-quality image from input bursts captured by a handheld device, and Deep-rep [3] extends it by applying reconstruction loss from image space to feature space. BIP-Net [11] obtains a restored and enhanced output image by aligning frames to use deformable convolution, exchanging information between burst features, and fusing the features progressively. However, all the methods mentioned above do not take input bursts with non-uniform exposures for practical scenario and assume that the first frame among the input bursts is a base frame.

Frame selection in an image sequence. Existing frame selection methods mainly focus on video object segmentation and action recognition in video. BubbleNet [16] chooses the best reference frame to propagate the predicted segmentation map to the other frames for video object segmentation. ATP [5] suggests using a single selected frame rather than all frames to construct video representation in action recognition. Dahary et al. [8] propose to learn exposure time in burst shots by leveraging that the long-exposed image obtains a high-SNR image and the short-exposed image is sharp. Unlike the previous methods, we cover a scenario of input bursts with dynamic exposure times and select the base frame for each input dynamically.

One important baseline for selecting the base frame is the Auto-Exposure (AE) algorithm. The common approach to AE is to measure optimal exposure based on the image’s histogram or entropy [40, 46]. Zhang *et al.* [46] selected the best exposure time based on the highest entropy of the histogram. Tedla *et al.* [40] combine the content-agnostic and semantic AE algorithms.

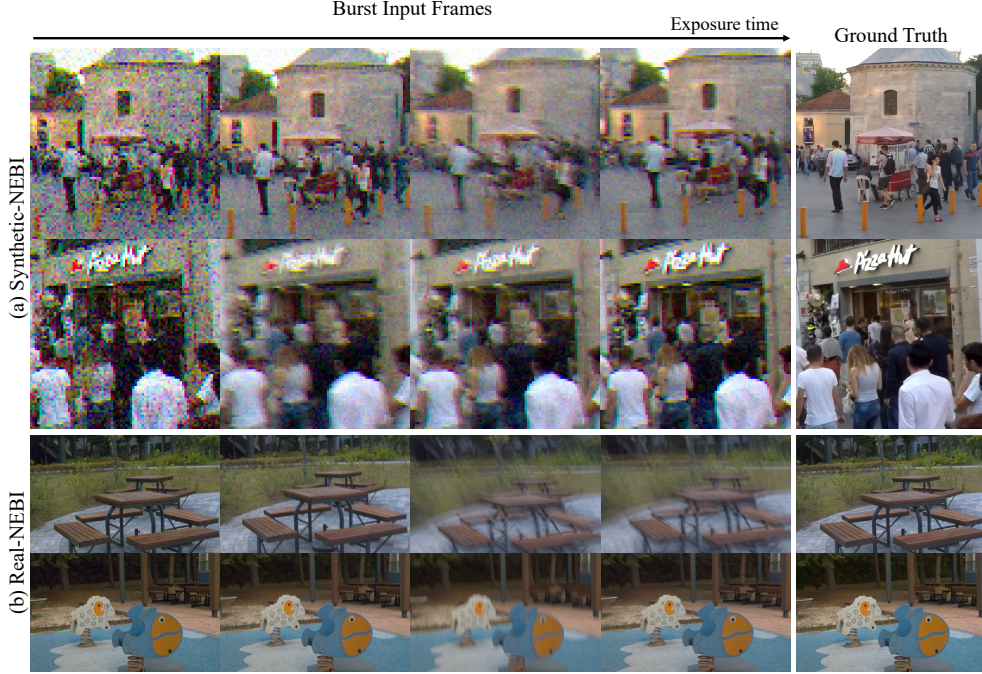


Figure 3. **Visualization of the Synthetic-/Real-NEBI dataset.** The left four columns display a subset of 14 input bursts, while the rightmost column presents the ground-truth image corresponding to the first input frame. From left to right in the burst sequence, the exposure time increases, resulting in decreased noise and increased blur. Best view to zoom.

3. Benchmark

We propose a new benchmark for burst image super-resolution with non-uniform camera settings (*i.e.*, varying the exposure time), named Non-uniformly Exposed Burst Image (NEBI). By acquiring the burst frames with various exposure times, our NEBI is able to capture a broader range of irradiance and motion characteristics within a scene for practical scenarios where the optimal exposure time is unknown. Specifically, the benchmark consists of two tracks: Synthetic-NEBI and Real-NEBI, where Synthetic-NEBI is used for training and testing, and Real-NEBI is only used for evaluation. Both benchmarks have heterogeneous image noise and motion blur, and there are misalignments between bursts that naturally occur from camera movement.

3.1. Synthetic-NEBI

We create Synthetic-NEBI to address the cost of obtaining sufficient pairs of burst frames and the corresponding ground-truth images in the real-world. To achieve this, we utilize sharp frames from a video dataset [31] and synthesize a dataset with realistic blur and noise. Specifically, we synthesize blur from gyro sensor data and inject realistic noise using the Poisson-Gaussian noise model. We simulate a longer exposure time by increasing the number of gyro sensor data and decreasing the noise amount [8, 20, 26, 48]. This process results in a sequence of frames that resemble a burst taken from short to long exposures, allowing us to

simulate non-uniform exposures and heterogeneous degradation that occur in real-world shooting environments.

We simulate various degradations, including motion blur, noise, and low-resolution, considering where they occur in the image acquisition process. During image acquisition, the scene radiance emitted from the target scene is first blurred by the relative motions between the camera and the scene, and then is blurred by the lens and sampled by the sensor. Then, it is mosaiced by the color filter array, and processed by the ISP. To mimic this process, we first apply the inverse ISP to obtain synthetic scene radiance, followed by applying blur synthesis, downsampling, and noise synthesis sequentially. Specifically, we first extract 14 consecutive images from the video dataset [31] and convert the images to the RAW color space through the unprocessing pipeline [4]: inverse gamma compression, inverse gain, and inverse CCM. We synthesize blur on the images by utilizing gyro sensor data collected from the Samsung Galaxy S22. Specifically, considering different exposure times, we randomly sample consecutive gyro sensor values matching each burst image’s exposure. Next, we interpolate these values and then calculate homographies. We obtain a blurred image by warping the ground truth sharp images corresponding to each burst using the previously derived homographies and subsequently averaging them. Then, we apply downsampling and synthesize realistic noise on the synthesized blurred images. We randomly sample shot noise λ_{shot} and read noise λ_{read} parameters of the burst images accord-

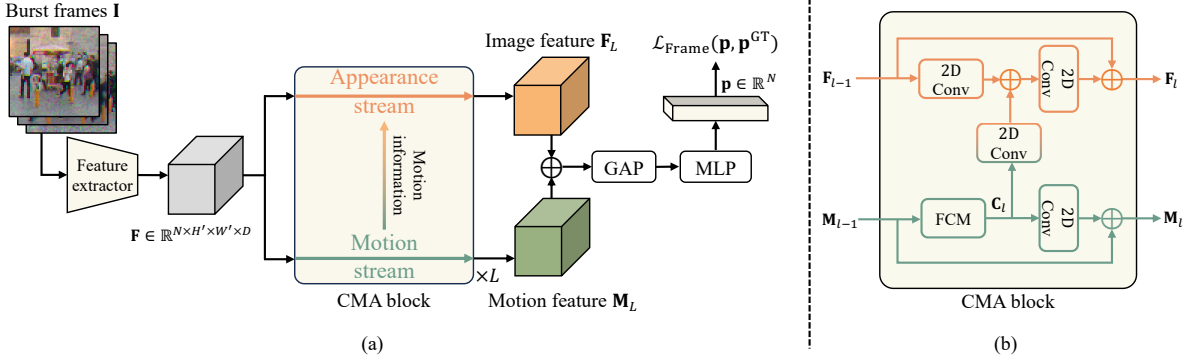


Figure 4. (a) **Overall architecture of frame selector.** The network begins with a feature extractor followed by L Contextual Motion Aggregation (CMA) blocks, each containing appearance and motion streams that produce image feature F_L and motion feature M_L . These features are summed and further refined through a Global-Averaging-Pooling (GAP) layer and an MLP layer to predict the likelihood of being the base frame. (b) **Feature Correlation Module (FCM).** Given the previous motion feature M_{l-1} , our FCM generates the motion information C_l by computing the local feature correlation. The resultant F_l and M_l are obtained by propagating C_l .

ing to those of exposure times. Noisy images are synthesized using the Poisson-Gaussian noise model [4] with the noise parameters. Finally, we convert the burst images to the Bayer pattern RAW images through mosaicing. Note that the input video frames naturally contain shifting between frames for hand motion, and this is sufficient to simulate the misalignment of the synthetic burst frames. We finally obtain a synthetic RAW burst sequence with heterogeneous image noise and motion blur. More details for synthesizing the dataset are provided in the supplementary.

The synthetic-NEBI consists of 2,750 synthetic non-uniformly exposed RAW bursts. We split the dataset into 2,200 bursts for training and 550 bursts for the test set, using GoPro [31] as a source dataset. Each burst sequence consists of 14 RAW frames center-cropped to 640×640 pixels, and the source frames are downsampled by 4 to produce $4 \times 80 \times 80$ input bursts. The size of the corresponding GT images is $3 \times 640 \times 640$ pixels. The images on the left of Fig. 3 (a) demonstrate the effect of the shorter exposure, which is sharp but noisy, while the images on the right demonstrate the effect of the longer exposure, which is high-SNR but blurry. Overall, the Synthetic-NEBI dataset provides a comprehensive representation of non-uniform exposures shooting environments.

3.2. Real-NEBI

We create Real-NEBI to evaluate the quality of the result of super-resolution captured in real-world. To collect the Real-NEBI dataset, we use a dual-camera system similar to [35, 36, 49, 50], which consists of two camera modules and a beam-splitter. The beam-splitter equally splits photons into two camera modules; the dual-camera system can simultaneously capture burst frames with different exposure and corresponding sharp frames at the same time. We install a 5% neutral density filter (ND filter) on the camera module used to capture the burst frames to ensure stable han-

dling of long exposure time [49]. Using the camera system, we capture real-world burst frames with non-uniform exposures and the corresponding sharp and high-SNR ground-truth frame.

Specifically, one of the two camera modules captures 14 frames with gradually increasing exposure times from 0.01 to 0.14 seconds, and the other camera module captures the corresponding sharp images with 0.005 seconds of exposure time. We inverse-proportionally decrease the gain value of each burst frame to maintain the same brightness level as the corresponding sharp frame. We collect all images in the RAW format and perform spatial and photometric alignments, as done in [36]. The burst frames are downsampled for the input of the enhancement pipeline, and the GT frames undergo only demosaicing.

The Real-NEBI dataset consists of 96 burst sequences with heterogeneous degradation by non-uniform exposure times. A burst set consists of 14 burst frames, and the input bursts and the ground-truth image have shapes of $4 \times 148 \times 238$ and $3 \times 1184 \times 1904$, respectively. Figure 3 (b) shows the examples of our Real-NEBI dataset. Detailed descriptions of our Synthetic- and Real-NEBI generation are in the supplementary material.

4. Frame Selection Network

The goal of BISR is to reconstruct a high-quality image by merging the information from successive frames. However, inevitable misalignments between frames occur as burst images are rapidly captured by handheld devices. This discrepancy hampers the fusion of their complementary information, resulting in ghosting and blurring artifacts in the output image [11, 43]. While existing methods primarily concentrate on developing burst alignment algorithms to address this issue, they focus on aligning the other frames with the first frame rather than determining which frame best aligns with the others. In this section, we present FSN that auto-

matically discerns the most suitable base frame that aligns best with the others.

Given raw burst frames $\mathbf{I} \in \mathbb{R}^{N \times H \times W \times 4}$, the objective of FSN is to predict an index k , indicating the position of the optimal base frame, where N , H , and W represent the burst size, height, and width of a burst frame, respectively. After the frame selection, the burst frames pass through a super-resolution network, producing an enhanced image by merging complementary information from the input burst frames into the selected base frame.

As depicted in Figure 4, FSN first constructs the image feature of each frame using Convolutional Neural Network (CNN). The constructed image feature is fed into our Correlation-based Motion Aware (CMA) blocks, which update the image feature based on motion information extracted from the burst frames. To extract motion information from the burst frames, we propose Feature Correlation Module (FCM), which computes the local correlation along both spatial and temporal axes. Finally, the output features from the last CMA block are then passed into the classifier to predict the likelihood of being the base frame. The remainder of this section presents the details of the feature extractor (Sec. 4.1), CMA block (Sec. 4.2), and FCM (Sec. 4.3) and the training procedure (Sec. 4.4).

4.1. Feature Extractor

Given raw burst frames \mathbf{I} , we use two-layer CNN to extract the image feature $\mathbf{F} = [\mathbf{f}_1; \dots; \mathbf{f}_N] \in \mathbb{R}^{N \times H' \times W' \times D}$, where $\mathbf{f}_i \in \mathbb{R}^{H' \times W' \times D}$ is the feature of each frame and $[\cdot; \cdot]$ is a concatenation operator. H' , W' , and D denote the height and width of the image feature and the feature dimension, respectively. The constructed image feature \mathbf{F} is fed into the first CMA block as input.

4.2. Correlation-based Motion Aware Block

To ascertain which frame aligns best with the others, the model must consider not only the degrees of degradation of each frame, as degradation contributes to misalignment issues, but also the motion information, as the alignment between frames captured by handheld devices should consider the motion trajectory.

To address this challenge, we propose a novel Correlation-based Motion Aware (CMA) block designed to enable the model to efficiently fuse noise information within a frame and motion information across frames. A CMA block has a two-stream architecture: an appearance stream and a motion stream. The appearance stream updates the image feature to consider the degradation of a frame. Simultaneously, the motion stream extracts the motion information in the frames using the Feature Correlation Module (FCM) (described in Sec. 4.3) and conveys motion information to the next block, independent of the image feature of the appearance stream. The motion information is then

propagated to the image feature, enabling the model to consider noise and motion information when selecting a base frame. Specifically, the l -th CMA block takes the previous results of image feature \mathbf{F}_{l-1} and motion feature \mathbf{M}_{l-1} , and outputs updated \mathbf{F}_l , and \mathbf{M}_l . In the appearance stream, the previous image feature \mathbf{F}_{l-1} is updated using correlation feature \mathbf{C}_l as follow:

$$g(\cdot) = \text{ReLU}(\text{Conv2D}(\cdot)) \quad (1)$$

$$\mathbf{F}_l = \mathbf{F}_{l-1} + g(g(\mathbf{F}_{l-1}) + \mathbf{C}_l), \quad (2)$$

$$\mathbf{C}_l = \text{FCM}(\mathbf{M}_{l-1}). \quad (3)$$

For the sake of simplicity, we omit the subscript that indicates different weights in g . In equation 3, the correlation feature \mathbf{C}_l is extracted from the previous motion feature \mathbf{M}_{l-1} using the FCM. Similar to the appearance stream, the motion feature is also updated by using \mathbf{C}_l as follows:

$$\mathbf{M}_l = \mathbf{M}_{l-1} + g(\mathbf{C}_l). \quad (4)$$

At the first CMA block, the image feature \mathbf{F} is utilized as \mathbf{F}_0 and \mathbf{M}_0 .

4.3. Feature Correlation Module

Traditional super-resolution methods typically utilize optical flow techniques to extract motion information for aligning input burst frames[2, 3]. However, these approaches use an off-the-shelf network pre-trained in the RGB space, resulting in inaccurate alignment when applied to RAW images, thus negatively impacting generated outputs and incurring high computational costs. To mitigate the issue, we utilize local feature correlation with spatial-temporal neighbors to extract motion information.

In detail, given a motion feature \mathbf{M} , feature correlation map $\mathbf{S} \in \mathbb{R}^{N \times H' \times W' \times U \times V \times W}$ is computed as follow:

$$\mathbf{S}_{t,x,y,u,v,w} = \text{sim}(\mathbf{M}_{t,x,y}, \mathbf{M}_{t+u,x+v,y+w}), \quad (5)$$

where $\text{sim}(\cdot, \cdot)$ is similarity function (e.g., cosine similarity). U , V , and W is the local window size of the time, height, and width axes, respectively. (t, x, y) is the position of a query point and (u, v, w) is a offset from the query point. For the sake of simplicity, we omit the subscript l in \mathbf{M} . Then, A correlation feature $\mathbf{C} \in \mathbb{R}^{T \times X \times Y \times D}$, which contains motion information, is extracted from \mathbf{S} using a series of 3D convolution along with U , V , and W axes as:

$$h(\cdot) = \text{ReLU}(\text{Conv3D}(\cdot)) \quad (6)$$

$$\mathbf{C} = (h \circ \dots \circ h)(\mathbf{S}). \quad (7)$$

We omit the subscript that indicates different weights in h for simplicity.

4.4. Base Frame Estimation and Network Training

\mathbf{F}_L and \mathbf{M}_L , the outputs of the last CMA block, are fed into MLP after the summation to predict the probability as the base frame $\mathbf{p} \in \mathbb{R}^N$ as follow:

$$\mathbf{p} = \sigma(\text{MLP}(\text{GAP}(\mathbf{F}_L + \mathbf{M}_L))), \quad (8)$$

where GAP is a global average pooling over spatial dimension and σ is a *softmax* operation. During inference, the index with the highest probability in \mathbf{p} is selected as a base frame.

To train our FSN, we employ the cross-entropy between the ground-truth \mathbf{p}^{GT} , which one-hot vector with ground-truth base frame index set to 1, and the predicted \mathbf{p} as follows:

$$\mathcal{L}_{\text{Frame}} = \sum (-\mathbf{p}^{\text{GT}} * \log(\mathbf{p})). \quad (9)$$

The ground-truth base frame index is determined by selecting the frame with the highest metric (*i.e.*, PSNR, SSIM, LPIPS) among the input burst frames. Unless otherwise stated, we select the ground-truth based on the PSNR.

5. Experiment

We demonstrate the effectiveness of the use of non-uniformly exposed burst shots in practical scenarios and our proposed frame selection network when integrated into existing super-resolution models [2, 3, 11].

5.1. Quantitative Results

Uniform vs. non-uniform exposure settings. Table 1 compares PSNR scores of BIPNet [11], Deep-sr [2], and Deep-rep [3] trained on uniformly-exposed burst images at various exposure times. The results show that image quality increases as the model utilizes an exposure time close to the optimal. It suggests that it is crucial to use optimally-exposed burst shots to generate high-quality images in image restoration. However, obtaining such optimally-exposed burst shots in real-world conditions poses inherent challenges due to the dynamic nature of lighting conditions. Interestingly, when utilizing non-uniformly-exposed burst shots, the model exhibits comparable results to those obtained with optimal exposure times. This demonstrates that bursts captured at non-uniform exposure settings are beneficial in practical scenarios where determining the optimal exposure time is challenging.

Effect of FSN on Synthetic-/Real-NEBI. Table 2 shows a quantitative comparison between existing super-resolution models and their variants with our FSN on Synthetic-/Real-NEBI. On the Synthetic-NEBI, the existing models exhibit subpar performance due to designating the first frame as the base frame. Compared to existing models, their variants incorporating our FSN consistently demonstrate performance improvements. These results indicate that select-

Model	PSNR				
	0.01	0.02	0.04	0.14	NEBI
BIPNet [11]	33.276	33.311	34.017	32.749	<u>33.774</u>
Deep-sr [2]	33.170	33.288	33.732	32.845	<u>33.516</u>
Deep-rep [3]	33.580	33.631	34.082	32.918	<u>34.059</u>

Table 1. **Performance comparison on uniform/non-uniform exposure settings.** The table shows PSNR scores for a vanilla BISR network [2, 3, 11], trained on the uniform exposure time from 0.01 to 0.14, and NEBI. When the model utilizes the burst shots close to the optimal exposure time, the quality of outputs is increased. When the optimal exposure time is unknown, it is beneficial to use the non-uniformly exposed burst shots.

Dataset	Model	Frame selection	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Synthetic-NEBI	BIPNet [11]	-	33.774	0.920	0.108
		FSN	33.878	0.922	0.106
	Deep-sr [2]	-	33.516	0.917	0.117
		FSN	33.872	0.921	0.107
	Deep-rep [3]	-	34.059	0.925	0.103
		FSN	34.313	0.928	0.099
Real-NEBI	BIPNet [11]	-	33.957	0.904	0.143
		AE	34.184	0.900	0.157
		FSN	34.367	0.901	0.157
	Deep-sr [2]	-	31.084	0.908	0.180
		AE	30.914	0.901	0.195
		FSN	31.149	0.907	0.180
	Deep-rep [3]	-	31.088	0.911	0.174
		AE	30.861	0.903	0.193
		FSN	31.308	0.904	0.194

Table 2. **Results on Synthetic-/Real- NEBI.** ‘AE’ and ‘FSN’ stands for Auto Exposure and Frame Selection Network, respectively. The best score is highlighted in bold.

ing an appropriate base frame contributes to the performance increase, facilitating improved alignment and fusion among the burst frames. For the Real-NEBI, we evaluate the performance of our FSN on Real-NEBI while training on Synthetic-NEBI as acquiring labeled data for super-resolution is challenging and costly. Similar to Synthetic-NEBI, Our FSN consistently improves the PSNR values of existing super-resolution models on the real-world burst examples of non-uniform exposure conditions, despite being trained on Synthetic-NEBI. These results show the importance of frame selection even in real-world scenarios, demonstrating the effectiveness of our FSN. Our FSN also shows comparable results on SSIM and LPIPS metrics.

Auto-Exposure (AE) vs. FSN. To demonstrate the effectiveness of our FSN, we also compare FSN with the AE algorithm. Following previous AE methods [40, 46], we select the base frame with the maximum entropy calculated based on the image’s histogram. In Table 2, we observe marginal PSNR improvement or even declines when using AE. This suggests that AE is sub-optimal in selecting the base frame, emphasizing the effectiveness of FSN.

Target frames for FSN training. As we train the FSN to select the frame that yields the highest PSNR on Synthetic-

NEBI, the FSN may exhibit sub-optimal results on other metrics, such as SSIM and LPIPS, when applied to Real-NEBI. In Table 3, we explore using different target frames in training FSN and evaluate its effect on performance in Real-NEBI. Specifically, we trained FSN to select the frame that yields the highest PSNR, SSIM, and LPIPS on the Synthetic-NEBI and evaluate the FSN on Real-NEBI. The results show that training based on SSIM yields high SSIM performance, while training based on LPIPS yields high LPIPS performance.

5.2. Qualitative results

We visualize super-resolution results and the selected base frame to verify the effectiveness of our FSN in Figure 5. The first and second columns show first frames, which are typically used as a base frame, and frames selected as base frames by FSN, respectively. We observe that our proposed method selects frames with less degradation. Additionally, we note increased qualities of super-resolution results when applying FSN to Deep-*rep* (third vs fourth columns). These results indicate that selecting an appropriate frame for the base frame increases the image quality in the super-resolution task. For example, when the base frame is noisy, the legs appear blurry in the first row. However, when selecting a frame with less degradation, the legs appear sharp compared to the Deep-*rep* [3].

5.3. Ablation Study of FSN

We explore the impact of each individual component of FSN. For the ablation study of FSN, we adopt Deep-*rep* [3] trained/tested on Synthetic-NEBI as a super-resolution network. Specifically, we attach variants of FSN to Deep-*rep* and evaluate their performance.

Number of CMA blocks. We utilize the CMA block to encode visual features with motion information. To investigate the impact of the number of CMA blocks, we gradually increase the number of CMA blocks to a vanilla FSN that has CMA blocks. Comparing the baseline (solo super-resolution network) and the baseline variants with the vanilla FSN, there is notable performance gain as shown in the first and second rows of Table 4. This result demonstrates that selecting an appropriate base frame contributes to the model’s ability to generate high-quality images. Furthermore, the addition of CMA blocks increases the performance gap between the baseline and the baseline with the FSN. we observed that employing two CMA blocks yields the best performance across all metrics.

We also increase the capacity of the vanilla FSN to have a similar number of parameters with the FSN with 2 CMA blocks (*i.e.*, 0*) to investigate whether the improvement is due to an increased number of parameters or the effectiveness of our CMA block. The results show that even though the capacity of the vanilla FSN is increased, the FSN with

Frame selection	PSNR↑	SSIM↑	LPIPS↓
-	33.957	0.904	0.143
FSN _P	34.367	0.901	0.157
FSN _S	34.675	0.907	0.155
FSN _L	34.053	0.905	0.143
oracle	35.650	0.912	0.142

Table 3. **Comparison of training strategies on Real-NEBI.** The table illustrates how performance varies when different metrics are employed to select the ground-truth frame in the Real-NEBI dataset. The subscription *P*, *S*, *L* denote frame selection strategies based on PSNR, SSIM, and LPIPS score respectively.

# of CMA block	PSNR↑	SSIM↑	LPIPS↓
-	34.059	0.925	0.103
0	34.174	0.926	0.101
0*	34.164	0.926	0.101
1	34.252	0.927	0.101
2	34.313	0.928	0.099
4	34.248	0.927	0.100

Table 4. **Ablation study on the number of CMA Blocks.** The first row represents the baseline, *i.e.*, vanilla super-resolution network, while the subsequent five rows explore the impact of varying numbers of CMA blocks. * denotes the balanced model size as the number of parameters when there are two of the CMA block.

Appear.	Motion	Motion Feat	PSNR↑	SSIM↑	LPIPS↓
-	-	-	34.174	0.926	0.101
✓	-	-	34.123	0.926	0.103
-	✓	PWCNet _{raw}	34.104	0.926	0.102
-	✓	PWCNet _{rgb}	34.088	0.925	0.102
-	✓	FCM	34.266	0.927	0.101
✓	✓	FCM	34.313	0.928	0.099

Table 5. **Impact of appearance & motion streams.** We compare the performance of our frame selector with different combinations of appearance and motion streams. The ‘Motion Feat’ column indicates the ways of extracting the motion feature.

the CMA blocks outperforms the vanilla FSN. These results demonstrate the effectiveness of our CMA block.

Appearance & motion streams. Our CMA block consists of the appearance stream to update the image feature and the motion stream to encode the motion information in the input burst frames. To investigate the impact of each stream, we gradually add each stream to the baseline, where neither appearance nor motion stream is applied (*i.e.*, vanilla FSN) in Table 5. The appearance stream in isolation fails to demonstrate discernible improvement compared to the baseline as there is a misalignment between the burst shots. However, the contribution of the motion stream is significant, indicating that motion information is crucial to the base frame selection. Notably, leveraging both appearance and motion streams provides the best performance as the image feature in the appearance stream is updated by using the correct sub-pixel information aligned by the FCM.

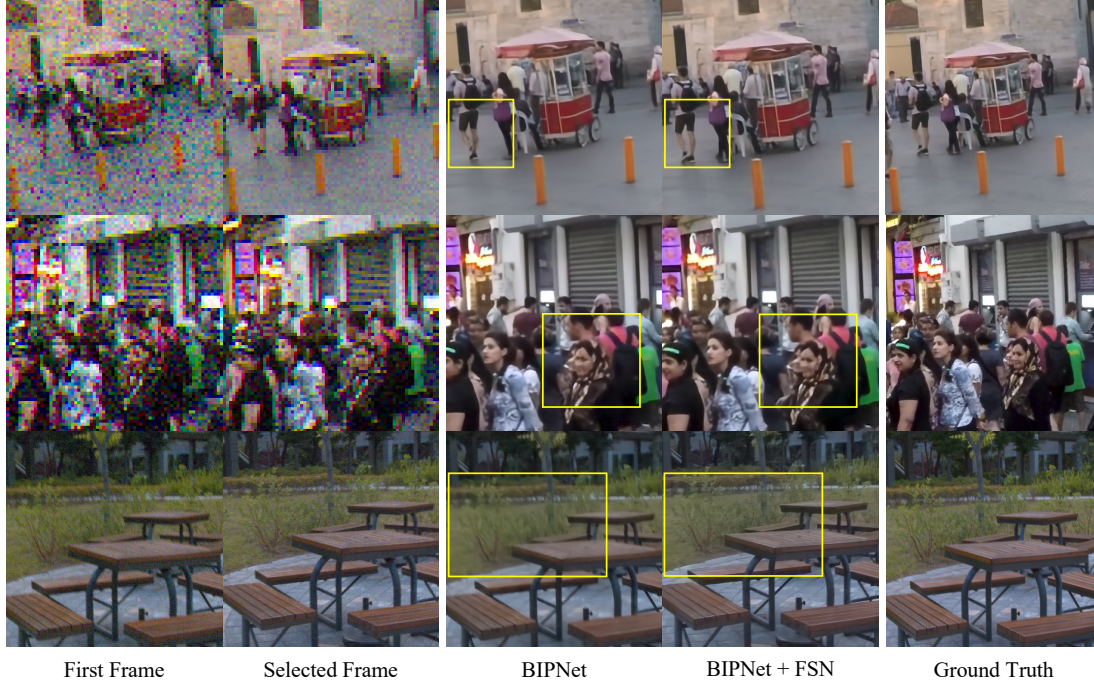


Figure 5. **Qualitative comparison on synthetic-NEBI (top two rows) and real-NEBI (bottom row).** BIPNet [11] with our frame selector enhances high-frequency image details by merging effectively complementary information from multiple frames into the selected base frame, while BIPNet [11] predicts blurry images due to the severe degradations in the first frame. Best view to zoom.

Model	Params.	FLOPs
BIPNet [11]	6.67M	1.67T
Deep-sr [2]	13.01M	544.86G
Deep-rep [3]	12.13M	1.28T
FSN	1.59M	18.53G

Table 6. **Parameter comparison between existing methods and FSN.** We analyze the number of parameters and FLOPs of FSN in contrast to previous burst enhancement models.

Impact of FCM. In Table 5, we replace the FCM with an off-the-shelf optical flow network, PWCNet, to investigate the effectiveness of our FCM. Following Deep-sr [2], we utilize the raw image to obtain the optical flow, PWCNet_{raw}. The result shows that utilizing optical flow for motion information extraction leads to a decrease in the model’s performance compared to utilizing our FCM. Since the optical flow network is trained on RGB, it may result in inaccurate predictions on raw images, negatively impacting the model’s performance. Therefore, we also utilize the RGB image to obtain the optical flow, PWCNet_{rgb}. Interestingly, the model’s performance is degraded compared to PWCNet_{raw}. These results indicate that optical flow may not be suitable for extracting motion features for frame selection, highlighting the effectiveness of FCM.

Parameter Analysis. Applying our FSN to existing super-resolution networks may increase computational costs, potentially impacting performance metrics such as FPS and memory usage. We conduct an analysis of the number of pa-

rameters and FLOPs in Table 6. FSN has a compact model size and small FLOPs compared to existing burst enhancement methods [2, 3, 11]. This guarantees plug-and-play integration of FSN into existing models with a low computational overhead.

6. Conclusion

We have explored using burst shots with non-uniform exposures for practical scenarios where the acquisition of an optimal exposure time is challenging. By using burst shots with non-uniform exposures, the model reconstructs a high-resolution image. As burst shots with non-uniform exposures present various degrees of degradation, we introduce a base frame selection model to improve the performance. Our evaluations on the Synthetic- and Real-NEBI datasets demonstrated the effectiveness of using non-uniformly exposed burst shots for practical scenarios and the importance of selecting a proper base frame. We believe that our proposed dataset and method can contribute to the advancement of super-resolution methods and pave the way for further research in this field.

Acknowledgements. This work was supported by Samsung Advanced Institute of Technology (SAIT), the NRF grant (No. 2023R1A2C200494611), and the IITP grants (No. 2022-0-00290, Visual Intelligence for Space-Time Understanding and Generation, No.2019-0-01906, AI Graduate School Program at POSTECH) funded by the Korea government.

References

- [1] Miika Aittala and Frédo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 731–747, 2018. 2
- [2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021. 1, 2, 5, 6, 8
- [3] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2460–2470, 2021. 2, 5, 6, 7, 8
- [4] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 3, 4
- [5] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the “video” in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927, 2022. 2
- [6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 2
- [7] Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis. Open high-resolution satellite imagery: The worldstrat dataset—with application to super-resolution. *Advances in Neural Information Processing Systems*, 35:25979–25991, 2022. 1
- [8] Omer Dahary, Matan Jacoby, and Alex M Bronstein. Digital gimbal: End-to-end deep image stabilization with learnable exposure times. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11936–11945, 2021. 2, 3
- [9] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*, 2020. 1, 2
- [10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1
- [11] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022. 2, 4, 6, 8
- [12] Michael Elad and Arie Feuer. Restoration of a single super-resolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997. 2
- [13] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004. 2
- [14] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *Proceedings of the European conference on computer vision (ECCV)*, pages 538–554, 2018. 2
- [15] Hayit Greenspan. Super-resolution in medical imaging. *The computer journal*, 52(1):43–63, 2009. 1
- [16] Brent A Griffin and Jason J Corso. Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8914–8923, 2019. 2
- [17] Russell Hardie. A fast image super-resolution algorithm using an adaptive wiener filter. *IEEE Transactions on Image Processing*, 16(12):2953–2964, 2007. 2
- [18] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28*, pages 387–395. Springer, 2021. 1
- [19] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 1, 2
- [20] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14781–14790, 2021. 3
- [21] Jithin Saji Isaac and Ramesh Kulkarni. Super resolution techniques for medical image processing. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, pages 1–6. IEEE, 2015. 1
- [22] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *IEEE Transactions on Image Processing*, 30:9372–9385, 2021. 2
- [23] Thomas Köhler. Multi-frame super-resolution reconstruction with applications to medical imaging. *arXiv preprint arXiv:1812.09375*, 2018. 2
- [24] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2370–2379, 2021. 2
- [25] Bruno Lecouat, Thomas Eboli, Jean Ponce, and Julien Mairal. High dynamic range and super-resolution from raw image bursts. *arXiv preprint arXiv:2207.14671*, 2022. 2
- [26] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *International Conference on Machine Learning*, pages 2965–2974. PMLR, 2018. 3

- [27] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T Barron, Dillon Sharlet, Ryan Geiss, et al. Handheld mobile photography in very low light. *ACM Trans. Graph.*, 38(6):164–1, 2019. 1, 2
- [28] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *ACM Transactions on Graphics (TOG)*, 33(6):1–9, 2014. 2
- [29] Dwarikanath Mahapatra, Behzad Bozorgtabar, and Rahil Garnavi. Image super-resolution using progressive generative adversarial networks for medical image analysis. *Computerized Medical Imaging and Graphics*, 71:30–39, 2019. 1
- [30] Yogendra Rao Musunuri, Oh-Seol Kwon, and Sun-Yuan Kung. Srodnet: Object detection network based on super resolution for autonomous vehicles. *Remote Sensing*, 14(24): 6270, 2022. 1
- [31] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 3, 4
- [32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1
- [33] Junhyug Noh, Wonho Bae, Wonhee Lee, Jinhwan Seo, and Gunhee Kim. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9725–9734, 2019. 1
- [34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arxiv 2022. *arXiv preprint arXiv:2204.06125*, 2022. 1
- [35] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 184–201. Springer, 2020. 4
- [36] Jaesung Rim, Geonung Kim, Jungeon Kim, Junyong Lee, Seungyong Lee, and Sunghyun Cho. Realistic blur synthesis for learning image deblurring. *arXiv preprint arXiv:2202.08771*, 2022. 4
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1
- [38] J Shermeyer and A Van Etten. The effects of super-resolution on object detection performance in satellite imagery. arxiv 2019. *arXiv preprint arXiv:1812.04098*. 1
- [39] Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on image processing*, 16(2):349–366, 2007. 2
- [40] SaiKiran Tedla, Beixuan Yang, and Michael S Brown. Examining autoexposure for challenging scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13076–13085, 2023. 2, 6
- [41] R Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984. 2
- [42] Patrick Wieschollek, Bernhard Schölkopf, Hendrik PA Lensch, and Michael Hirsch. End-to-end learning for image burst deblurring. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part IV 13*, pages 35–51. Springer, 2017. 2
- [43] Bartłomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4): 1–18, 2019. 1, 4
- [44] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision*, pages 492–511. Springer, 2020. 1
- [45] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 1
- [46] Chi Zhang, Zheng You, and Shijie Yu. An automatic exposure algorithm based on information entropy. In *Sixth International Symposium on Instrumentation and Control Technology: Signal Analysis, Measurement Theory, Photo-Electronic Technology, and Artificial Intelligence*, pages 152–156. SPIE, 2006. 2, 6
- [47] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2020. 1
- [48] Zhilu Zhang, RongJian Xu, Ming Liu, Zifei Yan, and Wangmeng Zuo. Self-supervised image restoration with blurry and noisy pairs. In *Advances in Neural Information Processing Systems*. 3
- [49] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 191–207. Springer, 2020. 4
- [50] Zhihang Zhong, Ye Gao, Yinqiang Zheng, Bo Zheng, and Imari Sato. Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network. 2022. 4