Efficient Burst Super-Resolution with One-step Diffusion

Kento Kawai Takeru Oba Kyotaro Tokoro Kazutoshi Akita Norimichi Ukita
Toyota Technological Institute
{sd21033,sd21502,sd24439,sd21501,ukita}@toyota-ti.ac.jp

Abstract

While burst Low-Resolution (LR) images are useful for improving their Super Resolution (SR) image compared to a single LR image, prior burst SR methods are trained in a deterministic manner, which produces a blurry SR image. Since such blurry images are perceptually degraded, we aim to reconstruct sharp and high-fidelity SR images by a diffusion model. Our method improves the efficiency of the diffusion model with a stochastic sampler with a high-order ODE as well as one-step diffusion using knowledge distillation. Our experimental results demonstrate that our method can reduce the runtime to 1.6 % of its baseline while maintaining the SR quality measured based on image distortion and perceptual quality.

1. Introduction

Super-Resolution (SR) is a task for super-resolving Low-Resolution (LR) images to their High-Resolution (HR) images. Among various SR methods, SR for super-resolving an LR image is called Single-Image Super-Resolution (SISR) [10, 14, 16, 19, 41]. However, SISR is not an easy task due to its ill-posed nature. That is, there are multiple appropriate HR images of each LR image. In addition to this ill-posed nature, various image degradations make SISR more difficult.

To improve the SR quality, SR methods can take multiple differently-degraded LR images instead of a single LR image to compensate for the degradations between these LR images. Such differently-degraded images can be easily captured by the burst shot mode of a smartphone. Burst SR [1–3, 11, 12, 15, 28–30] integrates these burst images into a single SR image.

Previous burst SR methods [1, 11, 12, 15, 28–30] are trained in a deterministic manner based mainly on a difference between the SR image and its ground truth HR image, such as L1 and L2 losses. However, such reconstruction losses lead to blurred SR images because these losses can be minimized by the average of HR images [25], each of which is an appropriate HR image of the input LR image.

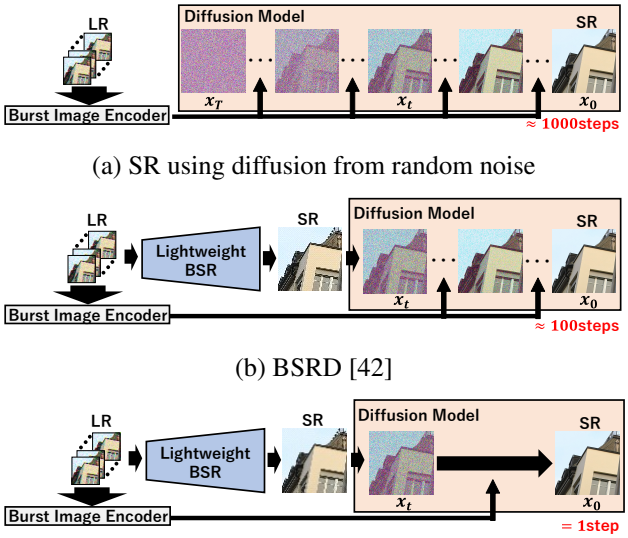


Figure 1. Comparison between prior SR model and our burst SR models.

To avoid this problem, this paper aims to improve the burst SR quality by probabilistic modeling. Probabilistic modeling allows us to represent the probabilistic distribution of sharp SR images of each LR image. Among such probabilistic models, diffusion models [21] are used by the recent burst SR method [42], as shown in (Fig. 1 (b)), because of their performance validated in various computer vision tasks.

To employ the diffusion models for burst SR, this paper focuses on the following two issues:

How to efficiently achieve burst SR. In diffusion models proposed earlier [21], a huge number of iterations (e.g., around 1,000) are required to generate realistic data such as natural images. Since such a huge number of iterations are computationally expensive, several approaches are proposed to reduce the iterative steps (e.g., model minimization using distillation [24, 40] and efficient sampling and training processes [22]). While these approaches are applied to SISR [6, 45, 48], they are not optimized for burst SR.

How to integrate differently-degraded LR images. Unlike random image generation [21], image enhancement and restoration, including SR, must condition a diffusion model for generating an image that fits with input images. For example, for diffusion-based SISR [37, 43], an LR image is used for conditioning the reverse diffusion process. Unlike SISR, however, burst SR takes multiple LR images with different degradations, including blur, noise, and displacement. If such differently-degraded images are directly fed into the diffusion model, the reverse process may reconstruct a blurry SR image by averaging these degraded images.

To cope with these two issues, CCDF [6] and BSRD [42] efficiently uses a diffusion model by skipping early diffusion steps. This skip is achieved by feeding an initial burst SR image reconstructed by a simple deterministic burst SR method into the intermediate step of a probabilistic diffusion model, as shown in Fig. 1 (b). Furthermore, unlike previous diffusion-based SISR methods [6, 45, 48], BSRD employs spatially-aligned multi-scale features extracted from input burst LR images for conditioning the diffusion model.

Our paper extends BSRD [42] by incorporating the following advancements:

- Instead of diffusion models using first-order differential equations used in [6] and [42], such as DDPM [21], DDIM [38], iDDPM (variance scheduling) [35], and VP/VE [39], a second-order differential equation-based model (i.e., Elucidating the Design Space of Diffusion-Based Generative Model (EDM) [22]) is used to reduce the number of diffusion steps as well as to improve the SR quality. While the original EDM removes a large amount of noise per step from random noise to reduce the number of diffusion steps, each step in our method removes only a small amount of noise by appropriately decreasing the noise given to an initial SR image, enabling fine-grained SR reconstruction even through a smaller number of diffusion steps.
- For further reducing the number of diffusion steps for efficient burst SR (i.e., between 5 and 100 in [6] and [42] to one step in our method), a distillation-based teacher-student model [24, 40] is employed. Since this distillation-based model degrades the quality of image synthesis if it begins from random noise, our method avoids this degradation by feeding a properly initialized SR image, which is reconstructed by simple deterministic Burst SR in our method, into the diffusion model.
- Several important parameters for EDM and the distillation model are optimized for our burst SR model.
- With the aforementioned contributions, the SR reconstruction time is reduced to 1.6 % of the baseline, i.e., BSRD [42].

2. Related Work

2.1. Burst SR

While a RAW image consisting of RGGB channels (4 channels) with high bits per pixel resolution (e.g., 14 bits, 16 bits) is captured by a standard digital camera, it is converted by an Image Signal Processing (ISP) to its 8-bit RGB image. General burst SR methods take a set of unprocessed RAW images instead of their processed RGB images. Burst SR methods generally consist of four processes, i.e., feature extraction, alignment, fusion, and reconstruction processes. The alignment process rectifies image features extracted by the feature extraction process so that the features are spatially consistent. The aligned features of all burst images are then merged by the fusion process. The fused features are fed into the reconstruction process to acquire the SR image.

Among the four processes, the alignment process is peculiar to burst SR. If non-aligned features are directly used for SR, the SR image may be blurred. In [15, 28], deformable convolution (DC) [9] is used for implicit spatial alignment. The two-step alignment with optical flow estimation and DC is also proposed in [29].

The reconstruction process is also essential to avoid blurry images. However, all burst SR methods introduced in Sec. 2.1 reconstruct SR images in a deterministic manner, which is prone to be blurry SR images.

2.2. SISR with Diffusion Models

The stochasticity of diffusion models allows us to reduce blur in reconstructed images. As with other image enhancement and restoration tasks [5, 7, 8, 23, 32, 33], an input degraded image (i.e., LR image in the SR task) is used for conditioning the diffusion model for SISR. For this conditioning, in SR3 [37], the LR image is upscaled by Bicubic interpolation and concatenated with images passing through diffusion steps. In LDMs [36], features extracted from the LR image are fed into the middle layers of U-Net in the reverse process through the cross attention mechanism. As well as the features extracted from the LR image, the number of the diffusion steps is also used for step-aware conditioning in Stable SR [43]. ResShift [49] employs the residuals between high-quality and low-quality images to balance efficiency and reconstruction quality. CCDF [6] further reduces the number of diffusion steps by better initialization than simple upscaling.

Unlike these SISR methods [6, 36, 37, 43, 49], this paper proposes the diffusion model conditioned with burst LR features.

2.3. Burst Super-Resolution with Diffusion Models

For multi-frame SR, such as video SR [13, 17, 18, 34] and burst SR, spatial alignment between frames is essen-

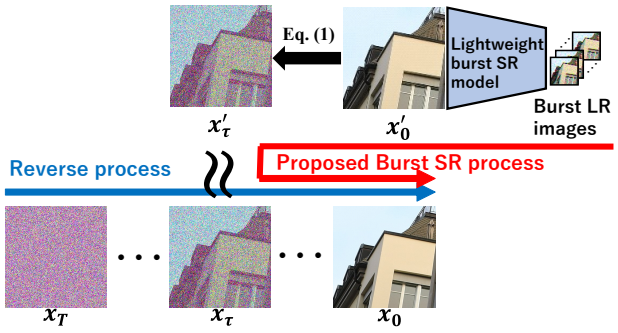


Figure 2. Reverse process from the intermediate step for the early-step skip. In BSRD, the initial burst SR image $\boldsymbol{x}'_0$ is appropriately noised by Eq. (1) and fed into the diffusion model from τ-th step to skip the early diffusion steps between T and τ.

Figure 3. Examples reconstructed by EDM. Images in the lower are reconstructed by EDM conditioned by those in the upper.

tial. Many burst SR methods [2, 3, 11, 12, 30] implement the alignment process in the feature domain because of its superiority compared to the image domain. In addition, the effectiveness of the hierarchical alignment is validated in many vision tasks [26, 27, 47, 51]. For burst SR using diffusion models also, BSRD [42] employs the hierarchical feature alignment process. As one of such hierarchical feature alignment and fusion frameworks, Burstormer [12] is employed in BSRD.

Burst Feature Conditioning for Reconstruction with Diffusion Model In BSRD, the feature map obtained from LR images, as described above and indicated by "Burst Image Encoder" in Fig. 1, is used to condition the diffusion model to reconstruct the SR image that fits with burst LR images. Since BSRD is implemented with U-Net, the feature map is rescaled to the xy dimensions of these hierarchical layers by Bicubic interpolation to smoothly condition all hierarchical layers in U-Net. This conditioning is achieved through Spatial Feature Transformation [44].

The aforementioned conditioning process is performed in all steps in the reverse process. This reverse process is regarded as the reconstruction process in BSRD.

Efficient and High-quality SR Reconstruction by the Reverse Process from Intermediate Steps To suppress the computational cost of the diffusion model, BSRD takes an initial burst SR image instead of random noise. To start the reverse process from the initial burst SR image, it is fed into an intermediate step, instead of T-th step, following SDEdit [31]. This intermediate step is denoted by τ. With this reverse process from the intermediate step, the number of execution steps is reduced, resulting in reducing computational costs. Furthermore, the reverse process can be trained only between τ-th and 1st steps to focus on fine details that are important for the SR task.

The aforementioned reverse process is illustrated in Fig. 2. x_t denotes the image at t-th step in the diffusion model. In BSRD, x_τ is replaced by x'_τ generated from the initial burst SR image denoted by x'_0. x'_0 can be provided by any burst SR method, "Lightweight burst SR model" in Fig. 2. While x'_0 corresponds to x_0 with no diffusion noise, x'_0 may be more blurred than x_0 because x'_0 is reconstructed in a deterministic manner. Furthermore, x'_τ computed from x'_0 must contain the diffusion noise included in x_τ in order to replace x_τ with x'_τ. We assume that x_τ can be approximated by x'_τ by giving the diffusion noise at τ-th step to x'_0. That is, small differences caused by the blur between x_0 and x'_0 can be drowned out by the diffusion noise if τ-th step is sufficiently apart from 0-th step. x'_τ is computed from x'_0 as follows:

$$x'_\tau = \sqrt{\bar{\alpha}_\tau} x'_0 + \sqrt{1-\bar{\alpha}_\tau}\epsilon, \quad (1)$$

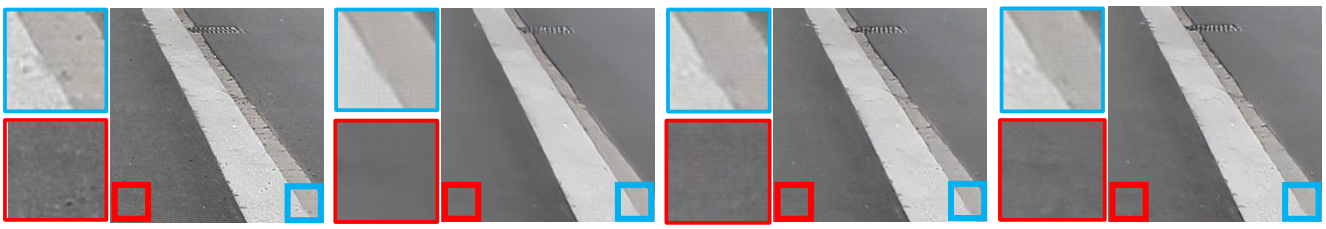
$$\bar{\alpha}_\tau = \prod_{s=1}^{\tau} \alpha_s, \quad (2)$$

$$\alpha_t = 1 - \beta_t, \quad (3)$$

where ϵ denotes the zero-mean Gaussian noise with $\sigma = 1$, and β_t is a constant between 0 and 1 that controls the noise level at t-th step.

3. E-BSRD: Efficient BSRD with One-step Diffusion

While BSRD introduced in Sec. 2.3 reduces diffusion steps used for SR reconstruction while improving the SR quality, in particular, the perceptual quality, its computational cost is still high for end users. To further reduce the diffusion steps to one step in our proposed method based on BSRD, Sec. 3 proposes to integrate the two schemes for diffusion-step reduction, introduced in Sec. 3.1. In our proposed method described in Sec. 3.2, these two schemes are optimized for BSRD so that their advantages compensate for the other's disadvantages for efficient one-step diffusion.



(a) Ground truth (b) $\tau = 1$ (c) $\tau = 5$ (d) $\tau = 40$

Figure 4. Examples of SR images reconstructed with different τ in BSRD using EDM on the BurstSR dataset.

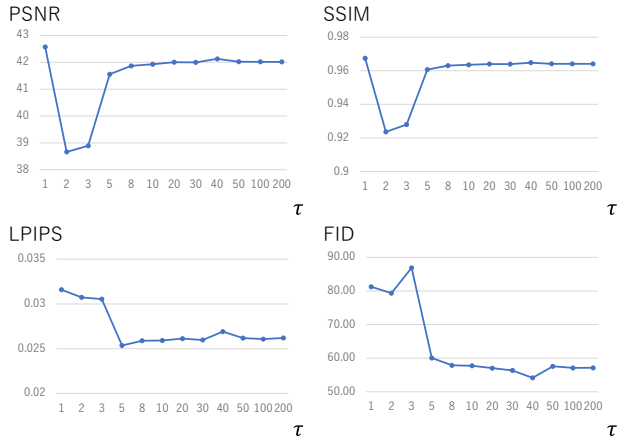


Figure 5. Effect of the number of diffusion steps, τ, in BSRD using EDM on the BurstSR dataset. The vertical and horizontal axes indicate the metric scores and τ (logarithmic scale).

3.1. Diffusion-step Reduction Schemes

EDM: High-order ODE-based Stochastic Sampler The first scheme for diffusion-step reduction is an improved sampling process. Our choice is EDM [22], in which a higher-order differential equation is used. EDM uses a second-order solver, which provides a good tradeoff between accuracy and efficiency, while many prior methods rely on Euler's method. In EDM, the preconditioning and training processes are also tweaked. EDM reduces the diffusion steps while maintaining data generation fidelity.

However, since EDM is tweaked for image generation from random noise, it is not appropriate to directly apply it to image restoration and enhancement, including burst SR, with image conditioning. Several examples reconstructed from such random noise are shown in Fig. 3. The images used for conditioning (shown in the upper of Fig. 3) and the image reconstructed with these conditioning inputs are different in terms of color and textures. The differences in color and textures are caused mainly during the early and late diffusion steps in the reverse process, respectively, as validated in [4].

CM: One-step Diffusion with Knowledge Distillation The second scheme for diffusion-step reduction is to reduce the maximum diffusion steps by distilling a large diffusion model as a teacher model into a small student diffusion model. The distilled knowledge about how to reach clean data from noisy data allows us to directly map noise to clean data without iterative diffusion steps, as proposed in the consistency model [40], CM.

However, the image generation quality of CM is clearly degraded from its original teacher model if image generation is achieved from random noise, as verified in [40]. In [40], an iterative process is also proposed for further improving the data generation quality, it is still insufficient to maintain the quality if the number of iterative steps (denoted by T_{CM}) is smaller.

3.2. E-BSRD: One-step Diffusion for Efficient Burst SR

Initial Burst SR Image for EDM While EDM can reduce diffusion steps, denoising from random noise is insufficient regarding SR quality, as validated in Fig. 3. In addition, denoising from random noise is inefficient for image restoration and enhancement tasks such as SR, as mentioned in Sec. 2.3. Therefore, as with DDPM used in BSRD, EDM begins the reverse process from an initial burst SR image by the early-step skipping process in our proposed method.

While many parameters in the preconditioning and training processes are tweaked in EDM, we empirically found that the most important one for the early-step skipping process used in BSRD (which is shown in Fig. 2) is the maximum amount of noise, $\sigma_{\max}$. This is because the default value of $\sigma_{\max}$ is proposed in the original EDM paper not for the early-step skipping process using an initial image but for data generation from random noise. The effect of $\sigma_{\max}$ is empirically verified later (in Table 2 and Table 4).

In addition to $\sigma_{\max}$, the number of diffusion steps is also crucial. $T = 1000$ and $\tau = 100$ are different for DDPM used in BSRD [42] because BSRD pretrains the diffusion model with T steps and then finetune it only from τ-th step to 1st step. For EDM, on the other hand, this pretraining may not be required because of the better convergence ability of EDM.

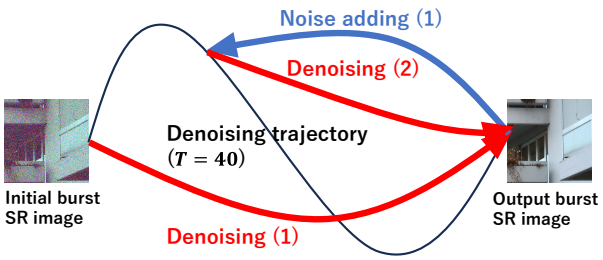


Figure 6. E-BSRD using CM. This figure shows an example with $T_{CM} = 2$. The figure in parentheses denotes the execution sequence of each process.

In our preliminary experiments, $\tau = 40$ diffusion steps of EDM allow us to maintain the high burst SR quality: given the initial burst SR image reconstructed by Burstormer [12], PSNR = 40.74 and 42.00 in "BSRD with $T = 1000$ and $\tau = 100$ steps of DDPM" and "BSRD with $\tau = 40$ steps of EDM," respectively, on the SyntheticBurst dataset [1] (i.e., SR from 32×32 pixels to 256×256 pixels).

To further verify (1) whether or not the pretraining with T steps is required for EDM and (2) if the pretraining is not required, how many steps (i.e., τ) are required for sufficient SR quality with EDM, the qualitative and quantitative effects of varying τ in BSRD using EDM are verified in Fig. 4 and Fig. 5, respectively, while $T = 40$ from random noise is proposed in the original EDM [22]. Note that all results shown in Fig. 4 and Fig. 5 are obtained with BSRD trained only with τ steps from an initial burst SR image given by Burstormer (i.e., without pretraining using T steps). In (a) the ground truth HR image of Fig. 4, we can see texture patterns on the road. In (b), on the other hand, such texture patterns almost disappear (i.e., oversmoothed). This is because only one diffusion step is insufficient to reconstruct such texture patterns, which are not observed in the initial burst SR image oversmoothed due to training using a deterministic image reconstruction error such as L1 and L2 losses. The texture patterns can be reconstructed as τ increases. In Fig. 5, we can see that image distortion-based scores (i.e., PSNR and SSIM) significantly decrease in the second and third steps, while they increase around five diffusion steps and are almost saturated. On the other hand, FID decreases gradually until around the 40th step. In our experiments, therefore, the default value of τ in EDM is 40.

Compensating Integration of EDM and Consistency Model with Early-step Skipping While EDM can reduce the diffusion steps, the inference cost decreases only almost in proportion to the number of the diffusion steps, τ. Therefore, the reduction rate only with EDM is insufficient: 37.1 sec (BSRD with 100 steps of DDPM) to 19.4 sec (BSRD with 40 steps of EDM).

Table 1. Comparison of runtime represented by secs/image.

Method	Steps	Runtime
Burstormer	NA	0.08
BIPNet	NA	0.40
BSRD	$\tau = 100$	27.2
E-BSRD w/o CM	$\tau = 40$	9.00
E-BSRD	$T_{CM} = 1$	0.44
E-BSRD	$T_{CM} = 3$	0.69
E-BSRD	$T_{CM} = 10$	1.80

On the other hand, CM can further reduce the diffusion steps, while its data generation quality sometimes declines significantly if the number of the diffusion steps is reduced to one for real-time use. Let BSRD-CM be CM distilled from BSRD modified to reconstruct the burst SR image not from an initial burst SR image but from random noise. In our preliminary experiments, for example, the mean PSNR scores of BSRD with 100 steps and BSRD-CM are 40.74 and 27.53 on the SyntheticBurst dataset [1] (i.e., burst SR from 32×32 pixels to 256×256 pixels).

However, the aforementioned problem is avoided by distillating BSRD, which reconstructs the burst SR image not from random noise but from an initial burst SR image. This BSRD is used as the teacher model for CM to achieve one-step diffusion. The diffusion model distilled from BSRD by using CM is called Efficient BSRD, E-BSRD in short.

E-BSRD using CM is illustrated in Fig. 6. CM allows for iterative steps to further improve the data generation quality, as introduced in Sec. 3.1. The number of the iterative steps is $T_{CM} = 2$ in Fig. 6. By iteratively reducing the uncertainty of the result of CM, the data generation quality can be improved.

4. Experiments

4.1. Details

Evaluation Metrics As standard evaluation metrics, LPIPS [50], FID [20], PSNR, and SSIM [46] are used. While PSNR and SSIM are for image distortion-based evaluation, LPIPS and FID are proposed for perceptual quality evaluation. Lower scores are better in LPIPS and FID, while higher scores are better in PSNR and SSIM.

Training details and Parameters We follow all training details and parameters of the original EDM [22].

Datasets The SyntheticBurst and BurstSR datasets are used [1].

In the SyntheticBurst dataset, each sRGB HR image is degraded to its LR image. This degradation process, including ISP, follows the one proposed in [1] as follows. Burst

Table 2. Comparison of burst SR images reconstructed with varying $\sigma_{\max}$ on the SyntheticBurst dataset. "+BS" and "+BIP" mean that initial burst SR images are given by Burstormer and BIPNet, respectively.

Method	$\sigma_{\max}$	PSNR↑	SSIM↑	LPIPS↓	FID↓
Burstormer	NA	43.21	0.971	0.030	70.76
BSRD (+BS)	NA	40.74	0.952	**0.027**	50.61
E-BSRD (+BS)	80	32.91	0.863	0.101	176.6
E-BSRD (+BS)	0.2	38.66	0.933	0.040	110.3
E-BSRD (+BS)	0.08	40.66	0.954	0.030	83.03
E-BSRD (+BS)	0.05	41.58	0.961	0.028	68.94
E-BSRD (+BS)	0.03	42.00	0.964	0.026	56.71
E-BSRD (+BS)	0.01	42.68	0.968	0.028	**51.90**
E-BSRD (+BS)	0.005	**42.95**	**0.970**	0.029	59.05
BIPNet	NA	42.61	0.968	0.032	64.17
BSRD (+BIP)	NA	40.53	0.951	**0.029**	50.26
E-BSRD (+BIP)	80	32.93	0.861	0.102	173.9
E-BSRD (+BIP)	0.2	38.47	0.932	0.042	112.7
E-BSRD (+BIP)	0.08	40.37	0.952	0.032	84.65
E-BSRD (+BIP)	0.05	41.22	0.959	0.030	68.43
E-BSRD (+BIP)	0.03	41.58	0.962	0.028	56.41
E-BSRD (+BIP)	0.01	42.17	0.966	0.030	**50.75**
E-BSRD (+BIP)	0.005	**42.39**	**0.967**	0.031	56.24

Table 3. Comparison between E-BSRD w/o CM and E-BSRD w/ CM on the SyntheticBurst dataset.

Methods	Steps	PSNR↑	SSIM↑	LPIPS↓	FID↓
E-BSRD w/o CM (+BS)	$\tau = 40$	42.00	0.964	0.026	56.71
E-BSRD (+BS)	$T_{CM} = 1$	42.00	0.964	0.026	57.39
E-BSRD (+BS)	$T_{CM} = 2$	42.02	0.964	0.026	58.36
E-BSRD (+BS)	$T_{CM} = 3$	**42.01**	0.964	0.026	**57.32**
E-BSRD (+BS)	$T_{CM} = 4$	41.99	0.964	0.026	58.13
E-BSRD (+BS)	$T_{CM} = 5$	41.89	0.963	0.026	60.03
E-BSRD (+BS)	$T_{CM} = 11$	41.65	**0.962**	0.026	62.47
E-BSRD w/o CM (+BIP)	$\tau = 40$	**41.58**	0.962	0.028	56.41
E-BSRD (+BIP)	$T_{CM} = 1$	**41.58**	0.962	0.028	58.21
E-BSRD (+BIP)	$T_{CM} = 2$	41.59	0.962	0.028	**58.06**
E-BSRD (+BIP)	$T_{CM} = 3$	**41.58**	0.962	0.028	58.07
E-BSRD (+BIP)	$T_{CM} = 4$	41.56	0.962	0.028	59.08
E-BSRD (+BIP)	$T_{CM} = 5$	41.48	0.961	0.028	59.44
E-BSRD (+BIP)	$T_{CM} = 11$	41.26	0.959	0.028	63.18

LR images, except for the reference frame, are randomly (1) translated up to 24 pixels along x and y axes and (2) rotated up to one degree. The SyntheticBurst dataset has 46,839 training images and 300 test images.

In the BurstSR dataset, both LR and HR images are real images. The LR images are captured using a burst shot mode of Samsung Galaxy S8. These LR images are subtly different from each other due to handshake effects. Each HR image is captured by CANON 5D Mark IV. This dataset has 5,405 training images and 882 test images.

In both of these two datasets, the center region of each original HR image is cropped out to 256×256 pixels.

Table 4. Comparison of burst SR images reconstructed with varying $\sigma_{\max}$ on the BurstSR dataset.

Methods	$\sigma_{\max}$	PSNR↑	SSIM↑	LPIPS↓	FID↓
Burstormer	NA	50.49	0.985	0.056	74.21
BSRD (+BS)	NA	49.45	0.915	0.050	48.51
E-BSRD (+BS)	0.1	48.36	0.979	0.057	85.83
E-BSRD (+BS)	0.08	48.58	0.908	0.056	81.95
E-BSRD (+BS)	0.05	49.21	0.982	**0.055**	78.92
E-BSRD (+BS)	0.03	49.21	0.982	**0.055**	63.70
E-BSRD (+BS)	0.01	49.88	**0.984**	**0.055**	65.03
E-BSRD (+BS)	0.005	**50.14**	**0.984**	**0.055**	**62.49**
BIPNet	NA	51.55	0.986	0.050	75.43
BSRD (+BIP)	NA	50.54	0.984	0.047	43.34
E-BSRD (+BIP)	0.1	48.94	0.980	0.050	82.10
E-BSRD (+BIP)	0.08	49.17	0.981	0.050	81.95
E-BSRD (+BIP)	0.05	49.97	0.983	0.049	73.82
E-BSRD (+BIP)	0.03	49.96	0.984	**0.048**	60.37
E-BSRD (+BIP)	0.01	50.73	**0.985**	**0.048**	**57.28**
E-BSRD (+BIP)	0.005	**51.07**	**0.985**	0.049	58.00

Table 5. Comparison between E-BSRD w/o CM and E-BSRD w/ CM on the BurstSR dataset.

Methods	Steps	PSNR↑	SSIM↑	LPIPS↓	FID↓
E-BSRD w/o CM (+BS)	$\tau = 40$	49.21	0.982	0.055	63.70
E-BSRD (+BS)	$T_{CM} = 1$	47.64	**0.979**	0.060	77.30
E-BSRD (+BS)	$T_{CM} = 2$	47.66	0.978	0.060	77.40
E-BSRD (+BS)	$T_{CM} = 3$	47.64	**0.979**	0.060	77.01
E-BSRD (+BS)	$T_{CM} = 4$	48.05	**0.979**	0.058	**76.68**
E-BSRD (+BS)	$T_{CM} = 5$	**48.09**	**0.979**	**0.057**	77.00
E-BSRD (+BS)	$T_{CM} = 11$	47.38	0.976	0.058	79.47
E-BSRD w/o CM (+BIP)	$\tau = 40$	49.96	0.984	0.048	60.37
E-BSRD (+BIP)	$T_{CM} = 1$	48.29	0.980	0.054	71.80
E-BSRD (+BIP)	$T_{CM} = 2$	48.31	0.980	0.053	72.59
E-BSRD (+BIP)	$T_{CM} = 3$	48.29	0.980	0.054	71.37
E-BSRD (+BIP)	$T_{CM} = 4$	48.77	**0.981**	0.051	**70.43**
E-BSRD (+BIP)	$T_{CM} = 5$	**48.78**	**0.981**	**0.050**	72.01
E-BSRD (+BIP)	$T_{CM} = 11$	48.03	0.978	**0.050**	75.26

Burst SR Condition In all experiments, the dimensions of LR and HR images are $32 \times 32 \times 4$ RAW images and $256 \times 256 \times 3$ RGB images, respectively. The number of a set of burst LR images is eight. For comparison, BIPNet [11], Burstormer [12], and BSRD [42] are evaluated. The weights of all these burst SR models are trained by the authors' codes.

Initial burst SR images given to BSRD and our method (E-BSRD) are provided by BIPNet and Burstormer.

While the number of diffusion steps of EDM in E-BSRD is $\tau = 40$, $T = 1000$ and $\tau = 100$ for DDPM in BSRD.

4.2. Runtime

Table 1 validates the efficient computational cost of E-BSRD. As mentioned in Sec. 3.2, the runtime decreases

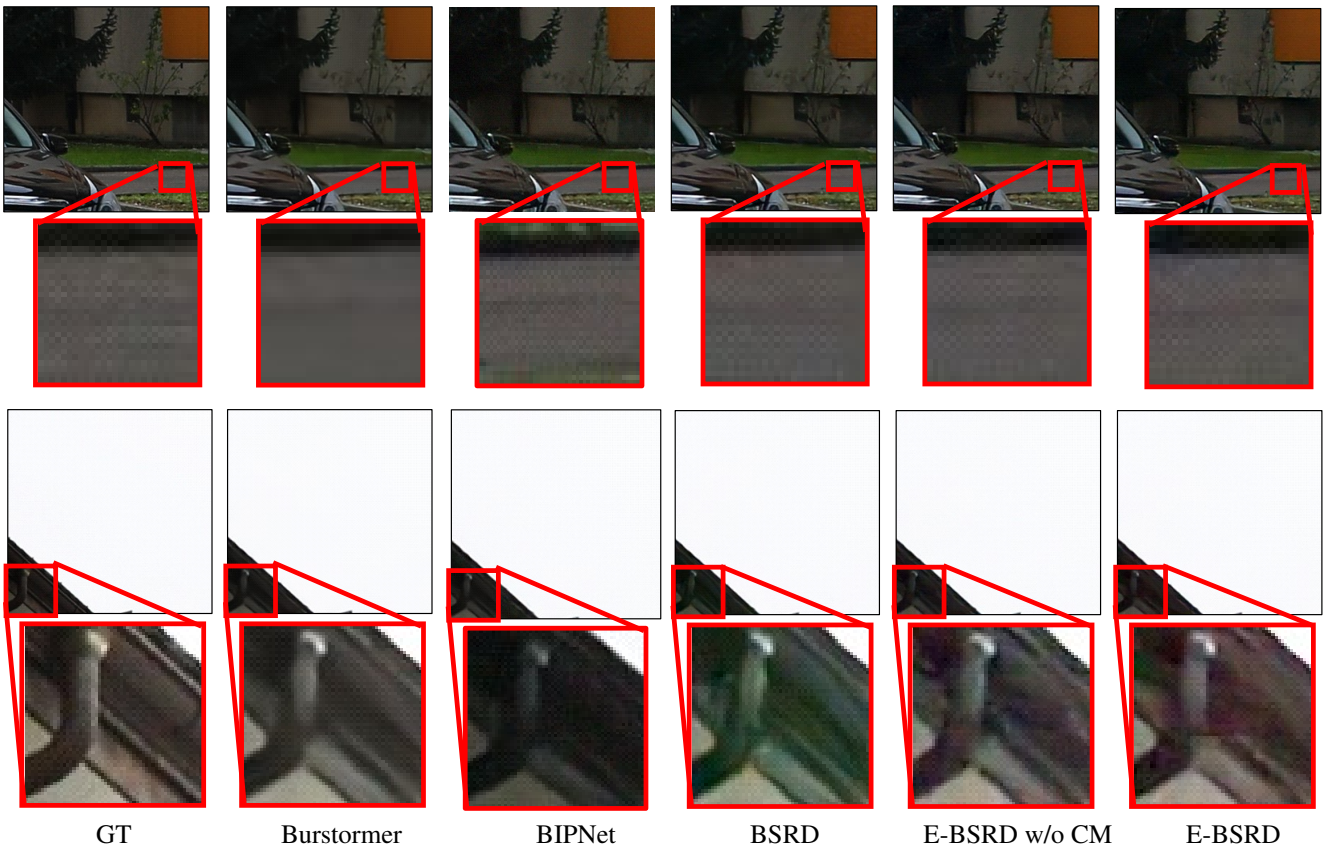


Figure 7. Visual results on the SyntheticBurst dataset. Best viewed with zoom-in. While clear textures, such as the boundary lines of texts, are generally used as the visual results of SR, smoother and finer textures are appropriate for validating the effectiveness of our method because our method emphasizes such smooth and fine textures.

almost in proportion to the diffusion steps from BSRD to E-BSRD w/o CM (i.e., from 27.2 to 9.00). In E-BSRD (i.e., E-BSRD with CM), the runtime is further significantly decreased by reducing the number of steps (i.e., $T_{CM} = \{1, 3, 10\}$). For example, the runtime decreases from 27.2 (BSRD with $\tau = 100$) to 0.44 (E-BSRD with $T_{CM} = 1$). That is, the runtime is reduced to 1.6 % of BSRD.

4.3. Results: SyntheticBurst Dataset

Effect of $\sigma_{\max}$ Table 2 shows the SR scores obtained with the varying amount of noise given to the initial burst SR image. Independent of the initial burst SR images, the SR qualities of different conditions have the same trend.

- The image distortion quality (i.e., PSNR and SSIM) is the highest in the original initial burst SR image reconstructed in a deterministic manner (i.e., without the diffusion model). Both PSNR and SSIM improve as $\sigma_{\max}$ decreases in our E-BSRD. This is natural because the effect of the diffusion model is reduced as $\sigma_{\max}$ decreases.
- Compared with Burstormer, which is a deterministic model, all results obtained by diffusion models are inferior in terms of the image distortion-based scores (PSNR and SSIM). On the other hand, the advantage of diffusion models [42] is to improve the perceptual quality (i.e., LPIPS and FID), as validated in our experiments as well as in [42]; i.e., LPIPS and FID of BSRD are the second-best and the best, respectively. While our E-BSRD reduces the runtime significantly as shown in Table 1, it maintains the advantage of BSRD, i.e., the best LPIPS and the second-best FID are observed in E-BSRD.

Effect of Iterations in CM Table 3 shows the SR scores obtained with the varying number of iterations of CM (i.e., T_{CM}). We can see that the SR quality is almost unchanged in all metrics, except for larger T_{CM}, such as $T_{CM} = 11$. Therefore, $T_{CM} = 1$ is sufficient for our E-BSRD.

Why is the data generation quality not changed by the iterations, unlike experiments in the original CM paper [40]? This may be because the difference between the initial data (i.e., initial burst SR image) and the final output is small, while the random noise is fed into CM in the original work.

The reconstructed burst SR images are shown in Fig. 7. For reference, the ground truth HR images are also shown. In the upper example, Burstormer oversmoothes the SR image, as we can see in the zoom-in region. Such over-

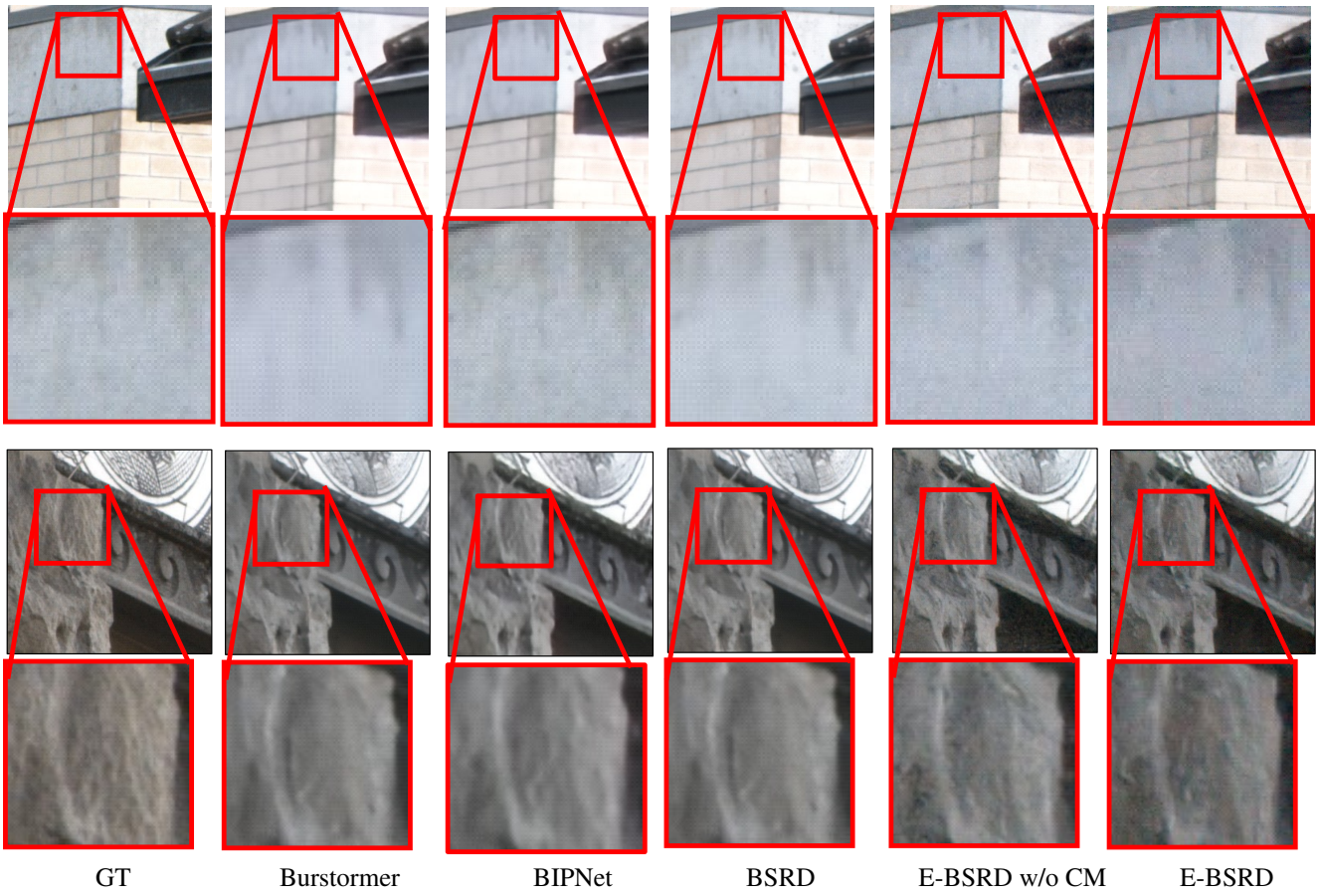


Figure 8. Visual results on the BurstSR dataset. Best viewed with zoom-in.

smoothed results are shown as typical bad examples of deterministic burst SR models in both Fig. 7 and Fig. 8. Another difficulty can be validated in the lower example. It is known that the global color structure can be changed by image generation using diffusion models [4]. In the results of BSRD and E-BSRD w/o CM, we can see that the global color structure is changed to greenish. On the other hand, E-BSRD can maintain the global color structure.

4.4. Results: BurstSR dataset

Table 4 shows the quantitative results obtained by varying σ_{max} in our E-BSRD. The overall trend of the performance measures in Table 4 is similar to Table 2, which shows the results on the SyntheticBurst dataset, as follows. The best scores in image distortion-based metrics (i.e., PSNR and SSIM) and perceptual metrics (i.e., LPIPS and FID) are acquired in the original deterministic burst SR methods (i.e., Bursotrmer and BIPNet) and BSRD, respectively. However, the gaps from these methods and our E-BSRD are not significant, while its runtime is significantly faster than these methods, as shown in Table 1. For example, several best and second-best scores are observed in E-BSRD with small σ_{max} both in E-BSRD (+BS) and E-BSRD (+BIP).

Table 5 shows the results with varying iterative steps, T_{CM}, in our E-BSRD. As with Table 4, Table 5 also shows the overall trend similar to Table 3, which shows the results on the SyntheticBurst dataset. While better scores are obtained by E-BSRD without CM, the performance is not decreased if T_{CM} is small in our E-BSRD with CM.

The examples of reconstructed burst SR images are shown in Fig. 8. As with the upper example shown in Fig. 7, we can see that Burstormer and BIPNet, each of which is a deterministic burst SR method, oversmoothes fine textures. Finer textures are reconstructed by E-BSRD.

5. Conclusion

This paper proposed a fast burst SR method using few-step diffusion. Even if the number of diffusion steps is reduced to one, the SR quality is comparable to the method using diffusion with many steps (e.g., 100 steps) [42]. Our proposed method, called E-BSRD, with one-step diffusion reduces the runtime to around 1.6 % of its baseline [42]. E-BSRD is enhanced with a high-order ODE (i.e., second-order ODE) and a teacher-student distillation approach.

This work is supported by JSPS KAKENHI 22H03618.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, 2021. 1, 5

[2] Goutam Bhat et al. NTIRE 2021 challenge on burst super-resolution: Methods and results. In *CVPRW*, 2021. 3

[3] Goutam Bhat et al. NTIRE 2022 burst super-resolution challenge. In *CVPRW*, 2022. 1, 3

[4] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, 2022. 4, 8

[5] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *NeurIPS*, 2022. 2

[6] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, 2022. 1, 2

[7] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *CVPR*, 2023. 2

[8] Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023. 2

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 2

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, 2016. 1

[11] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *CVPR*, 2022. 1, 3, 6

[12] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burstormer: Burst image restoration and enhancement transformer. In *CVPR*, 2023. 1, 3, 5, 6

[13] Dario Fuoli et al. AIM 2020 challenge on video extreme super-resolution: Methods and results. In *ECCV Workshop*, 2020. 2

[14] Shuhang Gu et al. AIM 2019 challenge on image extreme super-resolution: Methods and results. In *ICCVW*, 2019. 1

[15] Shi Guo, Xi Yang, Jianqi Ma, Gaofeng Ren, and Lei Zhang. A differentiable two-stage alignment scheme for burst image reconstruction with large shift. In *CVPR*, 2022. 1, 2

[16] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 1

[17] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, 2019. 2

[18] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Space-time-aware multi-resolution video enhancement. In *CVPR*, 2020. 2

[19] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projectinetworks for single image super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12): 4323–4337, 2021. 1

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 5

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2

[22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 1, 2, 4, 5

[23] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022. 2

[24] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *ICLR*, 2024. 1, 2

[25] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 3

[28] 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 *CVPR Workshops*, 2021. 1, 2

[29] Ziwei Luo, Youwei Li, Shen Cheng, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. BSRT: improving burst super-resolution with swin transformer and flow-guided deformable alignment. In *CVPR Workshops*, 2022. 2

[30] Nancy Mehta, Akshay Dudhane, Subrahmanyam Murala, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Adaptive feature consolidation network for burst super-resolution. In *CVPRW*, 2022. 1, 3

[31] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 3

[32] Xiangming Meng and Yoshiyuki Kabashima. Diffusion model based posterior sampling for noisy linear inverse problems. *CoRR*, abs/2211.12343, 2022. 2

[33] Naoki Murata, Koichi Saito, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Gibbsddrm: A partially collapsed gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In *ICML*, 2023. 2

[34] Seungjun Nah et al. NTIRE 2019 challenge on video super-resolution: Methods and results. In *CVPR Workshop*, 2019. 2

[35] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[37] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4713–4726, 2023. 2

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[40] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023. 1, 2, 4, 7

[41] Radu Timofte et al. Ntire 2018 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2018. 1

[42] Kyotaro Tokoro, Kazutoshi Akita, and Norimichi Ukita. Burst super-resolution with diffusion models for improving perceptual quality. In *IJCNN*, 2024. 1, 2, 3, 4, 6, 7, 8

[43] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *CoRR*, abs/2305.07015, 2023. 2

[44] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 3

[45] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. In *CVPR*, 2024. 1, 2

[46] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 5

[47] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H2FA R-CNN: holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *CVPR*, 2022. 3

[48] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2023. 1, 2

[49] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(1):116–130, 2025. 2

[50] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[51] Heyu Zhou, Weizhi Nie, Wenhui Li, Dan Song, and An-An Liu. Hierarchical instance feature alignment for 2d image-based 3d shape retrieval. In *IJCAI*, 2020. 3