

# Imaging Through the Atmosphere Using Turbulence Mitigation Transformer

Xingguang Zhang<sup>ID</sup>, Graduate Student Member, IEEE, Zhiyuan Mao<sup>ID</sup>, Nicholas Chimitt<sup>ID</sup>, Member, IEEE, and Stanley H. Chan<sup>ID</sup>, Senior Member, IEEE

**Abstract**—Restoring images distorted by atmospheric turbulence is a ubiquitous problem in long-range imaging applications. While existing deep-learning-based methods have demonstrated promising results in specific testing conditions, they suffer from three limitations: (1) lack of generalization capability from synthetic training data to real turbulence data; (2) failure to scale, hence causing memory and speed challenges when extending the idea to a large number of frames; (3) lack of a fast and accurate simulator to generate data for training neural networks. In this paper, we introduce the turbulence mitigation transformer (TMT) that explicitly addresses these issues. TMT brings three contributions: Firstly, TMT explicitly uses turbulence physics by decoupling the turbulence degradation and introducing a multi-scale loss for removing distortion, thus improving effectiveness. Secondly, TMT presents a new attention module along the temporal axis to extract extra features efficiently, thus improving memory and speed. Thirdly, TMT introduces a new simulator based on the Fourier sampler, temporal correlation, and flexible kernel size, thus improving our capability to synthesize better training data. TMT outperforms state-of-the-art video restoration models, especially in generalizing from synthetic to real turbulence data.

**Index Terms**—Atmospheric turbulence, video restoration, deep learning, transformer, multi-frame image processing, simulation.

## I. INTRODUCTION

ATMOSPHERIC turbulence mitigation methods aim at recovering images distorted by the random fluctuations of the refractive index in the atmosphere. Turbulence-distorted images suffer from spatially varying blurs and geometric warping. These problems, when presented to computer vision applications such as detection, recognition, and surveillance, will cause uncertainty about the object's shape, boundary, and

resolution. If they are further entangled with camera shake and sensor noise, restoring images would be challenging. This paper aims to present a new deep-learning method to overcome these challenges.

Image restoration methods for atmospheric turbulence have been studied for decades, yet three gaps remain:

- 1) Lack of an accurate and fast *simulator* to synthesize training data at a large scale. Existing turbulence simulators (for example [1], [2]) are either too slow or inaccurate for the purpose of generating a sufficient amount of training data to train a restoration neural network. The phase-to-space transform (P2S) [3] overcomes some difficulties, but it has a severe memory limitation, prohibiting dense-field turbulence simulation.
- 2) Lack of a restoration method that can *generalize* from synthetic data to real turbulence. Classical methods such as [4], [5], [6], [7], [8] are mostly based on optimization using heuristic priors. Their limited modeling capability makes them hard to generalize. Single frame deep-learning methods [9], [10], [11] are usually class-specific. Without a strong semantic prior, such as a face, these methods would be vulnerable to out-of-distribution samples.
- 3) Lack of a restoration method that offers both *speed and memory*. Among all the available methods in the literature, multi-frame deep-learning methods (such as [12]) have the greatest potential for the restoration task due to the incorporation of the lucky effect [13]. However, multi-frame methods have a huge memory requirement prohibiting us from using more frames.

In this paper, we present the *Turbulence Mitigation Transformer* (TMT). TMT is the first multi-frame image restoration transformer customized for atmospheric turbulence, with three components articulating the above-mentioned gaps.

- Recognizing the importance of the forward physical model, TMT explicitly uses the turbulence physics by (1) decoupling the restoration task into de-tilting and de-blurring steps, as opposed to generic video restoration transformers that are single-step networks; (2) introducing a multi-scale structure to supervise the training at different resolutions, as opposed to existing single-scale turbulence networks. These two changes significantly improve the *generalization* capability of TMT.
- Existing video restoration transformers have severe memory limitations. TMT introduces the concept of temporal-channel joint attention (TCJA), a self-attention module

Manuscript received 8 March 2023; revised 6 July 2023, 9 September 2023, and 7 November 2023; accepted 3 January 2024. Date of publication 16 January 2024; date of current version 29 January 2024. This work was supported in part by the Intelligence Advanced Research Projects Activity (IARPA) under Grant 2022-21102100004, and in part by the National Science Foundation under Grants CCSS-2030570 and IIS-2133032. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mauricio Delbracio. (Corresponding author: Xingguang Zhang.)

Xingguang Zhang, Nicholas Chimitt, and Stanley H. Chan are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: zhan3275@purdue.edu; nchimitt@purdue.edu; stanchan@purdue.edu).

Zhiyuan Mao was with Purdue University, West Lafayette, IN 47907 USA. The work was completed when he was a graduate student at Purdue. He is now with Samsung Research America, Plano, TX 75024 USA (e-mail: m940421@gmail.com).

Code, videos, and datasets are available at <https://xg416.github.io/TMT>.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TCI.2024.3354421>, provided by the authors.

Digital Object Identifier 10.1109/TCI.2024.3354421

along the temporal axis. TCJA allows TMT to process frames with much less memory, significantly improve the temporal horizon of usable frames, provide better out-sample generalization, and improve *speed*.

- To support the training of TMT, particularly the multi-scale supervision which requires a highly accurate dense-field simulator, we introduce a new turbulence simulator. The new simulator contains three elements: (1) a Fourier-based sampler to enable dense-field simulation without interpolation; (2) incorporation of the temporal correlation; (3) improvement of the kernel size flexibility.

## II. PROBLEM FORMULATION AND RELATED WORK

### A. Turbulence Degradation

Consider an object  $J(\mathbf{x}, t) \in \mathbb{R}^d$  with a coordinate  $\mathbf{x} \in \mathbb{R}^2$  and time stamp  $t \in \mathbb{R}$ . As light propagates from the object plane to the image plane, the random fluctuations of the refractive index in the atmosphere will cause turbulence distortion. The exact image formation process remains an open problem, although from Kolmogorov [14] to Fried [13], people have developed widely acceptable models and numerical simulators through wave propagation equations [1], [15], [16]. For brevity of this paper, we refer readers to tutorials such as [17] for discussions of this historical work.

For the purpose of this paper, we shall take an *image processing* perspective (as opposed to the wave propagation perspective) by considering the turbulence as a concatenation of two processes, tilt  $\mathcal{T}$  and blur  $\mathcal{B}$ :

$$\underbrace{I(\mathbf{x}, t)}_{\text{corrupted image}} = \underbrace{[\mathcal{B} \circ \mathcal{T}]}_{\text{tilt-then-blur } \mathcal{H}} \left( \underbrace{J(\mathbf{x}, t)}_{\text{clean image}} \right), \quad (1)$$

where  $\circ$  denotes the functional composition of the two operators, and  $\mathcal{H} \stackrel{\text{def}}{=} \mathcal{B} \circ \mathcal{T}$  is the overall distortion. The foundation of this tilt-blur decomposition is rooted in the Zernike decomposition in the phase space [15], where it can be shown that the first two Zernike coefficients of the phase are responsible for the pixel displacement, aka *tilt*, and the other Zernike coefficients are responsible for the high order aberrations, aka *blur*. The ordering of the tilt-then-blur does not commute and is justified in [18].

### B. Existing Turbulence Mitigation Methods

The problem of turbulence mitigation is to invert the process in (1). For small field-of-view objects such as stars, the turbulence is likely isoplanatic (i.e., spatially invariant), so standard deconvolution applies [19]. For larger fields of view, the turbulence becomes anisoplanatic (i.e., spatially varying). Classical methods typically use the lucky selection method [13], [20], and later a decoupling strategy of de-tilting and de-blurring [21]. During the 2010 s, numerous algorithms have been proposed, ranging from numerical optimizations [7], [22], [23], complex wavelets [4], Fourier burst accumulation [24], block matching [25], and other hybrid methods based on lucky fusion and deconvolution [5], [6], [26], [27], [28].

On the deep-learning side, both deterministic [9], [10], [29], [30] and generative methods [11], [31], [32] have been proposed

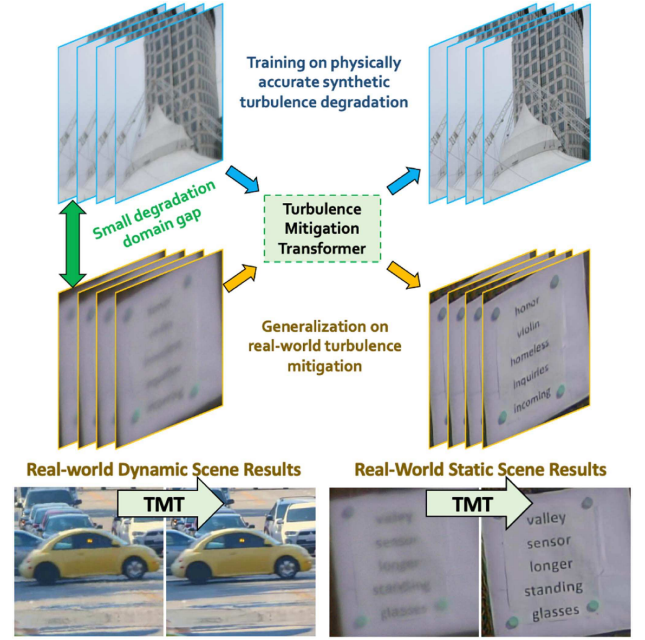


Fig. 1. Overall idea of our method for turbulence mitigation. Our main contribution includes a physically accurate, diverse, and fast turbulence synthesis and a specifically designed neural network called turbulence mitigation transformer (TMT) for blind multi-frame turbulence removal. The TMT is trained purely on synthetic data in a supervised framework yet shows good generalization and robustness in dealing with real-world turbulence degradation in dynamic and static scenes.

for single frame restoration. While generative methods or adversarial loss typically produce visually better images, they could also be more vulnerable to small perturbations on input [33], [34]. Among the above methods, similar to our work, [11], [29] consider removing tilt and blur progressively in two stages, while others mitigate turbulence degradation in one stage. [35], [36] introduced unsupervised methods to restore a single image from multiple instances. Although the large-scale dataset is not required, these works need several minutes or hours to converge and can only be applied to static scene sequences. [12] is the only work for general multi-frame turbulence mitigation in the literature; it is faster and covers dynamic scenes, but its generalization capability is limited.

### C. Learning-Based Image and Video Restoration Methods

Arguably, solving (1) shares many similarities with a video restoration problem. Classical algorithms have been used directly for turbulence [37]. Therefore, when moving to deep learning, it is fair to expect that existing state-of-the-art video restoration networks such as [38], [39] can be re-trained with appropriate data to perform the task.

Since turbulence degradation is spatially and temporally variant, the self-attention mechanism in vision transformers is well suited because of their adaptive weights based on the semantics of the image, while convolution-based networks use the same kernel on the entire image and feature map. Additionally, in single-frame restoration, transformers such as [40], [41], [42] have been introduced and shown superior performance.

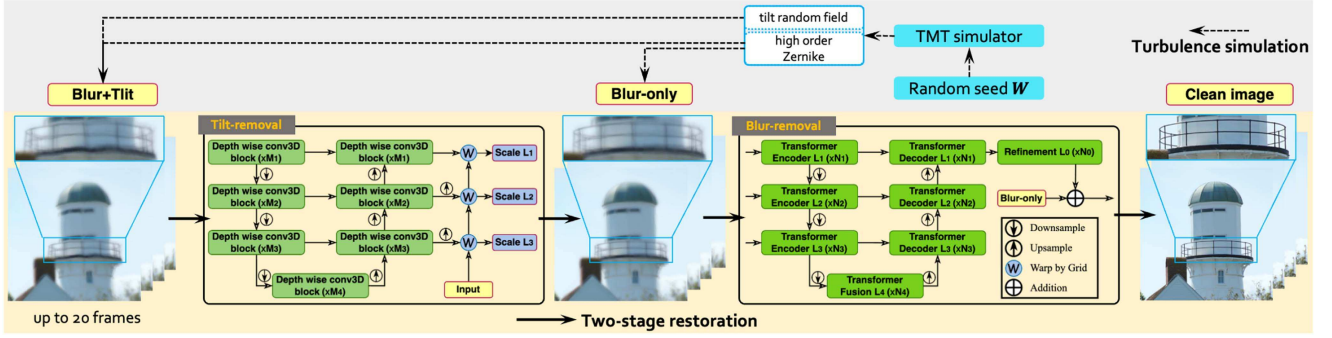


Fig. 2. Illustration of the simulation and restoration pipeline. The two-stage design of the restoration can take advantage of the physics-based simulation method. The lighthouse image is from the training set of our dataset.

However, when developing transformer-based models for multi-frame turbulence mitigation, the memory requirement is the biggest bottleneck for restoration methods. While extending the idea to the temporal dimension is possible [38], the computation and memory will make the implementation impractical. One of the key contributions of TMT is a new module, TCJA, to overcome this memory problem.

#### D. Existing Turbulence Simulators

When using deep learning based restoration methods, data is of the utmost importance. Since collecting real turbulence for *training* is nearly impossible due to the scale and variety of weather conditions we need, simulation is the best alternative. Turbulence simulation tools are mostly developed for physics and defense applications where precision is the most critical consideration. Split-step [1], [15], [43], [44] is often regarded as the “gold standard” although it is extremely slow. While faster simulations exist [2], [7], their accuracy is inadequate for our purpose. The simulator we present here is an improvement of the phase-to-space transform reported in [3], [45]. For detailed discussions of the latest development of turbulence simulations, we refer the readers to [17].

### III. TURBULENCE MITIGATION TRANSFORMER

#### A. Design Philosophy

To motivate the design of TMT, we start by looking at how vision transformers today are used in deconvolution problems. Vision transformers use self-attention to direct the neurons to focus on spatially correlated features [46], [47]. Transformers are particularly adaptive to *spatially varying* distortions because the self-attention can effectively construct *localized* kernels instead of a global kernel in a convolutional neural network [30].

When designing TMT, we need to revisit the forward model outlined in (1). The forward model says that the turbulence distortion is *compositional* — it contains tilt and blur. The ways to handle these two types of distortions are different:

- *Tilt*: Tilt is a temporal problem. Turbulent tilts follow a zero-mean Gaussian process [13]. Over a long period of *time*, pixels will experience the lucky phenomenon which is the foundation of many classical turbulence methods [5],

TABLE I  
COMPARISON BETWEEN EXISTING SOTA VIDEO RESTORATION AND TURBULENCE MITIGATION METHODS AND THE PROPOSED TMT

Existing video restoration and turbulence mitigation networks [12], [38], [39]	TMT (Proposed)
<ul style="list-style-type: none"> <li>• Single-stage</li> <li>• Agnostic to turbulence</li> <li>• Local spatial attention</li> <li>• Temporal modeling: N.A. / local window based / recurrent</li> <li>• Single-scale input</li> </ul>	<ul style="list-style-type: none"> <li>• Two-stage: tilt + blur</li> <li>• Customized for turbulence</li> <li>• Temporal-channel attention</li> <li>• Temporal modeling: Fully connected</li> <li>• Multi-scale loss (for tilt) and input (for blur)</li> </ul>

[6], [7], [26], [27], [28]. [36] shows that a small network is sufficient to rectify the images by measuring the distortion field and sampling pixels on its implicit “mean position”. To incorporate this idea, TMT introduces a *lightweight depth-wise 3D UNet*.

- *Blur*: Blur is a spatial *and* temporal problem where high-order aberrations in the phase cause the image to suffer from non-uniform blurs across the field of view. Spatial-temporal transformers are not easy to implement because of the huge memory consumption. To overcome the memory issue, TMT introduces a new temporal-channel joint attention (TCJA) module.

To summarize, TMT is a customized transformer-based restoration for *turbulence*. We summarize its key properties in Table I and its structure in Fig. 2.

#### B. Tilt-Removal Module

The tilt-removal module aims to align the images to the greatest extent. Since a turbulence scene contains both moving objects and pixel jittering, a direct temporal fusion (as in naively extending the spatial transformer to a spatial-temporal transformer) cannot extract features consistently without demanding a large memory consumption.

Our proposed tilt-removal module is based on a depth-wise 3D convolution. We chose this structure for two reasons:

- While classical video restoration methods (in the absence of turbulence) use optical flow [48] or deformable convolution [39], [49] to align the object motion, they are



TABLE II  
TILT-REMOVAL MODULE

Aspects	Encoder Level [1,2,3,4]	Decoder level [3,2,1]
Type of convolution	[Conv, DW, DW, DW]	[DW, DW, DW]
Size of kernels	[7, 7, 5, 3]	[3, 3, 3]
# output channels	[64, 128, 256, 512]	[256, 128, 64]
Down/Up-sampling	Max-pooling	Transposed Conv.

limited to *two adjacent frames*. Our 3D convolution does not have this limitation. As such, we can align multiple frames simultaneously.

- Recent works [36], [50] have shown that a complicated network is not needed for tilt-removal. A simple network is often enough to remove most pixel jitters due to the random tilt. Thus, we chose lightweight depth-wise 3D convolution inspired by these previous works.

The tilt-removal module is summarized in Table II. The basic structure is a UNet. We use a depth-wise (DW) 3D convolution and ReLU. It has four levels. Following the last three levels of the decoder, we output three sequences of warping grids to rectify the input image progressively. The warping grids from the 3rd and 2nd levels are upsampled to have the same spatial dimension as the input. To enlarge the perceptual field of the network, we set the kernel sizes of the first three encoder levels as 7, 7, and 5.

### C. Multi-Scale Loss

A critical component of TMT's tilt-removal module is the multi-scale loss. The multi-scale loss is also how we inject physics into the restoration model to improve generalization.

Assuming we have a faithful simulator, we will be able to generate *tilt-free* blur-only images at *multiple scales*:

$$\underbrace{\tilde{J}_\ell(\mathbf{x}, t)}_{\text{tilt-free at scale } \ell} = \underbrace{\mathcal{B}_\ell}_{\text{blur at scale } \ell} \left( \underbrace{J(\mathbf{x}, t)}_{\text{ground truth}} \right), \quad \ell = 1, \dots, L, \quad (2)$$

where  $L = 3$  is the number of scales. The difference between (2) and (1) is that there is no tilt in (2) but only blur. Therefore, we are preparing tilt+blur and blur-only pairs to train the tilt-removal module.

The tilt-removal module outlined in the previous subsection will return us, at every scale, an estimate of the tilt-free image:

$$\underbrace{\hat{J}_\ell(\mathbf{x}, t)}_{\text{estimated tilt-free at scale } \ell} = \underbrace{\mathcal{T}_\ell^{-1}}_{\text{tilt-removal}} \left( \underbrace{\{I(\mathbf{x}, t)\}}_{\text{fully distorted}} \right), \quad (3)$$

where  $\mathcal{T}_\ell^{-1}$  is a symbolic representation of the branch in the tilt-removal module that generates an image at scale  $\ell$ . The input to  $\mathcal{T}_\ell^{-1}$  is a stack of  $T$  frames  $\{I(\mathbf{x}, t)\}$  instead of a single frame at time  $t$ .

Once the tilt-free ground truths and the tilt-free estimates are determined, we can define the loss function as

$$\mathcal{L}_{\text{tilt}} = \sum_{\ell=1}^L \underbrace{\gamma_\ell}_{\text{weight}} \cdot \underbrace{\mathcal{L}_{\text{char}} \left( \hat{J}_\ell(\mathbf{x}, t), \tilde{J}_\ell(\mathbf{x}, t) \right)}_{\text{Charbonnier loss}}, \quad (4)$$

where  $\gamma_\ell$  denotes the weight and  $\mathcal{L}_{\text{char}}$  is the Charbonnier loss [51]. The choice of the weight is determined empirically as  $\gamma_1 = 0.6$ ,  $\gamma_2 = 0.3$  and  $\gamma_3 = 0.1$ .

The intuition of our multi-scale loss is that while tilts can be highly random at full resolution, they are much more structured and weaker at lower resolution. Therefore, tilt-removal is often easier at the lower scales and gradually becomes harder as we progress the scales. As a result, by *progressively* minimizing the loss across scales, we ensure that the tilt-removal is scale-consistent.

### D. Blur-Removal Module

The challenge of designing the blur-removal module is how to construct global temporal attention of all the frames without suffering from a high complexity and memory requirement. In conventional transformers, generating attention has a complexity that grows quadratically with the number of pixels in the video. Thus, conventional transformers adopt a window-based strategy to compute the attention locally [38].

TMT's deblurring module is shown on the right-hand side of the restoration pipeline in Fig. 2. The overall structure is similar to the tilt-removal module, except that basic operations are defined through a new transformer block outlined in Fig. 3. The key merits of this new transformer block are

- 1) In TMT's transformer, the spatial coordinates are connected purely via convolution layers. This smooths the interaction among neighboring pixels, avoiding inconsistent performance around the margin of local windows.
- 2) The core module of TMT is temporal-channel self-attention (TCJA). We compute the self-attention matrix for each pixel on the temporal and channel axes. Although the convolution operation is spatially invariant, the combination of the features of each pixel is spatially independent, enabling spatially varying restoration.
- 3) The dynamics of the blur and residue jitter largely follow a zero-mean random process. Therefore, by enabling full temporal connection over more frames, TMT can be more efficient than conventional transformers in capturing temporal dynamics.

On the right-hand side of Fig. 3, we present two variants of the proposed temporal channel joint attention (TCJA). As TCJA constructs attention along the temporal and channel dimensions, a standard way for self-attention should be flattening these two dimensions first, then using a linear projection to map it into the key, query, and value spaces. However, as we use a large chunk of frames (12 ~ 20 frames), the temporal-channel vector is large, and the projection will require  $O(\frac{HW C^2 T^2}{h})$  complexity, where  $h$  is the number of the head,  $H$ ,  $W$ ,  $T$  are height, width, and the number of frames. We propose two low-rank projection strategies to overcome this, as summarized below:

- In (a), we separate the temporal and channel axes and apply a linear projection on the temporal dimension. In this case, features in different channels of different frames do not communicate across frames before the self-attention step. This offers a slight boost in speed.
- In (b), we implement channel shuffle and partial connection inspired by the ShuffleNet [52]. We split the channels into groups and connect the temporal axis with the channels in

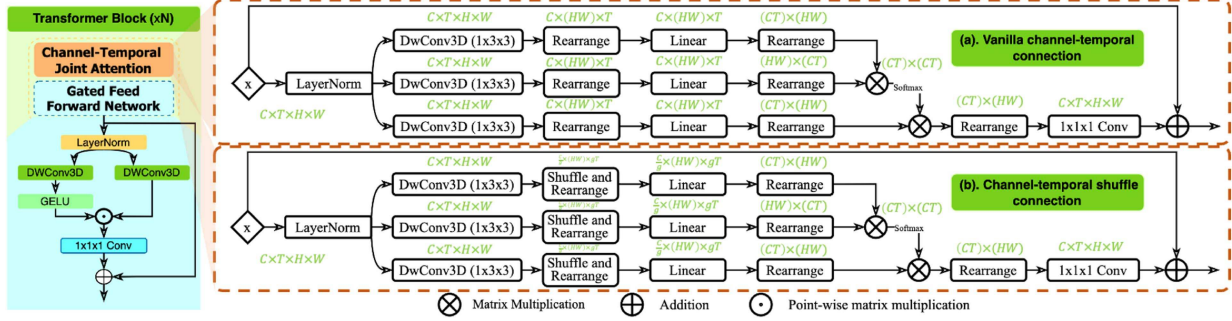


Fig. 3. Architecture of the transformer block. It consists of the channel-temporal joint attention (CTJA) module followed by a gated feed-forward network. As shown in the dashed box, we designed two variants (a) and (b) of the CTJA block for TMT<sub>a</sub> and TMT<sub>b</sub>, respectively. The only difference between those two is the channel shuffle operation.

each group, allowing them to communicate across several attention blocks.

#### IV. IMPROVEMENT FOR SIMULATOR

Like any deep learning algorithm, the success and failure of TMT are intimately related to how good our data is and how much data we have. For TMT, the source of *training* data is synthesized by our turbulence simulator modified from [3], [45], [53], with several important improvements.

##### A. Basic Principles

The starting point of our simulator is Zernike-space simulation [53], also known as the phase-over-aperture model [45]. For an input  $J(\mathbf{x}, t)$  at coordinate  $\mathbf{x}$  and time  $t$ , we directly generate the spatially varying point spread function (PSF) and compute the output  $I(\mathbf{x}, t)$  via the equation

$$I(\mathbf{x}, t) = \int \underbrace{h(\mathbf{x}, \mathbf{u}, t)}_{\text{PSF from } \mathbf{u} \text{ to } \mathbf{x} \text{ at time } t} J(\mathbf{u}, t) d\mathbf{u}, \quad (5)$$

where  $\mathbf{x}$  denotes the coordinate in the output space (i.e., the image plane) and the  $\mathbf{u}$  denotes the coordinate in the input space (i.e., the object plane). The PSF can, in theory, be obtained through

$$h(\mathbf{x}, \mathbf{u}, t) = \left| \text{Fourier} \left( e^{j\phi_{\mathbf{x},t}(\mathbf{u})} \right) \right|^2 \quad (6)$$

where  $\phi_{\mathbf{x},t}$  is the phase function at  $(\mathbf{x}, t)$ . Represented in the Zernike space, we can write  $\phi_{\mathbf{x},t}(\mathbf{u}) = \sum_{m=1}^M a_{\mathbf{x},t}(m) Z_m(\mathbf{u})$  with  $Z_m(\mathbf{u})$  being the  $m$ -th Zernike basis function, and  $a_{\mathbf{x},t}(m)$  is the corresponding Zernike coefficient [45].

The computational bottleneck in (5) is the generation of the Zernike coefficients  $a_{\mathbf{x},t}(m)$ , because once the Zernike coefficients are known, the PSF can be efficiently implemented through a basis decomposition as proposed in [3]:

$$h(\mathbf{x}, \mathbf{u}, t) = \sum_{k=1}^K \beta_{\mathbf{x},t}(k) \varphi_k(\mathbf{u}). \quad (7)$$

In (7), the basis function  $\varphi_k(\mathbf{u})$  can be determined through a principal component analysis (or a pre-defined set of basis functions). The basis coefficients  $\beta_{\mathbf{x},t}(k)$  are computed through

the phase-to-space transform (P2S) [3] where we use a shallow network to map the Zernike coefficients:

$$\{a_{\mathbf{x},t}(m)\} \xrightarrow{\text{P2S network}} \{\beta_{\mathbf{x},t}(k)\}. \quad (8)$$

Here, the bracket-to-bracket mapping means that we are transforming the entire set of Zernike coefficients  $\{a_{\mathbf{x},t}(m)\}$  to the set of basis coefficients  $\{\beta_{\mathbf{x},t}(k)\}$ .

The DF-P2S simulator [53] resolves another bottleneck issue, which is memory consumption. For the P2S transform in (8), we need to first generate the Zernike coefficients  $\{a_{\mathbf{x},t}(m)\}$ . However, generating these Zernike coefficients requires us to construct a 6-dimensional Gaussian covariance tensor (two dimensions for space and one for Zernike order, then duplicate to create the correlation). This has not yet included the time axis, which will add another two dimensions. Since this covariance matrix (technically a tensor) is extremely large, P2S is limited to generating the Zernike coefficients at a grid of points and interpolating otherwise.<sup>1</sup>

In [53], the memory bottleneck is mitigated by a few approximations to retain the homogeneity (aka wide-sense stationarity) of the covariance matrix. Doing so allows us to use Fast Fourier Transform to draw samples from the covariance matrix. As a result, we do not have to store the covariance matrix because homogeneity allows us to use one slice of the covariance matrix. In terms of speed, we have improved the speed of the covariance matrix generation from 4 minutes for a  $512 \times 512$  image to 17 ms.

##### B. Add Temporal Correlation

The first improvement is the addition of the temporal correlation. Recalling the phase-to-space equation in (8), the transformation from left to right requires us to start with the Zernike coefficients  $\{a_{\mathbf{x},t}(m)\}$ . For any fixed  $(\mathbf{x}, t)$ , we can define a Zernike coefficient vector  $\mathbf{a}_{\mathbf{x},t} = [a_{\mathbf{x},t}(1), \dots, a_{\mathbf{x},t}(M)]^T$  where  $M = 36$  is the typical number of coefficients we use. Enumerating over all the possible coordinates  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ , we can construct a very long vector  $\bar{\mathbf{a}}_t = [\mathbf{a}_{\mathbf{x}_1,t}; \mathbf{a}_{\mathbf{x}_2,t}; \dots; \mathbf{a}_{\mathbf{x}_N,t}]$  where  $N$  denotes the number of pixels of the image. Generation of  $\bar{\mathbf{a}}_t$  is

<sup>1</sup>The interpolation of phase-to-space is done in the Zernike space. This should be distinguished from interpolating the PSF in *space*, such as [1], [45].

done by multiplying a white noise vector  $\mathbf{w}_t \stackrel{\text{i.i.d.}}{\sim} \text{Gaussian}(\mathbf{0}, \mathbf{I})$  with a huge covariance matrix  $\Sigma$ :

$$\bar{\mathbf{a}}_t = \Sigma^{\frac{1}{2}} \mathbf{w}_t, \quad t = 1, 2, \dots, T \quad (9)$$

where  $\Sigma^{\frac{1}{2}}$  accounts for the spectral factorization of  $\Sigma$ . Using the sequence of approximation in [53], the spectral factorization can be done using the Fast Fourier Transform.

The temporal correlation of adjacent frames of a turbulence video can theoretically be modeled via Taylor's frozen hypothesis [54]. This would expand our 6-dimensional Zernike tensor to 8 dimensions. This is certainly doable in theory, but it is just computationally infeasible. To alleviate this computational difficulty, we adopt a surrogate approach by considering the auto-regressive model, where we define

$$\mathbf{w}_t = \alpha \mathbf{w}_{t-1} + \sqrt{1 - \alpha^2} \mathbf{z}, \quad (10)$$

where  $\mathbf{z} \stackrel{\text{i.i.d.}}{\sim} \text{Gaussian}(\mathbf{0}, \mathbf{I})$  and  $0 \leq \alpha \leq 1$ . It is easy to show that  $\mathbb{E}[\mathbf{w}_t] = \mathbf{0}$  because  $\mathbf{z}$  is zero-mean, and  $\mathbb{E}(\mathbf{w}_t \mathbf{w}_t^T) = \mathbf{I}$ . Furthermore, we can show that

$$\mathbb{E}[\mathbf{w}_t \mathbf{w}_{t-\tau}^T] = \alpha^\tau \mathbf{I} \quad (11)$$

where  $\tau$  is an integer. Thus, the correlation between frames decays as the frames are farther apart.

We remark that in TMT's simulator, the temporal correlation is realized through the *random seed* instead of the final samples. The impact on the final samples can be seen by substituting (10) into (9)

$$\begin{aligned} \bar{\mathbf{a}}_t &= \Sigma^{\frac{1}{2}} \left( \alpha \mathbf{w}_{t-1} + \sqrt{1 - \alpha^2} \mathbf{z} \right) \\ &= \alpha \bar{\mathbf{a}}_{t-1} + \sqrt{1 - \alpha^2} \Sigma^{\frac{1}{2}} \mathbf{z}. \end{aligned}$$

In other words, we introduce a recurrent relationship by regressing the current samples with the previous samples. The regression parameter  $\alpha$  is typically set as  $\alpha = 0.9$  for strong correlations and smaller values for weaker correlations.

### C. Kernel Size

Although the DF-P2S simulator [53] offers a significant improvement over the P2S version [3], both assume a *fixed* support (aka blur kernel size) of the PSF  $h(\mathbf{x}, \mathbf{u}, t)$ . More specifically, the kernel size is always  $33 \times 33$ , regardless of the system parameters. This is clearly a mismatch to reality, where the kernel size depends on the system.

To enable a wide range of kernel sizes, we have re-examined the PSF basis functions of [3]. We have observed that the basis functions  $\varphi_k$  from the P2S model may incur aliasing due to incorrect interpolation by the resizing process. To address this, we have regenerated  $\varphi_k$  at a higher resolution with more training examples. The PCA is, therefore, more representative of the distribution and contains additional high-frequency information. The result is a set of basis functions that may be resized without introducing any noticeable amount of aliasing into the PSF representation.

TMT's simulator includes kernel sizes from  $9 \times 9$  to  $33 \times 33$ . The choice of kernel size is randomly picked to ensure sufficient

coverage of the realistic turbulence conditions. In real-world static scene data such as [55], [56], the field of view is usually very narrow, and the corresponding blur kernels are usually large. In real-world dynamic scene data such as [12], [50], [55], turbulence conditions have more variety and most samples have a larger field of view with smaller blur kernels. Overall, we selected the kernel sizes based on all publicly available real-world datasets, and our experiments show that our synthetic data facilitates good generalization.

### D. TMT Dataset

A critical missing piece in the turbulence restoration literature is training data. The TMT simulator is able to simulate the data, but the source images and parameter space controlling the turbulence profile need to be clarified. The engineering question here is then *what* kind of scenarios should we simulate, and *how much* turbulence should we inject? Ultimately, can we create a training dataset for the community to use instead of re-running our simulator?

To this end, we introduce the TMT dataset, which consists of static and dynamic parts. We use the place dataset [57] for synthesizing static scenes. We randomly selected 9,017 images in the original dataset as input for the simulator. We generated 50 turbulence images and their associated distortion-free images for every single input, resulting in 9,017 pairs of image sequences of static scenes. We split them into 7,499 pairs and 1,518 pairs for training and testing, respectively.

For dynamic scenes, the TMT dataset contains many videos. The source datasets for our dynamic scene data are the Sports Video in the Wild (SVW) dataset [58] and all ground truth videos used in TSRWGAN [12]. These videos are mixed, generating 4,684 samples with a total number of frames of 1,979,564. We generated 4,684 pairs of full turbulence and distortion-free videos, then randomly split them into 3,500 videos for training and 1,184 for testing, keeping at most 120 frames per testing video.

For synthesizing turbulence data, we have identified key turbulence parameters shown in Table III. We partition turbulence parameters in three levels: weak, medium, and strong. From experience, we notice that a long distance requires a larger diameter of the aperture. Also, a smaller Fried parameter [13] implies stronger turbulence, which further requires a larger blur kernel to produce. We empirically set temporal correlations to match the visual perception of the existing real-world data for better generalizability.

Table IV summarizes our dataset. The TMT dataset is by far the largest and most comprehensive dataset for atmospheric turbulence deep learning research. Our dataset is open to the public at <https://xg416.github.io/TMT>.

## V. EXPERIMENTS

### A. Testing Data

Our testing data consists of two parts. For quantitative evaluation (which involves PSNR calculations), we use the testing portion of the TMT dataset. We remark that while the testing data



TABLE III  
PARAMETER SAMPLING SETTING OF THE DATA SYNTHESIS

Dataset	Proportion	Kernel size	Aperture (m)	$D/r_0$	Distance (m)	Temporal correlation
Static	20% (weak)	[33]	$U(0.001, 0.005)$	[0.5, 1, 1.2, 1.5]	$U(100, 400)$	$U(0.2, 0.6)$
	40% (medium)	[33]	$U(0.04, 0.1)$	[1, 1.5, 2]	$U(400, 800)$	$U(0.2, 0.6)$
	40% (strong)	[33]	$U(0.1, 0.2)$	[1.5, 2, 3]	$U(800, 1500)$	$U(0.2, 0.6)$
Dynamic	33% (weak)	[9, 13, 15, 21]	$U(0.001, 0.005)$	[0.3, 0.6, 1, 1.2]	$U(50, 400)$	$U(0.4, 0.8)$
	33% (medium)	[11, 17, 25, 33]	$U(0.04, 0.1)$	[0.3, 1, 1.5]	$U(400, 800)$	$U(0.8, 0.95)$
	33% (strong)	[15, 21, 27, 33]	$U(0.1, 0.2)$	[1, 1.5, 2, 2.5]	$U(800, 2000)$	$U(0.88, 0.95)$

$U(a, b)$  denotes uniform distribution in the range of  $(a, b)$ .

TABLE IV  
SPECIFICATION OF THE TMT DATASET, WHERE EACH SEQUENCE FOR THE STATIC SCENE DATA HAS 50 FRAMES

	Static	Dynamic
Source	Place [57]	Sports [58] and TSRWGAN [12]
Amount	9,017 sequences	4,684 videos (1,979,564 frames)
Training	7,499 sequences	3,500 videos
Testing	1,518 sequences	1,184 videos

is technically in-distribution, they are generated independently from the training data. Ground truth images used for testing are never seen during training.

To examine the generalization of TMT, we test TMT and all competing methods using data from different sources, including the static patterns from OTIS [55] dataset, dynamic scene videos from [12] and the CLEAR [50] dataset. These datasets together contain a wide range of turbulence conditions. We present samples testing results on each in Figs. 7, 8, and 9.

### B. Training Pipeline

We compare the proposed TMT model with the TSRWGAN [12], a recent multi-frame turbulence mitigation method and two state-of-the-art general video restoration methods VRT [38], and BasicVSR++ [39], [59].

All models are trained separately for static and dynamic scene datasets. We trained our model using the Adam optimizer with the Cosine Annealing scheduler [60]. The VRT, BasicVSR++, and TSRWGAN models are all trained under the same settings as the original paper. For VRT and BasicVSR++, we use their configurations for motion deblurring as it most closely resembles our task. For a fair comparison, we set the input and output as 12 frames for TMT, VRT, and BasicVSR++. We fine-tuned the pre-trained TSRWGAN model for 600 K iterations on our dataset to maximize the performance. Hence, we kept the original 15 input frames setting, which should not put the method at any disadvantage. For the BasicVSR++ [39], we trained it from scratch for 600 K iterations with batch size 4. The other training scheme is the same as the original paper.

We first trained the tilt-removal module for 400 K iterations, and the batch size was set to 4. We used Adam optimizer [61] and the learning rate was initialized as  $2 \times 10^{-4}$  then gradually decreased to  $1 \times 10^{-6}$  by the Cosine Annealing scheduler [60]. We fixed the tilt-removal module later and trained the deblurring module for 600 K iterations. The same batch size, optimizer, and learning rate scheduler were applied.

For all experiments, we used the same data augmentation strategy: random Gamma correction, random saturation,

TABLE V  
COMPARISON ON THE **STATIC** SCENE DATASET

Methods / # frames	PSNR	SSIM	CW-SSIM	LPIPS(↓)
TurbNet [30] / 1	22.7628	0.6923	0.8230	0.4012
BasicVSR++ [59] / 12	26.5055	0.8121	0.9189	0.2587
TSRWGAN [12] / 15	25.2888	0.7784	0.8982	0.2243
VRT [38] / 12	27.4556	0.8287	0.9338	0.1877
TMT [ours] / 12	27.7309	0.8341	0.9376	0.1815
TMT [ours] / 20	<b>28.4421</b>	<b>0.8580</b>	<b>0.9452</b>	<b>0.1693</b>

TABLE VI  
COMPARISON ON THE **DYNAMIC** SCENE DATASET

Methods / # frames	PSNR	SSIM	CW-SSIM	LPIPS(↓)
TurbNet [30] / 1	24.2229	0.7149	0.8072	0.4445
BasicVSR++ [59] / 12	27.0231	0.8073	0.8653	0.2492
TSRWGAN [12] / 15	26.3262	0.7957	0.8596	0.2606
VRT [38] / 12	27.6114	0.8300	0.8691	0.2485
TMT [ours] / 12	27.8816	0.8318	0.8705	0.2475
TMT [ours] / 20	<b>28.0124</b>	<b>0.8352</b>	<b>0.8741</b>	<b>0.2412</b>

and random cropping, followed by a random combination of horizontal/vertical flipping and 90° rotation. Since real turbulence images are often captured using telephoto lenses with large F-number, a high ISO is required to improve the image quality, inevitably leading to non-negligible image noise. Therefore, we also injected Gaussian random noise from  $\mathcal{N}(0, \beta \mathbf{I})$ , where  $\beta \sim \text{uniform}(0, 0.0004)$  during training to enhance the robustness of the models.

### C. Quantitative Comparison With Other Networks

The neural networks we trained are evaluated on our synthetic testing set. We measured PSNR, SSIM, complex-wavelet SSIM (CW-SSIM) [62] and perceptual quality LPIPS [63] for quantitative comparison. The testing results are shown in Tables V and VI. The TMT outperforms the multi-frame turbulence mitigation network TSRWGAN [12] and 2 SOTA video restoration networks BasicVSR++ [39] and VRT [38] by a large margin. We noticed the CW-SSIM scores are closer to perceptual quality, and all metrics are mostly correlated when we do not use perceptual loss. We also provide a visual comparison in Fig. 4 to show the advance of our method.

We present the inference-time computing budget of all models we trained in Table VII. The measurement is based on a single NVIDIA 2080 Ti GPU. If the full size cannot fit into the GPU, we split the input into the largest possible patches. The splitting should be overlapped to reduce artifacts. Although VRT has the closest performance in terms of PSNR and SSIM to TMT, TMT requires much less computation and memory requirement; hence, it is much more efficient.

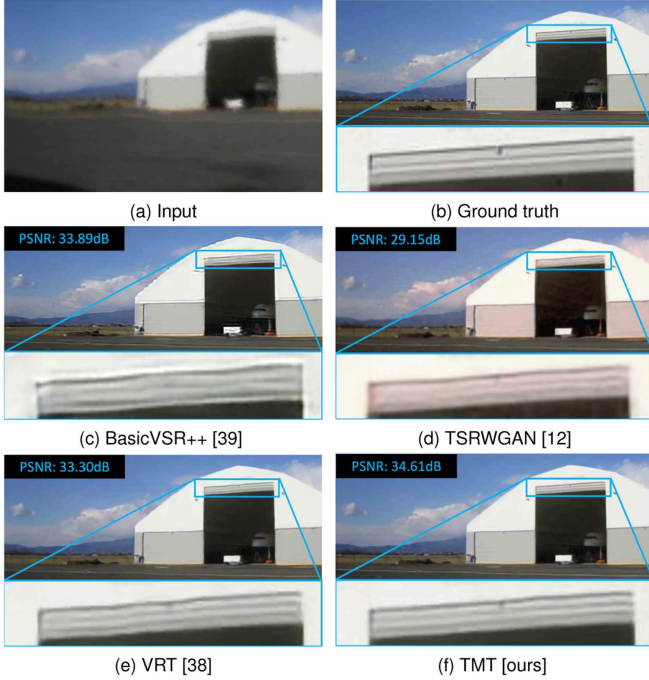


Fig. 4. Example of testing results on our synthetic static scene dataset. (a) Input (b) Ground truth (c) Output of **TMT [ours]** (d) The output of TSRWGAN [12] (e) The output of VRT [38] (f) Output of BasicVSR++ [39].

TABLE VII  
COMPARISON OF COMPUTATIONAL CONSUMPTION IN INFERENCE

Methods	# parameters (M)	FLOPs/frame (G)	speed (s)
Mao et al. [5]	-	-	~ 5500
CLEAR [4]	-	-	~ 20
BasicVSR++ [59]	9.76	338.4	0.08
TSRWGAN [12]	46.28	2471	1.15
VRT [38]	18.32	7756	5.88
TMT [ours]	26.04	1826	1.52

The speed is measured per frame on  $540 \times 960$  resolution images. The first two rows are conventional methods tested on CPU.

#### D. Ablation Study

*a) Influence of number of input frames:* To restore the clean image sequence under turbulence, the network must perceive a large enough area in the spatial-temporal domain to estimate the property of the degradation field, especially for samples having high temporal correlation. If the correlation factor is 1, all frames have the same degradation, the multi-frame reconstruction will collapse into a single-frame problem, which is extremely hard to solve, as revealed by [64]. To understand how the number of frames would affect the quality of restoration, we trained our TMT<sub>b</sub> using 4, 8, 12, and 20 input frames with our synthetic dataset. The testing performance is shown in Fig. 5. The results show the performance of the network can be improved with more input frames. We also note that to make a fair comparison with a strong baseline such as [38], which requires large GPU memory space so that it is hard to deploy with more than 12 frames, we kept a 12-frame setting in this paper. However, we can easily train with 20 frames to get a much better performance than VRT with the same hardware.

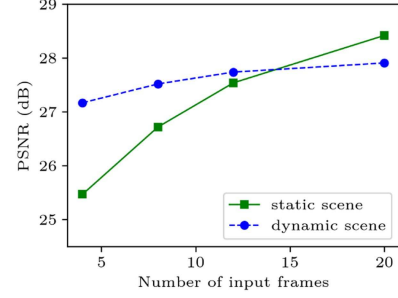


Fig. 5. Number of frames and performance.

*b) Two-stage design and the tilt-removal module:* To evaluate the value of the tilt-removal module, we trained our TMT backbone to restore images directly without removing the tilt. The result is shown in Table VIII. As it shows, the tilt-removal module can boost the overall performance by 0.02 ~ 0.08 dB. Tilt-removal is more effective in static scenes, as distinguishing true object motion from turbulence distortions in dynamic scenes introduces uncertainty. Note the tilt-removal module is very lightweight. If we scale up one-stage TMT, we can only add 10% more channels on the one-stage TMT and finally get around 0.01 dB improvement, much less than the two-stage model. Moreover, instead of more parameters, more frames have a more significant impact on the performance. The two-stage model can be scaled up to take more frames because they are applied sequentially. The one-stage model has a lower frame number limitation than the two-stage model, further restricting the performance.

*c) Influence of channel shuffle operation:* The purpose of the channel shuffle operation is to facilitate the communication of channels across different frames in the TCJA module. The comparison in Table VIII demonstrates the effectiveness of the channel shuffle operation. The performance benefits 0.02 ~ 0.03 dB from the shuffle operation compared to the plain connection.

*d) Influence of multi-scale input in the TMT backbone:* For general U-shaped design [65], only a single input is required to feed into the encoder. Several works [66], [67], [68] suggest feeding inputs in multiple scales/resolutions could help the network to learn more efficiently. In the proposed paper, we utilize multi-scale input by simply downsampling the raw input into lower resolutions. To demonstrate the advantage of this operation, we also trained our TMT with single-scale input: the input dimension of each level in the encoder remains the same, only the input feature is changed from the concatenation of outputs produced by the last level, and the 3D convolution layer to solely from the last level. The comparison in Table VIII demonstrates the effectiveness of the multi-scale input. The PSNR of networks with multi-scale input is ~ 0.15 dB higher than that with single-scale input, while the consumption complexity does not increase.

#### E. Generalization to Real-World Data

With our physically-based simulation process, samples of real-world turbulence can be viewed as the interpolation of samples generated by our simulation. Other simulation methods are



TABLE VIII  
QUANTITATIVE COMPARISON ON THE SYNTHETIC TESTING SET FOR DIFFERENT DESIGN CHOICES OF THE TMT

Dataset	Static Scenes				Dynamic Scenes				Computation Consumption	
Methods	PSNR	SSIM	PSNR <sub>Y</sub>	SSIM <sub>Y</sub>	PSNR	SSIM	PSNR <sub>Y</sub>	SSIM <sub>Y</sub>	# Params (M)	FLOPs (G)
SS-TMT <sub>a</sub> w.o. warp	27.2782	0.8221	28.7316	0.8398	27.5635	0.8258	29.0644	0.8459	<b>22.75</b>	1490
SS-TMT <sub>b</sub> w.o. warp	27.3092	0.8235	28.7626	0.8411	27.5864	0.8265	29.0886	0.8465	22.95	1514
SS-TMT <sub>a</sub>	27.3432	0.8257	28.8013	0.8431	27.5779	0.8282	29.0863	0.8487	24.87	1676
SS-TMT <sub>b</sub>	27.3836	0.8266	28.8396	0.8440	27.6051	0.8281	29.1110	0.8485	25.07	1700
MS-TMT <sub>a</sub> w.o. warp	27.4718	0.8291	28.9243	0.8452	27.6637	0.8301	29.1685	0.8501	23.70	<b>1206</b>
MS-TMT <sub>b</sub> w.o. warp	27.5003	0.8301	28.9602	0.8468	27.6841	0.8310	29.1894	0.8506	23.92	1304
MS-TMT <sub>a</sub>	27.5215	0.8307	28.9781	0.8469	27.7239	<b>0.8329</b>	29.2337	<b>0.8526</b>	25.82	1392
MS-TMT <sub>b</sub> [Proposed]	<b>27.5422</b>	<b>0.8320</b>	<b>29.0011</b>	<b>0.8487</b>	<b>27.7419</b>	0.8323	<b>29.2510</b>	0.8520	26.04	1490

PSNR<sub>Y</sub> and SSIM<sub>Y</sub> indicate PSNR and SSIM Measured in the YCbCr space. TMT<sub>a</sub> and TMT<sub>b</sub> denote the TMT model with variants a and b of the attention module, respectively. “warp” means the tilt-removal module, “SS” and “MS” are aberrations of single-scale input and multi-scale input. FLOPs are measured with an input size of  $12 \times 208 \times 208$ . All networks are trained with batch size 1 and 800 k iterations.



Fig. 6. Test on sequences of real-world text data [56]. The three lines of images from top to bottom are the input and output of the 2nd, 24th, and 96th sequences in that dataset. Column (a). Input sequences (b). The output of BasicVSR++ [39] (c). The output of TSRWGAN [12] (d). VRT [38] (e). TMT [ours]

either much more time-consuming or inaccurate [3]. Two recent data-driven works, TSRWGAN [12] and complex CNN [50], also used synthetic data to train neural networks, but their generalization capability is limited. Quantitative comparisons among models on real-world images are hard because most real-world image sequences don’t have ground truth. Despite this problem, direct visual comparison can still illustrate how our synthetic data helps generalization.

*a) Visual comparison among models trained on our datasets:* In the CVPR 2022 UG2+ Challenge, a new long-range turbulence dataset is released for benchmarking turbulence mitigation algorithms [56]. This dataset consists of 100 image sequences of text patterns captured from 300 meters away in hot weather. We tested our trained models on this dataset. All models have a certain generalization capability, but our network performs better than others under heavy turbulence. Several samples are given in Fig. 6.

Besides the turbulence text dataset, we tested all models on an earlier OTIS dataset [55] and TSRWGAN’s real-world test set. The OTIS has 16 sequences of static patterns in different

scales and turbulence levels. The TSRWGAN’s real-world test set has 27 dynamic scenes in relatively mild turbulence strength. Since the ground truth is unavailable, we only show some visual comparisons in Figs. 7 and 8, from which one can easily conclude that models trained on our dataset generalize well on a broad range of real-world turbulence conditions and among them, our model could restore images in better visual quality than others.

*b) Impact of a better simulator:* The simulation tool that TSRWGAN [12] used is [69], which generates physically valid tilts, but the blur is spatially invariant. [12] also produces synthetic data by artificial heat sources to create turbulence effects in a short distance. However, this approach tends to generate highly correlated degradation with a weak blur. We observed that their released model does not generalize well on CLEAR’s real-world dataset, OTIS [55] and the text dataset [56] which contain stronger turbulence effect captured at longer range. We fine-tuned TSRWGAN on our synthetic data, showing a significant improvement of the TSRWGAN model on those out-of-distribution datasets in Fig. 9.

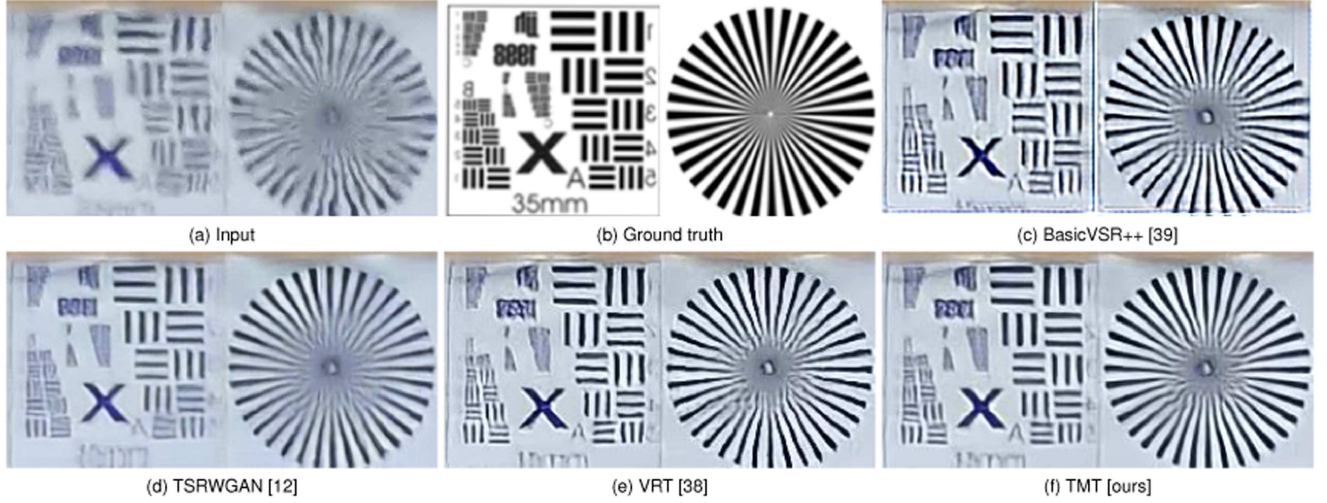


Fig. 7. Visual comparison on the OTIS dataset [55]. We show the result of Patterns 15 and 16 which contain the strongest turbulence effect.

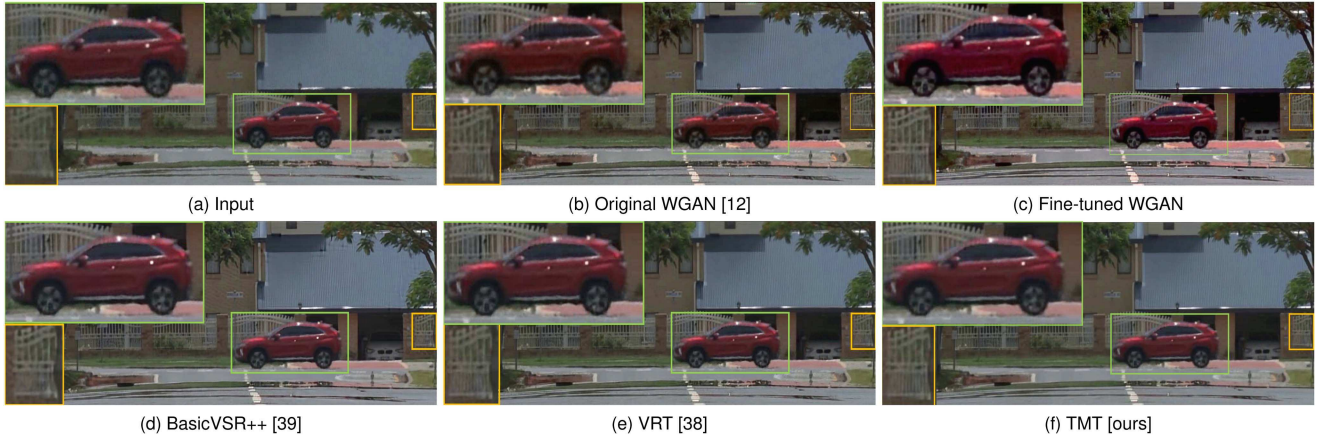


Fig. 8. Visual comparison on real turbulence data from TSRWGAN's dataset [12]. Model of (b) is provided by authors of [12], and the model for (c) is fine-tuned on our data. From the comparison between (b) and (c), we find the robustness to relatively strong turbulence has been enhanced using our data, despite the artifacts of railing on the window still existing. Notice (f) has the least turbulence left while preserving the most details.

*c) Compare with the simulation method used in [50]:* Recent learning-based extension [50] of CLEAR [4] employed nine predefined PSF of atmospheric turbulence introduced in [8] to simulate the degradation. Those PSFs are not physically grounded and lack diversity, so the restoration performance on real-world data is not ideal. Since their trained model was not released, we could not train it on our dataset to demonstrate the potential improvement. Despite this, we can compare the output from our tilt-removal module with the result of [50] on their real-world videos. As shown in Fig. 11, our lightweight tilt-removal module outperforms a more complex model, which suggests the generalization capability is from our synthetic dataset.

#### F. Comparison With Conventional Methods

We make a comparison with two state-of-the-art conventional turbulence mitigation methods [4], [5]. Due to the slow processing speed of the conventional methods, we only compare the

TABLE IX  
QUANTITATIVE COMPARISON WITH CONVENTIONAL TURBULENCE MITIGATION METHODS ON THE 100 SCENES DATASET

Methods	PSNR	SSIM	CW-SSIM	LPIPS(↓)
Mao et al. [5]	27.4154	0.8174	0.9262	0.1998
CLEAR [4]	26.9969	0.8295	0.9293	0.1986
BasicVSR++ [59]	28.0811	0.8410	0.9368	0.1852
TSRWGAN [12]	25.8139	0.7977	0.9006	0.2315
VRT [38]	28.3454	0.8512	0.9411	0.1727
TMT-12f [ours]	<b>28.4543</b>	<b>0.8539</b>	<b>0.9448</b>	<b>0.1640</b>

The first two rows are conventional methods.

methods on a random subset of 100 scenes from our synthetic static dataset. The comparison of the 100 scenes is shown in Table IX. Conventional methods could still work because our synthetic data has similar properties compared to real-world turbulence-degraded images. For example, the lucky effect can be synthesized with this simulator [3]. It is worth mentioning



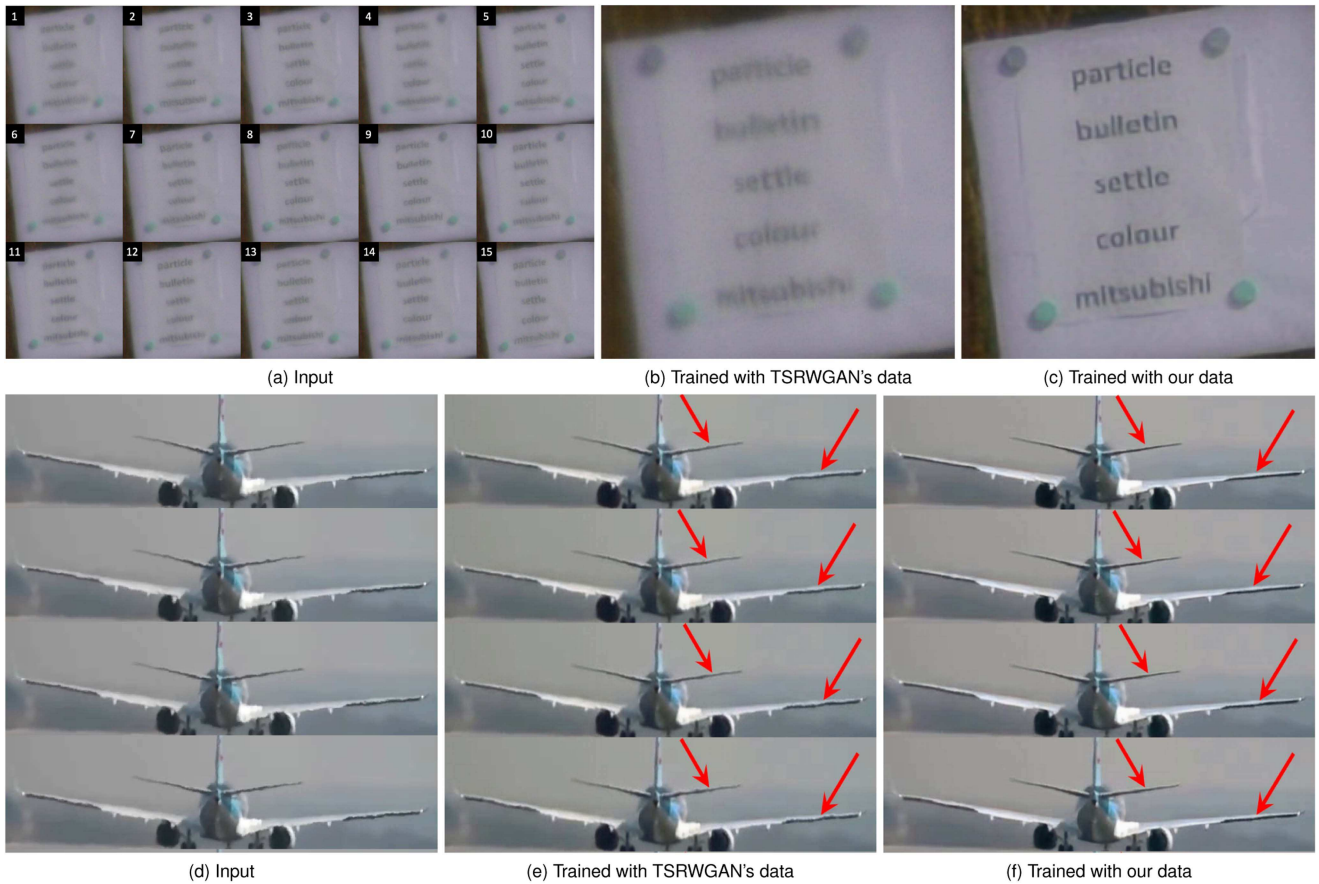


Fig. 9. Test TSRWAGN's generator on real-world images. (a)–(c). On text dataset [56] (d)–(f). On CLEAR's dataset (a). Input sequence (d). Single frame sample of input (b)–(e). Restoration result of TSRWAGN trained with their original data (c)–(f). Restoration result of TSRWAGN trained with our synthetic datasets. Clearly, our datasets directly facilitate the generalization of real-world images

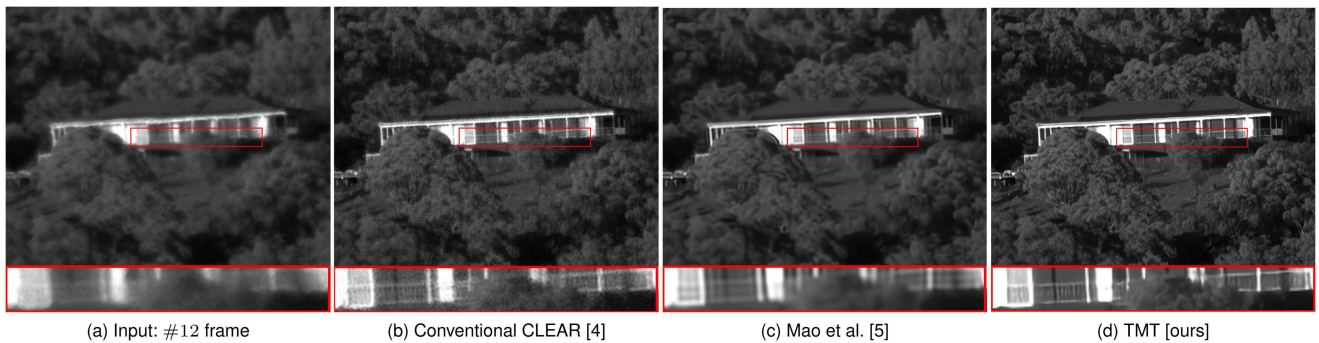


Fig. 10. Comparison with representative conventional methods on the real-world image sequence. Our model is trained on the synthetic static scene dataset. The presented input frame is the 12th in the whole sequence. Zoom in for a better view of details.

that we used 50 frames as input for conventional methods and 12/15 for learning-based methods.

Additionally, we compare our proposed method with [4] and [5] on real-world images. A qualitative comparison is presented in Fig. 10. CLEAR [4] used a subsample mechanism to select high-quality images among all 75 frames of the input sequence, then used a complex wavelet-based algorithm to fuse selected frames. After those operations, the output is still

subjected to explicit denoising, while our method blindly uses the first 12 frames as input, and the noise is mostly compressed. Mao et al. [5] is based on the lucky fusion algorithm, a popular framework in conventional turbulence mitigation tasks. It has an explicit denoising stage after the fusion, making the result more blurry. It can be seen that our method produced more natural and high-contrast reconstruction without further loss of details.





Fig. 11. Application of our tilt-removal module on real-world data from [50]. Images on the top are full-resolution images; below are the same patches in 3 frames starting from them.

## VI. CONCLUSION

In this paper, we presented a complete and generalizable solution for a multi-frame blind atmospheric turbulence mitigation problem. We refined a physics-based simulation in [3] by approximating the spatially varying field using a wide sense stationary field. We further proposed an efficient transformer-based network to restore image sequences degraded by turbulence, and our method can adapt to various scenes and turbulence conditions. Our model surpassed the state-of-the-art models designed for general video restoration tasks. Compared to conventional methods of turbulence mitigation, the proposed method has superior performance in visual quality, speed, and robustness.

## ACKNOWLEDGMENT

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] R. C. Hardie, J. D. Power, D. A. LeMaster, D. R. Droege, S. Gladysz, and S. Bose-Pillai, "Simulation of anisoplanatic imaging through optical turbulence using numerical wave propagation with new validation analysis," *SPIE Opt. Eng.*, vol. 56, 2017, Art. no. 071502.
- [2] C. P. Lau and L. M. Lui, "Subsampled turbulence removal network," *Mathematics, Computation Geometry Data*, vol. 1, no. 1, pp. 1–33, 2021.
- [3] Z. Mao, N. Chimitt, and S. H. Chan, "Accelerating atmospheric turbulence simulation via learned phase-to-space transform," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14 759–14 768.
- [4] N. Anantrasirichai, A. Achim, N. G. Kingsbury, and D. R. Bull, "Atmospheric turbulence mitigation using complex wavelet-based fusion," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2398–2408, Jun. 2013.
- [5] Z. Mao, N. Chimitt, and S. H. Chan, "Image reconstruction of static and dynamic scenes through anisoplanatic turbulence," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1415–1428, 2020.
- [6] C. P. Lau, Y. H. Lai, and L. M. Lui, "Restoration of atmospheric turbulence-distorted images via RPCA and quasiconformal maps," *Inverse Problems*, vol. 35, no. 7, 2019, Art. no. 074002.
- [7] X. Zhu and P. Milanfar, "Removing atmospheric turbulence via space-invariant deconvolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 157–170, Jan. 2013.
- [8] M. Hirsch, S. Sra, B. Schölkopf, and S. Harmeling, "Efficient filter flow for space-variant multiframe blind deconvolution," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 607–614.
- [9] R. Yasarla and V. M. Patel, "Learning to restore images degraded by atmospheric turbulence using uncertainty," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 1694–1698.
- [10] N. G. Nair and V. M. Patel, "Confidence guided network for atmospheric turbulence mitigation," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 1359–1363.
- [11] C. P. Lau, C. D. Castillo, and R. Chellappa, "ATFaceGAN: Single face semantic aware image restoration and recognition from atmospheric turbulence," *IEEE Trans. Biom., Behav., Ident. Sci.*, vol. 3, no. 2, pp. 240–251, Feb. 2021.
- [12] D. Jin et al., "Neutralizing the impact of atmospheric turbulence on complex scene imaging via deep learning," *Nature Mach. Intell.*, vol. 3, pp. 876–884, 2021.
- [13] D. L. Fried, "Probability of getting a lucky short-exposure image through turbulence," *J. Opt. Soc. Amer.*, vol. 68, no. 12, pp. 1651–1658, Dec. 1978.
- [14] A. N. Kolmogorov, "The local structure of turbulence in incompressible viscous fluid for very large reynolds' numbers," *Akademiia Nauk SSSR Doklady*, vol. 30, pp. 301–305, 1941.
- [15] M. C. Roggemann, B. M. Welsh, D. Montera, and T. A. Rhoadarmer, "Method for simulating atmospheric turbulence phase effects for multiple time slices and anisoplanatic conditions," *Appl. Opt.*, vol. 34, no. 20, pp. 4037–4051, Jul. 1995.
- [16] J. D. Schmidt, *Numerical Simulation of Optical Wave Propagation: With Examples in MATLAB*. Bellingham, WA, USA: SPIE Press, Jan. 2010.

- [17] S. H. Chan and N. Chimitt, "Computational imaging through atmospheric turbulence," *Found. Trends Comput. Graph. Vis.*, vol. 15, no. 4, pp. 253–508, 2023.
- [18] S. H. Chan, "Tilt-then-blur or blur-then-tilt? Clarifying the atmospheric turbulence model," *IEEE Signal Process. Lett.*, vol. 29, pp. 1833–1837, 2022.
- [19] J. Primot, G. Rousset, and J. C. Fontanella, "Deconvolution from wavefront sensing: A new technique for compensating turbulence-degraded images," *J. Opt. Soc. Amer. A*, vol. 7, no. 9, pp. 1598–1608, 1990.
- [20] M. A. Vorontsov and G. W. Carhart, "Anisoplanatic imaging through turbulent media: Image recovery by local information fusion from a set of short-exposure images," *J. Opt. Soc. Am. A*, vol. 18, no. 6, pp. 1312–1324, Jun. 2001.
- [21] M. Shimizu, S. Yoshimura, M. Tanaka, and M. Okutomi, "Super-resolution from image sequence under influence of hot-air optical turbulence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
- [22] Y. Lou, S. H. Kang, S. Soatto, and A. L. Bertozzi, "Video stabilization of atmospheric turbulence distortion," *Inverse Problems Imag.*, vol. 7, no. 3, pp. 839–861, 2013.
- [23] O. Oreifej, X. Li, and M. Shah, "Simultaneous video stabilization and moving object detection in turbulence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 450–462, Feb. 2012.
- [24] J. Gilles and S. Osher, "Wavelet burst accumulation for turbulence mitigation," *J. Electron. Imag.*, vol. 25, no. 3, 2016, Art. no. 033003.
- [25] R. C. Hardie, M. A. Rucci, A. J. Dapone, and B. K. Karch, "Block matching and Wiener filtering approach to optical turbulence mitigation and its application to simulated and real imagery with quantitative error analysis," *Proc. SPIE*, vol. 56, 2017, Art. no. 071503.
- [26] M. Aubailly, M. A. Vorontsov, G. W. Carhart, and M. T. Valley, "Automated video enhancement from a stream of atmospherically-distorted images: The lucky-region fusion approach," *Proc. SPIE*, vol. 7463, pp. 104–113, 2009.
- [27] Y. Mao and J. Gilles, "Non rigid geometric distortions correction-application to atmospheric turbulence stabilization," *Inverse Problems Imag.*, vol. 6, no. 3, pp. 531–546, 2012.
- [28] Y. Xie, W. Zhang, D. Tao, W. Hu, Y. Qu, and H. Wang, "Removing turbulence effect via hybrid total variation and deformation-guided kernel regression," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4943–4958, Oct. 2016.
- [29] R. Yasarla and V. M. Patel, "CNN-Based restoration of a single face image degraded by atmospheric turbulence," *IEEE Trans. Biom., Behav., Ident. Sci.*, vol. 4, no. 2, pp. 222–233, Apr. 2022.
- [30] 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 *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 430–446.
- [31] S. N. Rai and C. Jawahar, "Removing atmospheric turbulence via deep adversarial learning," *IEEE Trans. Image Process.*, vol. 31, pp. 2633–2646, 2022.
- [32] N. G. Nair, K. Mei, and V. M. Patel, "AT-DDPM: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3434–3443.
- [33] J.-H. Choi, H. Zhang, J.-H. Kim, C.-J. Hsieh, and J.-S. Lee, "Evaluating robustness of deep image super-resolution against adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 303–311.
- [34] J.-H. Choi, H. Zhang, J.-H. Kim, C.-J. Hsieh, and J.-S. Lee, "Deep image destruction: Vulnerability of deep image-to-image models against adversarial attacks," in *Proc. IEEE 26th Int. Conf. Pattern Recognit.*, 2022, pp. 1287–1293.
- [35] B. Y. Feng, M. Xie, and C. A. Metzler, "TurboGAN: An adversarial learning approach to spatially-varying multiframe blind deconvolution with applications to imaging through turbulence," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 3, pp. 543–556, Sep. 2022.
- [36] N. Li, S. Thapa, C. Whyte, A. W. Reed, S. Jayasuriya, and J. Ye, "Unsupervised non-rigid image distortion removal via grid deformation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2522–2532.
- [37] S. H. Chan, R. Khoshabeh, K. B. Gibson, P. E. Gill, and T. Q. Nguyen, "An augmented Lagrangian method for total variation video restoration," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3097–3111, Nov. 2011.
- [38] J. Liang et al., "VRT: A video restoration transformer," 2022, Accessed: Jan. 20, 2024, [arXiv:2201.12288](https://arxiv.org/abs/2201.12288). [Online]. Available: <https://arxiv.org/abs/2201.12288>
- [39] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR : Improving video super-resolution with enhanced propagation and alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5972–5981.
- [40] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17683–17693.
- [41] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 1833–1844.
- [42] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5728–5739.
- [43] J. P. Bos and M. C. Roggemann, "Technique for simulating anisoplanatic image formation over long horizontal paths," *SPIE Opt. Eng.*, vol. 51, no. 10, 2012, Art. no. 101704.
- [44] J. D. Schmidt, *Numerical Simulation of Optical Wave Propagation: With Examples in MATLAB*. Bellingham, WA, USA: SPIE Press, 2010.
- [45] N. Chimitt and S. H. Chan, "Simulating anisoplanatic turbulence by sampling correlated Zernike coefficients," in *Proc. IEEE Int. Conf. Comput. Photogr.*, 2020, pp. 1–12.
- [46] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, Accessed: Jan. 20, 2024. [Online]. Available: <https://openreview.net/forum?id=YichFdnNTTy>
- [47] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.
- [48] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4161–4170.
- [49] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
- [50] N. Anantrasirichai, "Atmospheric turbulence removal with complex-valued convolutional neural network," *Pattern Recognit. Lett.*, vol. 171, pp. 69–75, 2023.
- [51] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 298–311, Feb. 1997.
- [52] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.
- [53] N. Chimitt, X. Zhang, Z. Mao, and S. H. Chan, "Real-time dense field phase-to-space simulation of imaging through atmospheric turbulence," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 1159–1169, 2022.
- [54] M. C. Roggemann and B. M. Welsh, *Imaging Through Atmospheric Turbulence, Ser. Laser & Optical Science & Technology*. New York, NY, USA: Taylor & Francis, 1996.
- [55] J. Gilles and N. B. Ferrante, "Open turbulent image set (OTIS)," *Pattern Recognit. Lett.*, vol. 86, pp. 38–41, 2017.
- [56] "Bridging the gap between computational photography and visual recognition: 5thUG2+ prize challenge," track 3, 2022. [Online]. Available: [http://cvpr2022.ug2challenge.org/dataset22\\_t3.html](http://cvpr2022.ug2challenge.org/dataset22_t3.html)
- [57] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
- [58] S. M. Safdarnejad, X. Liu, L. Udpa, B. Andrus, J. Wood, and D. Craven, "Sports videos in the wild (SVW): A video dataset for sports analysis," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2015, pp. 1–7.
- [59] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "On the generalization of BasicVSR to video deblurring and denoising," 2022, Accessed: Jan. 20, 2024, [arXiv:2204.05308](https://arxiv.org/abs/2204.05308). [Online]. Available: <https://arxiv.org/abs/2204.05308>
- [60] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.
- [61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, Accessed: Jan. 20, 2024. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [62] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2385–2401, Nov. 2009.
- [63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
- [64] D. Vint, G. D. Caterina, J. Soraghan, R. Lamb, and D. Humphreys, "Analysis of deep learning architectures for turbulence mitigation in long-range imagery," *Proc. SPIE*, vol. 11543, 2020, Art. no. 1154303.

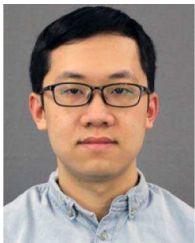


- [65] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
- [66] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3883–3891.
- [67] B. Ji and A. Yao, "Multi-scale memory-based video deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1919–1928.
- [68] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 821–14 831.
- [69] E. Repasi and R. Weiss, "Computer simulation of image degradations by atmospheric turbulence for horizontal views," *Proc. SPIE*, vol. 8014, 2011, Art. no. 80140U.



**Xingguang Zhang** (Graduate Student Member, IEEE) received the B.E. degree in optoelectronic information science and engineering from Zhejiang University, Hangzhou, China, in 2018, and the M.S. degree in electrical and computer engineering from Purdue University in 2021. In 2021, he joined the i2Lab, as a Ph.D. Student. His research interests include turbulence modeling and mitigation, computational imaging, object detection, and machine learning for video restoration. His research works have been published at top computer vision conferences,

such as CVPR/ICCV, and journals such as the IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING. He was an Organizer of the 6th UG2+ workshop at the CVPR 2023.



**Zhiyuan Mao** received the B.Eng. degree (with the first class honor) in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2017, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2023. He is currently a Senior Research Engineer with Samsung Research America. His research mainly focuses on computational imaging. During his Ph.D., he developed image restoration methods for imaging through atmospheric turbulence and photon-limited imaging. he also works at the intersection

of computational imaging and computer vision by using computational imaging methods to improve the performance of visual recognition systems. His research works were published at top computer vision conferences, such as CVPR/ICCV/ECCV, and IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING. He regularly was Reviewer for international conferences and journals, including CVPR, ICCV, ECCV, ICIP, ICASSP, ICME, IEEE TCI, IEEE TIP, and Optics Express. He was an organizer of the 5th and 6th UG2+ workshops at the CVPR 2022 and 2023.



**Nicholas Chimitt** (Member, IEEE) received the B.S. degree in electrical engineering from Purdue University Northwest, Hammond, IN, USA, in 2017, and the Ph.D. degree from Purdue University, West Lafayette, IN, in 2023 where he is currently a post-doctoral research fellow. He is the co-author of the book *Computational Imaging through Atmospheric Turbulence*, Now Publishers 2023, and has given tutorials on imaging through turbulence at CVPR and ICIP. His research interests include computational imaging, wave propagation through random media, and statistical image modeling. His research works have been published at top computer vision conferences, such as CVPR and ICCV, as well as IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING and *SPIE Optical Engineering*.



**Stanley H. Chan** (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from the University of Hong Kong, Hong Kong, in 2007, the M.A. degree in mathematics, and the Ph.D. degree in electrical engineering from the University of California San Diego, San Diego, CA, USA, in 2009 and 2011, respectively. From 2012 to 2014, he was a Post-doctoral Research Fellow with Harvard University, Cambridge, MA, USA. In 2014, he joined Purdue University, West Lafayette, IN, where he is currently an Elmore Associate Professor of ECE. He is the

author of a popular undergraduate textbook *Introduction to Probability for Data Science*, Michigan Publishing 2021 and a co-author of the book *Computational Imaging through Atmospheric Turbulence*, Now Publisher 2023. His research interests include single-photon imaging, imaging through atmospheric turbulence, and computational photography. He was the recipient of the 2022 IEEE Signal Processing Society Best Paper Award. He is the Senior Area Editor of IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING. He also was an Associate Editor for the OSA *Optics Express* during 2016–2018, and an Elected Member of the IEEE Signal Processing Society Technical Committee in Computational Imaging during 2015–2020.