Single Frame Atmospheric Turbulence Mitigation: A Benchmark Study and A New Physics-Inspired Transformer Model

Zhiyuan Mao[1*], Ajay Jaiswal[2*], Zhangyang Wang[2], and Stanley H. Chan[1]

[1] Purdue University, West Lafayette IN 47907, USA
[2] University of Texas at Austin, Austin TX 78712, USA

Abstract. Image restoration algorithms for atmospheric turbulence are known to be much more challenging to design than traditional ones such as blur or noise because the distortion caused by the turbulence is an entanglement of spatially varying blur, geometric distortion, and sensor noise. Existing CNN-based restoration methods built upon convolutional kernels with static weights are insufficient to handle the spatially dynamical atmospheric turbulence effect. To address this problem, in this paper, we propose a physics-inspired transformer model for imaging through atmospheric turbulence. The proposed network utilizes the power of transformer blocks to jointly extract a dynamical turbulence distortion map and restore a turbulence-free image. In addition, recognizing the lack of a comprehensive dataset, we collect and present two new real-world turbulence datasets that allow for evaluation with both classical objective metrics (e.g., PSNR and SSIM) and a new task-driven metric using text recognition accuracy. The code and datasets are available at github.com/VITA-Group/TurbNet.

Keywords: atmospheric turbulence mitigation, image restoration

1 Introduction

In long-range imaging systems, atmospheric turbulence is one of the main sources of distortions that causes geometric displacements of the pixels and blurs. If unprocessed, the distorted images can have significant impacts on all downstream computer vision tasks such as detection, tracking, and biometric applications. The atmospheric turbulence effects are substantially harder to model and mitigate compared to the commonly seen image degradations such as deconvolution, as the turbulence is an entanglement of pixel displacement, blur, and noise. As a result, a dedicated image restoration pipeline is an essential element for long-range computer vision problems.

Image processing algorithms for mitigating the atmospheric turbulence effect have been studied for decades [1, 21, 12, 14, 40, 36, 10, 20, 23, 18]. However, many of them have limitations that prohibit them from being launched to practical

[*] Equal contribution.

systems: 1) Many of the existing algorithms [1, 21, 12, 40, 1, 10] are based on the principle of *lucky imaging* that requires multiple input frames. These methods often have a strong assumption that both the camera and the moving objects are static, which can easily become invalid in many real applications. 2) The conventional algorithms are often computationally expensive, making them unsuitable for processing large-scale datasets to meet the need of the latest computer vision systems. 3) Existing deep learning solutions [36, 23, 14] are not utilizing the physics of the turbulence. Many of them are also tailored to recovering faces instead of generic scenes. The generalization is therefore a question. 4) The algorithms may not be properly evaluated due to the absence of a widely accepted real large-scale benchmarking dataset.

To articulate the aforementioned challenges, in this paper we make three contributions:

1. We present a comprehensive benchmark evaluation of deep-learning based image restoration algorithms through atmospheric turbulence. We tune a sophisticated physics-grounded simulator to generate a large-scale dataset, covering a broad variety of atmospheric turbulence effects. The highly realistic and diverse dataset leads to exposing shortages of current turbulence mitigation algorithms.
2. Realizing the existing algorithms' limitations, we introduce a novel physics-inspired turbulence restoration model, termed *TurbNet*. Built on a transformer backbone, *TurbNet* features a modularized design that targets modeling the spatial adaptivity and long-range dynamics of turbulence effects, plus a self-supervised consistency loss.
3. We present a variety of evaluation regimes and collect two large-scale real-world turbulence *testing* datasets, one using the heat chamber for classical objective evaluation (e.g., PSNR and SSIM), and one using real long-range camera for optical text recognition as a semantic "proxy" task. Both of the new testing sets will be released.

2 Related Works

Turbulence mitigation methods. The atmospheric turbulence mitigation methods have been studied by the optics and vision community for decades. To reconstruct a turbulence degraded image, conventional algorithms [1, 21, 12, 40, 10, 8, 34, 9] often adopt the multi-frame image reconstruction strategy. The key idea is called "lucky imaging", where the geometric distortion is first removed using image registration or optical flow techniques. Sharper regions are then extracted from the aligned frames to form a lucky frame. A final blind deconvolution is usually needed to remove any residue blur. These methods are usually very computationally expensive. The time required to reconstruct a 256×256 image may range from a few seconds to tens of minutes. Despite the slow speed that prohibits them from being applied in real-world applications, the performance of conventional methods is often consistent across different image contents.

Recent deep learning methods adopt more dynamic strategies. Li et al. [18] propose to treat the distortion removal as an unsupervised training step. While it can effectively remove the geometric distortions induced by atmospheric turbulence, its computational cost is comparable to conventional methods, as it needs to repeat the training step for each input image. There are also several works that focus on specific types of images, such as face restoration [36, 23, 14]. They are usually based on a simplified assumption on atmospheric turbulence where they assume the blur to be spatially invariant. Such assumption cannot extend to general scene reconstruction, where the observed blur can be highly spatially varying due to a wide field of view.

There also exists general image processing methods, such as [38, 37]. They have demonstrated impressive performance on restoration tasks, including denoising, deblurring, dehazing, etc. However, whether they can be extended to turbulence mitigation remains unclear as turbulence evolves more complicated distortions.

Available datasets. Despite recent advances in turbulence mitigation algorithms, there is a very limited amount of publicly available datasets for atmospheric turbulence. The most widely used testing data are two images, the *Chimney* and *Building* sequences released in [10]. Besides, authors of [21, 1, 18] have released their own testing dataset, each of which often consists of less than 20 images. These data are then seldom used outside the original publications. Additionally, the scale of these datasets is not suitable for evaluating modern learning-based methods.

Due to the nature of the problem, it is very difficult to obtain aligned clean and corrupted image pairs. Existing algorithms are all trained with synthetic data. The computationally least expensive synthesis technique is based on the random pixel displacement + blur model [15, 13]. In the optics community, there are techniques based on ray-tracing and wave-propagation [25, 27, 7]. A more recent physics-based simulation technique based on the collapsed phase-over-aperture model and the phase-to-space transform is proposed in [3, 22]. Our data synthesis scheme is based on the *P2S* model provided by authors of [22].

3 Restoration Model

3.1 Problem Setting and Motivation

Consider a clean image $\mathbf{I}$ in the object plane that travels through the turbulence to the image plane. Following the classical split-step wave-propagation equation, the resulting image $\widetilde{\mathbf{I}}$ is constructed through a sequence of operations in the phase domain:

$$\mathbf{I} \rightarrow \text{Fresnel} \rightarrow \text{Kolmogorov} \rightarrow \cdots \text{Fresnel} \rightarrow \text{Kolmogorov} \rightarrow \widetilde{\mathbf{I}}, \tag{1}$$

where "Fresnel" represents the wave propagation step by the Fresnel diffraction, and "Kolmogorov" represents the phase distortion due to the Kolmogorov power spectral density [11].

Certainly, Eqn. 1 is implementable as a forward equation (ie for simulation) but it is nearly impossible to be used for solving an inverse problem. To mitigate this modeling difficulty, one computationally efficient approach is to approximate the turbulence as a composition of two processes:

$$\widetilde{\mathbf{I}} = \Big(\underbrace{\mathcal{H}}_{\text{blur}} \circ \underbrace{\mathcal{G}}_{\text{geometric}} \Big)(\mathbf{I}) + \mathbf{N}, \tag{2}$$

where $\mathcal{H}$ is a convolution matrix representing the spatially *varying* blur, and $\mathcal{G}$ is a mapping representing the geometric pixel displacement (known as the tilt). The variable $\mathbf{N}$ denotes the additive noise / model residue in approximating Eqn. 1 with a simplified model. The operation "$\circ$" means the functional composition. That is, we first apply $\mathcal{G}$ to $\mathbf{I}$ and then apply $\mathcal{H}$ to the resulting image.

We emphasize that Eqn. 2 is only a mathematically convenient way to derive an approximated solution for the inverse problem but not the true model. The slackness falls into the fact that the pixel displacement in $\mathcal{G}$ across the field of view are correlated, so do the blurs in $\mathcal{H}$. The specific correlation can be referred to the model construction in the phase space, for example [3]. In the literature, Eqn. 2 the shortcoming of this model is recognized, although some successful algorithms can still be derived [1, 40].

The simultaneous presence of $\mathcal{H}$ and $\mathcal{G}$ in Eqn. 2 makes the problem hard. If there is only $\mathcal{H}$, the problem is a simple deblurring. If there is only $\mathcal{G}$, the problem is a simple geometric unwrapping. Generic deep-learning models such as [36, 23] adopt network architectures for classical restoration problems based on conventional CNNs, which are developed for one type of distortion. Effective, their models treat the problem as

$$\widetilde{\mathbf{I}} = \mathcal{T}(\mathbf{I}) + \mathbf{N}, \tag{3}$$

where $\mathcal{T} = \mathcal{G} \circ \mathcal{H}$ is the overall turbulence operator. Without looking into how $\mathcal{T}$ is constructed, existing methods directly train a generic restoration network by feeding it with noisy-clean training pairs. Since there is no physics involved in this generic procedure, the generalization is often poor.

Contrary to previous methods, in this paper, we propose to jointly estimate the physical degradation model $\mathcal{T}$ of turbulence along with reconstruction of clean image from the degraded input $\widetilde{\mathbf{I}}$. Such formulation explicitly forces our model to focus on learning a generic turbulence degradation operator independent of image contents, along with the reconstruction operation to generate clean output. Moreover, our network training is assisted by high-quality, large-scale, and physics-motivated synthetic training data to better learn the key characteristics of the atmospheric turbulence effect. The detailed model architecture will be presented in the following subsection.

3.2 Model Architecture

Turbulence and limitation of CNNs: CNNs have been *de facto* choice by most of the previous image restoration algorithms, yet they are limited by two primary issues: 1) The convolutional filters cannot adapt to image content during

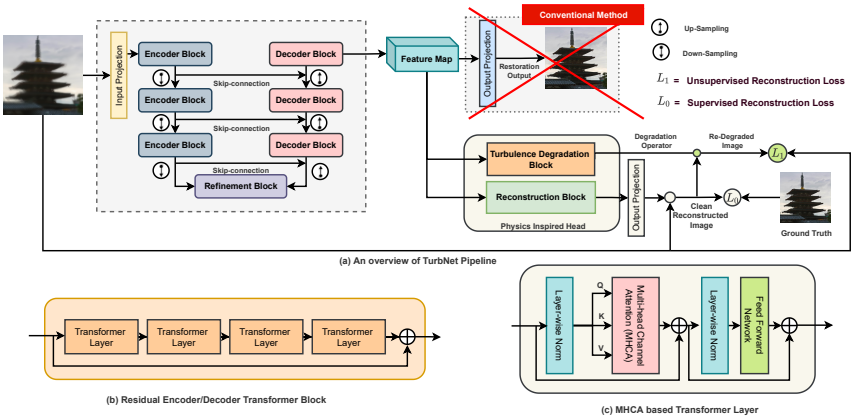

Fig. 1. Architecture of the proposed method. (a) The overall architecture consists: (i) a transformer to pull the spatially dynamical features from the scene; (ii) instead of directly constructing the image, we introduce a physics-inspired model to estimate the turbulence while reconstructing the image. (b) The structure of the residual encoder/decoder transformer block. (c) The details of each transformer layer.

inference due to their static weights. 2) The local receptive fields cannot model the long-range pixel dependencies. A key characteristic of the atmospheric turbulence effect is the "lucky effect" [6], meaning that **image regions** or **frames** with less degradation will randomly occur due to the distortions being spatially varying. Previous restoration methods treat turbulence restoration as a regression problem using CNNs but ignore the fact that turbulence is highly location adaptive and should not be represented as static fixed kernel applied to all locations. It is not difficult to see that applying static weight convolutions to regions with drastically different distortions will lead to sub-optimal performance.

The *self-attention* mechanism proposed in recent work [30, 31, 5] can be a powerful alternative, as it can capture context-dependent global interactions by aggregating information across image regions. Leveraging the capability of multi-head self-attention, we propose the ***TurbNet***, a transformer-based end-to-end network for restoring turbulence degraded images. Transformer-based architecture allows the creation of input-adaptive and location-adaptive filtering effect using *key, query,* and *weight*, where *key* and *query* decide content-adaptivity while *weight* brings location-adaptivity. Our design, as shown in Figure 1, is composed of several key building blocks:

Transformer Backbone: Our proposed network consists of a transformer-based backbone that has the flexibility of constructing an input-adaptive and location-adaptive unique kernel to model spatially- and instance-varying turbulence effect. Inspired by the success of [26, 32, 37] in various common image restoration tasks (e.g., denoising, deblurring, etc.), TurbNet adopts a U-shape

encoder-decoder architecture due to its hierarchical multi-scale representation while remaining computationally efficient. As shown in Figure 1 (b), the residual connection across the encoder-decoder provides an identity-based connection facilitating aggregation of different layers of features. Our backbone consists of three modules: input projection, deep encoder and decoder module. Input project module uses convolution layers to extract low frequency information and induces dose of convolutional inductive bias in early stage and improves representation learning ability of transformer blocks [33]. Deep encoder and decoder modules are mainly composed of a sequential cascade of Multi-head channel attention (MHCA) based transformer layers. Compared to prevalent CNN-based turbulence mitigation models, this design allows content-based interactions between image content and attention weights, which can be interpreted as spatially varying convolution [4].

The primary challenge of applying conventional transformer blocks for image restoration task comes from the quadratic growth of key-query dot product interactions, i.e., $\mathcal{O}(W^2H^2)$, for images with $W \times H$ pixels. To alleviate this issue, we adopt the idea of applying self-attention across channels instead of spatial dimension [37], and compute cross-covarience across channels generating attention map. Given *query* ($\mathbf{Q}$), *key* ($\mathbf{K}$), and *value* ($\mathbf{V}$), we reshape $\mathbf{Q}$ and $\mathbf{K}$ such that their dot-product generates a transposed-attention map $\mathbf{A} \in \mathbb{R}^{C\times C}$, instead of conventional $\mathbb{R}^{HW\times HW}$ [5]. Overall, the MHCA can be summarized as:

$$\mathbf{X}' = \mathbf{W_p}\ \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathbf{X} \tag{4}$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \cdot \text{softmax}\left\{ \frac{\mathbf{K} \cdot \mathbf{Q}}{\alpha} \right\} \tag{5}$$

where $\mathbf{X}'$ and $\mathbf{X}$ are input and output feature maps, $\mathbf{W}_\mathbf{p}^{(\cdot)}$ is the 1×1 point-wise convolution, and α is a learnable scaling parameter to control the magnitude of $(\mathbf{K} \cdot \mathbf{Q})$ before applying softmax.

Image Reconstruction Block: To further enhance deep features generated by the transformer backbone, TurbNet uses the reconstruction block. The primary job of the reconstruction block is to take deep features corresponding to turbulence degraded input image $\widetilde{\mathbf{I}}$ by the transformer backbone, further enrich it at high spatial resolution by encoding information from spatially neighboring pixel positions. Next, the enriched features pass through an output projection module with 3×3 convolution layers to project it back low dimension feature map corresponding to the reconstructed clean image $\mathbf{J}$. The design of the reconstruction block is very similar to the encoder block having MHCA, with an introduction of Locally-Enhanced Feed Forward Network (LoFFN) [32].

Precisely, the work of Reconstruction module can be summarized as:

$$\underbrace{\mathbf{F}_{\widetilde{\mathbf{I}}}}_{\substack{\text{Deep Features of degraded} \\ \text{Input Image } \widetilde{\mathbf{I}}}} \rightarrow \text{Reconstruction Module} \rightarrow \underbrace{\mathbf{J}_{\widetilde{\mathbf{I}}}}_{\substack{\text{Reconstructed Clean} \\ \text{Output Image}}} \tag{6}$$

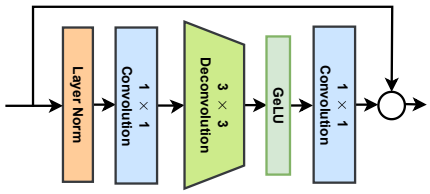

Fig. 2. Locally-Enhanced Feed Forward Network (LoFFN) used in the image reconstruction block and the turbulence degradation block.

Turbulence Degradation Block: In TurbNet, the turbulence degradation module learns the physical turbulence degradation operator $\mathcal{T}$ from the input synthetic training data. The primary job of turbulence degradation module is to take clean reconstructed image $\mathbf{J}_{\widetilde{\mathbf{I}}}$ corresponding to degraded input image $\widetilde{\mathbf{I}}$, apply the learned degradation operator $\mathcal{T}$, to construct back the **re-degraded** input image $\widetilde{\mathbf{I}}_{\mathcal{T}}$. This formulation enriches the training set by incorporating additional latent degradation images ($\widetilde{\mathbf{I}}_{\mathcal{T}}$), in addition to synthesized degraded images ($\widetilde{\mathbf{I}}$), during the training process. Additionally, this module facilitates self-supervised learning without the availability of ground truth. The architecture of this module is the same as Image Reconstruction Block with LoFFN.

Precisely, the work of Degradation Block can be summarized as:

$$\underbrace{\mathbf{J}_{\widetilde{\mathbf{I}}}}_{\substack{\text{Reconstructed Clean}\\ \text{Output Image}}} \rightarrow \text{Degradation Operator } \mathcal{T}(\cdot) \rightarrow \underbrace{\widetilde{\mathbf{I}}_{\mathcal{T}}}_{\substack{\text{Re-degraded}\\ \text{Output Image}}} \tag{7}$$

Loss Function: TurbNet optimization requires the joint optimization of reconstruction operation and the turbulance degradation operation. Given the synthetic training pair of degraded input $\widetilde{\mathbf{I}}$, and corresponding ground truth image $\mathbf{I}$, we formulate following two losses:

$$\underbrace{\mathcal{L}_0}_{\text{Supervised Reconstruction Loss}} = ||\mathbf{J}_{\widetilde{\mathbf{I}}} - \mathbf{I}||_1 \tag{8}$$

$$\underbrace{\mathcal{L}_1}_{\text{Self-supervised Reconstruction Loss}} = ||\widetilde{\mathbf{I}}_{\mathcal{T}} - \widetilde{\mathbf{I}}||_1 \tag{9}$$

where, $\mathcal{L}_0$ is responsible for constructing a clean image $\mathbf{J}_{\widetilde{\mathbf{I}}}$ given the degraded input image $\widetilde{\mathbf{I}}$, $\mathcal{L}_1$ helps to ensure degradation operator $\mathcal{T}$ can reconstruct the original the original input $\widetilde{\mathbf{I}}$ from the reconstructed clean image $\mathbf{J}_{\widetilde{\mathbf{I}}}$.

Eventually, the overall loss $\mathcal{L}$ to train TurbNet can be summarized as:

$$\mathcal{L} = \alpha \times \mathcal{L}_0 + (1 - \alpha) \times \mathcal{L}_1 \tag{10}$$

Overall Pipeline As shown in Figure 1(a), TurbNet utilizes a U-shape architecture built upon transformer blocks to extract deep image features. As suggested in [33], an initial convolution-based input projection is used to project the input image to higher dimensional feature space, which can lead to more stable optimization and better results. After obtaining the feature maps, TurbNet jointly learns the turbulence degradation operator ($\mathcal{T}$) along with the reconstructed image ($\mathbf{J}_{\tilde{\mathbf{I}}}$), in contrary to general image restoration methods [32, 19, 2, 37] that directly reconstruct the clean image. This design facilitates spatial adaptivity and long-range dynamics of turbulence effects, plus a self-supervised consistency loss.

Synthetic-to-Real Generalization: With a pre-trained TurbNet model $\mathcal{M}(\cdot)$ using the synthetic data, TurbNet design allows an effective way of generalizing $\mathcal{M}(\cdot)$ on unseen real data (if required) with the help of degradation operator $\mathcal{T}(\cdot)$ in a self-supervised way. Starting from model $\mathcal{M}(\cdot)$, we create a generalization dataset by incorporating unlabelled real data with the synthetic data to fine-tune $\mathcal{M}(\cdot)$. For input images with no ground truth, $\mathcal{M}(\cdot)$ is optimized using Equation (9), while for input images from labeled synthetic data $\mathcal{M}(\cdot)$ is optimized using Equation (8, and 9). Note that we incorporate synthetic data into the fine-tuning process to mitigate the issue of catastrophic forgetting during generalization.

4 Large-Scale Training and Testing Datasets

4.1 Training Data: Synthetic Data Generating Scheme

Training a deep neural network requires data, but the real clean-noisy pair of turbulence is nearly impossible to collect. A more feasible approach here is to leverage a powerful turbulence simulator to synthesize the turbulence effects.

Turbulence simulation in the context of deep learning has been reported in [36, 23, 14]. Their model generates the geometric distortions by repeatedly smoothing a set of random spikes, and the blur is assumed to be spatially invariant Gaussian [13]. We argue that for the face images studied in [36, 23, 14], the narrow field of view makes their simplified model valid. However, for more complex scenarios, such a simplified model will fail to capture two key phenomena that could cause the training of the network to fail: (1) The instantaneous distortion of the turbulence can vary significantly from one observation to another even if the turbulence parameters are fixed. See Figure 3(a) for an illustration from a real data. (2) Within the same image, the distortions are spatially varying. See Figure 3(b).

In order to capture these phenomena, we adopt an advanced simulator [22] to synthesize a large-scale *training* dataset for atmospheric turbulence. The clean data used by the simulator is the *Places* dataset [39]. A total of 50,000 images are generated, and the turbulence parameters are configured to cover a wide range of conditions. The details of the simulation can be found in the supplementary material. We remark that this is the first attempt in the literature to systematically generate such a comprehensive and large-scale training dataset.

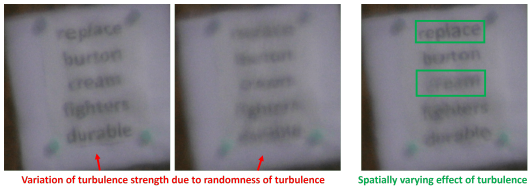

Fig. 3. Key turbulence effects requiring attention while designing synthetic dataset.

4.2 Testing Data: Heat Chamber and Text Datasets

Our real benchmarking dataset consists of two parts: the *Heat Chamber Dataset* and the *Turbulent Text Dataset*. Although this paper focuses on single frame restoration, both our benchmarking datasets contain 100 static turbulence degraded frames for each scene. We believe that by doing so, researchers in the field working on multi-frame reconstruction can also benefit from our dataset. Both datasets will be made **publicly available**.

Heat Chamber Dataset. The *Heat Chamber Dataset* is collected by heating the air along the imaging path to artificially create a stronger turbulence effect. The setup for collecting the heat chamber dataset is shown in 4. Turbulence-free ground truth images can be obtained by shutting down the heat source. The images are displayed on a screen placed 20 meters away from the camera.

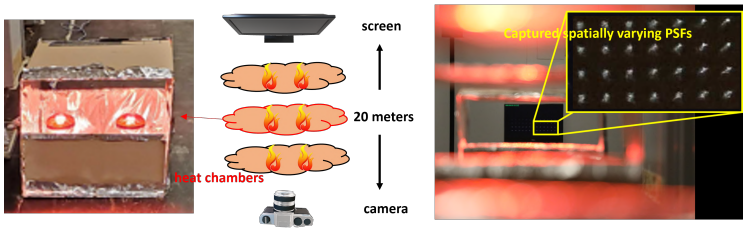

Fig. 4. The setup of heat chamber data collection. We evenly placed three heat chambers along the imaging path. Our dataset captures better spatially varying effect.

We remark that while similar datasets have been collected in [10, 1], our data has a clear improvement: we use a long path and more evenly distributed heat so that the turbulence effect is closer to the true long-range effect. The captured images have a better anisoplanatic (spatially varying) effect such that an almost distortion-free frame is less likely to occur compared with the dataset in [10, 1]. In addition, our dataset is much large in scale. It contains 2400 different images, which allows for a better evaluation of the learning-based model. Sample images of the *Heat Chamber Dataset* can be found in Figure 5.

Turbulence Text Dataset. Due to the nature of the problem, it is extremely difficult, if not impossible, to capture ground truth clean images in

Fig. 5. Sample turbulence degraded images (top) and corresponding ground truth (bottom) from our *Heat Chamber Dataset*. The D/r_0 is estimated to be around 3.

truly long-range settings. Therefore, we adopt the idea of using the performance of high-level vision task as an evaluation metric for image restoration [17, 16]. Specifically, we calculate the detection ratio and longest common subsequence on the output of an OCR algorithm [29, 28] as the evaluation metrics. The terms will be defined in section 5.4.

There are several advantages of using text recognition: 1) The degradation induced by atmospheric turbulence, the geometric distortion and the loss of resolution, can be directly reflected by the text patterns. Both types of degradation need to be removed for the OCR algorithms to perform well. 2) The OCR is a mature application. The selected algorithms should be able to recognize the text patterns as long as the turbulence is removed. Other factors such as the domain gap between the training and testing data will not affect the evaluation procedure as much as other high-level vision tasks. 3) An important factor to consider when designing the dataset is whether the difficulty of the task is appropriate. The dataset should neither be too difficult such that the recognition rate cannot be improved by the restoration algorithms nor too easy making all algorithms perform similarly. We can easily adjust the font size and contrast of text patterns to obtain a proper difficulty level.

The *Turbulence Text Dataset* consists of 100 scenes, where each scene contains 5 text sequences. Each scene has 100 static frames. It can be assumed that there is no camera and object motion within the scene, and the observed blur is caused by atmospheric turbulence. The text patterns come in three different scales, which adds variety to the dataset. We also provide labels to crop the individual text patterns from the images. Sample images from the dataset are shown in Figure 6.

5 Experiment Results

Implementation Details: TurbNet uses a 4-staged symmetric encoder-decoder architecture, where stage 1, 2, 3, and 4 consist of 4, 6, 6, and 8 MHCA-based transformer layers respectively. Our Reconstruction block and Turbulence Degradation block consist of 4 MHCA-transformer layers enhanced with LoFFN. TurbNet is trained using $50,000$ synthetic dataset generated using a physics-based

(a) Data collection site. (b) Sample collected images

Fig. 6. Data collection site of the *Turbulence Text Dataset*. The distance between the camera and the target is 300 meters. The D/r_0 is estimated to be in range of 2.5 to 4 (varies due to the temperature change during the collection process). The collected text patterns are in 3 different scales.

Table 1. Performance comparison of state-of-art restoration baselines with respect to TurbNet on synthetic and *Heat Chamber* dataset.

	TDRN[35]	MTRNN[24]	MPRNet[38]	Uformer[32]	Restormer[37]	**TurbNet**
Synthetic Dataset						
PSNR	21.35	21.95	21.78	22.03	22.29	**22.76**
SSIM	0.6228	0.6384	0.6410	0.6686	0.6719	**0.6842**
HeatChamber Dataset						
PSNR	18.42	18.12	18.68	19.12	19.01	**19.76**
SSIM	0.6424	0.6379	0.6577	0.6840	0.6857	**0.6934**

stimulator [22] and MIT Places dataset [39] while synthetic evaluation results are generated on $5,000$ synthetic images. Due to resource constraint, our synthetic training uses a batch size of 8 with Adam optimizer. We start our training with learning rate of $1e-4$, and use the cosine annealing scheduler to gradually decrease the learning rate over the span of 50 epcohs. During training, to modulate between the loss Equation 8 and 9, we have use α to be 0.9. All the baselines method used in our evaluation has been trained with exactly same settings and same dataset using their official GitHub implementation for fair comparison. Additional implementation details are provided in supplementary materials.

5.1 Synthetic and *Heat Chamber* Dataset Results

We first conduct an experiment on a synthetic testing dataset generated with the same distribution as testing data. In Figure 7, we show a qualitative comparison between our restored images with ground truth. It can be seen that our results are accurately reconstructed with the assist from estimated turbulence map.

We then compare our results qualitatively with the existing algorithms on both synthetic and *Heat Chamber* dataset. A Visual comparison on the synthetic dataset can be found in Figure 8. It can be observed that the transformer-based methods generally perform better than the CNN-based methods due to their ability to adapt dynamically to the distortions. The proposed method achieves

authentic reconstruction due to its ability to explicitly model the atmospheric turbulence distortion. Table 1 presents the quantitative evaluation of TurbNet wrt. other baselines. TurbNet achieves the best results in both PSNR and SSIM. Note that Uformer[32], and Restomer[37] (designed for classical restoration problems like deblurring, deraining, etc.) uses transformer-based encoder decoder architecture, but their performance is significantly low than TurbNet, which validates the importance of our decoupled (reconstruction and degradation estimation) design.

5.2 *Turbulence Text Dataset* Results

Evaluation Method: In order to evaluate the performance of TurbNet on our real-world turbulence text dataset, we use publicly available OCR detection and recognition algorithms [29, 28]. We propose the following two evaluation metrics - Average Word Detection Ratio (**AWDR**), and Average Detected Longest Common Subsequence (**AD − LCS**) defined as follows:

$$\textbf{AWDR} = \frac{\sum_{Scene=1}^{N} \frac{\text{Word Detected}_{scene}}{\text{Word Count}_{scene}}}{N}, \tag{11}$$

$$\textbf{AD} - \textbf{LCS} = \frac{\sum_{Scene=1}^{N} \sum_{Word=1}^{K} \mathcal{LCS}(DetectedString, TrueString)}{N}, \tag{12}$$

where $\mathcal{LCS}$ represents the Longest Common Subsequence, $TrueString$ represents the ground truth sequence of characters corresponding to a word i in the image, $DetectedString$ represents a sequence of characters recognized by OCR algorithms for word i, and N is the total number of scenes in the test dataset.

Table 2. Performance comparison of state-of-art restoration baselines with respect to TurbNet on our *Turbulence Text Dataset*.

	Raw Input	TDRN[35]	MTRNN[24]	MPRNet[38]	Restormer[37]	**TurbNet**
AWDR	0.623	0.617	0.642	0.633	0.702	**0.758**
AD-LCS	5.076	5.011	5.609	5.374	6.226	**7.314**

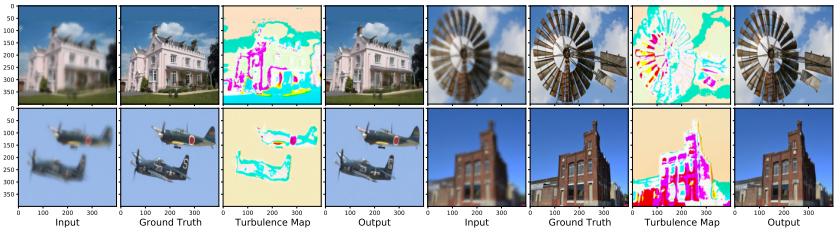

Fig. 7. Qualitative Performance comparison of TurbNet wrt. the ground truth.

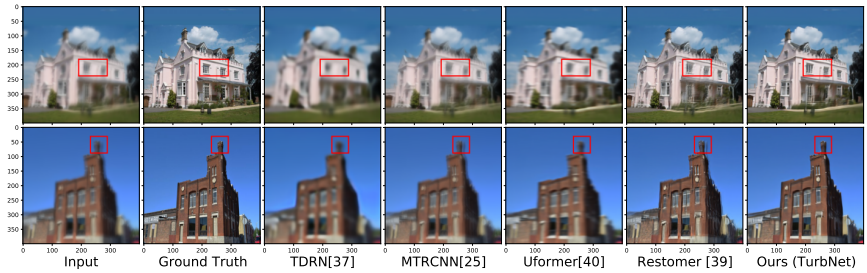

Fig. 8. Qualitative Performance comparison of TurbNet wrt. other SOTA methods.

Discussion: Figure 9 represents the performance of OCR on the real turbulence impacted images and images restored by TurbNet. It is evident that our restoration model significantly helps in improving the OCR performance by identifying comparatively more words with higher confidence. Table 2 presents the performance gain by TurbNet over the real turbulence degraded text images and their restored version by various state-of-the-art methods. OCR algorithms achieve massive improvements of +0.135 (AWDR) and +2.238 (AD-LCS) when used on images restored by TurbNet compared to being used directly on real images from our proposed test dataset.

5.3 Experimental Validity of the Proposed Model

We conduct two additional experiments to validate the proposed model. The first experiment is an ablation study, where we demonstrate the impact of replacing transformer as feature encoder with U-Net [26] and removing the turbulence map estimation part. The result is reported in 3, where we observe a significant performance drop in both cases. The second experiment is to prove the effectiveness of the extracted turbulence map. We extract a turbulence map from a simulated frame and apply the map back to the ground-truth image. We calculate the PSNR of this re-corrupted image w.r.t. the original turbulence frame. We tested on 10K turbulence frames and the average PSNR is **39.89 dB**, which is a strong evidence that our turbulence map can effectively extract the turbulence information embedded in the distorted frames. A visualization of the experiment can be found in Figure 10.

Table 3. Ablation on Heat Chamber Dataset

Model type	PSNR	SSIM
TurbNet [Ours]	19.76	0.6934
TurbNet - Turbulance Map	19.03 (↓)	0.6852 (↓)
TurbNet - Transformer	18.62 (↓)	0.6481 (↓)

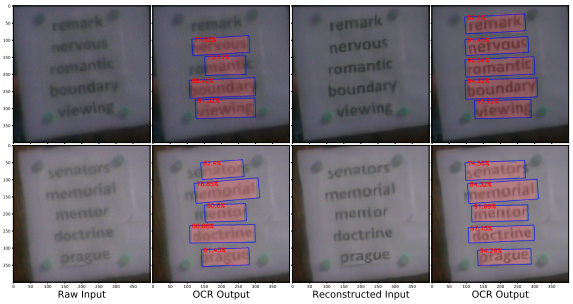

Fig. 9. OCR performance of our reconstruction algorithm for *Turbulance Text Dataset*

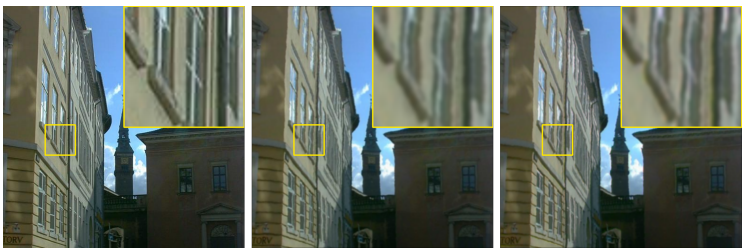

Fig. 10. Validation of our turbulence map. Left: groundtruth. Middle: original turbulence frame. Right: groundtruth re-corrupted with the extracted turbulence map.

6 Conclusions

In this work, identifying the short-come of existing image restoration algorithms, we propose a novel physics-inspired turbulence restoration model (TurbNet) based on transformer architecture to model spatial adaptivity and long-term dynamics of turbulence effect. We present a synthetic data generation scheme for tuning a sophisticated physics-grounded simulator to generate a large-scale dataset, covering a broad variety of atmospheric turbulence effects. Additionally, we introduce two new large-scale testing datasets that allow for evaluation with classical objective metrics and a new task-driven metric with optical text recognition. Our comprehensive evaluation on realistic and diverse datasets leads to exposing limitations of existing methods and the effectiveness of TurbNet.

Acknowledgement

The research is based upon work supported in part by the Intelligence Advanced Research Projects Activity (IARPA) under Contract No. 2022-21102100004, and in part by the National Science Foundation under the grants CCSS-2030570 and IIS-2133032. 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 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.

References

1. Anantrasirichai, N., Achim, A., Kingsbury, N.G., Bull, D.R.: Atmospheric turbulence mitigation using complex wavelet-based fusion. IEEE Transactions on Image Processing **22**(6), 2398–2408 (June 2013). https://doi.org/10.1109/TIP.2013.2249078
2. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12294–12305 (2021)
3. Chimitt, N., Chan, S.H.: Simulating anisoplanatic turbulence by sampling intermodal and spatially correlated Zernike coefficients. Optical Engineering **59**(8), 1 – 26 (2020). https://doi.org/10.1117/1.OE.59.8.083101, `https://doi.org/10.1117/1.OE.59.8.083101`
4. Cordonnier, J.B., Loukas, A., Jaggi, M.: On the relationship between self-attention and convolutional layers. arXiv preprint arXiv:1911.03584 (2019)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Fried, D.L.: Probability of getting a lucky short-exposure image through turbulence*. J. Opt. Soc. Am. **68**(12), 1651–1658 (Dec 1978). https://doi.org/10.1364/JOSA.68.001651, `http://www.osapublishing.org/abstract.cfm?URI=josa-68-12-1651`
7. Hardie, R.C., Power, J.D., LeMaster, D.A., Droege, D.R., Gladysz, S., Bose-Pillai, S.: Simulation of anisoplanatic imaging through optical turbulence using numerical wave propagation with new validation analysis. Optical Engineering **56**(7), 1 – 16 (2017). https://doi.org/10.1117/1.OE.56.7.071502, `https://doi.org/10.1117/1.OE.56.7.071502`
8. Hardie, R.C., Rucci, M.A., Dapore, A.J., Karch, B.K.: Block matching and wiener filtering approach to optical turbulence mitigation and its application to simulated and real imagery with quantitative error analysis. Optical Engineering **56**(7), 071503 (2017)
9. He, R., Wang, Z., Fan, Y., Feng, D.: Atmospheric turbulence mitigation based on turbulence extraction. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1442–1446 (March 2016)
10. Hirsch, M., Sra, S., Schölkopf, B., Harmeling, S.: Efficient filter flow for space-variant multiframe blind deconvolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 607–614 (June 2010)
11. Kolmogorov, A.N.: The Local Structure of Turbulence in Incompressible Viscous Fluid for Very Large Reynolds' Numbers. Akademiia Nauk SSSR Doklady **30**, 301–305 (1941)
12. Lau, C.P., Lai, Y.H., Lui, L.M.: Restoration of atmospheric turbulence-distorted images via RPCA and quasiconformal maps. Inverse Problems (Mar 2019). https://doi.org/10.1088/1361-6420/ab0e4b
13. Lau, C.P., Lui, L.M.: Subsampled turbulence removal network. Mathematics, Computation and Geometry of Data **1**(1), 1–33 (2021). https://doi.org/10.4310/MCGD.2021.v1.n1.a1
14. Lau, C.P., Souri, H., Chellappa, R.: Atfacegan: Single face semantic aware image restoration and recognition from atmospheric turbulence. IEEE Transactions on Biometrics, Behavior, and Identity Science pp. 1–1 (2021). https://doi.org/10.1109/TBIOM.2021.3058316

15. Leonard, K.R., Howe, J., Oxford, D.E.: Simulation of atmospheric turbulence effects and mitigation algorithms on stand-off automatic facial recognition. In: Proc. SPIE 8546, Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII. pp. 1–18 (Oct 2012)
16. Li, B., Peng, X., Wang, Z., Xu, J.Z., Feng, D.: Aod-net: All-in-one dehazing network. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
17. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2019). https://doi.org/10.1109/TIP.2018.2867951
18. Li, N., Thapa, S., Whyte, C., Reed, A.W., Jayasuriya, S., Ye, J.: Unsupervised non-rigid image distortion removal via grid deformation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2522–2532 (October 2021)
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. ArXiv **abs/2103.14030** (2021)
20. Lou, Y., Ha Kang, S., Soatto, S., Bertozzi, A.: Video stabilization of atmospheric turbulence distortion. Inverse Problems and Imaging **7**(3), 839–861 (Aug 2013). https://doi.org/10.3934/ipi.2013.7.839
21. Mao, Z., Chimitt, N., Chan, S.H.: Image reconstruction of static and dynamic scenes through anisoplanatic turbulence. IEEE Transactions on Computational Imaging **6**, 1415–1428 (2020). https://doi.org/10.1109/TCI.2020.3029401
22. Mao, Z., Chimitt, N., Chan, S.H.: Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14759–14768 (October 2021)
23. Nair, N.G., Patel, V.M.: Confidence guided network for atmospheric turbulence mitigation. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1359–1363 (2021). https://doi.org/10.1109/ICIP42928.2021.9506125
24. Park, D., Kang, D.U., Kim, J., Chun, S.Y.: Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In: European Conference on Computer Vision. pp. 327–343. Springer (2020)
25. Roggemann, M.C., Welsh, B.M.: Imaging through Atmospheric Turbulence. Laser & Optical Science & Technology, Taylor & Francis (1996)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
27. Schmidt, J.D.: Numerical simulation of optical wave propagation: With examples in MATLAB. SPIE Press (Jan 2010). https://doi.org/10.1117/3.866274
28. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2017)
29. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: European conference on computer vision. pp. 56–72. Springer (2016)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

31. Wang, X., Girshick, R.B., Gupta, A.K., He, K.: Non-local neural networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7794–7803 (2018)
32. Wang, Z., Cun, X., Bao, J., Liu, J.: Uformer: A general u-shaped transformer for image restoration. arXiv preprint arXiv:2106.03106 (2021)
33. Xiao, T., Dollar, P., Singh, M., Mintun, E., Darrell, T., Girshick, R.: Early convolutions help transformers see better. Advances in Neural Information Processing Systems **34** (2021)
34. Xie, Y., Zhang, W., Tao, D., Hu, W., Qu, Y., Wang, H.: Removing turbulence effect via hybrid total variation and deformation-guided kernel regression. IEEE Transactions on Image Processing **25**(10), 4943–4958 (Oct 2016)
35. Yasarla, R., Patel, V.M.: Learning to restore a single face image degraded by atmospheric turbulence using cnns. arXiv preprint arXiv:2007.08404 (2020)
36. Yasarla, R., Patel, V.M.: Learning to restore images degraded by atmospheric turbulence using uncertainty. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1694–1698 (2021). https://doi.org/10.1109/ICIP42928.2021.9506614
37. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. arXiv preprint arXiv:2111.09881 (2021)
38. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14821–14831 (2021)
39. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
40. Zhu, X., Milanfar, P.: Removing atmospheric turbulence via space-invariant deconvolution. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 157–170 (Jan 2013). https://doi.org/10.1109/TPAMI.2012.82