NeRT: Implicit Neural Representations for Unsupervised Atmospheric Turbulence Mitigation

Weiyun Jiang Vivek Boominathan Ashok Veeraraghavan
Rice University
{wyjiang, vivekb, vashok}@rice.edu

Method	CLEAR [1]	Mao et al. [10]	TurbuGAN [6]	TurbNet [12]	NDIR [9]	TSR-WGAN [8]	NeRT (Ours)
Physically grounded?	✔	✔	✔	✔	✗	✗	✔
Correct tilt-then-blur model?	✗	✗	✔	✔	✗	✗	✔
Unsupervised?	✔	✔	✔	✗	✔	✗	✔
Generalizable?	✔	✔	✗	✗	✔	✗	✔
Time-efficient?	~5 secs	10 secs~1 hr	~3 hrs	~1 sec	Anytime*	~1 sec	Anytime*

Figure 1. **How is our method, NeRT, different from other SOTA?** NeRT is the first generalizable unsupervised physically grounded model. *Section 5 discusses the anytime convergence property of NeRT.

Abstract

The atmospheric turbulence mitigation problem has emerged as a challenging inverse problem in the communities of computer vision and optics. However, current methods either rely heavily on the quality of the training dataset or fail to generalize over various scenarios, such as static scenes, dynamic scenes, and text reconstructions. We propose a novel implicit neural representation for unsupervised atmospheric turbulence mitigation (NeRT). NeRT leverages the implicit neural representations and the physically correct tilt-then-blur turbulence model to reconstruct the clean and undistorted image, given only dozens of distorted images. Further, we show that NeRT outperforms the state-of-the-art through various qualitative and quantitative evaluations. Lastly, we incorporate NeRT into continuously captured video sequences and demonstrate $48\times$ *speedup.*

1. Introduction

Atmospheric turbulence inevitably exists in long-range ground-based passive imaging systems. This unwanted phenomenon happens when light propagates in the form of waves through media with a nonuniform index of refraction [15]. If light waves simply propagate through free space with a uniform index of refraction, the imaging system will always capture clean and sharp images. On the other hand, when light propagates from a dense medium into a sparse medium, the light path will be refracted according to Snell's law. At first glance, it seems to be a very easy problem to solve. One might use Snell's law and ray tracing to simulate the entire light path, reconstructing the undistorted and sharp scene. However, the distribution of the index of refraction in the nonuniform medium is unknown, making it very hard to use such an approach. Many factors, such as distance, temperature, altitude, humidity, wind, and so on, might affect the degree of distortions. As a result, atmospheric turbulence is a highly challenging problem to solve with spatially and temporally varying blurring and tilting.

Existing deep learning approaches [6, 8, 10, 24] require a huge amount of training dataset. It is a challenging task, in the first place, to build a physically accurate and fast simulator [11]. In addition, by leveraging domain-specific priors, these existing deep learning approaches inevitably have dataset biases [5, 20, 23] and poor performance for out-of-domain distributions.

Classical non-deep learning approaches [1, 10, 17] do not require a huge amount of training dataset. However, they need to rely on optical flow or non-rigid registration techniques, such as a B-spline function, to model the grid deformation under atmospheric turbulence. Both optical flow and B-spline methods require a reference frame to start with.

The reference frames selected by these methods do not accurately represent the original sharp image. Mao et al. [10] select the sharpest frame, which still contains blurring and tilting, as the reference frame for the optical flow algorithm. Additionally, Shimizu et al. [17] select the naive average of all the distorted frames, which also contain noise, blurring, and tilting, as the reference frame.

To address the above challenges, we design a novel implicit neural representation for unsupervised atmospheric turbulence mitigation (NeRT), removing temporally and spatially tilting and blurring. The key idea is to constrain the network to learn the physically correct forward model, the tilt-then-blur model [2]. Inspired by NDIR [9], NeRT can model temporally and spatially tilting using deformed grids and implicit neural representations. For instance, if the implicit neural representations take uniform undistorted coordinates, they render clean and sharp images. If the implicit neural representations take distorted coordinates, they will output the corresponding distorted images under atmospheric turbulence.

The overall architecture of our learning framework, NeRT, is depicted in Figure 2. The network contains three major components, grid deformers $\mathcal{G}$ that estimate the spatially and temporally varying tilting at each pixel location, an image generator $\mathcal{I}$ that outputs pixel values at corresponding coordinates, and shift-varying blurring $\mathcal{P}$ that approximates the spatially blurring at each pixel location.

We perform extensive experiments on both real and synthetic atmospheric turbulence datasets. We show that NeRT outperforms the state-of-the-art supervised and unsupervised methods. Our specific contributions include as follows:

- We are the first to propose an unsupervised and physically grounded deep learning method for atmospheric turbulence mitigation. The pipeline follows the physically correct forward turbulence model, tilt-then-blur model [2].
- Our unsupervised algorithm is highly generalizable as it can recover clean and distortion-free images without domain-specific priors such as distorted-clean image pairs.
- We successfully deploy our method on real-time continuously captured video footage and achieve rapid convergence within 10 seconds on the latest frame.

2. Related work

Implicit neural representations. Implicit neural representations, which use multi-layer perceptions (MLPs) as the backbone networks, store 2D images [18, 19] and 3D shapes [13, 14] as continuous functions. The inputs of implicit neural representations are 2D or 3D coordinates, while the outputs are the corresponding signal. This kind of continuous representation shows not only extraordinary results in overfitting a single image or multiple images but also exceeds other state-of-the-art architectures in solving inverse problems, such as single-image superresolution [3], medical image reconstruction [16] and medical image registration [21]. Our work, NeRT, uses implicit neural representations as 2D image functions to render distorted images under atmospheric turbulence and clean, undistorted images.

Atmospheric turbulence mitigation. Many works have been proposed to undistort the effects of atmospheric turbulence. Some of the recent works have demonstrated the uses of transformer architectures as supervised turbulence removal networks for single frame [12] and mutiframe [24] atmospheric turbulence mitigation tasks. These proposed supervised transformer architectures must rely on fast and physically accurate simulators to generate a huge collection of paired distorted-clean image pairs for training datasets. Although they have fast inference speed, they are hard to generalize over out-of-the-domain datasets. Our proposed unsupervised learning architecture, NeRT, does not require any datasets for pretraining and, thus, can generalize over all kinds of datasets. TurbuGAN [6] proposes a self-supervised approach for imaging through turbulence by leveraging an adversarial learning framework and a fast turbulence simulator [11]. This approach requires no paired training datasets; however, it is hard to generalize domain-specific priors over out-of-the-domain distributions. NDIR [9] is the closest to our work. This method exploits convolutional neural networks to model non-rigid distortion. However, it relies on an off-the-shelf physically incorrect spatially invariant deblurring algorithm [22]. Our method, NeRT, incorporates a physically grounded, spatially and temporally varying deblurring approach to restoring the sharp image.

3. Physically grounded restoration network

3.1. Forward atmospheric turbulence model

Imagine the light reflected from a scene, represented as a clean and sharp image $\mathbf{J}$, travels through space with a spatially and temporally varying index of refraction. The light finally arrives at a passive imaging device, such as a digital single-lens reflex (DSLR) camera, forming many distorted images $\mathbf{I}$ over time. Each $\mathbf{I}$ has stochastic distortion at different pixel locations and time stamps. The generalized forward atmospheric turbulence model can be written as [2]:

$$\mathbf{I}(x, y, t) = \mathcal{H}_t(\mathbf{J}(x, y, t)), \tag{1}$$

where $\mathcal{H}$ is a general linear distortion operator. However, we desire to decompose $\mathcal{H}$ into simpler operations for the computational tractability of inverting the distortion. For-

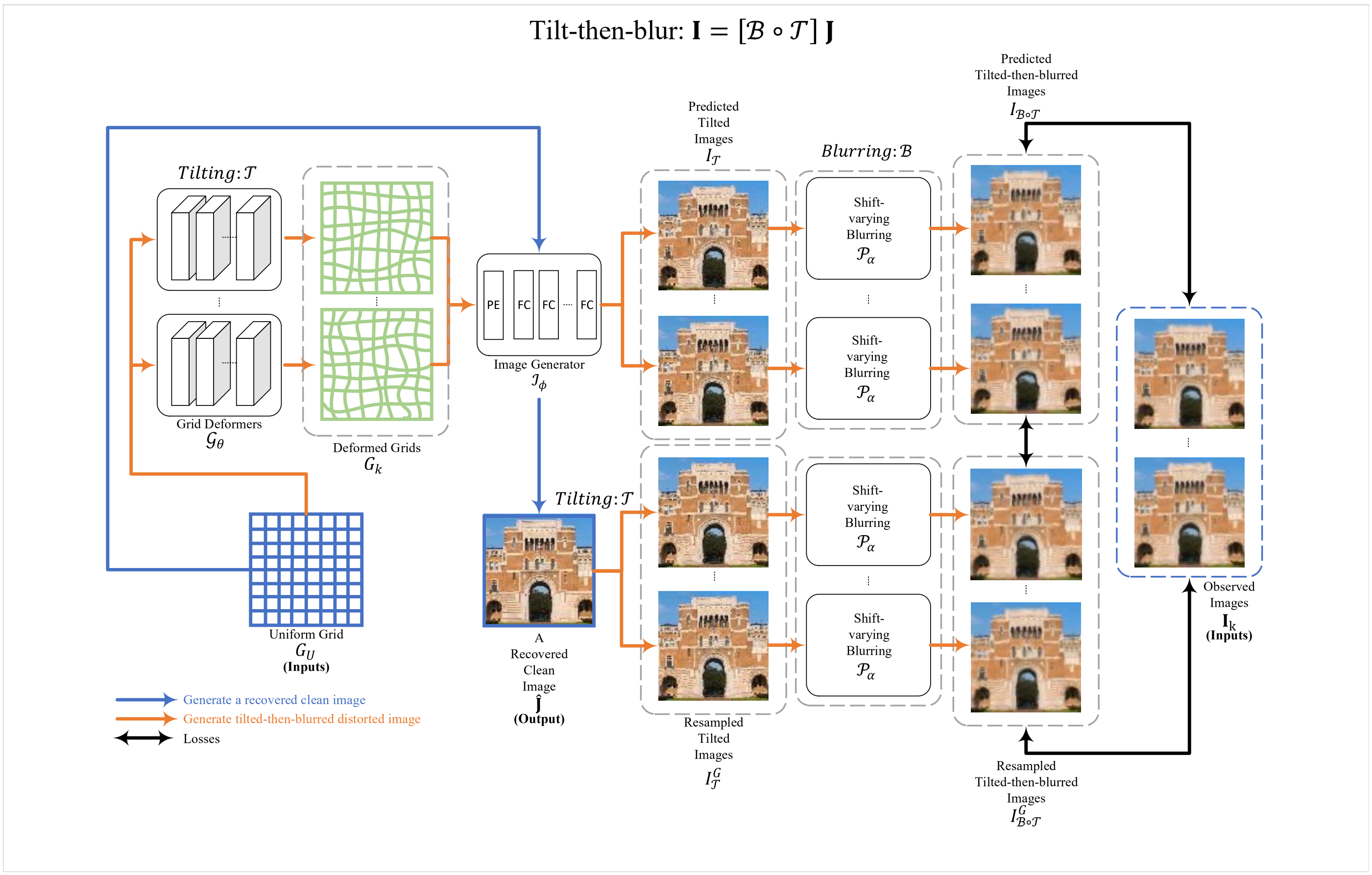

Figure 2. **The overall architecture of NeRT.** The network predicts a clean and sharp image $\hat{\mathbf{J}}$, given a series of observed atmospheric turbulence distorted images $\mathbf{I}$. We compute L_1 loss between predicted tilted-then-blurred images $I_{\mathcal{B}\circ\mathcal{T}}$, resampled tilted-then-blurred images $I^G_{\mathcal{B}\circ\mathcal{T}}$, and observed images $\mathbf{I}$ during optimization to update the parameters in image generators $\mathcal{I}_\phi$, grid deformers $\mathcal{G}_\theta$, and shift-varying blurring $\mathcal{P}_\alpha$.

tunately, representation using Zernike bases allows us to do so.

The distortion from atmospheric turbulence can be parameterized by coefficients of Zernike polynomials in the phase space [4]. The representation in the Zernike space allows us to decouple the distortion into interpretable operations of tilting operation ($\mathcal{T}$) and blurring operation ($\mathcal{B}$). The tilt $\mathcal{T}$ is encoded by the first two Zernike bases (barring the constant term) and is the most significant contributor to shifting the center of mass of the distortion spread. The blur $\mathcal{B}$ is encoded by the remaining Zernike bases and describes the distortion spread. The question now is how to compose the decoupled operations to accurately describe the true atmospheric distortion. There are two possible options:

$$\text{Blur-then-tilt: } \mathbf{I} = [\mathcal{T} \circ \mathcal{B}]\mathbf{J},$$
$$\text{Tilt-then-blur: } \mathbf{I} = [\mathcal{B} \circ \mathcal{T}]\mathbf{J},$$

where $\circ$ is a functional composition operator, and the composition is read from right to left.

Many of the previous works [1,10,25] opt to use the blur-then-tilt model, whose inversion is to untilt first and then to deblur. They choose this route because it is relatively easier to estimate the tilt first using well-known computations such as optical flow [10]. Then, after untilting, an off-the-shelf deblurring algorithm is used to deblur.

However, Chan [2] showed, by careful analysis, that the blur-then-tilt model is inaccurate and that the tilt-then-blur model is physically more accurate. In our work, we use the correct model of tilt-then-blur to represent our forward model, which is given as follows [2]:

$$\mathbf{I}(x, y, t) = [\mathcal{B} \circ \mathcal{T}]\mathbf{J}(x, y, t), \tag{2}$$

where $\mathcal{B}$ denotes a temporally and spatially varying blurring operator, $\mathcal{T}$ represents a temporally and spatially varying tilting operator and $\circ$ is a functional composition operator. We apply spatially and temporally varying tilting operators to the clean image $\mathbf{J}(x, y)$ first to obtain multiple tilted images $I_{\mathcal{T}}$ at different time stamps. Then, we apply spatially varying blurring operators to those tilted images $I_{\mathcal{T}}$ to render corresponding final tilted-then-blurred images $\mathbf{I} = I_{\mathcal{B}\circ\mathcal{T}}$ under atmospheric turbulence.

3.2. Why choose tilt-then-blur?

Although the two compositions $\mathcal{B}\circ\mathcal{T}$ (tilt-then-blur) and $\mathcal{T}\circ\mathcal{B}$ (blur-then-tilt) are analytically different, their impacts on images of natural scenes tend to be similar, and the differences might be imperceptible [2]. However, the errors in the incorrect $\mathcal{T}\circ\mathcal{B}$ model can quickly accumulate at the edges and high-resolution regions, as shown in the analysis below.

For simplicity, let us assume that the blurring is spatially invariant. We can write the equations of the two models as [2]:

$$I_{\mathcal{T}\circ\mathcal{B}} = \sum_{j=1}^{N} g(\boldsymbol{x}_i - \boldsymbol{u}_j)\boldsymbol{J}(\boldsymbol{u}_j - \boldsymbol{t}_i), \tag{3}$$

$$I_{\mathcal{B}\circ\mathcal{T}} = \sum_{j=1}^{N} g(\boldsymbol{x}_i - \boldsymbol{u}_j)\boldsymbol{J}(\boldsymbol{u}_j - \boldsymbol{t}_j), \tag{4}$$

where g is the spatially invariant blur, and t is the tilt. Note that the subtlety is captured in the indexing of t. More details regarding the derivation can be found in [2].

We may now evaluate the difference between the correct tilt-then-blur model and the incorrect blur-then-tilt model as

$$\begin{aligned} &I_{\mathcal{T}\circ\mathcal{B}} - I_{\mathcal{B}\circ\mathcal{T}} \\ &\quad= \sum_{j=1}^{N} g(\boldsymbol{x}_i - \boldsymbol{u}_j)[\boldsymbol{J}(\boldsymbol{u}_j - \boldsymbol{t}_i) - \boldsymbol{J}(\boldsymbol{u}_j - \boldsymbol{t}_j)] \\ &\quad\approx \sum_{j=1}^{N} g(\boldsymbol{x}_i - \boldsymbol{u}_j)\nabla\boldsymbol{J}(\boldsymbol{u}_j^T)(\boldsymbol{t}_i - \boldsymbol{t}_j), \end{aligned} \tag{5}$$

where $\nabla\boldsymbol{J}(\boldsymbol{u}_j^T)$ stands for the image gradient, and $\boldsymbol{t}_i - \boldsymbol{t}_j$ represents the random tilt. For natural scene images, the image gradients $\nabla\boldsymbol{J}(\boldsymbol{u}_j^T)$ are typically sparse, and the error between the two models is close to zero for most of the image regions. However, the image gradients are strong at edges and high-resolution regions, and there will be a significant error between the two models. Thus, it is sub-optimal to solve atmospheric mitigation problems following the incorrect blur-then-tilt model. Using the correct tilt-then-blur model gives us the opportunity to recover edges and high-resolution details from a time window of dynamically distorted frames.

3.3. Network Structure

Figure 2 demonstrates the architecture of NeRT. Our model has three major components, grid deformers, image generators, and shift-varying blurring.

Grid deformers $\mathcal{G}_\theta$ take uniform grid $G_U \in \mathbb{R}^{2\times m\times n}$ as inputs, where m and n are the image sizes, and output deformed grid, $G_{\mathcal{T}} \in \mathbb{R}^{2\times m\times n}$. Like NDIR [9], each grid deformer $\mathcal{G}_\theta$ consists of four convolutional layers with 256 channels and ReLU activation layers. Similarly, we have a dedicated grid deformer $\mathcal{G}_\theta$ for each distorted image $\mathbf{I} \in \mathbb{R}^{3\times m\times n}$.

Image generator $\mathcal{I}_\phi$ take deformed 2D pixel coordinates $G_{\mathcal{T}} \in \mathbb{R}^{m\times n\times 2}$ as inputs, and output 3D RGB pixel value that corresponds to the tilted images $I_{\mathcal{T}} \in \mathbb{R}^{m\times n\times 3}$. It can also take uniform 2D pixel coordinates $G_U \in \mathbb{R}^{m\times n\times 2}$ as inputs and output colored pixel value of the clean image $\mathbf{J} \in \mathbb{R}^{m\times n\times 3}$. We build our image generator as an implicit representation, which consists of five layers of fully connected layers of hidden size 256 with ReLU activation and positional encoding. We implement our coordinate-based MLPs following previous work, SIREN [18] and Fourier feature network [19]. Specifically, we reshape the 2D pixel coordinates $G \in \mathbb{R}^{m\times n\times 2}$ as $G \in \mathbb{R}^{m\cdot n\times 2}$ to parse into the coordinate-based MLPs. Additionally, we reshape the output of the coordinate-based MLPs, corresponding 3D RGB pixel value, $\mathbf{J} \in \mathbb{R}^{m\cdot n\times 3}$ as $\mathbf{J} \in \mathbb{R}^{m\times n\times 3}$.

Shift-varying blurring $\mathcal{P}_\alpha$ take generated tilted images $I_{\mathcal{T}} \in \mathbb{R}^{m\times n\times 3}$ as inputs and output generated tilted-then-blurred images $I_{\mathcal{B}\circ\mathcal{T}} \in \mathbb{R}^{m\times n\times 3}$. Shift-varying blurring leverages the Phase-to-Space (P2S) transform [11] to apply pixel-wise spatially and temporally varying blurring. We initialize per-pixel correlated Zernike coefficients $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ by multiplying independent and identically distributed Gaussian vectors with pre-computed correlation matrices. In addition, we apply P2S transform network during the optimization to convert Zernike coefficients $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ to PSF basis coefficients $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_K]$. Together with the pre-computed PSF basis, we are able to use the converted basis coefficient $\boldsymbol{\beta}$ to compute spatially and temporally varying PSF for the shift-varying blurring operation. Note that we don't consider the first two Zernike bases since they are already accounted for as tilt by the grid deformers.

3.4. Parameter initialization

Since our method is unsupervised, parameter initialization is rather important. A good initial point could help our model avoid saddle points and local minimums during the optimization. First, D/r_0, which characterizes the strength of the atmospheric turbulence and typically ranges from 1.0 to 5.0, determines the variance of the i.i.d. Gaussian vector during α initialization. Higher D/r_0 means stronger atmospheric turbulence. When the observed images have relatively high turbulence strength, our model tends to converge better with a higher D/r_0. Second, *corr* [11] refers to how correlated these nearby PSFs are and typically ranges from

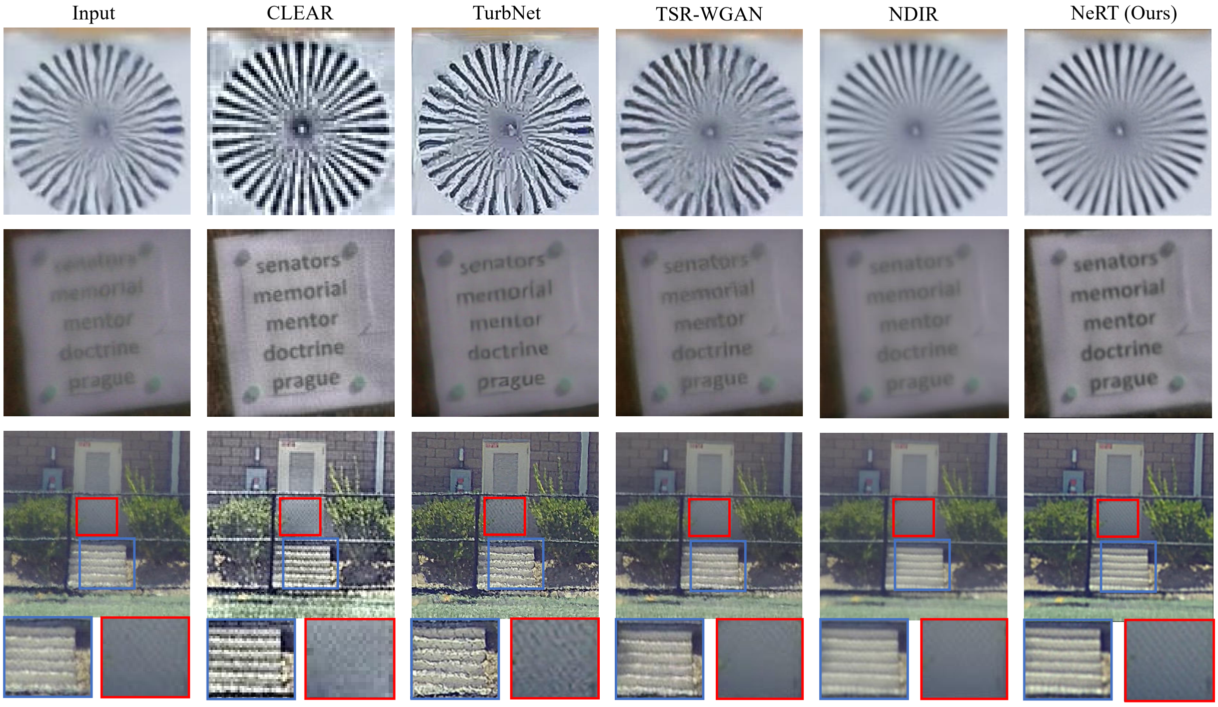

Figure 3. **Qualitative results from the static scene test datasets, including Siemens star dataset [7], text dataset [12], and door dataset [7].** We compare NeRT with other supervised [12] and unsupervised [1, 9] SOTA. NeRT is able to achieve high spatial resolution, recover high-contrast text, and reconstruct fine details, such as wire fences. CLEAR [1], TSR-WGAN [8], and TurbNet [12] fails to mitigate the atmospheric turbulence, while NDIR [9] fails to preserve the wire fences.

-5 to -0.01. A higher value means a stronger correlation. When the observed images have relatively high turbulence strength, our model tends to converge better with a higher $corr$. Third, the kernel size of the PSF basis should vary as the size of the image varies. Large image dimensions usually require a large kernel size of the PSF basis.

3.5. Two-step optimization

We follow the network initialization in NDIR [9]. During the first initialization step, the grid deformers $\mathcal{G}_\theta$ are constrained to learn an identity mapping from uniform grid $\mathbf{G_U}$ to be close to the uniform grid $\mathbf{G_U}$. In this way, we can limit the grid deformation from extreme pixel mixing. The image generator is forced to learn an average of all the distorted input images. The loss function of the first initialization step is formulated as

$$\min_{\theta,\phi} \sum_k \|\mathcal{G}_\theta^k(G_U) \ - G_U\|_1 + \|\mathcal{I}_\phi(G_U) - \mathbf{I}_k\|_1. \tag{6}$$

During the second iterative optimization step, we initialize the $\boldsymbol{\alpha}$ in shift-varying blurring choosing appropriate D/r_0, $corr$, and PSF kernel size. The loss function is then formulated as

$$\begin{aligned}\min_{\theta,\phi,\alpha} \sum_k \|\mathcal{P}_\alpha(\mathcal{I}_\phi(\mathcal{G}_\theta^k(G_U))) - \boldsymbol{I}_k\|_1 \\ +\|\mathcal{P}_\alpha(I_T^G) - \boldsymbol{I}_k\|_1 \\ +\|\mathcal{P}_\alpha(\mathcal{I}_\phi(\mathcal{G}_\theta^k(G_U))) - \mathcal{P}_\alpha(I_T^G)\|_1,\end{aligned} \tag{7}$$

where I_T^G is a resampled tilted image given deformed grids G_k. We would like to enforce consistency between predicted tilted-then-blurred images, observed images, and resampled tilted-then-blurred images.

4. Experiments and results

In this section, we compare NeRT with other state-of-the-art supervised and unsupervised methods, such as CLEAR [1], TurbNet [12], TSR-WGAN [8] and NDIR [9] on both real and synthetic atmospheric turbulence mitigation datasets. CLEAR [1] is an optimization-based multi-frame restoration method. TurbNet [12] is a deep learning-based single-frame restoration method. TSR-WGAN [8] and NDIR [9] are deep learning-based multi-frame restoration methods. We show that NeRT exhibits superior performance in both qualitative and quantitative assessments

Table 1. **Quantitative performance on synthetic dataset created using simulator [11].** ↑ means the higher the better.

Strength	Metric	CLEAR [1]	TurbNet [12]	TSR-WGAN [8]	NDIR [9]	NeRT (Ours)
Weak	PSNR ↑ (dB)	20.164	20.532	20.428	21.366	**22.109**
($D/r_0 = 1.5$)	SSIM ↑	0.704	0.601	0.600	0.716	**0.766**
Medium	PSNR ↑ (dB)	19.341	18.220	18.811	19.606	**20.576**
($D/r_0 = 3$)	SSIM ↑	0.611	0.440	0.457	0.603	**0.659**
Strong	PSNR ↑ (dB)	17.715	17.786	17.198	18.812	**19.311**
($D/r_0 = 4.5$)	SSIM ↑	0.488	0.512	0.347	0.544	**0.567**

when compared to the state-of-the-art unsupervised and supervised methods.

4.1. Implementation details

We implement our model in Pytorch with one NVIDIA A100 80GB GPU. We use Adam optimizer with a learning rate of 1×10^{-4} to update parameters in grid deformers $\mathcal{G}_\theta$, image generator $\mathcal{I}_\phi$ and shift-varying blurring $\mathcal{P}_\alpha$. We use 1000 epochs for both the first initialization step and the second iterative optimization step. We empirically choose $D/r_0 = 5.0$, $corr = -5.0$, and a PSF kernel size of 11 for all experiments. We resize the dimensions of all the distorted input images to 256 × 256. We randomly choose 20 distorted images as input for the experiments.

4.2. Evaluation on synthetic datasets

We use the P2S atmospheric turbulence simulator [11] to create our synthetic distorted video sequences for evaluation. We choose three different levels of turbulence strength. We use $D/r_0 = 1.5$ as weak turbulence, $D/r_0 = 3$ as medium turbulence, and $D/r_0 = 4.5$ as strong turbulence. Table 1 demonstrates the quantitative comparison between our method, NeRT, and other SOTA. NeRT outperforms other methods in terms of peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) and is robust to different turbulence strengths.

4.3. Evaluation on real datasets

Static scenes. We include three real datasets for static scenes, the Siemens start dataset [7], the text dataset [12], and the door dataset [7]. Figure 3 shows the qualitative results of these static scenes. NeRT achieves the best overall performance in terms of spatial resolution, high-contrast text reconstruction, and fine details recovery. Further, NeRT is able to preserve the fine details, such as the wire fences, while suppressing the blurring and tilting caused by atmospheric turbulence.

Dynamic scenes. We present the moving car dataset [1] for dynamic scenes. The distorted input image sequences depict a car moving from back to front and from left to right. Figure 4 presents the qualitative results from the dynamic scene. Again we compare with other SOTA. NeRT is able to recover a high-contrast license plate with higher fidelity. To handle the dynamic scene, the image generator $\mathcal{I}_\alpha$ in our unsupervised model converges to a reference frame as a starting point during the first initialization step. During the second iterative optimization step, the clean image generated by the image generator $\mathcal{I}_\alpha$ is further optimized given dozens of distorted images.

5. Anytime reconstruction of continuous video frames

Imagine some ground-based imaging systems that capture long-range video sequences continuously. Every second, these passive imaging systems would capture some

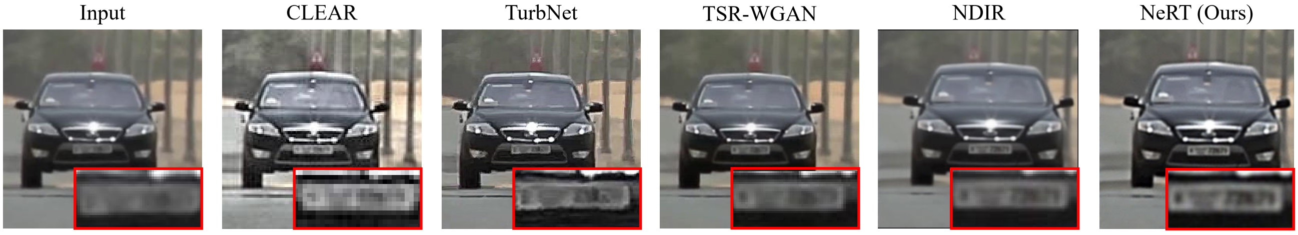

Figure 4. **Qualitative results from the dynamic scene test dataset, moving car dataset [1].** We compare NeRT with other supervised [8, 12] and unsupervised [1, 9] SOTA. We are able to recover high-contrast and fine details of the license plate while other methods show blurry and low-contrast license plate numbers. NDIR [9], TSR-WGAN [8], and NeRT choose to recover the clean image based on the newest frame while CLEAR [1] and TurbNet [12] decide to recover the clean image based on the oldest frame.

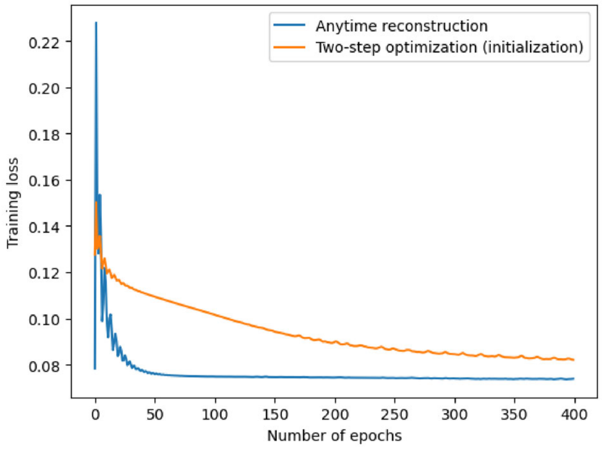

Figure 5. **NeRT converges $\sim$ 48× faster after two-step optimization (initialization).** It takes a total of 2000 epochs ($\sim$ 8 minutes) to converge during the two-step optimization stage while it only takes 60 epochs ($\sim$ 10 seconds) to converge during anytime reconstruction.

latest video frames under the effect of atmospheric turbulence. NeRT is optimal at recovering these continuously captured video sequences because it can leverage the tiltings and blurrings from previously captured video frames to reconstruct the latest captured scene. We name the use of the newest video frame for atmospheric turbulence mitigation, together with all the previous video frames, ”anytime” reconstruction.

Our method NeRT, like NDIR [9], has a separate grid deformer $\mathcal{G}_\theta$ and a separate shift-varying blurring $\mathcal{P}_\alpha$ for each distorted input image, and shares a single image generator $\mathcal{I}_\phi$ across all the distorted images. During anytime reconstruction, we simply initialize a new separate grid deformer $\mathcal{G}_\theta$ and shift-varying blurring $\mathcal{P}_\alpha$ for the most recent frame that is captured. All the other grid deformers $\mathcal{G}_\theta$, shift-varying blurring $\mathcal{P}_\alpha$ and the image generator $\mathcal{I}_\phi$ can contain all the information of the previously observed distorted frame, speeding up the anytime reconstruction.

Figure 5 demonstrates 48× speedup of our method during anytime convergence. It takes about 8 minutes to complete the two-step initialization step. However, it only takes about 10 seconds to converge for every newly captured video frame.

6. Conclusions and discussions

We have proposed the first unsupervised and physically grounded model for atmospheric turbulence mitigation. Given multiple observed distorted images, our model leveraged the physically correct tilt-then-blur model to reconstruct a clean and undistorted image. Our model could generalize and outperform other SOTA in various scenarios, such as static scenes, dynamic scenes, and text reconstructions. Our method converged 48× faster on the latest captured frame after the two-step initialization.

Limitations and future directions. Our shift-varying deblurring did not have any regularization. Thus, the reconstructed clean image inevitably consisted of some noise due to blind deconvolution. A more sophisticated shift-varying deblurring process remains a future research direction. Additionally, one might also leverage the implicit neural network for image superresolution. As we know, the implicit neural network is a continuous representation of the image. More pixel coordinates queried into the implicit image function lead to higher resolution images generated.

Acknowledgements. This work was supported by NSF CAREER: IIS-1652633 and NSF Expeditions in Computing: IIS-1730574.

References

[1] Nantheera Anantrasirichai, Alin Achim, and David Bull. Atmospheric turbulence mitigation for sequences with moving objects using recursive image fusion. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 2895–2899. IEEE, 2018. 1, 3, 5, 6

[2] Stanley H Chan. Tilt-then-blur or blur-then-tilt? clarifying the atmospheric turbulence model. *IEEE Signal Processing Letters*, 29:1833–1837, 2022. 2, 3, 4

[3] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 2

[4] Nicholas Chimitt and Stanley H Chan. Simulating anisoplanatic turbulence by sampling intermodal and spatially correlated zernike coefficients. *Optical Engineering*, 59(8):083101–083101, 2020. 3

[5] Elizabeth Cole, Qingxi Meng, John Pauly, and Shreyas Vasanawala. Learned compression of high dimensional image datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1748–1752, 2022. 1

[6] Brandon Y Feng, Mingyang Xie, and Christopher A Metzler. Turbugan: An adversarial learning approach to spatially-varying multiframe blind deconvolution with applications to imaging through turbulence. *IEEE Journal on Selected Areas in Information Theory*, 2023. 1, 2

[7] Jérôme Gilles and Nicholas B Ferrante. Open turbulent image set (otis). *Pattern Recognition Letters*, 86:38–41, 2017. 5, 6

[8] Darui Jin, Ying Chen, Yi Lu, Junzhang Chen, Peng Wang, Zichao Liu, Sheng Guo, and Xiangzhi Bai. Neutralizing the impact of atmospheric turbulence on complex scene imaging via deep learning. *Nature Machine Intelligence*, 3(10):876–884, 2021. 1, 5, 6

[9] Nianyi Li, Simron Thapa, Cameron Whyte, Albert W Reed, Suren Jayasuriya, and Jinwei Ye. Unsupervised non-rigid

image distortion removal via grid deformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2522–2532, 2021. 2, 4, 5, 6, 7

[10] Zhiyuan Mao, Nicholas Chimitt, and Stanley H Chan. Image reconstruction of static and dynamic scenes through anisoplanatic turbulence. *IEEE Transactions on Computational Imaging*, 6:1415–1428, 2020. 1, 2, 3

[11] Zhiyuan Mao, Nicholas Chimitt, and Stanley 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. 1, 2, 4, 6

[12] Zhiyuan Mao, Ajay Jaiswal, Zhangyang Wang, and Stanley 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. 2, 5, 6

[13] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2

[14] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[15] Michael C Roggemann, Byron M Welsh, and Bobby R Hunt. *Imaging through turbulence*. CRC press, 1996. 1

[16] Liyue Shen, John Pauly, and Lei Xing. Nerp: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[17] Masao Shimizu, Shin Yoshimura, Masayuki Tanaka, and Masatoshi Okutomi. Super-resolution from image sequence under influence of hot-air optical turbulence. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1, 2

[18] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 2, 4

[19] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2, 4

[20] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 1

[21] Jelmer M Wolterink, Jesse C Zwienenberg, and Christoph Brune. Implicit neural representations for deformable image registration. In *International Conference on Medical Imaging with Deep Learning*, pages 1349–1359. PMLR, 2022. 2

[22] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1107–1114, 2013. 2

[23] Tuo Zhang, Tiantian Feng, Samiul Alam, Sunwoo Lee, Mi Zhang, Shrikanth S Narayanan, and Salman Avestimehr. Fedaudio: A federated learning benchmark for audio tasks. *arXiv preprint arXiv:2210.15707*, 2022. 1

[24] Xingguang Zhang, Zhiyuan Mao, Nicholas Chimitt, and Stanley H Chan. Imaging through the atmosphere using turbulence mitigation transformer. *arXiv preprint arXiv:2207.06465*, 2022. 1, 2

[25] Xiang Zhu and Peyman Milanfar. Removing atmospheric turbulence via space-invariant deconvolution. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):157–170, 2012. 3