

Atmospheric Turbulence Correction via Variational Deep Diffusion

Xijun Wang¹, Santiago López-Tapia², Aggelos K. Katsaggelos²

¹Dept. of Computer Science, Northwestern University, Evanston, IL, USA

²Dept. of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA
xijunwang2022@u.northwestern.edu

Abstract

Atmospheric Turbulence (AT) correction is a challenging restoration task as it consists of two distortions: geometric distortion and spatially variant blur. Diffusion models have shown impressive accomplishments in photo-realistic image synthesis and beyond. In this paper, we propose a novel deep conditional diffusion model under a variational inference framework to solve the AT correction problem. We use this framework to improve performance by learning latent prior information from the input and degradation processes. We use the learned information to further condition the diffusion model. Experiments are conducted in a comprehensive synthetic AT dataset. We show that the proposed framework achieves good quantitative and qualitative results.

1. Introduction

Atmospheric Turbulence is an issue in real-life long-range imaging caused by slight perturbations in atmospheric conditions (e.g., temperature), and it can cause severe blur and perceptual degradation. This, in turn, could severely effect performance in the subsequent downstream vision tasks, such as detection, recognition, and so on. Unlike other imaging inverse problems, atmospheric turbulence degradation contains a mixture of geometrical distor-

tion, spatially variant blur, and noise, which makes AT more challenging to mitigate.

Earlier works in AT correction mainly focus on optics and lucky imaging algorithms [21]. These algorithms are often computationally expensive. In recent years, with the development of deep-learning (DL) algorithms for solving various inverse problems [18], some works have proposed DL-based AT removal methods [9, 13]. With [8] proposing a fast AT simulation algorithm, large-scale data-driven DL training for AT correction becomes possible [9]. In this paper, we also adopt the simulation method in [8] to construct our training and testing datasets.

Recently, deep diffusion models have been proposed and developed for image generation [5]. As a likelihood-based algorithm, it is more stable during training than generative adversarial networks (GAN) and does not suffer from mode collapse. Diffusion models have shown significant success in various vision problems, like image synthesis [14] and super-resolution [14, 16]. In a very recent work, diffusion models are used for the atmospheric turbulence restoration of faces [12]. However, no published work has addressed the use of diffusion models in generic scenes AT correction. In this paper, we propose a diffusion model to remove atmospheric turbulence in generic scenes, producing results with great visual quality. In addition, we refer to a variational inference image restoration framework [17] to learn the latent features related to task-specific prior information from the input and the degradation process; we then inject this learned knowledge as a condition into the diffusion models. Therefore, the diffusion model is trained to adjust its behavior according to both the input degraded image and the task-specific prior information, which further enhances its performance.

In summary, our main contributions are: 1) We are the first to use diffusion models to solve the AT correction problem in generic scenes. 2) We propose to include a variational inference framework to provide a task-specific condition to the diffusion models. 3) We show that our proposed AT **variational deep diffusion** (AT-VarDiff) model generates results with outstanding visual quality evaluated both

© 20XX IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100007]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

quantitatively and qualitatively.

2. Method

In this section, we present the proposed AT-VarDiff model. We explain our model’s conditional denoising diffusion process in Section 2.1. In Section 2.2 we introduce the variational framework used to obtain the condition encoding task-specific information. Finally, in Section 2.3 we illustrate the inference during testing.

2.1. Conditional Diffusion Model

Our training dataset contains N image pairs $\{y_i, x_i\}_{i=1}^N$, where y_i represents the AT degraded image and x_i the corresponding ground-truth image. As shown in Figure 1, our model aims at learning the data distribution $p(x|y, c)$ by a stochastic iterative refinement process, which maps the input degraded image y and the learned latent prior information c to the ground-truth image x . The forward/diffusion process (from right to left) gradually adds Gaussian noise, denoted by $q(x_t|x_{t-1})$. Our goal is to reverse the diffusion process (from left to right) by gradually recovering the image from the input Gaussian noise with conditions, which corresponds to learning the reverse process of a fixed Markov Chain of length T conditioned on y and c . More specifically, starting from a pure Gaussian noise image $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the model learns the conditional transition distribution $p_\theta(x_{t-1}|x_t, y, c)$ and iteratively denoises the image for T steps, generating the target image x_0 in the end, such that $x_0 \sim p(x|y, c)$.

The overall training framework of the AT-VarDiff model is shown in Figure 2. Following the model design in denoising diffusion probabilistic model (DDPM) [5], the architecture of our conditional diffusion module is a U-Net [15] based on a wide ResNet [19], denoted as ϵ_θ . Training is performed by optimizing the usual variational bound on the negative log-likelihood, and the corresponding objective function can be simplified to [5, 14]:

$$\mathcal{L}_{diff} = \mathbb{E}_{x,y,c,\varepsilon,t} [\|\varepsilon - \epsilon_\theta(x_t, t, y, c)\|_2^2], \quad (1)$$

with t is uniformly sampled from $\{1, \dots, T\}$. According to Equation 1, ϵ_θ takes as input the noisy target image x_t , time step t , the AT degraded image y , and the learned latent prior information c to provide an estimate of the noise ε . The details of obtaining c are discussed in the following Section.

2.2. Variational Inference Framework

The current conditional diffusion models used for solving image restoration tasks like super-resolution [16, 14], and face AT correction [12] only use the input degraded

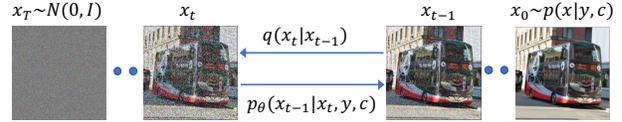


Figure 1. Conditional denoising diffusion process.

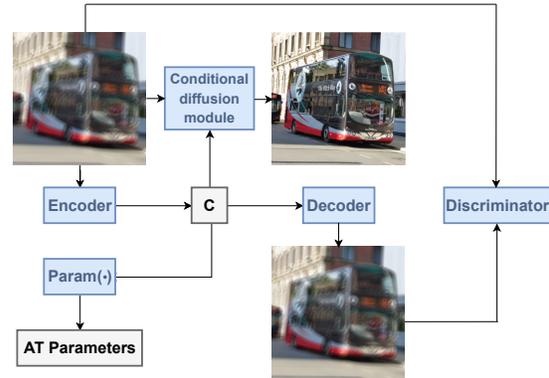


Figure 2. Training framework of AT-VarDiff model.

image as the condition. No work in the literature has employed any other task-specific prior information or domain-knowledge to further enhance the conditioning progress. According to [17], providing additional information can be interpreted as dividing a complex distribution into simpler sub-distributions that will eventually make network training easier and the results more accurate, since the number of possible solutions would be reduced. In this paper, we propose to use a variational inference framework to extract the latent task-specific prior information from the input and the degradation process and use the extracted feature as an additional condition to guide the diffusion model.

As shown in Figure 2, we refer to a variational autoencoder (VAE) based framework [17] to learn the latent feature c from the input degraded image y and the AT degradation parameters. To achieve this goal, the objective we use here contains three parts: the VAE loss, the adversarial loss, and the AT degradation parameters’ loss.

The VAE loss contains the fidelity term and the reconstruction term, that is,

$$\mathcal{L}_{vae} = D_{KL}(q_{e_\psi}(c|y)||p(c)) + \|y - \hat{y}\|_2^2. \quad (2)$$

The first term is the fidelity term; it measures the fidelity of c extracted from the encoder e_ψ , whose input is the degraded image y . It is represented as the KL divergence of the approximate posterior $q_{e_\psi}(c|y)$ from the prior $p(c)$. We select the prior $p(c)$ as a standard Gaussian distribution.

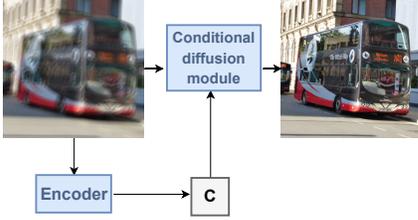


Figure 3. Testing framework of AT-VarDiff model.

The second term is the reconstruction term, and we adopt the pixel-wise mean squared error (MSE) distance between the input degraded image y and the output \hat{y} of the decoder d_φ . In addition, we utilize a GAN [3] to better learn the input degraded image distribution, in which an additional discriminator is jointly trained to discriminate the generated \hat{y} and the true degraded image y . Therefore, we also include an adversarial loss

$$\mathcal{L}_{adv} = -\log(D(\hat{y})), \quad (3)$$

and the corresponding loss for the discriminator D is

$$\mathcal{L}_{disc} = -\log(D(y)) - \log(1 - D(\hat{y})). \quad (4)$$

Finally, we would like the latent feature c to contain knowledge from the AT degradation process. Therefore, we add a degradation loss defined as:

$$\mathcal{L}_{degrad} = \|\phi_{at} - \hat{\phi}_{at}\|_2^2, \quad (5)$$

where ϕ_{at} represents the ground-truth AT degradation parameters from the pre-trained AT simulator [8]. $\hat{\phi}_{at} = Param(c)$ represents the estimated AT degradation parameters, and is the output of a small network (parameter estimation module) $Param(\cdot)$ taking c as input, as shown in Figure 2.

Therefore, the final objective used for training our AT-VarDiff model is defined as

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda_1 \mathcal{L}_{vae} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{degrad}, \quad (6)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters.

2.3. Inference

The testing framework of our model is shown in Figure 3. During testing, we followed the DDPM’s denoising sampling procedure (Algorithm 2 in [5]) conditioned on the input degraded image y and the learned task-specific latent feature c to generate the output restored image. During both training and testing, we perform the conditioning via concatenation, and we set $T = 1000$ for all the experiments.

Table 1. LPIPS & FID metrics comparison on simple-DDPM, AT-VarDiff, and AT-DDPM [12].

	AT-DDPM [12]	Simple-DDPM	AT-VarDiff (Ours)
LPIPS ↓	0.2150	0.1923	0.1094
FID ↓	80.05	60.87	32.69

3. Experiments

3.1 Experimental Settings

We use the AT simulator in [8] to simulate the effect of AT on the REDS dataset [11], and the hyper-parameter of the simulator (D/r_0) is chosen randomly in the range [0.5, 2.0]. Our synthetic training dataset has one million AT degraded and clean image pairs. We use another 2500 synthetic AT degraded images as the testing dataset.

For our encoder module, we use 5 2D-convolution (2D-conv) layers with ReLU activation and one down-sampling layer after the first conv layer. For the decoder module, we use 5 2D-conv layers with ReLU activation and one up-sampling layer after the first conv layer. For our parameter estimation module, we simply use 2 2D-conv layers with LeakyReLU activation. The discriminator is formed by 11 2D-conv layers with LeakyReLU activation and spectral normalization [10].

During training, we augment the training data by random cropping (160×160), random vertical and horizontal flips, and random transposing. We train our model for 200 epochs with 1500 iterations per epoch, and we set the batch size to 16. We use the Adam optimizer [6] with a weight decay of 0, and we set the initial learning rate to $1e - 4$ and gradually reduced it to $5e - 6$ during training utilizing the cosine annealing schedule [7]. The hyper-parameters λ_1, λ_2 , and λ_3 used in our final training objective (Equation 6) are set to 0.1, 0.1, and 0.5, respectively.

3.2 Results

To evaluate our model, we use the Fréchet Inception Distance (FID) [4] and the Learned Perceptual Image Patch Similarity (LPIPS) [20] metrics, which are measures of similarity between two sets of images. These two metrics are shown to correlate well with the human judgment of visual quality. In Table 1, we show the quantitative results of our proposed AT-VarDiff model and compare it to using a pure conditional DDPM-based diffusion model like the approach used in [12], i.e., this simple-DDPM model is only built with the conditional diffusion module and is only conditioned on the input degraded image y . We can see that our AT-VarDiff model improves on both metrics, demonstrating the effectiveness of our proposed variational conditional diffusion framework. We also compare with the pre-trained AT-DDPM model from [12] in the table. As can

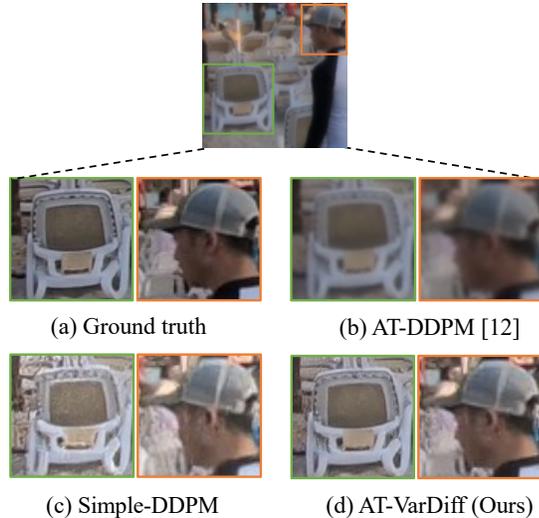


Figure 4. Visual comparisons of AT-DDPM [12], simple-DDPM, and AT-VarDiff.

be seen in Figure 4, our proposed approach achieves much better visual clarity, far fewer artifacts, and higher quality.

4. Conclusions

In this paper, we propose the variational deep diffusion model AT-VarDiff to restore images degraded by atmospheric turbulence. We propose to use the diffusion process to remove AT in generic scenes, and we use a variational inference framework to extract the latent task-specific prior information from the input and the AT degradation. We further inject extracted features as an additional condition to guide the diffusion model. We show that the proposed method achieves good results and outstanding visual quality, outperforming the current state-of-art. In the future, we will use more advanced diffusion techniques to further enhance the performance.

References

- [1] I. M. Author. Some related article I wrote. *Some Fine Journal*, 99(7):1–100, January 1999.
- [2] A. N. Expert. *A Book He Wrote*. His Publisher, Erewhon, NC, 1999.
- [3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [4] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [5] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [7] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [8] Z. Mao, N. Chimitt, and S. H. Chan. Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14759–14768, 2021.
- [9] Z. Mao, A. Jaiswal, Z. Wang, and S. H. Chan. Single frame atmospheric turbulence mitigation: A benchmark study and a new physics-inspired transformer model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 430–446. Springer, 2022.
- [10] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.
- [11] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [12] N. G. Nair, K. Mei, and V. M. Patel. At-ddpm: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3434–3443, 2023.
- [13] S. N. Rai and C. Jawahar. Removing atmospheric turbulence via deep adversarial learning. *IEEE Transactions on Image Processing*, 31:2633–2646, 2022.
- [14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [15] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [16] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [17] J. W. Soh and N. I. Cho. Variational deep image restoration. *IEEE Transactions on Image Processing*, 31:4363–4376, 2022.
- [18] G. Wang, J. C. Ye, and B. De Man. Deep learning for tomographic image reconstruction. *Nature Machine Intelligence*, 2(12):737–748, 2020.
- [19] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

- [20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [21] X. Zhu and P. Milanfar. Removing atmospheric turbulence via space-invariant deconvolution. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):157–170, 2012.