BUFF: Bayesian Uncertainty Guided Diffusion Probabilistic Model for Single Image Super-Resolution

Zihao He[1], Shengchuan Zhang[1], Runze Hu[2], Yunhang Shen[3], Yan Zhang[1,*]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China.

[2]Department of Information Science, Tsinghua university.

[3]Tencent Youtu Lab, Shanghai 200233, China.

36920221153081@stu.xmu.edu.cn, {zsc_2016, bzhy986}@xmu.edu.cn

Abstract

Super-resolution (SR) techniques are critical for enhancing image quality, particularly in scenarios where high-resolution imagery is essential yet limited by hardware constraints. Existing diffusion models for SR have relied predominantly on Gaussian models for noise generation, which often fall short when dealing with the complex and variable texture inherent in natural scenes. To address these deficiencies, we introduce the Bayesian Uncertainty Guided Diffusion Probabilistic Model (BUFF). BUFF distinguishes itself by incorporating a Bayesian network to generate high-resolution uncertainty masks. These masks guide the diffusion process, allowing for the adjustment of noise intensity in a manner that is both context-aware and adaptive. This novel approach not only enhances the fidelity of super-resolved images to their original high-resolution counterparts but also significantly mitigates artifacts and blurring in areas characterized by complex textures and fine details. The model demonstrates exceptional robustness against complex noise patterns and showcases superior adaptability in handling textures and edges within images. Empirical evidence, supported by visual results, illustrates the model's robustness, especially in challenging scenarios, and its effectiveness in addressing common SR issues such as blurring. Experimental evaluations conducted on the DIV2K dataset reveal that BUFF achieves a notable improvement, with a +0.61 increase compared to baseline in SSIM on BSD100, surpassing traditional diffusion approaches by an average additional +0.20dB PSNR gain. These findings underscore the potential of Bayesian methods in enhancing diffusion processes for SR, paving the way for future advancements in the field.

Introduction

Image super-resolution (SR), the art and science of enhancing the resolution of images, has undergone a remarkable evolution over the past decades. From its inception, where basic interpolation techniques like bicubic and Lanczos resampling were the norm, to the advent of deep learning, which has radically transformed the landscape, SR has been pivotal in fields ranging from satellite imaging and surveillance to medical imaging and entertainment.

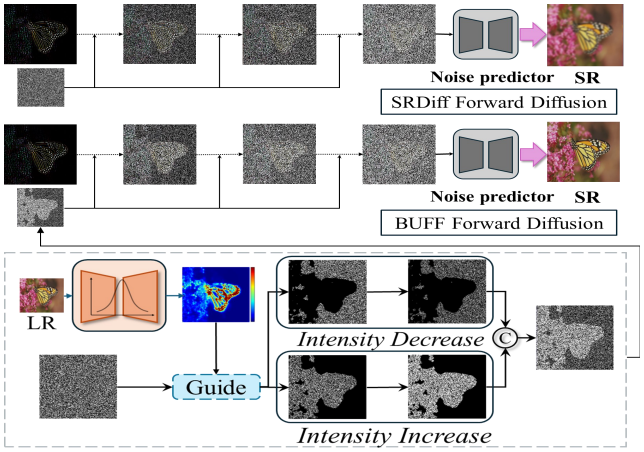

Figure 1: Comparison of the noise addition strategies in the forward diffusion phase between SRDiff(Li et al. 2022a) and BUFF approaches, with BUFF utilizing a Bayesian model to generate uncertainty masks that guide noise intensity adjustments across different regions of the image for SR results.

The breakthrough comes with the introduction of Convolutional Neural Networks (CNNs) into the SR domain. Early models such as SRCNN (Dong et al. 2015) opened the floodgates to a new era of deep learning-based SR, offering significant improvements over traditional methods. These CNN-based models (Dai et al. 2019; Zhang et al. 2018; Lim et al. 2017) excel in mapping low-resolution (LR) to high-resolution (HR) images, learning a direct end-to-end transformation. Despite their success, they often fall short in capturing the perceptual nuances of images (Lepcha et al. 2023; Bashir et al. 2021; Wang, Chen, and Hoi 2020), leading to outcomes that, while technically accurate, lack the rich textures and details that make images lifelike (Chauhan et al. 2023; He and Zhang 2024). Transformers, known for their prowess in handling sequential data, make their mark on image SR with their ability to capture long-range dependencies within an image. By focusing on global information, Transformer-based SR models such as (Chen et al. 2023b,a; Liang et al. 2021) offer a promising approach to reconstructing images. Yet, their computational intensity and substantial

*Corresponding Author.

memory footprint make them less viable for real-time applications, limiting their widespread adoption (Lu et al. 2022). Generative Adversarial Networks (GANs) shifted the focus towards generating images that not only are high resolution but also perceptually convincing (Tian et al. 2022; Wang, Chen, and Hoi 2020). By employing a dual network system, one to generate images and another to critique them, GAN-based SR models like ESRGAN(Wang et al. 2018),BebyGAN(Li et al. 2022b) and Real-ESRGAN(Wang et al. 2021b) have pushed the boundaries of what's possible in terms of image quality. However, the adversarial nature of these models can lead to instability during training, sometimes resulting in artifacts or overly stylized images that detract from the realism of the output (Wang, Bayram, and Sertel 2022).

Diffusion-based super-resolution (SR) models, such as those referenced in (Li et al. 2022a; Niu et al. 2024; Shang et al. 2024; Wang et al. 2024; Gao et al. 2023), are at the forefront of image reconstruction, showcasing impressive capabilities in enhancing image quality. These models simulate the natural diffusion process through iterative noise introduction and attenuation to refine image details. However, despite their effectiveness in rendering rich textures, they assume noise is independent and identically distributed (Li et al. 2022a; Niu et al. 2024), potentially overlooking the distinct data distributions across different image regions. This can compromise structural integrity, leading to inconsistent textures and noticeable artifacts. While (Wang et al. 2024) attempts to address this by modulating Gaussian noise with a SAM-generated mask (Kirillov et al. 2023), it still struggles with precise pixel restoration, particularly along edges. Additionally, managing noise in diffusion processes, though innovative, introduces significant computational burdens and parameter sensitivity, posing challenges for practical deployment.

In this context, Bayesian models offer a promising complementary perspective. Known for their role in probabilistic image generation, Bayesian methods (Wang and Yeung 2020; van de Schoot et al. 2021) facilitate image creation and enable the quantification of uncertainty in predictions (Gao and Zhuang 2022). Despite the potential synergy, integrating Bayesian approaches with diffusion-based generative models remains largely unexplored. While diffusion models excel at transforming noise into coherent structures, they typically lack a formalized method to evaluate or leverage the inherent uncertainty of this process. Incorporating a Bayesian framework into diffusion SR models could enhance both the fidelity and precision of synthesized images, particularly in regions of high uncertainty (Liu, Cheng, and Tan 2023; Luo et al. 2023). However, directly integrating Bayesian principles into diffusion models is challenging due to the computational intensity and instability of Bayesian training, compounded by the difficulty of achieving convergence while maintaining the accuracy of the noise predictor.

In response to these challenges,depicted in Figure.1, this paper introduces a novel framework, BUFF (Bayesian Uncertainty Guided Diffusion Model). Drawing on the structure of the (Li et al. 2022a) diffusion model, BUFF ingeniously modifies the noise distribution by embedding a Bayesian model to guide the noise prediction with pixel-level uncertainty insights. By doing so, BUFF injects structural information into the diffusion process more judiciously, which is critical for advancing the super-resolution performance of the diffusion model. Specifically, for each low-resolution (LR) image in the training set, our Bayesian model generates a corresponding uncertainty map for the high-resolution (HR) version. This map, delineating areas of high and low uncertainty, undergoes a refinement process that scales the uncertainties with designated multipliers, resulting in a modulation mask. During training, this mask directs the diffusion model's attention, specifically the noise predictor (U-Net), to areas of higher uncertainty, ensuring that these regions receive heightened focus. The modulated noise, obtained by applying the modulation mask to Gaussian noise, is then used alongside the LR image as part of the conditional inputs to the model. The masks for the training samples are pre-generated by the Bayesian model and can be reused, enhancing efficiency across training epochs. Through extensive experimentation on various standard image SR benchmarks, BUFF has been shown to surpass current diffusion-based methodologies. Our method's versatility is further demonstrated through additional experiments in tasks such as deblurring and face super-resolution, confirming the scalability of BUFF and its practicality in diverse multimedia applications.

Related Work

Uncertainty in Bayesian Deep Learning

The incorporation of uncertainty into deep learning, especially within the Bayesian framework, has gained substantial attention for its potential to enhance model robustness and interpretability (Wang and Yeung 2020). Bayesian Deep Learning(BDL) offers a principled approach to quantify uncertainty, distinguishing between aleatoric (data) and epistemic (model) uncertainties (Gawlikowski et al. 2023). Pioneering research by (van de Schoot et al. 2021) and (Wang and Yeung 2020) has shown how BDL can be applied across various tasks, including vision (Zhao et al. 2023) and natural language processing (Ruz, Henríquez, and Mascareño 2020), to improve decision-making under uncertainty. In the context of SR, understanding and modeling uncertainty can significantly enhance the quality of reconstructed images (Upadhyay et al. 2022; Gao and Zhuang 2022; Liu, Cheng, and Tan 2023). By identifying areas of high uncertainty, models can adaptively refine these regions, potentially leading to higher quality reconstructions (Marinescu, Moyer, and Golland 2020). Recent works, such as those by (Wu et al. 2023; Narnhofer et al. 2021), have explored the application of BDL in medical image reconstruction, demonstrating how uncertainty can guide the recovery of fine details while assessing the confidence in the model's predictions. This research underscores the value of incorporating uncertainty into SR models, suggesting avenues for improving SR techniques through a better understanding of where models are most and least certain in their predictions.

Diffusion-Based Super Resolution

Diffusion-based models mark an innovative paradigm in generative modeling, closely emulating the ebb and flow of noise

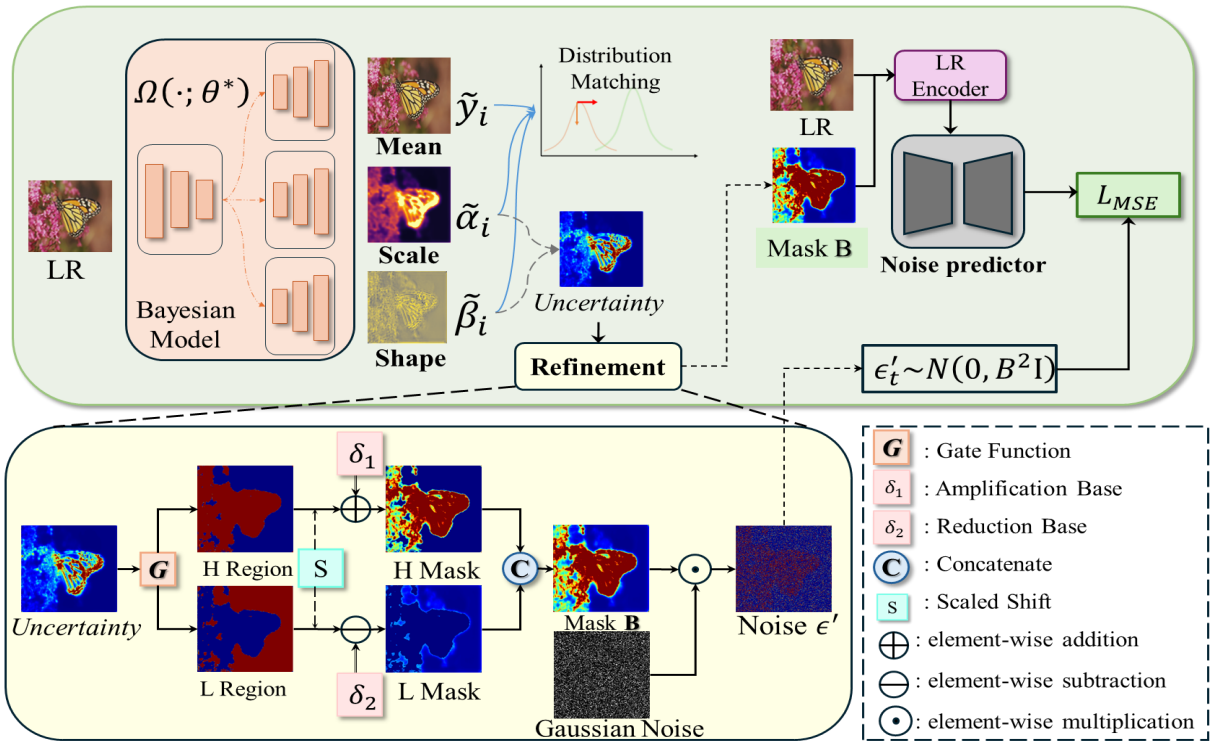

Figure 2: Process diagram showcasing the Bayesian modeling for uncertainty estimation in image super-resolution, where a Bayesian neural network refines uncertainty measures to guide noise modulation, enhancing the LR to HR reconstruction process.

addition and subtraction to fabricate or restore images. These models derive their methodology from the physical diffusion process, involving a forward phase of progressive noise application and a backward phase aiming to revert the noised data to its original form (Croitoru et al. 2023). Across various domains, diffusion methodologies, as pioneered by (San-Roman, Nachmani, and Wolf 2021; Chen 2023), stand out as a robust alternative to conventional generative techniques, facilitating the synthesis of high-fidelity images by the meticulous orchestration of noise. Distinct from GANs, which are prone to generating synthetic artifacts (Wang et al. 2018), diffusion models are celebrated for their aptitude in crafting images replete with realistic textures and finer details (Yang et al. 2023). Despite their promise, the computational demands and the intricate optimization required for specific applications, such as SR tasks, present substantial challenges. Nevertheless, recent innovations that meld diffusion models with supplementary architectures have made strides in enhancing training efficiency and elevating SR performance. Pioneering work by (Li et al. 2022a; Wang et al. 2024; Shang et al. 2024) has underscored the prowess of diffusion-based models in the SR arena, setting new benchmarks in image clarity and lifelikeness. This burgeoning research underpins a keen and expanding interest in diffusion-based strategies as a potent mechanism for image reconstruction, particularly in the realm of SR tasks.

Method

In this study, we present an innovative Bayesian-guided diffusion process specifically designed for image super-resolution (SR). This method enhances traditional diffusion-based SR models by integrating a Bayesian network's uncertainty estimates directly into the diffusion sequence, tailoring noise addition from each input image based on per-pixel uncertainty. This allows for a stochastic process that adapts to the image's inherent confidence levels, facilitating a more intelligent and dynamic reconstruction approach.

Bayesian Training and Inference for Uncertainty Estimation in SR. The key component of our model is the estimation of uncertainty for each pixel in low-resolution (LR) images. This estimation is grounded in Bayesian inference, which provides a robust mechanism for assessing confidence and managing the inherent ambiguities in the SR process. Beforehand, based on (Upadhyay et al. 2022), we discuss a model initialized from scratch (training the network from an uninitialized state rather than using pretrained weights) to tackle the target task and estimate uncertainty, denoted as $\mathbf{\Psi}s(\cdot;\zeta) : \mathbb{R}^m \rightarrow \mathbb{R}^n$, with ζ being its trainable parameters and $\mathbf{\Psi}$ refers to the set of parameters modeling the uncertainty distribution. This model aims to estimate the parameters of the output distribution $\mathcal{P}_{Y|X}$ (representing the predicted probability distribution over possible super-resolved outputs) to account for aleatoric uncertainty. For an input x_i, it generates parameters $\hat{\mathbf{y}}_i, \hat{\nu}i \ldots \hat{\rho}i$ that define $\mathcal{P}Y|X(\mathbf{y}; \hat{\mathbf{y}}_i, \hat{\nu}i \ldots \hat{\rho}i)$, optimizing these through likelihood maximization. The distribution $\mathcal{P}Y|X$ is chosen to allow

uncertainty estimation via a closed-form solution, dependent on the network's parameters.

As Figure 2 shows, we employ a Bayesian neural network, denoted as $\Omega(\cdot;\theta)$, to predict the per-pixel mean $(\hat{y}_i)$ alongside the uncertainty parameters scale $(\hat{\alpha}_i)$ and shape $(\hat{\beta}_i)$ that characterize the per-pixel uncertainty. The scale parameter $(\hat{\alpha}_i)$ indicates the expected deviation of each prediction, while the shape parameter $(\hat{\beta}_i)$ adapts the distribution's tails to the presence of outliers or other irregularities, thereby capturing the heteroscedastic nature of the data. As they are all trainable parameters($i.e.$ $\{\hat{\mathbf{y}}_i, \hat{\alpha}_i, \hat{\beta}_i\} := \Omega(\mathbf{x}_i;\theta)$,), we can describe the optimization problem as the following equation:

$$\begin{aligned}\theta^* :&= \operatorname{argmax} \prod_{i=1}^{N} \mathcal{P}_{Y|X}(\mathbf{y}_i; \{\hat{\mathbf{y}}_i, \hat{\alpha}_i, \hat{\beta}_i\}) \\ &= \operatorname{argmax} \prod_{i=1}^{N} \frac{\hat{\beta}_i}{2\hat{\alpha}_i \Gamma(\frac{1}{\hat{\beta}_i})} e^{-(|\hat{\mathbf{y}}_i - \mathbf{y}_i|/\hat{\alpha}_i)^{\hat{\beta}_i}} \\ &= \operatorname{argmin} \sum_{i=1}^{N} \left(\frac{|\hat{\mathbf{y}}_i - \mathbf{y}_i|}{\hat{\alpha}_i} \right)^{\hat{\beta}_i} - \log \frac{\hat{\beta}_i}{\hat{\alpha}_i} + \log \Gamma(\frac{1}{\hat{\beta}_i})\end{aligned} \tag{1}$$

After training the Bayesian model, we can set the the predicted variance itself as uncertainty in the prediction, $i.e.$, the uncertainty mask M_{Bayes} that we need to use for our subsequent processes:

$$M_{Bayes} = \frac{\hat{\alpha}_i^2 \Gamma(\frac{3}{\hat{\beta}_i})}{\Gamma(\frac{1}{\hat{\beta}_i})} \tag{2}$$

Refinement of Uncertainty Masks Using Nonlinear Transformation. Upon obtaining the initial Bayesian uncertainty mask M_{Bayes} , our method applies a nonlinear transformation to refine this mask, creating a more discriminating variable, denoted B. This refined mask B is critical for modulating the noise profile in the subsequent diffusion process, allowing for a tailored noise addition that is more attuned to the nuances of image content and uncertainty. This transformation hinges on a sigmoidal function, traditionally used in neural networks to introduce nonlinearity. The sigmoid function is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, mapping real-valued inputs into a (0, 1) range, providing a smooth transition between two states. In the context of our model, the sigmoid function is employed to differentiate between regions that require noise amplification versus those that need noise suppression within the super-resolution process. For each pixel indexed by i, the adjustment factor A_i is computed using the sigmoid function applied to the mask M_{Bayes} subtracted by a threshold α , and scaled to ensure a steep transition around this threshold. Mathematically, the adjustment factor is given by: $A_i = \sigma((M_{Bayes,i} - \alpha) \cdot k)$, where k is a scaling constant set to provide a sensitive response to the deviations from α, which is typically chosen as 10 to amplify the effect.

The amplification and reduction factors for each pixel, are then adjusted according to the calculated adjustment factor based on predefined base values amplification base δ_1 and reduction base δ_2, These factors ensure that areas with uncertainty levels above α undergo noise amplification, while regions below this threshold experience noise reduction. The final transformation applied to M_{Bayes} to produce the refined mask B is as follows:

$$B = \begin{cases} N_i \cdot [(\delta_1 + (A_i - 0.5) \cdot \gamma)] & \text{if } M_{Bayes,i} > \alpha \\ N_i \cdot [(\delta_2 - (0.5 - A_i) \cdot \gamma)] & \text{if } M_{Bayes,i} \leq \alpha \end{cases} \tag{3}$$

Here, N_i is the Gaussian noise at the i-th pixel, A_i is the sigmoid-transformed adjustment factor for the i-th pixel, δ is the base value for noise modulation, and γ is the amplification or reduction intensity which scales the contribution of the adjustment factor. This formulation allows for a unified base value δ to be modulated by an intensity factor γ, enhancing the model's flexibility to calibrate noise levels effectively across the image.

Integration of Uncertainty in the Diffusion Process. In our framework, we use a Bayesian refinement mask B to modulate the heteroscedastic generalized Gaussian noise used in the original SRDiff by applying B to $\boldsymbol{\epsilon}_t$. Then the sampling of x_t becomes:

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}(\boldsymbol{\epsilon}_t \odot B) \tag{4}$$

Let $\alpha_t = 1 - \beta_t$ and iteratively apply Equation 4, it comes:

$$x_t(x_0, \epsilon) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}(\epsilon \odot B), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{5}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, Our novel diffusion equation modulates noise variance at each pixel according to the corresponding uncertainty:

$$q(x_t|x_0, B) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, ((1-\bar{\alpha}_t)B^2 I)\right) \tag{6}$$

To achieve the SR image from restoration of an LR image, learning the reverse of the forward diffusion process is essential, characterized by the posterior distribution $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, B)$. However, the intractability arises due to the known marginal distributions $p(\boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t)$.This challenge is addressed by incorporating $\boldsymbol{x}_0$ into the condition. Employing Bayes' theorem, the posterior distribution $p(\boldsymbol{x}_{t1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \mathrm{B})$ can be formulated as:

$$\begin{aligned}&\tilde{\mu}(x_t, t, B) := \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, B)\right), \\ &\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \\ &p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, B) = \mathcal{N}(\boldsymbol{x}_{t-1}; \tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0, B), \tilde{\beta}_t B^2 \mathbf{I}),\end{aligned} \tag{7}$$

where ϵ_θ is a noise predictor.

Training. In the Training Phase, we train BUFF using LR-HR image pairs from the dataset over T total diffusion steps. The conditional noise predictor ϵ_θ begins with random initialization, while the RRDB-based LR encoder D is pretrained using an L1 loss function alongside our Bayesian model Ω. For each training iteration, we select a mini-batch of LR-HR pairs, compute the residual image x_r, and process LR images through the Bayesian model to produce a modification mask B. This mask is used to modulate Gaussian noise and, after merging with LR, is encoded and input into ϵ_θ with t and x_t. We modulate sampled Gaussian noise for each t within $1, \cdots, T$ and iteratively optimize the noise predictor through

Method	Set14		Urban100		BSD100		Manga109		General100		DIV2K	
	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)
SRCNN	24.45	0.6432	21.95	0.6457	24.69	0.6365	20.72	0.7008	22.19	0.6432	24.70	0.6929
EDSR	26.72	0.7428	23.49	0.7233	25.78	0.6808	29.42	0.8798	27.25	0.7886	29.29	0.8027
RCAN	26.81	0.7440	23.56	0.7241	25.69	0.6797	29.09	0.8746	27.18	0.7861	29.32	0.8033
ESRGAN	25.27	0.6801	22.99	0.6940	24.65	0.6374	28.60	0.8553	26.03	0.7449	27.18	0.7709
Real-ESRGAN	25.18	0.7098	22.12	0.6869	25.11	0.6712	26.73	0.8639	25.64	0.7607	27.56	0.7893
BSRGAN	25.05	0.6746	22.37	0.6628	24.95	0.6365	26.09	0.8272	25.23	0.7309	27.32	0.7577
BebyGAN	25.73	0.6994	23.36	0.7113	24.75	0.6527	29.35	0.8775	26.19	0.7549	28.62	0.7904
SwinIR	26.77	0.7269	25.06	0.7488	26.11	0.6913	28.94	0.8687	27.83	0.8015	28.19	0.7727
HAT	27.09	0.7482	25.38	0.7642	26.41	0.6998	29.33	0.8839	30.18	0.8297	29.21	0.8066
IDM	26.53	0.7255	24.87	0.7479	26.04	0.6938	29.11	0.8697	27.79	0.8005	28.46	0.7898
ACDMSR	26.81	0.7398	25.15	0.7589	25.98	0.6875	29.21	0.8842	30.24	0.8337	28.87	0.7956
SAM-DiffSR	27.01	0.7456	25.46	0.7621	26.39	0.7003	29.36	0.8899	30.06	0.8353	28.84	0.8009
ResDiff	26.73	0.7457	25.21	0.7629	26.32	0.6951	29.23	0.8739	30.11	0.8241	28.77	0.8023
SRDiff(Baseline)	26.69	0.7287	25.13	0.7582	25.79	0.6813	28.82	0.8725	29.79	0.8241	28.76	0.7912
BUFF (Ours)	27.11	0.7487	25.49	0.7634	26.40	0.7011	29.38	0.8861	30.21	0.8349	29.35	0.8078
Δ	**+0.42**	**+0.0200**	**+0.36**	**+0.0052**	**+0.61**	**+0.0198**	**+0.56**	**+0.0136**	**+0.42**	**+0.0108**	**+0.59**	**+0.0166**

Table 1: Results on test sets of several public benchmarks and the validation set of DIV2K. The first 12 rows report the results achieved by MSE-based, GAN-based, Flow-based and diffusion-based approaches. Δ represents performance improvements over the diffusion-based baseline SRDiff. (↑) and (↓) indicate that a larger or smaller corresponding score is better, respectively.

gradient steps, streamlining the process while ensuring the training's effectiveness and efficiency.

Inference. A T-step BUFF inference begins by taking a low-resolution (LR) image x_L as input. Initially, we draw a latent variable x_T from a standard Gaussian distribution and upscale x_L using a bicubic kernel. Simultaneously, we employ a Bayesian model to generate a corresponding mask B, mirroring the training process. Moreover, the LR image x_L and mask B are encoded into x_e by the LR encoder—a step executed just once before iteration commencement—to expedite the inference process. Iterations initiate from $t = T$, with each iteration yielding a residual image characterized by progressively diminishing noise levels as t decreases. For iterations where $t > 1$, Gaussian noise z is sampled and modulated, and x_{t-1} is computed employing the noise predictor ϵ_θ with inputs x_t, x_e, and t. Upon reaching $t = 1$, we set $z = 0$, and x_0 emerges as the final residual prediction. The super-resolution (SR) image is reconstructed by adding the residual image x_0 to the upscaled LR image, denoted as up(x_L).

Experiment

Experimental Setup

Datasets and benchmarks. We train both the Bayesian model and BUFF on the DF2K dataset, a combination of DIV2K (Timofte et al. 2017) and Flickr2K, comprising 3450 (800 + 2650) high-quality images. After Bayesian model's training, for all images in the training set, we adopt the Bayesian model to obtain their corresponding masks. We then adopt a patch size settings of 160×160 to crop each image and its corresponding mask. During testing, we evaluate our models using PSNR and SSIM on six publicly available benchmark datasets: Set5 (Bevilacqua et al. 2012), Set14 (Zeyde, Elad, and Protter 2012), B100 (Arbelaez et al. 2011), Urban100 (Huang, Singh, and Ahuja 2015), Manga109 (Matsui et al. 2015) and DIV2K (Timofte et al. 2017). For face SR, we train the models at $16 \times 16 \rightarrow 128 \times 128$ on Flickr-Faces-HQ (FFHQ) dataset, which includes 70k images in total, and we sample 400 images from CelebA-HQ dataset for evaluation. Both objective and subjective metrics are used in our experiment. To evaluate the perceptual quality, we also adopt Frechet inception distance (FID) (Heusel et al. 2017) as the subjective metric, which measures the fidelity and diversity of generated images. PSNR and SSIM results are calculated on the **Y** channel in the YCbCr color space.

Training implementation details. We trained the Bayesian model for 50k rounds using a batchsize of 16, and the training strategy uses the NLL loss mentioned in Section., and the initial learning rate of $\eta = 1 \times 10^{-4}$, which was decayed by a factor of 0.5 every 2×10^5 iterations. We employed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, without weight decay. We train the diffusion model for 400K iterations with a batch size of 16, and adopt Adam as the optimizer. The initial learning rate is $\eta = 2 \times 10^{-4}$ and the cosine learning rate decay is adopted. All experiments were conducted on a system equipped with an NVIDIA RTX 3090 GPU, and the models were implemented using the PyTorch framework.

Performance of Image SR

Our proposed BUFF method showcases a remarkable performance. Table 1, which compiles our extensive benchmarking results, indicates that BUFF surpasses the diffusion-based baseline SRDiff across nearly all metrics. Notably, BUFF achieves impressive enhancements in PSNR and SSIM across multiple datasets. While there's a marginal decrease in the SSIM on Set14 and BSD100, this is offset by significant gains in image fidelity elsewhere. Visual evidence of BUFF's superior reconstruction capabilities is provided in Figure 3.

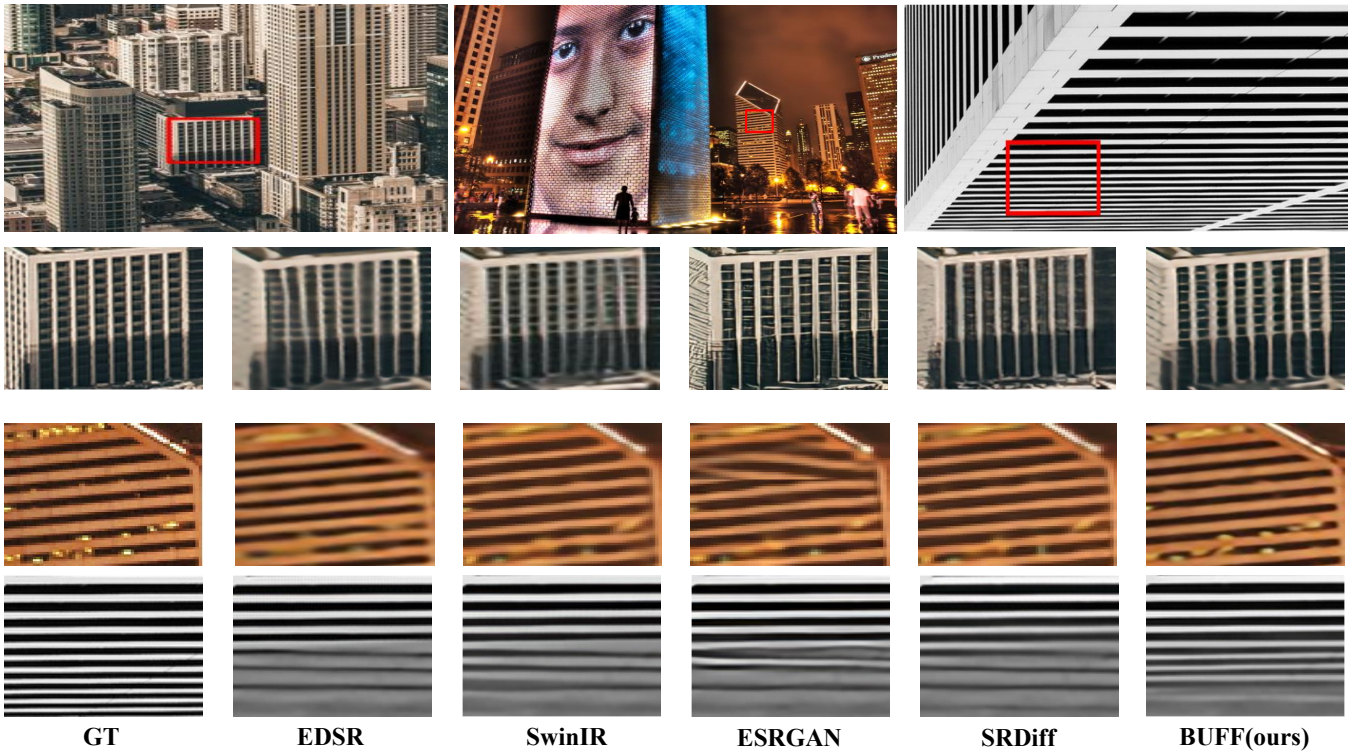

Figure 3: Visualization of restored images generated by different methods. Our BUFF surpasses other approaches in terms of both higher reconstruction quality and fewer artifacts. Additional visualization results can be found in our supplementary material.

Here, BUFF-generated images exhibit notably sharper details and more faithful textures when compared to baseline methods. The comparison highlights BUFF's ability to reduce artifacts, those unwanted distortions that often accompany SR processes, thereby ensuring a cleaner and more accurate rendition of the original scene.

Figure 4: Visual comparisons on CelebA-HQ dataset for 8× face SR. Zoom in for a better view.

Performance of Deblurring

In an effort to benchmark the deblurring efficacy of our BUFF, we subjected it to a series of evaluations against several state-of-the-art methodologies on the DIV2K100 and ImageNet-1K datasets. The high-resolution (HR) images were synthetically degraded using random anisotropic Gaussian kernels to simulate a range of blur conditions. BUFF's performance was measured against seven contemporary deblurring methods—ranging from DASR(Wang et al. 2021a), DAN(Huang et al. 2020), and DCLS(Luo et al. 2022), to BSRGAN(Zhang et al. 2021), AdaTarget(Jo et al. 2021), DARSR(Zhou et al. 2023), and KDSR(Xia et al. 2022). Quantitative results presented in Table 2 reveal BUFF's superior performance, even amidst the generalized and demanding conditions of the tests. These improvements prove that BUFF also has a very great advantage and potential handling complex deblurring tasks compared to SRDiff and general deblurring models.

Method	DIV2K100			ImageNet-1K		
	LPIPS	FID	PSNR	LPIPS	FID	PSNR
DASR	0.4476	149.11	25.46	0.4116	100.66	26.22
DAN	0.3597	96.63	26.74	0.3272	68.52	27.33
DCLS	0.3085	69.98	28.31	0.2791	54.59	29.02
BSRGAN	0.3526	98.39	24.90	0.3546	80.95	25.60
AdaTarget	0.2923	77.04	28.25	0.3249	56.81	27.58
DARSR	0.4956	148.34	24.05	0.4618	107.79	24.22
KDSR	0.4328	144.25	25.82	0.4035	101.22	26.48
SRDiff	0.3041	56.12	26.63	0.2772	64.12	27.41
BUFF (ours)	0.2946	51.62	26.75	0.2741	52.42	28.29

Table 2: Quantitative results on DIV2K100 and ImageNet-1K datasets.

Mask quality (AUSE)	Urban100			DIV2K		
	PSNR (↑)	SSIM (↑)	FID (↓)	PSNR (↑)	SSIM (↑)	FID (↓)
0.308	25.01	0.7582	4.93	29.01	0.7991	0.45
0.217	25.23	0.7605	4.87	29.12	0.8013	0.44
0.121	**25.49**	**0.7634**	**4.61**	**29.35**	**0.8078**	**0.41**

Table 3: Comparison of uncertainty masks with different qualities.

Amplification Intensity	Urban100			DIV2K		
	PSNR (↑)	SSIM (↑)	FID (↓)	PSNR (↑)	SSIM (↑)	FID (↓)
1.30	25.41	0.7631	4.60	29.28	0.8070	0.45
1.10	25.32	0.7631	4.64	29.17	0.8045	0.47
1.20(ours)	**25.49**	**0.7634**	**4.61**	**29.35**	**0.8078**	**0.41**

Table 4: Comparison of different intensity for noise amplification.

Performance of Face SR

Figure 4 showcases the performance of our model on the CelebA-HQ dataset for $8\times$ face super-resolution (SR). The comparison spans several methods: bicubic interpolation, PULSE(Menon et al. 2020), SRDiff, and our BUFF model, against the high-resolution (HR) benchmarks. Our BUFF model exhibits a remarkable improvement in recovering fine details and producing lifelike textures, as seen in the clarity of facial features. This visual assessment underscores BUFF's advanced capabilities in enhancing image quality, affirming its effectiveness in generating high-fidelity face SR images.

Ablation study

Quality of Uncertainty mask. In our BUFF study, we examined how uncertainty masks of different qualities, produced by Bayesian models at varying training levels, affect super-resolution performance. From Table.3, these qualities, reflected by Area Under the Sparsification Error (AUSE) scores (0.10, 0.20, 0.30), relate to the precision of uncertainty estimation—the lower the AUSE, the better the estimation. We categorized the masks into "Low," "Medium," and "High" qualities based on their AUSE scores and integrated each into the diffusion model. Evaluating the model on Urban100 and DIV2K datasets revealed a clear pattern: better mask quality leads to improved model performance, evidenced by metrics like PSNR, SSIM, and FID. This highlights the crucial role of accurate uncertainty estimation in super-resolution, with high-quality masks enhancing image fidelity and detail by aligning noise modulation with actual errors more effectively. We also present the visual results of the ablation experiment in Fig. 5, which verifies the effect of uncertainty masks on super-resolution output.

Comparison of Amplification Intensity. Our investigation into noise amplification within super-resolution tasks in Table.4 shows that overly high amplification factors ($e.g.$, 1.30) adversely affect both training efficiency and image quality, as demonstrated by performance drops in the Urban100 and DIV2K datasets. This indicates that excessive noise can destabilize the super-resolution model, complicating training and diminishing output quality. On the other hand, lower amplification levels, like 1.10, fail to significantly improve performance and introduce unnecessary complexity compared to the baseline SRDiff model. An amplification setting of 1.20, however, finds a sweet spot by improving key metrics such as PSNR and SSIM, and by reducing the FID.

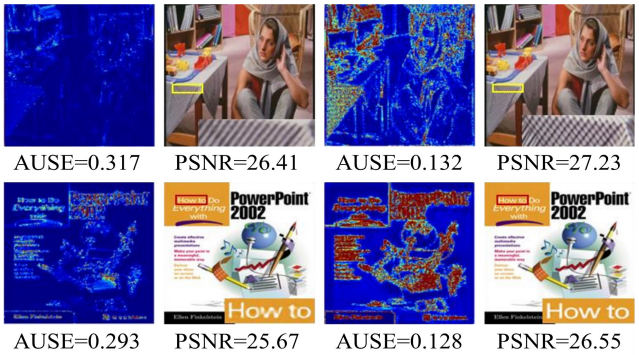

Figure 5: Visualization of the effect of uncertainty Mask Quality.

Method	BG	BE	PSNR↑	SSIM↑	DI↑	LPIPS↓
a	✓		26.21	0.7004	15.31	0.2834
b		✓	26.01	0.6992	15.15	0.2958
c	✓	✓	**26.40**	**0.7011**	**15.36**	**0.2741**

Table 5: Comparison of model configurations and their performance metrics for $\times 4$ SR task on BSD100 dataset.

Directly integrating Mask into diffusion model. In our study, we performed an ablation test on the BSD100 dataset for a $4\times$ super-resolution (SR) task to assess the impact of Bayesian Guided (BG) and Bayesian Embedding (BE) components in our SR model depicted in Table.5. The baseline 'a' applies the BG method, using Bayesian-generated masks for noise modulation, while 'b' employs BE, adding refined masks to the noise predictor with encoded low-resolution inputs. Model 'c' combines BG and BE, enhancing both noise modulation and prediction accuracy. Results show BG and BE independently improve PSNR and SSIM metrics, with BUFF achieving the best scores across all metrics, including the lowest LPIPS, indicating superior image quality.

Conclusion

In this work, we have presented BUFF, a novel framework that augments diffusion-based image super-resolution models by integrating Bayesian-derived uncertainty masks to refine structure-level detail enhancement. Our method effectively injects structural information into the diffusion process, tuning the noise profile at a pixel-level based on localized uncertainty. This targeted modulation of noise leads to a marked improvement in the delineation of structural details and a concurrent reduction in image artifacts. The performance of our approach has been rigorously validated through comprehensive testing on standard image super-resolution benchmarks, confirming its efficacy and potential applicability in advanced imaging tasks.

Acknowledgments

This work was supported by National Science and Technology Major Project (No. 2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. U23A20383, No. U21A20472, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No. 2021J06003, No.2022J06001).

References

Arbelaez, P.; Maire, M.; Fowlkes, C.; and Malik, J. 2011. Contour Detection and Hierarchical Image Segmentation. *PAMI*, 33(5): 898–916.

Bashir, S. M. A.; Wang, Y.; Khan, M.; and Niu, Y. 2021. A comprehensive review of deep learning-based single image super-resolution. *PeerJ Computer Science*, 7: e621.

Bevilacqua, M.; Roumy, A.; Guillemot, C.; and line Alberi Morel, M. 2012. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In *BMVC*, 135.1–135.10.

Chauhan, K.; Patel, S. N.; Kumhar, M.; Bhatia, J.; Tanwar, S.; Davidson, I. E.; Mazibuko, T. F.; and Sharma, R. 2023. Deep learning-based single-image super-resolution: a comprehensive review. *IEEE Access*, 11: 21811–21830.

Chen, T. 2023. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023a. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22367–22377.

Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yang, X.; and Yu, F. 2023b. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12312–12321.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11065–11074.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10021–10030.

Gao, S.; and Zhuang, X. 2022. Bayesian image super-resolution with deep modeling of image statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1405–1423.

Gawlikowski, J.; Tassi, C. R. N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589.

He, Z.; and Zhang, S. 2024. ESR-DDLN: Enhanced Single Image Super-Resolution Via Dual-Domain Learning Network. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Huang, J.-B.; Singh, A.; and Ahuja, N. 2015. Single Image Super-Resolution From Transformed Self-Exemplars. In *CVPR*, 5197–5206.

Huang, Y.; Li, S.; Wang, L.; Tan, T.; et al. 2020. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, 33: 5632–5643.

Jo, Y.; Oh, S. W.; Vajda, P.; and Kim, S. J. 2021. Tackling the ill-posedness of super-resolution through adaptive target generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16236–16245.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lepcha, D. C.; Goyal, B.; Dogra, A.; and Goyal, V. 2023. Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 91: 230–260.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022a. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

Li, W.; Zhou, K.; Qi, L.; Lu, L.; and Lu, J. 2022b. Best-buddy gans for highly detailed image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1412–1420.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136–144.

Liu, T.; Cheng, J.; and Tan, S. 2023. Spectral Bayesian uncertainty for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18166–18175.

Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; and Zeng, T. 2022. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 457–466.

Luo, G.; Blumenthal, M.; Heide, M.; and Uecker, M. 2023. Bayesian MRI reconstruction with joint uncertainty estimation using diffusion models. *Magnetic Resonance in Medicine*, 90(1): 295–311.

Luo, Z.; Huang, H.; Yu, L.; Li, Y.; Fan, H.; and Liu, S. 2022. Deep constrained least squares for blind image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17642–17652.

Marinescu, R. V.; Moyer, D.; and Golland, P. 2020. Bayesian image reconstruction using deep generative models. *arXiv preprint arXiv:2012.04567*.

Matsui, Y.; Ito, K.; Aramaki, Y.; Yamasaki, T.; and Aizawa, K. 2015. Sketch-based Manga Retrieval using Manga109 Dataset. *arXiv preprint arXiv:1510.04389*.

Menon, S.; Damian, A.; Hu, S.; Ravi, N.; and Rudin, C. 2020. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2437–2445.

Narnhofer, D.; Effland, A.; Kobler, E.; Hammernik, K.; Knoll, F.; and Pock, T. 2021. Bayesian uncertainty estimation of learned variational MRI reconstruction. *IEEE transactions on medical imaging*, 41(2): 279–291.

Niu, A.; Pham, T. X.; Zhang, K.; Sun, J.; Zhu, Y.; Yan, Q.; Kweon, I. S.; and Zhang, Y. 2024. ACDMSR: Accelerated conditional diffusion models for single image super-resolution. *IEEE Transactions on Broadcasting*.

Ruz, G. A.; Henríquez, P. A.; and Mascareño, A. 2020. Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. *Future Generation Computer Systems*, 106: 92–104.

San-Roman, R.; Nachmani, E.; and Wolf, L. 2021. Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*.

Shang, S.; Shan, Z.; Liu, G.; Wang, L.; Wang, X.; Zhang, Z.; and Zhang, J. 2024. Resdiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8975–8983.

Tian, C.; Zhang, X.; Lin, J. C.-W.; Zuo, W.; Zhang, Y.; and Lin, C.-W. 2022. Generative adversarial networks for image super-resolution: A survey. *arXiv preprint arXiv:2204.13620*.

Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; and Zhang, L. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results. In *CVPR Workshops*.

Upadhyay, U.; Karthik, S.; Chen, Y.; Mancini, M.; and Akata, Z. 2022. Bayescap: Bayesian identity cap for calibrated uncertainty in frozen neural networks. In *European Conference on Computer Vision*, 299–317. Springer.

van de Schoot, R.; Depaoli, S.; King, R.; Kramer, B.; Märtens, K.; Tadesse, M. G.; Vannucci, M.; Gelman, A.; Veen, D.; Willemsen, J.; et al. 2021. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1): 1.

Wang, C.; Hao, Z.; Tang, Y.; Guo, J.; Yang, Y.; Han, K.; and Wang, Y. 2024. SAM-DiffSR: Structure-Modulated Diffusion Model for Image Super-Resolution. *arXiv preprint arXiv:2402.17133*.

Wang, H.; and Yeung, D.-Y. 2020. A survey on Bayesian deep learning. *ACM computing surveys (csur)*, 53(5): 1–37.

Wang, L.; Wang, Y.; Dong, X.; Xu, Q.; Yang, J.; An, W.; and Guo, Y. 2021a. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10581–10590.

Wang, P.; Bayram, B.; and Sertel, E. 2022. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232: 104110.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021b. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1905–1914.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0.

Wang, Z.; Chen, J.; and Hoi, S. C. 2020. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3365–3387.

Wu, Z.; Chen, X.; Xie, S.; Shen, J.; and Zeng, Y. 2023. Super-resolution of brain MRI images based on denoising diffusion probabilistic model. *Biomedical Signal Processing and Control*, 85: 104901.

Xia, B.; Zhang, Y.; Wang, Y.; Tian, Y.; Yang, W.; Timofte, R.; and Van Gool, L. 2022. Knowledge distillation based degradation estimation for blind super-resolution. *arXiv preprint arXiv:2211.16928*.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.

Zeyde, R.; Elad, M.; and Protter, M. 2012. On Single Image Scale-Up Using Sparse-Representations. In *Curves and Surfaces*, 711–730.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4791–4800.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, 286–301.

Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023. DDFM: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8082–8093.

Zhou, H.; Zhu, X.; Zhu, J.; Han, Z.; Zhang, S.-X.; Qin, J.; and Yin, X.-C. 2023. Learning Correction Filter via Degradation-Adaptive Regression for Blind Single Image Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12365–12375.