MLVU Project:
Image Deblurring with Generative Diffusion Process

Ki-Ung Song
Department of Mathematical Sciences
sk851@snu.ac.kr

Jiwon Kang*
Department of Mechanical Engineering
2007jiwon@snu.ac.kr

Junsu Leem*
Department of Energy Resources Engineering
sooyalim@snu.ac.kr

Dennis Ledwon*
Department of Computer Science
dennis.ledwon@rwth-aachen.de

Abstract

Photographs that do not satisfy quality standards are a common problem. But deblurring the given image is an ill-defined inverse problem that is hard to solve. State-of-the-art network architectures for image deblurring are generative model frameworks including GANs and encoder-decoder networks. The Generative Diffusion Process, which is also called as diffusion model, is a new branch of a generative model framework that is in the spotlight in various image restoration tasks such as Super Resolution. However, there are still insufficient experimental attempts to see if the generative diffusion framework without task-specific design can be applied to other image restoration tasks as well. In this work, we first investigated whether the generative diffusion framework without task-specific design is suitable for the image deblurring task compared to the existing approaches. Unlike an image classification task which has considered "Train-Test Resolution Discrepancy", the evaluation of the model performance considering the resolution discrepancy has not received attention in the image deblurring task. Thus we compared the diffusion model's deblurring performance to the existing models in a more general image deblurring setting where the train-test resolution discrepancy exists rather than a standard-setting where two resolutions are the same. And by inspecting the iterative steps of a diffusion-based deblurring process, it has been attempted to visually understand this generation process.

1. Introduction

Photographs that do not satisfy quality standards are a common problem: Shakes of the camera, fast object motions and depth variation introduce blur to the image or the resolution of the image is unsatisfactory, making it difficult to perceive details. The additional details that a deblurred, high-resolution photo provides are useful in many applications, including object detection [7] and medical imaging [14]. Image blur and low resolution can be modeled as follows:

$$I_B = (K * I_S) \downarrow_s + N \tag{1}$$

where I_B, I_S are the blurry and sharp images, respectively, K is the blur kernel, $\downarrow_s$ is a downsampling operator with a scaling factor s with $s \geq 1$ and N is additive noise [1,9,24]. Both image deblurring and Super-Resolution (SR) are concerned with restoring I_B from I_S, either with (non-blind) or without (blind) sharp image data I_S. Thus, image deblurring and SR are ill-defined inverse problems, *i.e*. challenges arise because multiple sharp images match the same blurry, low resolution image. As a result, image priors are required to restrict the solution space.

Recently, CNNs (convolutional neural networks) have been successfully used for both tasks without explicitly estimating, effectively learning image priors from large-scale image data [1, 17]. State-of-the-art (SOTA) network architectures for image deblurring are generative model frameworks including GANs (generative adversial networks) and encoder-decoder networks. GAN-based approaches formulate the deblurring task as a minmax-game between a generator and a discriminator network [8]. The generator takes the blurred image and a noise vector as input and tries to create a plausible deblurred image, that the discriminator cannot distinguish from the true sharp image [15,16,26]. However, GANs are often difficult to train with mode collapsing being a well-known issue [2]. Encoder-decoder models downsample the blurred input image while generating features that encode broad contextual information, and then use upsampling operations to restore a sharp image of the same

*Equal contribution

size as the input [3,25]. [3] is currently the most competitive image deblurring network.

Another framework of generative models, the generative diffusion process that is also called diffusion models, has recently entered the spotlight in the related task of SR [6, 19]. The diffusion models are a class of likelihood-based models that learn a reverse diffusion process that starts at a pure white-noise image $x_T \sim \mathcal{N}(0, I)$ and gradually removes noise to arrive at a sharp image x_0 according to learned conditional transition distributions. According to [6], the diffusion models achieve SOTA performance on SR tasks, which was previously held by GANs. Additionally, diffusion models are easier to train and capture more diversity than GANs [6]. Due to its SOTA performance, the generative diffusion process is under active research, but several experimental attempts are still lacking.

Both SR and image deblurring tasks aim to have high-resolution images from given low-resolution images. But in the case of a recent SR task, we focus on the aspect of "generation" from the image of static scenes of lower resolution. And in the case of a deblurring task, we want to focus on the aspect of "restoration" from the images which was obtained from the more dynamic scene such as camera shake. Thus, the existing SR and Deblurring models are designed task-specifically and there is no universal framework that performs both tasks generally. Based on the similarity of SR and image deblurring task, we investigated whether the diffusion model without task-specific design is suitable for image deblurring as well and its deblurring performance was compared to both SR and deblurring models in a more general deblurring setting.

The traditional approach before the era of deep learning estimated the explicit kernel for deblurring, *e.g*. motion kernel for motion blurred image [20], which partially explains the deblurred result. But the existing SOTA models for the image deblurring task is an almost black-box model which we cannot interpret. Since the diffusion model performs the deblurring process through iterative steps, by inspecting the iterative steps of a diffusion-based deblurring process, it has been attempted to visually understand this generation process.

2. Related works

Image deblurring task is traditionally treated as a deconvolution problem, which can be treated either in a blind or non-blind manner. Since the era of deep learning, various techniques have boosted performance on image deblurring tasks. However, the image deblurring model which can be applied to the blurry image of all situations still has a long way to go. To handle this, recent studies propose models for blurred images limited to more specific situations: for instance, Chen *et al*. [4] proposed to use Non-Blind Deblurring Network (NBDN), which is composed of fidelity term and prior term, and Hyper Parameter Estimation Unit (HPEU) to deblur night blurry images. The proposed method can deal with different blur level images using HPEU and generate fewer artifacts that are produced due to pixel saturation. However, for NBDN, as the network deals with saturated pixels, it may not work well with inputs having both day images and night images. Also, as it is a non-blind deblurring process, the method is highly ill-posed.

2.1. Generative adversarial network

Commonly used network architecture for image deblurring task is GANs. GANs, which put the generator into competition with discriminator, are widely applied to the image to image translation problems. Kupyn *et al*. [16] proposed DeblurGAN-v2 using Feature Pyramid Network (FPN) as a generator and relativistic discriminator with a least-square loss as a discriminator. Especially, they applied FPN to deblurring tasks for the first time to deal with multiple blur levels. Also, for its backbone frame, they could choose between Inception-ResNet-v2 backbone and MobileNet Depth-wise Separatable Convolutions (DSC) which have a trade-off between performance and efficiency.

As datasets used for training in the deblurring tasks are synthetically made, Zhang *et al*. [26] proposed GAN architecture where two models are used: learning to Blur GAN (BGAN) and learning to DeBlur GAN (DBGAN). BGAN takes Real-World Blurred Image (RWBI) dataset as input to evaluate the discrepancy between synthetic blur images and real blur images using Realistic Blur Loss (RBL), and generate realistic blur images. These images are then input to DBGAN and DBGAN trains how to deblur blurry images.

2.2. Encoder-decoder network

Another commonly used network architecture for image deblurring task is encoder-decoder networks. Zamir *et al*. [25] proposed Multi-stage Progressive image Restoration (MPR) Net. They use a combination of encoder-decoder architecture and single-scale feature pipeline architecture by adapting U-Net architecture and Original Resolution Subnetwork (OSRNet). As the method is composed of multiple stages, the connection between stages is done by applying Cross Stage Feature Fusion (CSFF) and Supervised Attention Module (SAM).

Chen *et al*. [3] adopted normalization and surpassed the SOTA on various image restoration tasks. As image restoration tasks use a small mini-batch size, it is hard to adapt batch normalization (BN) but instead, they used instance normalization (IN) in order to build denoiser. The proposed method is mainly composed of two U-net architecture-shaped stages with HIN blocks inside, so-called HIN Net. HIN (Half Instance Normalization) block is a process that applies IN to half of the channels and the rest half is trans-

ferred as their identity. Two stages are connected using CSFF and SAM.

2.3. Generative Diffusion Process

The generative diffusion process which is also called diffusion model is a new branch of a generative model framework based on a stochastic process and is in the spotlight recently. Saharia *et al*. [19] used a diffusion model framework to handle the SR task mentioned above and achieved competitive performance. Dhariwal *et al*. [6] demonstrated that the diffusion models even beat GANs on image synthesis tasks. And Kawar *et al*. [13] used the conditional diffusion model to reach the competitive performance in the image denoising task. Although the diffusion models are showing good performance in various image restoration tasks, experimental attempts regarding the diffusion models are still lacking. Therefore, it seems worth trying to investigate whether the generative diffusion model without any task-specific design is suitable for the image deblurring task as well.

3. Method

The Generative Diffusion Process is elaborately based on a stochastic process. A further formal description is given in the following subsection. We may assume that the blurring process (1) can be approximated by the forward diffusion process. Thus by training the reverse diffusion process, which is the inverse of forward diffusion process, we may hope to recover the desired deblurred image.

3.1. Background

According to the work of Jonathan Ho *et al*. [10], consider the forward diffusion process q which is Markovian process. The forward diffusion process can be viewed as noising process, in other words it perturbs the given image x_{t-1} at a time step $t-1$ to the image x_t at a time step t with Gaussian noise where α_t controls the amount of noise perturbation at a time step t.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I) \tag{2}$$

Then we can iterate this process up to sufficiently large T so that x_T becomes almost pure Gaussian noise. And by marginalizing the intermediate step, we can get the following distribution for x_t

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\beta_t}x_0, (1-\beta_t)I) \tag{3}$$

where $\beta_t = \prod_{i=1}^{t} \alpha_t$. To have actual generative process, we need the reverse process of this noising forward process, *i.e*. reverse diffusion process $q(x_{t-1}|x_t)$ of equation (2). It cannot be derived directly, but according to [10], if we condition the variable x_0, we can derive the posterior distribution $q(x_{t-1}|x_t, x_0)$ instead.

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \mu, \sigma^2 I) \tag{4}$$

$$\mu = \frac{\sqrt{\beta_{t-1}}(1-\alpha_t)x_0 + \sqrt{\alpha_t}(1-\beta_{t-1})x_t}{1-\beta_t} \tag{5}$$

$$\sigma^2 = \frac{(1-\beta_{t-1})(1-\alpha_t)}{1-\beta_t} \tag{6}$$

However, in the actual generation phase, we do not have an access to the original image x_0. But note that from the equation (3), we can put as

$$x_t = \sqrt{\beta_t}x_0 + \sqrt{1-\beta_t}z \tag{7}$$

where $z \sim \mathcal{N}(0, 1)$. Thus when given x_t only, we can conversely approximate x_0 as

$$x_0 \approx \frac{x_t - \sqrt{1-\beta_t}z}{\sqrt{\beta_t}} \tag{8}$$

Therefore, by plugging (8) in to equation (5), we can approximate the reverse diffusion process $q(x_{t-1}|x_t)$ which corresponds to denoising process, which can be used for data generation.

There is also another perspective on diffusion models. Yang Song and Stefano Ermon [21] suggested the score matching framework which also perturbs the original image similar to the forward diffusion process of [10]. At later work of Yang Song *et al*. [22], these two perspectives were integrated as a one stochastic process framework with Stochastic Differential Equation (SDE). The integrated framework of the forward diffusion process which perturbs image can be modeled as the solution of Ito SDE

$$dx = f(x,t)dt + g(t)dw \tag{9}$$

where f is vector-valued function, g is scalar function and w is Wiener process. If we reverse this process, we can generate the desired image, *i.e*. starting from noise x_T, we can obtain sample $x_0 \sim p_0$ where p_0 is original data distribution. The desired reverse process of (9) is also formulated as SDE and given as follows:

$$dx = [f(x,t) - g(t)^2 \nabla_x \log p_t(x)]\, dt + g(t)\, d\bar{w} \tag{10}$$

where $\bar{w}$ is standard Wiener process when time flows backwards and the estimate of $\log p_t(x)$ is called score. With this framework, the diffusion model can be trained by minimizing the following objective:

$$\mathbf{E}\left[\frac{\lambda(t)}{2}||s_\theta(x_t,t) - \nabla_{x_t} \log p_t(x_t|x_0)||_2^2\right] \tag{11}$$

with proper weighting function $\lambda(t)$. With above mentioned SDE framework, inference of the trained model, *i.e*. generation process of the data, becomes equivalent to solving the

equation (11). Since there are various ways of solving SDE, it means that the data generation process can have various methods [12, 22] to be selected. And it may be possible to visually analyze the pattern in which data is generated when different inference methods are selected.

3.2. Diffusion Process with Conditional Input

For actual training, the U-net shape model architecture is used to approximate the reverse diffusion process. By simply using conditional U-net architecture, we can use the conditional input to the model. There are already some approaches [13, 19] trying to solve the image restoration problem by simply utilizing the diffusion process with conditional input and without any task-specific loss term. They have shown descent performance in the SR and the image denoising task respectively. However, most of the papers that applied the conditional diffusion model to a specific image restoration task focused on showing the excellent performance through final visual results but did not talk about how the generation process went through in detail.

Thus, similar to [13, 19], the conditional diffusion model was trained without any task-specific design in this work. Simply utilizing the conditional U-net architecture, the diffusion-based deblurring process could be trained in a conditional diffusion model framework. The existing deblurring approaches based on GANs including [16, 26] are usually trained with an additional loss term with task-specific design. This approach seems more flexible but it often causes the training more unstable, which requires heavy experiments to find the proper weighting hyper-parameter of the additional loss term. It means that a model framework that can perform the desired task properly without any task-specific design seems to have more generalization power. The conditional diffusion model was trained to test this idea, *i.e*. whether the generative diffusion framework can generalize well to another kind of image restoration task, image deblurring without any task-specific still.

3.3. Visual Understanding

Combined with the fact that various experiments to explain the generation process of the diffusion model are still lacking, we not only trained the conditional diffusion model to perform the image deblurring task but also attempted to visually understand the diffusion deblurring process. The most important hyper-parameters in the training of diffusion model are the start and end values of beta that determine the degree of noise at each step, and the number of total time steps. In this work by visualizing the intermediate results from iterative steps of a diffusion-based deblurring process, we attempted to further the understanding of this restoration process. Based on the results, the additional experiments were conducted by changing the number of time steps, and as a result, the effect of the time step, which is an important factor in the training of the diffusion model, could be interpreted indirectly in the diffusion deblurring process.

Method	DDPM	HAN	RCAN	MIMO-Unet
PSNR (64x64)	22.17	25.17	25.18	28.8
PSNR (128x128)	22.48	24.1	24.03	28.42
Parameters (M)	20.74	15.74	15.3	9.91

Table 1. The average PSNR, and the number of model parameters. Every models were trained on 64x64, and 128x128 size image patches respectively and tested on 64x64 size image patches to compute PSNR.

And by introducing the self-attention layers inside the model, we can try to visualize the inside of the model while performing the deblurring task. Given the L embeddings, it is well known that the naive self-attention mechanism has quadratic complexity regarding the number of embeddings *i.e*. $O(L^2)$. In the case of spatial embeddings, since $L = HW$, the self-attention has a quartic complexity regarding image size. Due to its high computational complexity, the naive self-attention layer cannot be fully utilized inside the U-net architecture. To deal with this problem, we tried 2 approaches: 1) Using a self-attention layer only to sufficiently downsampled feature resolution and 2) Using a linearized self-attention layer that has linear complexity.

3.4. Train-Test Resolution Discrepancy

In addition, let's consider the actual situation where the image deblurring should be performed by using the trained model. There is no guarantee that the resolution of the image used for training and the resolution of the image to be deblurred are the same. Thus, the performance of the deblurring models concerning the resolution discrepancy between the train and test image should be considered. The performance difference depending on the discrepancy between train and test resolution was investigated in the image classification task [23], but such approaches were lacking in the image generation area including the image restoration tasks. The existing deblurring methods are trained with a fixed size of the image patches and also tested with the same size of image patches qualitatively and even quantitatively, which does not consider the train-test resolution discrepancy. Thus, we tried to evaluate the diffusion model's deblurring performance in a more general situation where the train-test resolution discrepancy does exist and compared it to the existing SR and deblurring models.

4. Results

To train and test the image deblurring model in a non-blind setting. The dataset should consist of blurry and sharp images pairs (I_B, I_S). Chen *et al*. [4] utilized Flickr's copyright-free night images to make blurry and sharp im-

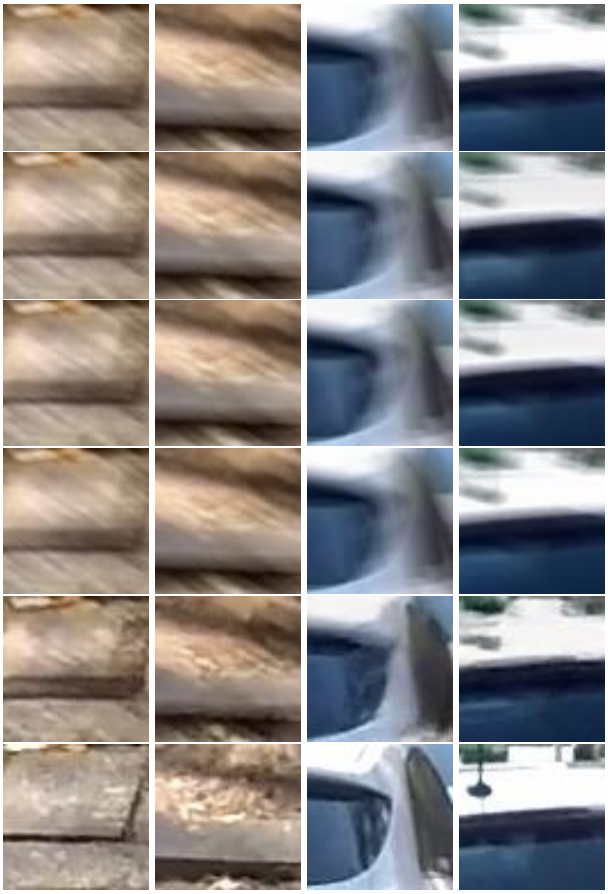


Figure 1. Qualitative result from the models trained on 128x128 size image patches and tested with smaller 64x64 size image patches. From the top row to the bottom, images in each rows are the blurry input images, results from HAN, results from RCAN, results from MIMO-Unet, results from DDPM, and the sharp images respectively

ages pairs. But most non-blind image deblurring models test their performance with already existing datasets such as GOPRO dataset [17] and RealBlur dataset [11]. Thus, the GOPRO dataset [17] was used in this work.

4.1. Experiment Setting and Baselines

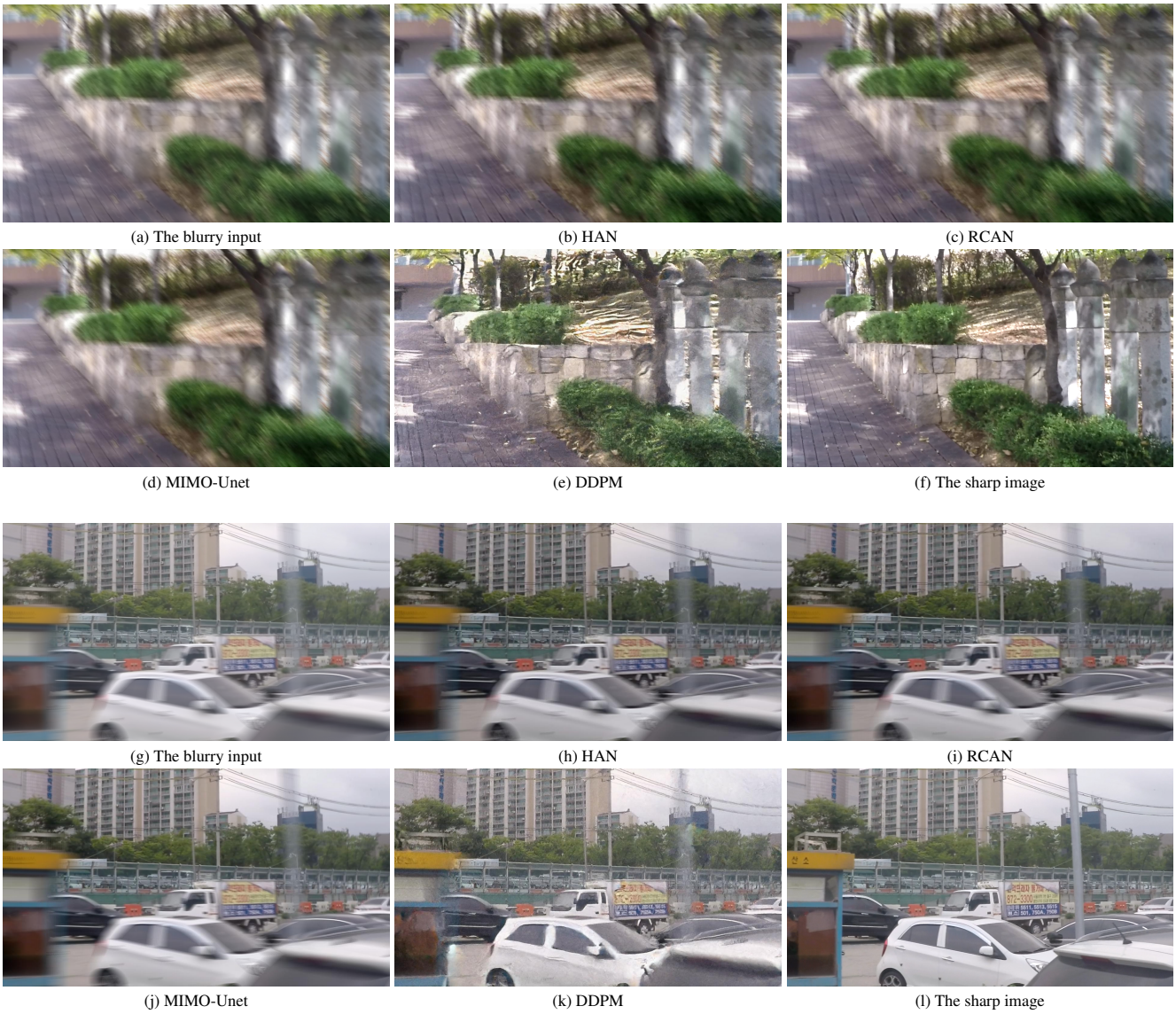
To train the deblurring model, we used a fixed size of randomly cropped image patches of the train data, which is the same setting as the existing deblurring approaches. Both baseline models and the diffusion models were trained with 64x64, and 128x128 size of patches respectively. The diffusion model used in this work is the DDPM [10]. And since the DDPM model can be trained in two different ways: the original framework in [10] and the SDE framework in [22], the DDPM models were trained with both frameworks for more extensive experiences. Although the SDE framework has much more various inference methods than the original one, the resulting deblurred image quality was almost similar. Thus the original DDPM was majorly used for the performance evaluation. The number of total time steps, *i.e.* total iteration steps is the important hyper-parameter of the training of the diffusion models. And it was set to 2000, 1000, and 400 respectively in this work to understand its impact in the diffusion deblurring process.

The baseline models were selected from both SR and deblurring tasks: 2 SR models (HAN [18], and RCAN [27]) and one of the SOTA deblurring model (MIMO-Unet [5]). The SR models were trained by simply using the paired GOPRO dataset, instead of the paired SR dataset. Since the diffusion model has shown its great performance in SR task [19], the diffusion model's generalization ability to another task without any task-specific design was compared with those of the SR-specific models. And beyond its generalization ability, the trained diffusion model was compared to one of the SOTA deblurring models to investigate whether the diffusion model can achieve the competitive deblurring performance, not just a relatively better performance compared to SR models.

4.2. Quantitative Evaluation Result

Due to the limit of computational resources, the trained deblurring model was quantitatively evaluated by using the 64x64 size randomly cropped fixed image patches with PSNR(Peak Signal-to-Noise Ratio) metric. Since the diffusion models were trained with both 64x64, and 128x128 size image patches, both cases were evaluated quantitatively. In the case of 64x64 size training, the evaluation was performed in a standard situation where there is no train-test resolution discrepancy. And in the case of 128x128 size training, we can say the evaluation was performed in a more general deblurring situation where the train-test resolution discrepancy exists.

See Table 1 for a quantitative evaluation result. For a standard case, the result shows that there is some PSNR difference between SR-specific models and the deblurring-specific model, which means that task-specific design made a significant difference in PSNR. And the diffusion model trained without task-specific design showed the lowest performance on PSNR. There was a significant PSNR difference from the SR models, so it showed a more PSNR difference from the deblurring model. Both the original and SDE framework DDPM showed low PSNR performance compared to the baselines. For a train-test resolution discrepancy scenario, we could check the PSNR change differences. Since 64x64 size image patches contain little information of the whole image compared to 128x128 size image patches, when the model was trained with the larger size of patches, the performance of the model should increase. However, in this train-test resolution discrepancy situation, it was confirmed that the PSNR performance of



(a) The blurry input (b) HAN (c) RCAN

(d) MIMO-Unet (e) DDPM (f) The sharp image

(g) The blurry input (h) HAN (i) RCAN

(j) MIMO-Unet (k) DDPM (l) The sharp image

Figure 2. Qualitative Result of Models trained on 128x128 size image patches and tested with original 1280x720 size image.

the baseline models decreased slightly. From this, we can think that the existence of the train-test resolution discrepancy deteriorated the actual performance of the baseline models. However, the diffusion model showed some slight increase in PSNR performance, although its absolute PSNR performance still lags behind baselines.

4.3. Qualitative Evaluation Result

See Fig. 1 for a second quantitative evaluation case where the train-test resolution discrepancy exists, *i.e*. the deblurred results from the models trained on 128x128 size image patches and qualitatively evaluated with 64x64 size image patches. The figure shows that although there was a significant PSNR difference between the existing SR and deblurring models, the actual deblurred results from all baselines were the mere reconstruction of the input blurry images, which does not show any visual distinctness. The quantitative evaluation of the diffusion model was the lowest according to the previous subsection, but the figure shows that the result from the diffusion model was the most relatively deblurred one. It means that the quantitative metric, PSRN, could not properly reflect the the significance of image deblurring task, and the diffusion model was trained to perform the image deblurring in a more general situation.

Next, to check this situation more visually, the deblurring models trained with 128x128 size patches were qualitatively evaluated by using the full size of 1280x720 images. See Fig. 2 for a qualitative evaluation result with the size of

(a) The blurry input (b) Result from 50th step (c) Result from 1600th step

(d) Result from 1750th step (e) Result from 1850th step (f) Result from 1900th step

(g) Result from 1950th step (h) Deblurred Result (i) The sharp image

Figure 3. Visualization of the Iterative steps of the diffusion-based deblurring process.

(a) Result from the models with 2000 time steps (b) Result from the models with 1000 time steps (c) Result from the models with 400 time steps

Figure 4. Deblurred results from the diffusion models trained with the different number of steps. From the top row to the bottom, the result is from the model trained with 2000, 1000, and 400 steps respectively

1280x720. In this case, a large train-test resolution discrepancy exists. The figure shows that actual the results from all baselines were the mere reconstruction of the input blurry images, similar to the former case. And the results from the diffusion models were the most relatively deblurred results, although there are still blurred regions. Since evaluated with 1280x720 size images, the degree of improvements of deblurring can be more clearly and visually confirmed. Therefore, we can say that the generative diffusion framework was able to be trained to perform the image deblurring in a more general situation where the train-test resolution discrepancy exists. Considering that the MIMO-Unet was trained by using 256x256 size image patches [5], a situation where the models are trained with 128x128 size image patches may limit the learning of deblurring. Thus, training the diffusion model with the larger size of image patches and evaluating its deblurring performance in a general situation needs to be addressed as future work.

Apart from the discussion above, the diffusion-based model has an additional advantage. Previously mentioned [3,5] architectures have only deterministic deblurred results once they are trained. It means that if the deblurring process

was not properly implemented, there is no way to use that model further. However, since the diffusion-based model has stochastic property, it can generate various deblurred results corresponding to the blurry image. And the generative steps can be more precisely controlled by using the SDE framework of the diffusion model.

4.4. Visualization of the iterative steps

Due to previous works of unconditional diffusion models, it has been understood that the generating process of the diffusion model occurs evenly and gradually during iterative time steps. In this work, we visualized the iterative steps of the diffusion-based deblurring process to check whether this belief is still true for the conditional diffusion models. See figure Fig. 3 for visualization of iterative steps of the deblurring process. The figure shows that until the 1600th step, the non-interpretable noise states were maintained, which accounts for 80% of the total 2000 steps. And significant generation process which can be visually perceived occurred only at the last 400 steps, which only accounts for 20% of the total 2000 steps. In every process of performing diffusion-based image deblurring, the same generation pattern as the above result was confirmed and this pattern was even agnostic to the image size.

At first glance, the meaningful generation process seems to only occur for the last 400 time steps. Therefore, it can be considered that an excessively large number of time steps was set for the training of the diffusion model. Since removing unnecessary time steps means the faster inference time of the diffusion model. To test the effect of the time step and the role of the non-interpretable noise states, the diffusion model was trained with the number of time steps of 1000 and 400 respectively. In order to only consider the impact of time steps, start and end values of the beta were also controlled to match the values of beta during the last 400 time steps.

See Fig. 4 for the deblurred result from the model trained with different time steps respectively. The input blurry image was the (a) of figure Fig. 2. The figure shows that the result from the model trained with 1000 steps could not match the overall tone of the original input image. By using the SDE framework [22], the inference process of the trained model can be more carefully controlled to match the overall tone of the input image, but it was still impossible to visually match the overall tone of the input image. When the model was trained with 400 steps only, it not only could not match the overall tone of the input image and but also could not perform image deblurring properly. Although most of the steps were meaningless noise visually, proper deblurred results could not be obtained when attempting to train the model by shortening the number of time steps. It suggests that most of the iterative steps that appear to be non-interpretable simple noise states are essential preparations to match the overall image tone and active generation in the late stage of the deblurring process.

4.5. Additional Approaches

In an attempt to further interpretation of the above mentioned intermediate noise states, we introduced the attention layers inside the U-net architecture with 2 different approaches: Using a self-attention layer only to sufficiently downsampled feature resolution and using a linearized self-attention layer. However, the usage of these two approaches dropped the overall deblurring performance of the diffusion models. It seems that the number of parameters of the model was insufficient to utilize self-attention layers. But due to the limitation of computational resources, experiments in larger models could not proceed. Thus, this approach for interpretation of the noise state of conditional generation needs to be addressed in future work.

5. Conclusions

The generative diffusion process is in the big spotlight nowadays and is likely to be applied in various image restoration tasks. In this work, we investigated whether the generative diffusion framework without any task-specific design is suitable for image deblurring as well as for other image restoration tasks. To this end, we trained the diffusion model to perform the image deblurring task and tried to compare its performance to the existing models in a more general deblurring setting where the train-test resolution discrepancy exists. In the resolution discrepancy situation, the conditional diffusion model could output the most relatively deblurred results while the existing models cannot perform the image deblurring properly, merely reconstructing the blurry input images. It has been understood that the generating process of the diffusion model occurs evenly and gradually during iterative time steps with the figures presented in several existing papers. However, by visually observing the iterative steps of the diffusion-based deblurring process, it was confirmed that the significant visually perceivable generation process occurred only at the last 400 steps out of the total 2000 steps. Although most of the steps were non-interpretable noise states, the further experiment confirmed that these steps were essential preparations to match the overall image tone and active generation in the late stage of the deblurring process.

References

[1] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey, 2020. 1

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. 1

[3] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 2, 7

[4] Liang Chen, Jiawei Zhang, Jinshan Pan, Songnan Lin, Faming Fang, and Jimmy S Ren. Learning a non-blind deblurring network for night blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10542–10550, 2021. 2, 4

[5] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring, 2021. 5, 7

[6] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021. 2, 3

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, 2016. 1

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1

[9] Ankit Gupta, Neel Joshi, C. Lawrence Zitnick, Michael Cohen, and Brian Curless. Single image deblurring using motion density functions. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 171–184, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 1

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 3, 5

[11] Jucheol Won Sunghyun Cho Jaesung Rim, Haeyun Lee. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 5

[12] Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021. 4

[13] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Stochastic image denoising by sampling from the posterior distribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1866–1875, October 2021. 3, 4

[14] D. Kouame and M. Ploquin. Super-resolution in medical imaging : An illustrative approach through ultrasound. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 249–252, 2009. 1

[15] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks, 2018. 1

[16] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019. 1, 2, 4

[17] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 5

[18] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network, 2020. 5

[19] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021. 2, 3, 4, 5

[20] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. *Acm transactions on graphics (tog)*, 27(3):1–10, 2008. 2

[21] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020. 3

[22] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3, 4, 5, 8

[23] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *arXiv preprint arXiv:1906.06423*, 2019. 4

[24] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 491–498, 2010. 1

[25] 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. 2

[26] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020. 1, 2, 4

[27] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 5