LEARNING GEOMETRY-STABILIZED INTERMEDIATE REPRESENTATIONS FOR REAL-WORLD ATMOSPHERIC TURBULENCE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Atmospheric turbulence introduces geometric distortions and spatially varying blur, making it challenging to restore images captured over long distances. Although learning-based methods have shown promising results, their generalization to real-world scenes is often limited by reliance on synthetic training data that may not fully capture actual imaging conditions. Motivated by the geometry-stabilizing effect of temporal integration, we introduce a geometry-stabilized intermediate representation that can be learned directly from real turbulent videos while retaining residual blur. This representation enables a short-to-representation-to-clean framework that decouples single-frame restoration into geometric stabilization and detail restoration, with the latter learned from physics-guided blurred/clean pairs. We further propose a training-free algorithm that registers and refines the intermediate representations, extending the framework to a flexible number of input frames while keeping the networks fixed. Across five real-world benchmarks, our model achieves strong performance in perceptual quality, image sharpness, temporal stability, and downstream text recognition. Our code will be publicly released.

1 INTRODUCTION

Atmospheric turbulence is a major source of degradation in long-range optical imaging (Fried, 1966). Turbulent mixing of air with different temperatures and densities produces refractive-index fluctuations that distort propagating wavefronts, giving rise to geometric distortions and spatially varying blur (Chan & Chimitt, 2023; Jin et al., 2021). These degradations obscure fine details and cause apparent image motion across frames, making distant objects and text difficult to interpret and recognize (Mao et al., 2022). Restoring clear and geometrically stable imagery is therefore important for applications such as remote sensing and long-range surveillance, as well as for supporting reliable downstream visual analysis (Gal et al., 2014; Anantrasirichai et al., 2013).

Recent deep learning methods have achieved promising results for turbulence restoration, but their ability to generalize depends heavily on how well the training data represent real-world degradations. However, obtaining paired turbulent and clean images under real imaging conditions is challenging, particularly at the scale required for training (Zeng et al., 2026). This difficulty has motivated continued advances in turbulence simulation to generate paired training data (Mao et al., 2021; Zhang et al., 2024a; Hu et al., 2026). Yet faithfully capturing the diversity of real turbulence remains difficult: its stochastic distortions vary across space and time, with characteristics that depend on atmospheric conditions, propagation paths, and camera settings. Consequently, synthetic training data may leave substantial gaps in the degradation patterns encountered in practice (Zhang et al., 2024a; Zeng et al., 2026). In this work, our focus is not on designing a specialized network architecture or improving turbulence simulation. Instead, we aim to learn directly from real turbulent observations to achieve robust generalization, even when only a single frame is available at inference.

Although turbulence-induced distortions vary stochastically, temporal integration in the imaging system reveals a useful regularity: for static scenes, increasing exposure duration tends to reduce frame-to-frame geometric fluctuations while increasing blur (Zeng et al., 2026). Figure 1(a) illustrates this transition, and the quantitative analysis in Figure 1(b) shows the corresponding reduction

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

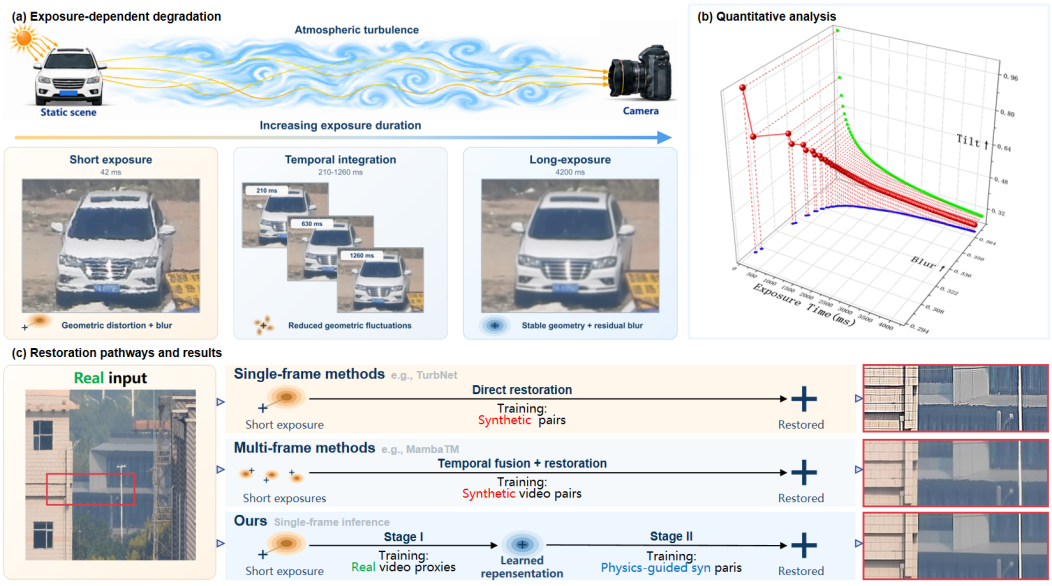

Figure 1: **Temporal-integration motivation and geometry-stabilized restoration.** (a) Longer exposure stabilizes geometry while retaining blur. (b) Tilt–blur trajectory across 800 scenes (detailed in Appendix B.2). (c) Single- and multi-frame methods versus our two-stage framework.

in geometric distortion and increase in blur. These observations motivate decoupling turbulence restoration into geometric stabilization and detail recovery through a learned intermediate representation. We characterize this representation as an image with reduced frame-dependent geometric variation and residual blur, designed to support subsequent detail restoration. Notably, this representation can be learned directly from real degraded observations: temporal integration provides the design cue, while real video sequences supply the supervision targets.

Specifically, for each static-scene training video, we first average multiple short-exposure frames to construct a scene-level temporal-average proxy shared by all observations of that scene. Although finite averages can retain residual distortions (Chan & Chimitt, 2023), the proxy provides a common content target; consistency between predictions from different frames further discourages observation-dependent variations. At inference, the representation is predicted from an individual frame rather than formed by temporal aggregation. Second, we construct physics-guided blurred/clean pairs to train the detail-restoration stage. Fried's long-exposure theory provides a compact frequency-domain description of centered blur under Kolmogorov turbulence assumptions (Fried, 1966), which serves as a physical reference. To accommodate deviations between this idealized response and the residual blur of learned representations, we extend the blur family to include Gaussian blur and intermediate frequency-decay profiles. Together, these stages connect real-video and physics-guided supervision in a short-to-representation-to-clean framework (Figure 1(c)). Furthermore, when additional observations are available, a training-free algorithm registers and refines their representations, further suppressing residual geometric distortions and improving temporal stability while keeping the networks fixed.

Our contributions are as follows:

- **A geometry-stabilized intermediate representation.** We learn from real turbulent videos a single-frame mapping that reduces observation-dependent geometric variation while retaining scene content and residual blur for subsequent restoration.
- **A representation-level decomposition of turbulence restoration.** The learned intermediate domain separates geometric stabilization from detail recovery, enabling real-video supervision and physics-guided restoration to be optimized independently.
- **A training-free multi-frame extension.** We register and refine learned representations to support flexible input frame counts and improve temporal stability without retraining.

2 Related Work

2.1 Turbulence Modeling and Simulation

Physics-based turbulence simulation provides paired training data by modeling random wavefront aberrations and their effects on image formation. To reduce the cost of numerical wave propagation, Chimitt & Chan (2020) formulate simulation as sampling spatially correlated Zernike coefficients. Mao et al. (2021) further accelerate image generation through a learned phase-to-space (P2S) transform. Subsequent developments improve the diversity and realism of generated degradations: ATSyn incorporates temporal correlation for video simulation (Zhang et al., 2024a), while D^{2}Turb incorporates scene depth to generate depth-dependent turbulence degradations (Hu et al., 2026). Exposure duration also shapes the observed degradation by determining how temporal fluctuations are integrated during image formation. Fried (1966) established a statistical description of optical resolution in the long- and short-exposure limits, while ET-MTF models blur continuously as a function of exposure time (Zeng et al., 2026). Despite these advances, faithfully reproducing both the diverse geometric distortions and blur of real-world turbulence within a unified simulation remains challenging. Rather than simulating the full degradation process, our framework learns geometric stabilization directly from real turbulent videos and restricts physics-guided simulation to residual blur in the learned intermediate domain.

2.2 Atmospheric Turbulence Restoration

Atmospheric turbulence restoration methods recover scene content from either a single observation or multiple complementary frames. For single-frame restoration, AT-Net uses estimated uncertainty to guide reconstruction (Yasarla & Patel, 2021), while TurbNet employs a physics-inspired transformer to jointly estimate turbulence distortion and restore image content (Mao et al., 2022). Multi-frame methods exploit temporal information through different aggregation mechanisms: TMT introduces temporal-channel joint attention (Zhang et al., 2024b), DATUM combines recurrent aggregation with deformable and temporal-channel attention (Zhang et al., 2024a), and MambaTM uses selective state space models guided by learned latent phase distortion (Zhang et al., 2025). CDSP combines a frequency-aware registration reference with low-rank refinement (Xu et al., 2024), while Turb-Seg-Res separates dynamic foregrounds from static backgrounds before restoration (Saha et al., 2024). However, existing methods typically couple geometric correction and detail recovery within a single restoration process, or rely on multiple observations for temporal stabilization. In contrast, we introduce an explicit geometry-stabilized intermediate representation that decouples geometric stabilization from detail recovery and provides a common interface for both single- and multi-frame restoration.

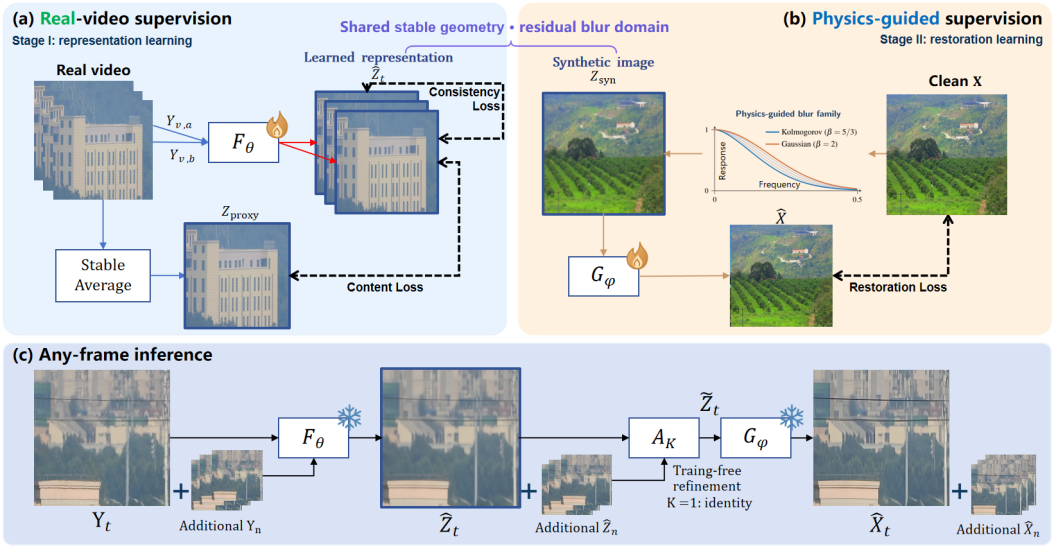

Figure 2: **A geometry-stabilized restoration interface.** (a) Real-video proxies and consistency train F_θ; (b) physics-guided pairs train G_ϕ; (c) $\mathcal{A}_K$ refines representations.

3 METHOD

As illustrated in Figure 2, our framework decomposes atmospheric turbulence restoration into two stages connected by a geometry-stabilized intermediate representation. Given a turbulence-degraded frame Y_t, the first network F_θ maps it to a geometry-stabilized representation $\widehat{Z}_t$ by suppressing frame-dependent geometric fluctuations while leaving residual blur for subsequent detail restoration. The second network G_ϕ then removes the residual blur and recovers the clean estimate $\widehat{X}_t$:

$$\widehat{Z}_t = F_\theta(Y_t), \qquad \widehat{X}_t = G_\phi(\widehat{Z}_t). \tag{1}$$

The two networks are trained using real-video supervision and physics-guided blurred/clean pairs, respectively. For multi-frame inference, F_θ processes the target frame and its neighboring frames independently, after which a training-free algorithm registers and refines their intermediate representations before G_ϕ restores the target image. Section 3.1 develops the physical motivation; Sections 3.2–3.4 present representation learning, detail restoration, and multi-frame refinement.

3.1 TEMPORAL INTEGRATION FOR GEOMETRY STABILIZATION

This section analyzes how temporal integration suppresses turbulence-induced geometric fluctuations and characterizes the resulting residual blur.

From a single observation to temporal integration. For a clean static scene X, we model an individual turbulent observation under the short-exposure approximation as the composition of geometric displacement and residual aberration blur:

$$Y_t(\mathbf{x}) = \mathcal{B}_t[\mathcal{T}_t(X)]\,(\mathbf{x}), \qquad \mathcal{T}_t(X)(\mathbf{x}) = X(\mathbf{x} + \mathbf{u}_t(\mathbf{x}))\,. \tag{2}$$

Here, $\mathcal{T}_t$ is the geometric warping operator induced by the displacement field $\mathbf{u}_t$, which arises mainly from tip/tilt, and $\mathcal{B}_t$ describes the remaining optical blur. We use angular image coordinates $\mathbf{x}$. When the exposure duration is much shorter than the atmospheric correlation time τ_{atm}, the image approximately captures one realization of the turbulence-induced distortion. As the exposure duration increases, multiple temporally varying turbulence realizations are integrated. The exposure-normalized signal over duration T is

$$Z_{t,T}(\mathbf{x}) = \frac{1}{T}\int_t^{t+T} \mathcal{B}_s[\mathcal{T}_s(X)]\,(\mathbf{x})\,ds. \tag{3}$$

As illustrated in Figure 1(a), this integration changes both the position and the spread of the effective imaging response.

Temporal integration converts geometric fluctuations into spatial spread. To quantify how temporal integration changes geometric distortion, we separate the displacement into a residual centroid shift and a within-exposure fluctuation component:

$$\mathbf{u}_s = \overline{\mathbf{u}}_{t,T} + \widetilde{\mathbf{u}}_{s,T}, \qquad \overline{\mathbf{u}}_{t,T} = \frac{1}{T}\int_t^{t+T} \mathbf{u}_s\,ds. \tag{4}$$

The mean displacement $\overline{\mathbf{u}}_{t,T}$ determines the remaining centroid offset, whereas $\widetilde{\mathbf{u}}_{s,T}$ describes fluctuations around that centroid. For locally stationary, zero-mean displacement, let $c_u(\tau) = \mathbb{E}[\mathbf{u}_{s+\tau}^\top \mathbf{u}_s]$ denote the temporal autocorrelation at lag τ, with the spatial argument suppressed for brevity. In particular, $c_u(0) = \mathbb{E}\|\mathbf{u}_s\|^2$ is the instantaneous displacement energy. The expected squared centroid offset D_T and the expected within-exposure displacement spread S_T satisfy

$$\begin{aligned} D_T &= \mathbb{E}\|\overline{\mathbf{u}}_{t,T}\|^2 = \frac{2}{T^2}\int_0^T (T-\tau)c_u(\tau)\,d\tau, \\ S_T &= \mathbb{E}\left[\frac{1}{T}\int_t^{t+T} \|\widetilde{\mathbf{u}}_{s,T}\|^2 ds\right] = c_u(0) - D_T. \end{aligned} \tag{5}$$

For $T \ll \tau_{\mathrm{atm}}$, $D_T \approx c_u(0)$ and $S_T \approx 0$: the image largely retains an instantaneous displacement. With integrable temporal covariance, $D_T \to 0$ and $S_T \to c_u(0)$ as $T \to \infty$. Temporal integration therefore suppresses centroid fluctuations while converting the sampled displacements into spatial spread. Here, S_T accounts for the displacement-induced contribution to blur; residual aberrations contribute additional blur through $\mathcal{B}_s$. These relations describe statistical behavior rather than monotonic changes in every finite observation. The derivation is provided in Appendix A.

Long-exposure limit and residual blur. Within a locally shift-invariant patch, the centered blur has optical transfer function (OTF) $B_s(\boldsymbol{\nu})$ at angular spatial frequency $\boldsymbol{\nu}$. The Fourier shift theorem separates centroid displacement from the centered, integrated response:

$$\mathcal{F}[Z_{t,T}](\boldsymbol{\nu}) = e^{i2\pi\boldsymbol{\nu}^\top \overline{\mathbf{u}}_{t,T}} \left[\frac{1}{T} \int_t^{t+T} B_s(\boldsymbol{\nu}) e^{i2\pi\boldsymbol{\nu}^\top \widetilde{\mathbf{u}}_{s,T}} ds \right] \mathcal{F}[X](\boldsymbol{\nu}). \tag{6}$$

Under stationary, ergodic turbulence with zero-mean displacement, the centroid factor approaches one and the bracketed response converges to the ensemble OTF. For Gaussian phase increments with Kolmogorov structure function $D_\phi(\boldsymbol{\rho}) = 6.88(\|\boldsymbol{\rho}\|/r_0)^{5/3}$, Fried's theory gives (Fried, 1966)

$$H_{\mathrm{LE}}(\boldsymbol{\nu}) = \mathbb{E}\Big[B_s(\boldsymbol{\nu}) e^{i2\pi\boldsymbol{\nu}^\top \mathbf{u}_s}\Big] = H_{\mathrm{dif}}(\boldsymbol{\nu}) \exp\left[-3.44 \left(\frac{\lambda\|\boldsymbol{\nu}\|}{r_0} \right)^{5/3} \right], \tag{7}$$

where r_0 is the Fried coherence diameter, λ the wavelength, and H_{dif} the diffraction-limited OTF. The real, even atmospheric factor describes centered blur. The limiting response defines the ideal long-exposure image $Z_{\mathrm{LE}} = \mathcal{F}^{-1}[H_{\mathrm{LE}}\mathcal{F}[X]]$, with stabilized geometry and centered residual blur. Appendix A provides the derivation and pixel-frequency conversion.

3.2 Learning Geometry-Stabilized Representations from Real Videos

Stage I learns geometry-stabilized representations directly from real static-scene turbulent videos. The ideal image Z_{LE} corresponds to the statistical long-exposure limit, whereas real videos provide a finite set of temporally sampled turbulence observations. We therefore use a finite-frame temporal average from each video as a practical supervision proxy. For a real static-scene video v corrected for global translational drift, we average a fixed set $\mathcal{I}_v$ of N_v frames to form $Z_v^{\mathrm{proxy}} = N_v^{-1} \sum_{t\in\mathcal{I}_v} Y_{v,t}$. All training frames from video v share this scene-specific proxy. Although finite-frame averaging reduces geometric variation, the resulting proxy Z_v^{proxy} may still retain residual fluctuations and is therefore used as a scene-level anchor rather than an exact Z_{LE}.

Using this proxy as supervision, we aim to learn a geometry-stabilized representation $\widehat{Z}_t = F_\theta(Y_t)$ that preserves scene content while further reducing frame-dependent geometric variation. Under the stationary model in Section 3.1, different turbulent observations of the same static scene share a common statistical reference. We therefore impose inter-frame consistency on the outputs of F_θ through the representation-learning loss $\mathcal{L}_{\mathrm{I}}$.

Specifically, we sample two neighboring frame indices a and b from each training sequence and process the two inputs independently with F_θ (Figure 2(a)). Let $F_\theta^{(q)}$ denote the output at scale $q \in \mathcal{S} = \{1, 1/2, 1/4\}$ and D_q bilinear resizing to that scale. We minimize

$$\begin{aligned} \mathcal{L}_{\mathrm{I}} = \frac{1}{2} \sum_{i\in\{a,b\}} \sum_{q\in\mathcal{S}} \mathcal{L}_{\mathrm{Char}}\Big(F_\theta^{(q)}(Y_{v,i}), D_q Z_v^{\mathrm{proxy}}\Big) \\ + \lambda_{\mathrm{con}} \Big\| F_\theta^{(1)}(Y_{v,a}) - F_\theta^{(1)}(Y_{v,b}) \Big\|_1 . \end{aligned} \tag{8}$$

The Charbonnier term anchors predictions to the proxy's content at multiple scales, while the full-resolution consistency term, weighted by λ_{con}, suppresses variations across observations. Preprocessing, loss definitions, and training settings are given in Appendix B.

3.3 Physics-Guided Detail Restoration

We train Stage II on physics-guided blurred/clean pairs generated from clean images X in Places365 (Zhou et al., 2018). To accommodate residual-blur variations in the learned representations, we extend the frequency-decay model in Eq. (7) as

$$H_{\gamma,\beta}(\mathbf{f}) = \exp\big(-\gamma\|\mathbf{f}\|^\beta\big), \qquad \mathcal{B}_{\gamma,\beta}(I) = \mathcal{C}\,\mathcal{F}^{-1}[H_{\gamma,\beta}\mathcal{F}(\mathcal{P}I)]\,. \tag{9}$$

Here, $\mathbf{f}$ is frequency in cycles per pixel, $\gamma > 0$ controls blur strength, and $\mathcal{P}, \mathcal{C}$ denote reflection padding and cropping, respectively. The exponent $\beta = 5/3$ retains the Kolmogorov decay, whereas $\beta = 2$ corresponds to Gaussian blur. Sampling $\beta \in [5/3, 2]$ spans these reference profiles and

the intermediate frequency decays. Turbulence-induced blur is also spatially varying. To model smooth changes in local blur strength, we mix two responses with a shared exponent β and different strengths using a spatial weight map $w \in [0,1]$, generated by upsampling and smoothing a coarse random field:

$$Z_{\text{lin}} = w \odot \mathcal{B}_{\gamma_1,\beta}(\Gamma(X)) + (1-w) \odot \mathcal{B}_{\gamma_2,\beta}(\Gamma(X)), \tag{10}$$

where Γ denotes an approximate inverse-gamma transform that maps encoded image intensities to a linear domain, and $\odot$ denotes pointwise multiplication. Clipping to $[0,1]$, applying Γ^{-1}, and quantizing to eight bits yield Z_{syn} and the training pair (Z_{syn}, X). Using these pairs, we supervise outputs $G_\phi^{(q)}$ at the same scales q (Figure 2(b)) with spatial and complex-spectrum losses:

$$\mathcal{L}_{\text{II}} = \sum_{q \in \mathcal{S}} \Big[\Big\| G_\phi^{(q)}(Z_{\text{syn}}) - D_q X \Big\|_1 + \lambda_{\text{fft}} \Big\| \Psi\Big(G_\phi^{(q)}(Z_{\text{syn}})\Big) - \Psi(D_q X) \Big\|_1 \Big], \tag{11}$$

where $\Psi(I) = [\operatorname{Re}\mathcal{F}(I), \operatorname{Im}\mathcal{F}(I)]$ retains both real and imaginary Fourier components, including phase information; λ_{fft} balances the terms. Appendix B gives the blur-strength sampling, spatial mixing, and optimization settings used for Stage-II pair generation and training.

3.4 Training-Free Multi-Frame Refinement

For multi-frame inference, geometry-stabilized representations from the same sequence are used to construct a temporal reference for correcting residual geometric variations before detail restoration. For target t, we define a temporal window $\mathcal{W}_t$ containing up to K frames, including the target itself, and shift the window near sequence boundaries to preserve context. With $\widehat{Z}_k = F_\theta(Y_k)$, the training-free operator $\mathcal{A}_K$ refines these representations before detail restoration (Figure 2(c)):

$$\widetilde{Z}_t = \mathcal{A}_K(\{\widehat{Z}_k\}_{k \in \mathcal{W}_t}; t), \qquad \widehat{X}_t = G_\phi(\widetilde{Z}_t). \tag{12}$$

Both networks remain fixed for any frame count; with one observation, $\mathcal{A}_K$ is the identity.

Algorithm 1 summarizes $\mathcal{A}_K$. We first form the pixelwise median $R_t^{(0)}$ from $\{\widehat{Z}_k\}_{k \in \mathcal{W}_t}$; sparse correspondences to this reference yield partial-affine motion estimates via RANSAC. Large translations or rotations trigger a bypass, $\widetilde{Z}_t = \widehat{Z}_t$.

Otherwise, ECC with a Euclidean model (Evangelidis & Psarakis, 2008) aligns the inputs to $R_t^{(0)}$, yielding $Z_{k,t}^{\text{g}}$ and their median R_t. This median guides the target's local geometric correction.

Farnebäck flow (Farnebäck, 2003) from R_t to $Z_{t,t}^{\text{g}}$ is spatially smoothed and magnitude-clipped to obtain $\mathbf{f}_t$. The local warp $A_t(\mathbf{x}) = Z_{t,t}^{\text{g}}(\mathbf{x} + \mathbf{f}_t(\mathbf{x}))$ expresses the target in reference coordinates.

Algorithm 1: Training-Free Multi-Frame Restoration

Input: sequence $\{Y_k\}$, index t, size K, fixed F_θ, G_ϕ.
Output: restored target $\widehat{X}_t$.

Select $\mathcal{W}_t$; compute $\widehat{Z}_k \leftarrow F_\theta(Y_k)$ for $k \in \mathcal{W}_t$.
if $|\mathcal{W}_t| = 1$ **then return** $G_\phi(\widehat{Z}_t)$.
$R_t^{(0)} \leftarrow \text{Med}(\{\widehat{Z}_k\})$.
$\{m_k\} \leftarrow \text{RANSAC}(\{\widehat{Z}_k\}, R_t^{(0)})$.
if $\text{MotionGate}(\{m_k\}; t)$ **then**
  $\widetilde{Z}_t \leftarrow \widehat{Z}_t$.
else
  $Z_{k,t}^{\text{g}} \leftarrow \text{ECC}(\widehat{Z}_k, R_t^{(0)})$; $R_t \leftarrow \text{Med}(\{Z_{k,t}^{\text{g}}\})$.
  $\mathbf{f}_t \leftarrow \text{SmoothClip}(\text{Flow}(R_t, Z_{t,t}^{\text{g}}))$.
  $A_t \leftarrow \text{Warp}(Z_{t,t}^{\text{g}}, \mathbf{f}_t)$; compute $\widetilde{Z}_t$ by Eq. (13).
end if
return $\widehat{X}_t \leftarrow G_\phi(\widetilde{Z}_t)$.

Finally, we correct the target's low-frequency component using the registered reference:

$$\widetilde{Z}_t = \mathcal{Q}\big[A_t + \alpha\big(B_\sigma(R_t) - B_\sigma(A_t)\big)\big], \tag{13}$$

where B_σ is Gaussian filtering, α controls correction strength, and $\mathcal{Q}$ clips and quantizes. Before this last operation, the high-frequency residual $A_t - B_\sigma(A_t)$ is retained, while low frequencies are blended; $\alpha = 1$ fully replaces them with those of R_t.

This correction stabilizes low-frequency structure while preserving available high-frequency details, after which the refined representation is passed to G_ϕ for detail restoration. Registration thresholds, flow smoothing, and reference-correction parameters for Algorithm 1 are provided in Appendix B.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. Five real-world benchmarks cover scene restoration and text recognition: RLR-AT (Xu et al., 2024) (65 RLR-AT-Static and 60 RLR-AT-Dynamic test scenes), UoB-Turb (Hill & Anantrasirichai, 2024) (38 outdoor sequences), CLEAR (Anantrasirichai et al., 2013; 2018) (13 turbulence-degraded sequences), and TurbText (Mao et al., 2022) (100 scenes with five annotated text regions each).

Metrics. Without paired clean references, we assess perceptual quality using LIQE (Zhang et al., 2023), MUSIQ (Ke et al., 2021), NIMA (Talebi & Milanfar, 2018), MANIQA (Yang et al., 2022), Q-Align (Wu et al., 2024), ARNIQA (Agnolucci et al., 2024), and HyperIQA (Su et al., 2020); sharpness is measured by CPBD (Narvekar & Karam, 2011) and LPC-SI (Hassen et al., 2013). Temporal stability is measured by nTII↓ and nTSS↑, defined in Appendix B.3. On TurbText, we report word recognition accuracy (%) with fixed CRNN (Shi et al., 2017), DAN (Wang et al., 2020), and ASTER (Shi et al., 2019) recognizers.

Compared methods. We compare against single-frame methods AT-Net (Yasarla & Patel, 2021) and TurbNet (Mao et al., 2022), and multi-frame methods Turb-Seg-Res (Saha et al., 2024), CDSP (Xu et al., 2024), TMT (Zhang et al., 2024b), DATUM (Zhang et al., 2024a), and MambaTM (Zhang et al., 2025). For fair comparison, all baselines use official code or released weights; Appendix B details the training, inference, and evaluation protocols.

4.2 SINGLE-FRAME AND MULTI-FRAME RESTORATION

Single-frame restoration. Table 1 shows advantages across three dimensions: our method leads 26 of 28 perceptual comparisons, achieves the highest LPC-SI sharpness scores on all four benchmarks, and improves temporal stability, reducing nTII by 8.2–12.8% and increasing nTSS by 6.5–13.2% over the best respective baselines. In Figure 3, our results show straighter window frames, more regular character strokes, and cleaner aircraft contours. The representation column shows stabilized geometry before deblurring; Stage II then restores edge definition while preserving the stabilized structures. Notably, AT-Net and TurbNet show strong local contrast around text and aircraft boundaries, but exhibit jagged edge artifacts and residual warping. These differences illustrate the benefit of stabilizing geometry before restoring details.

Table 1: Single-frame restoration on four real-world benchmarks. Best and second-best scores within each dataset are **bold** and underlined.

		Perceptual quality							Sharpness		Temporal stability	
Dataset	Method	LIQE↑	MUSIQ↑	NIMA↑	MANIQA↑	Q-Align↑	ARNIQA↑	HyperIQA↑	CPBD↑	LPC-SI↑	nTII↓	nTSS↑
RLR-AT-Static	AT-Net	1.372	56.515	4.185	0.333	2.283	0.516	0.396	0.640	0.959	0.961	38.559
	TurbNet	1.978	66.921	4.498	0.374	2.276	**0.651**	0.564	**0.855**	0.973	1.223	29.585
	TMT	1.801	62.051	4.561	0.405	2.134	0.644	0.514	0.790	0.966	1.225	29.500
	DATUM	1.768	61.767	4.547	0.386	2.303	0.624	0.486	0.788	0.959	1.099	33.400
	MambaTM	1.755	58.503	4.421	0.371	2.467	0.550	0.469	0.777	0.947	0.975	37.793
	Ours	**2.539**	**67.356**	**4.859**	**0.427**	**2.672**	0.604	**0.656**	0.834	**0.975**	**0.838**	**43.646**
RLR-AT-Dynamic	AT-Net	1.242	56.450	3.913	0.290	2.378	0.537	0.383	0.619	0.962	0.953	38.725
	TurbNet	1.530	65.761	4.149	0.320	2.355	**0.641**	0.530	**0.844**	**0.974**	1.201	30.242
	TMT	1.705	60.185	4.279	0.335	2.517	0.631	0.459	0.743	0.963	1.075	34.173
	DATUM	1.602	56.888	4.158	0.320	2.566	0.576	0.418	0.703	0.957	1.013	36.315
	MambaTM	1.634	58.879	4.242	0.328	2.665	0.560	0.431	0.747	0.957	0.978	37.619
	Ours	**2.127**	**66.743**	**4.722**	**0.369**	**2.771**	0.605	**0.631**	0.785	**0.974**	**0.867**	**42.103**
UoB-Turb	AT-Net	1.593	41.579	3.697	0.245	2.159	0.420	0.299	0.331	0.944	1.026	35.986
	TurbNet	1.823	45.328	3.871	0.245	2.284	0.499	0.337	**0.554**	0.957	1.170	31.160
	TMT	1.777	43.605	4.015	0.285	2.394	0.530	0.336	0.503	0.958	1.119	32.828
	DATUM	1.530	37.840	3.808	0.243	2.297	0.385	0.311	0.303	0.939	0.983	37.444
	MambaTM	1.634	40.119	3.859	0.250	2.364	0.379	0.311	0.308	0.943	0.942	39.070
	Ours	**2.356**	**55.577**	**4.633**	**0.325**	**2.910**	**0.585**	**0.466**	0.416	**0.969**	**0.865**	**42.375**
CLEAR	AT-Net	1.419	42.311	4.178	0.209	2.052	0.492	0.321	0.344	0.965	1.005	36.896
	TurbNet	1.238	34.467	4.055	0.171	1.842	0.420	0.258	0.318	0.955	1.114	33.981
	TMT	1.590	39.764	4.329	0.180	2.067	0.476	0.393	**0.591**	0.965	1.219	29.801
	DATUM	1.622	42.250	4.323	0.204	2.251	0.481	0.338	0.451	0.964	1.030	35.905
	MambaTM	1.048	34.809	4.649	0.231	1.639	0.466	0.334	0.311	0.957	0.965	40.174
	Ours	**1.671**	**48.515**	**4.666**	**0.242**	**2.403**	**0.515**	**0.409**	0.485	**0.968**	**0.861**	**42.799**

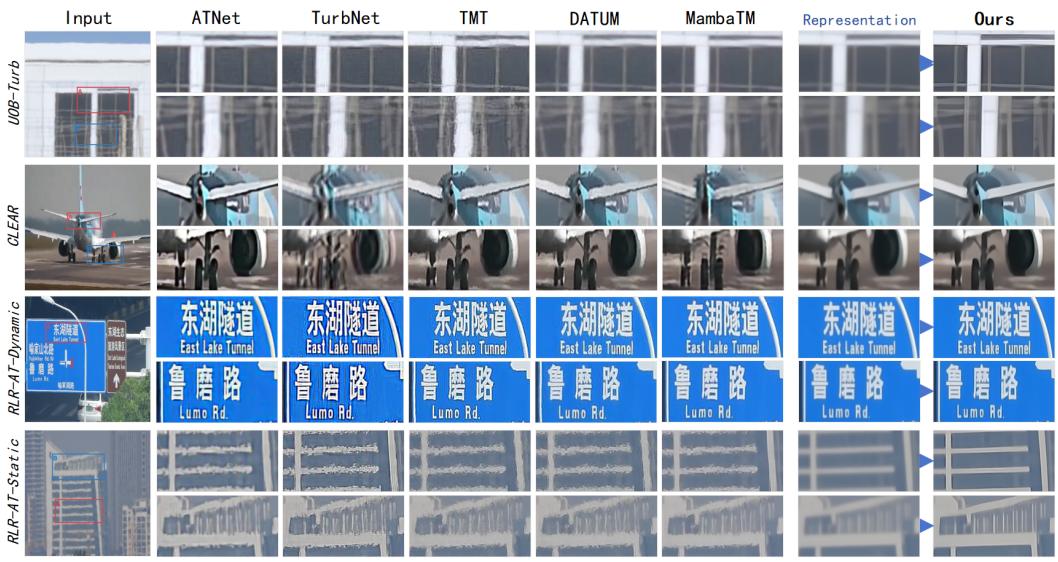

Figure 3: **Single-frame comparison.** Boxes mark crops; representations are Stage-I outputs.

Table 2: Twelve-frame restoration on RLR-AT-Static and word recognition accuracy (%) on Turb-Text. Best/second-best scores are **bold**/underlined.

	Perceptual quality							Sharpness		Temporal stability		Text Recognition		
Method	LIQE↑	MUSIQ↑	NIMA↑	MANIQA↑	Q-Align↑	ARNIQA↑	HyperIQA↑	CPBD↑	LPC-SI↑	nTII↓	nTSS↑	CRNN↑	DAN↑	ASTER↑
Turb-Seg-Res	1.668	65.154	4.602	0.404	2.336	0.600	0.461	0.796	0.964	0.991	37.378	19.312	24.864	28.246
CDSP	1.916	63.746	4.502	0.366	2.138	0.626	0.554	0.722	0.969	0.832	43.778	32.150	40.186	43.702
TMT	2.288	**67.940**	4.609	0.401	2.396	**0.656**	0.596	0.836	0.972	0.753	47.250	**78.058**	**80.580**	**85.172**
DATUM	2.077	66.292	4.636	0.406	2.431	0.634	0.542	0.838	0.967	0.780	45.927	30.174	39.988	48.376
MambaTM	1.999	66.233	4.622	0.399	2.462	0.615	0.540	**0.861**	0.960	0.818	44.268	42.330	68.698	77.100
Ours	**2.537**	67.571	**4.889**	**0.435**	**2.684**	0.604	**0.657**	0.831	**0.976**	**0.574**	**56.585**	68.400	71.600	77.200

Multi-frame restoration. In the twelve-frame setting, our method leads five of seven perceptual metrics and LPC-SI on RLR-AT-Static (Table 2). Relative to TMT, it reduces nTII by 23.8% and increases nTSS by 19.8%, consistent with the continuous space–time traces in Figure 4. Given the temporally varying nature of turbulence, we highly recommend that readers consult Appendix C.2 and watch the supplementary videos. On TurbText, Ours ranks second across all three recognizers. While "porcelain" is clearly restored, "banana" suggests a limitation under severe input blur: Stage I may yield a geometrically stabilized yet blurrier representation, obscuring details needed by Stage II. The final text remains less legible than TMT, consistent with the lower OCR accuracies in Table 2.

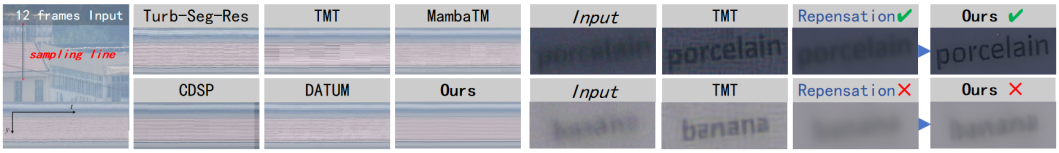

Figure 4: Temporal stability (left) and successful/failed text restoration (top/bottom right).

4.3 Ablation Study

Intermediate representation. We assess the learned representation as a restoration interface. First, Figure 5 shows that the median edge-based effective MTF lies within the blur-family envelope in Eq. (9) across sampled frequencies, with substantial overlap between the envelope and the 10th–90th percentile band. This supports frequency-response compatibility with the Stage-II blur family. Second, using the same fixed Stage-II model, we compare restorations obtained from individual turbulent frames, 5–50-frame temporal averages, and learned single-frame representations (Table 3). Outputs restored from learned representations outperform those from 50-frame averages across all six metrics, raising MUSIQ from 58.602 to 67.356 and CPBD from 0.753 to 0.834. In Figure 6's input row, the learned representation has more regular building edges than the 50-frame average, which retains geometric distortions; the restored row shows straighter, clearer structures. These results show that, compared with finite temporal averages, the learned representation provides a more effective input to the fixed Stage-II model, supporting its role as a restoration-oriented intermediate representation.

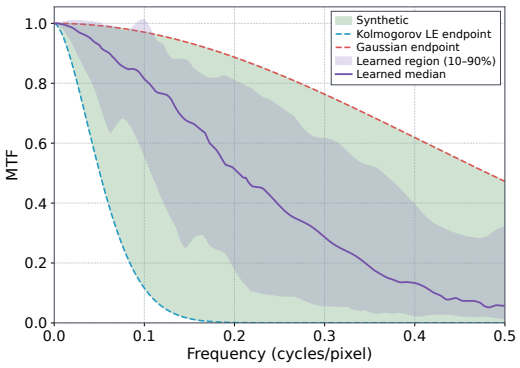

Figure 5: Learned effective MTFs (purple median and 10th–90th percentiles) and synthetic reference profiles (dashed); see Appendix B.2.

Table 3: Restoration with a fixed Stage II.

Input (frames)	LIQE↑	MUSIQ↑	NIMA↑	Q-Align↑	CPBD↑	LPC-SI↑
Short (1)	1.622	56.126	4.353	2.269	0.709	0.944
Average (5)	1.750	56.989	4.365	2.357	0.718	0.943
Average (10)	1.811	57.876	4.389	2.391	0.733	0.945
Average (25)	1.859	58.537	4.419	2.438	0.748	0.947
Average (50)	1.869	58.602	4.437	2.462	0.753	0.948
Learned (1)	2.539	67.356	4.859	2.672	0.834	0.975

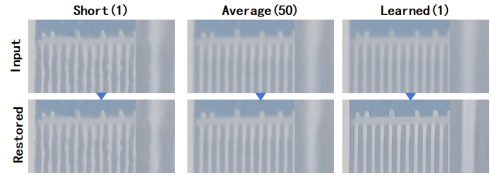

Figure 6: Real inputs and fixed Stage-II restorations for the same crop.

Restoration framework. Table 4 progressively ablates the full model. Removing inter-frame consistency weakens Stage I, suggesting that proxy fidelity alone leaves view-dependent variation. Replacing Charbonnier fidelity with ℓ_1 has less impact; switching to P2S training (Mao et al., 2021) under matched losses causes the largest MUSIQ drop, underscoring the value of real degradations. Removing spatial blur variation or fixing β weakens Stage II, consistent with varied residual blur in learned representations. Figure 7 shows that increasing α strengthens low-frequency correction while retaining visible facade detail, alongside rising nTSS. More frames improve stability at fixed and adaptive temporal strides, indicating suppression across multiple time scales.

Table 4: Two-stage framework ablation on RLR-AT-Static. Real: ✓ real, ✗ synthetic data. Charb.: ✗ denotes ℓ_1. $\mathcal{U}$: uniform $\beta \in [5/3, 2]$. Shaded: full model.

Stage I			Stage II		Perceptual							Sharpness	
Real	Charb.	Consist.	β	Spatial	LIQE↑	MUSIQ↑	NIMA↑	MANIQA↑	Q-Align↑	ARNIQA↑	HyperIQA↑	CPBD↑	LPC-SI↑
✓	✓	✓	$\mathcal{U}$	✓	2.539	67.356	4.859	0.427	2.672	0.604	0.656	0.834	0.975
✓	✓	✗	$\mathcal{U}$	✓	2.179	63.291	4.620	0.379	2.623	0.583	0.585	0.810	0.963
✓	✗	✗	$\mathcal{U}$	✓	2.130	62.277	4.606	0.371	2.646	0.569	0.568	0.806	0.963
✗	✗	✗	$\mathcal{U}$	✓	1.663	57.004	4.250	0.342	2.394	0.526	0.510	0.747	0.939
✓	✓	✓	$\mathcal{U}$	✗	2.225	63.151	4.707	0.403	2.594	0.574	0.617	0.799	0.966
✓	✓	✓	2	✗	1.994	60.710	4.605	0.385	2.491	0.551	0.593	0.775	0.962
✓	✓	✓	5/3	✗	2.016	60.639	4.603	0.381	2.542	0.553	0.576	0.774	0.961

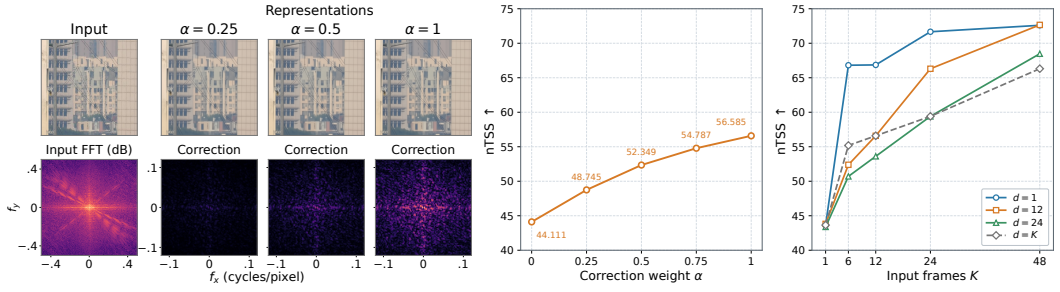

Figure 7: Training-free refinement. Left: input, refined representations, and spectral corrections relative to $\alpha = 0$ at $K = 12$ (input spectrum in dB; corrections on a shared linear scale). Center: nTSS versus correction weight α. Right: nTSS versus input count K at $d = 1, 12, 24, K$.

Limitations. Multi-frame refinement improves temporal stability, while gains in restoration quality and sharpness remain modest (Tables 1 and 2). The current extension uses training-free registration and reference-guided refinement of the intermediate representations. Future work could incorporate learned temporal aggregation within the geometry-stabilized domain to better exploit cross-frame dependencies and complementary details, particularly in dynamic scenes.

5 Conclusion

We present a two-stage turbulence restoration framework that decouples geometric stabilization from detail restoration through a learned geometry-stabilized representation. It bridges real-video supervision with physics-guided training, supports single- and multi-frame inference, and demonstrates strong performance across five real-world benchmarks.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. ARNIQA: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 189–198, 2024.

Nantheera Anantrasirichai, Alin Achim, Nick G. Kingsbury, and David R. Bull. Atmospheric turbulence mitigation using complex wavelet-based fusion. *IEEE Transactions on Image Processing*, 22(6):2398–2408, 2013. doi: 10.1109/TIP.2013.2249078.

Nantheera Anantrasirichai, Alin Achim, and David Bull. Atmospheric turbulence mitigation for sequences with moving objects using recursive image fusion. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 2895–2899, 2018. doi: 10.1109/ICIP.2018.8451755.

Stanley H. Chan and Nicholas Chimitt. Computational imaging through atmospheric turbulence. *Foundations and Trends in Computer Graphics and Vision*, 15(4):253–508, 2023. doi: 10.1561/0600000103. URL https://arxiv.org/abs/2411.00338.

Nicholas Chimitt and Stanley H. Chan. Simulating anisoplanatic turbulence by sampling intermodal and spatially correlated Zernike coefficients. *Optical Engineering*, 59(8):083101, 2020. doi: 10.1117/1.OE.59.8.083101. URL https://arxiv.org/abs/2004.11210.

Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Focal network for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13001–13011, 2023.

Georgios D. Evangelidis and Emmanouil Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008. doi: 10.1109/TPAMI.2008.113.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis (SCIA)*, pp. 363–370, 2003. doi: 10.1007/3-540-45103-X_50.

D. L. Fried. Optical resolution through a randomly inhomogeneous medium for very long and very short exposures. *Journal of the Optical Society of America*, 56(10):1372–1379, 1966. doi: 10.1364/JOSA.56.001372.

Ronen Gal, Nahum Kiryati, and Nir Sochen. Progress in the restoration of image sequences degraded by atmospheric turbulence. *Pattern Recognition Letters*, 48:8–14, 2014. doi: 10.1016/j.patrec.2014.04.007.

Joseph W. Goodman. *Introduction to Fourier Optics*. W. H. Freeman, Macmillan Learning, 4 edition, 2017. ISBN 9781319119164.

Rania Hassen, Zhou Wang, and Magdy M. A. Salama. Image sharpness assessment based on local phase coherence. *IEEE Transactions on Image Processing*, 22(7):2798–2810, 2013. doi: 10.1109/TIP.2013.2251643.

Paul Hill and Nantheera Anantrasirichai. Atmospheric turbulence dataset. Zenodo, 2024.

Zixiao Hu, Tianyu Li, Guoqing Wang, Wei Li, Guoguo Xin, Xun Liu, and Peng Wang. D^2Turb: Depth-aware simulation and decoupled learning for single-frame atmospheric turbulence mitigation. *arXiv preprint arXiv:2605.27460*, 2026. doi: 10.48550/arXiv.2605.27460. URL https://arxiv.org/abs/2605.27460.

Darui Jin, Ying Chen, Yi Lu, Junzhang Chen, Peng Wang, Zichao Liu, Sheng Guo, and Xiangzhi Bai. Neutralizing the impact of atmospheric turbulence on complex scene imaging via deep learning. *Nature Machine Intelligence*, 3(10):876–884, 2021. doi: 10.1038/s42256-021-00392-1.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Zhiyuan Mao, Nicholas Chimitt, and Stanley H. Chan. Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14759–14768, 2021.

Zhiyuan Mao, Ajay Jaiswal, Zhangyang Wang, and Stanley H. Chan. Single frame atmospheric turbulence mitigation: A benchmark study and a new physics-inspired transformer model. In *European Conference on Computer Vision*, pp. 430–446, 2022. URL https://arxiv.org/abs/2207.10040.

Niranjan D. Narvekar and Lina J. Karam. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011. doi: 10.1109/TIP.2011.2131660.

Ripon Kumar Saha, Dehao Qin, Nianyi Li, Jinwei Ye, and Suren Jayasuriya. Turb-Seg-Res: A segment-then-restore pipeline for dynamic videos with atmospheric turbulence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25286–25296, 2024.

Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2017. doi: 10.1109/TPAMI.2016.2646371.

Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. ASTER: An attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2035–2048, 2019. doi: 10.1109/TPAMI.2018.2848939.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3667–3676, 2020.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. doi: 10.1109/TIP.2018.2831899.

Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12216–12224, 2020. doi: 10.1609/aaai.v34i07.6903.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029, 2024.

Shengqi Xu, Run Sun, Yi Chang, Shuning Cao, Xueyao Xiao, and Luxin Yan. Long-range turbulence mitigation: A large-scale dataset and a coarse-to-fine framework. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2407.08377.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Rajeev Yasarla and Vishal M. Patel. Learning to restore images degraded by atmospheric turbulence using uncertainty. In *IEEE International Conference on Image Processing*, pp. 1694–1698, 2021. doi: 10.1109/ICIP42928.2021.9506614.

Junwei Zeng, Dong Liang, Sheng-Jun Huang, Kun Zhan, and Songcan Chen. Continuous exposure-time modeling for realistic atmospheric turbulence synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26678–26687, 2026. URL https://arxiv.org/abs/2603.01398.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023.

Xingguang Zhang, Nicholas Chimitt, Yiheng Chi, Zhiyuan Mao, and Stanley H. Chan. Spatio-temporal turbulence mitigation: A translational perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2889–2899, 2024a. URL https://arxiv.org/abs/2401.04244.

Xingguang Zhang, Zhiyuan Mao, Nicholas Chimitt, and Stanley H. Chan. Imaging through the atmosphere using turbulence mitigation transformer. *IEEE Transactions on Computational Imaging*, 10:115–128, 2024b. doi: 10.1109/TCI.2024.3354421.

Xingguang Zhang, Nicholas Chimitt, Xijun Wang, Yu Yuan, and Stanley H. Chan. Learning phase distortion with selective state space models for video turbulence mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2504.02697.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

A Physical Analysis of Temporal Integration

We provide the displacement, optical-transfer, and convergence arguments supporting the physical motivation in Section 3.1.

Centroid and displacement spread. For a fixed scene location, write $\overline{\mathbf{u}}_{t,T} = T^{-1}\int_t^{t+T} \mathbf{u}_s\, ds$ and $\widetilde{\mathbf{u}}_{s,T} = \mathbf{u}_s - \overline{\mathbf{u}}_{t,T}$. For a zero-mean stationary process, let $c_u(\tau) = \mathbb{E}[\mathbf{u}_{s+\tau}^\top \mathbf{u}_s]$. Its scalar autocovariance is even, so the expected squared centroid displacement is

$$\begin{aligned} D_T &= \frac{1}{T^2}\int_t^{t+T}\int_t^{t+T} \mathbb{E}[\mathbf{u}_s^\top \mathbf{u}_r]\, ds\, dr \\ &= \frac{1}{T^2}\int_{-T}^{T}(T-|\tau|)c_u(\tau)\, d\tau = \frac{2}{T^2}\int_0^T (T-\tau)c_u(\tau)\, d\tau. \end{aligned} \tag{A.1}$$

Expanding the squared centered displacement gives the pathwise identity

$$\frac{1}{T}\int_t^{t+T} \|\mathbf{u}_s - \overline{\mathbf{u}}_{t,T}\|^2 ds = \frac{1}{T}\int_t^{t+T} \|\mathbf{u}_s\|^2 ds - \|\overline{\mathbf{u}}_{t,T}\|^2. \tag{A.2}$$

Taking expectations yields $S_T = c_u(0) - D_T$. For $T \ll \tau_{\mathrm{atm}}$, $c_u(\tau) \approx c_u(0)$ over the integration interval, giving $D_T \approx c_u(0)$ and $S_T \approx 0$. If $\int_0^\infty |c_u(\tau)|\, d\tau < \infty$, then

$$0 \le D_T \le \frac{2}{T}\int_0^\infty |c_u(\tau)|\, d\tau \longrightarrow 0, \qquad S_T \longrightarrow c_u(0). \tag{A.3}$$

This partition describes displacement-induced spread, with higher-order aberration blur contained separately in the PSF. The limits concern expected behavior and do not imply monotonic convergence in every realization.

Local shift–blur factorization. Freezing the spatial dependence within a locally shift-invariant patch, let $H_{\mathrm{blur},t,T}$ denote the bracketed factor in Eq. (6), and define $Z_{t,T}^{\mathrm{blur}} = \mathcal{F}^{-1}[H_{\mathrm{blur},t,T}\mathcal{F}[X]]$. The Fourier shift theorem gives, within this local model,

$$\begin{aligned} Z_{t,T}(\mathbf{x}) &= Z_{t,T}^{\mathrm{blur}}(\mathbf{x} + \overline{\mathbf{u}}_{t,T}) \\ &= Z_{t,T}^{\mathrm{blur}}(\mathbf{x}) + \nabla Z_{t,T}^{\mathrm{blur}}(\mathbf{x})^\top \overline{\mathbf{u}}_{t,T} + \mathcal{O}(\|\overline{\mathbf{u}}_{t,T}\|^2). \end{aligned} \tag{A.4}$$

The patch-local factorization separates integrated blur from the residual centroid shift, which vanishes as the centroid converges to zero.

Pupil representation of the PSF and OTF. Let $P(\boldsymbol{\xi})$ be the real pupil transmission and $A_P = \int P^2(\boldsymbol{\xi})\, d\boldsymbol{\xi}$. For one scene direction, decompose the phase as $\phi_t(\boldsymbol{\xi}) = (2\pi/\lambda)\boldsymbol{\alpha}_t^\top \boldsymbol{\xi} + \phi_t^{\mathrm{ho}}(\boldsymbol{\xi})$, after removing its intensity-invariant constant component. Fourier optics gives the intensity-normalized PSF (Goodman, 2017; Chan & Chimitt, 2023):

$$h_t(\boldsymbol{\vartheta}) = \frac{1}{\lambda^2 A_P}\left|\int P(\boldsymbol{\xi}) e^{i\phi_t(\boldsymbol{\xi})} e^{-i2\pi \boldsymbol{\xi}^\top \boldsymbol{\vartheta}/\lambda}\, d\boldsymbol{\xi}\right|^2. \tag{A.5}$$

Substitution shifts the Fourier kernel by $\boldsymbol{\alpha}_t$, giving $h_t(\boldsymbol{\vartheta}) = h_t^{\mathrm{ho}}(\boldsymbol{\vartheta} - \boldsymbol{\alpha}_t)$. The image-warp convention in Eq. (2) gives $\mathbf{u}_t = -\boldsymbol{\alpha}_t$, while h_t^{ho} supplies the residual blur. The Fourier transform of Eq. (A.5) gives the pupil autocorrelation

$$H_t(\boldsymbol{\nu}) = \frac{1}{A_P}\int P(\boldsymbol{\xi})P(\boldsymbol{\xi} + \lambda\boldsymbol{\nu}) e^{i\Delta_{\lambda\boldsymbol{\nu}}\phi_t}\, d\boldsymbol{\xi}. \tag{A.6}$$

Here $\Delta_{\lambda\boldsymbol{\nu}}\phi_t = \phi_t(\boldsymbol{\xi}) - \phi_t(\boldsymbol{\xi} + \lambda\boldsymbol{\nu})$. Under homogeneous phase-increment statistics, the expectation of the exponential is independent of $\boldsymbol{\xi}$. It therefore factors out of the pupil integral, leaving $H_{\mathrm{dif}}(\boldsymbol{\nu}) = A_P^{-1}\int P(\boldsymbol{\xi})P(\boldsymbol{\xi} + \lambda\boldsymbol{\nu})\, d\boldsymbol{\xi}$. For zero-mean Gaussian phase increments, $\mathbb{E}[e^{i\Delta_{\lambda\boldsymbol{\nu}}\phi_t}] = e^{-D_\phi(\lambda\boldsymbol{\nu})/2}$. Combining this identity with the Kolmogorov structure function in Eq. (7) gives the stated long-exposure OTF. The resulting OTF is local to a scene direction.

Angular and pixel frequency conventions. The main-text derivation uses angular image coordinates, so $\boldsymbol{\nu}$ is measured in cycles per radian and $\lambda\boldsymbol{\nu}$ is a separation in the pupil plane. For an angular pixel pitch δ_θ and a digital frequency $\mathbf{f}$ in cycles per pixel, $\boldsymbol{\nu} = \mathbf{f}/\delta_\theta$. The Fried attenuation then has the form $\exp[-\gamma\|\mathbf{f}\|^{5/3}]$, with $\gamma = 3.44[\lambda/(r_0\delta_\theta)]^{5/3}$. The training family samples this attenuation scale and generalizes the exponent, rather than estimating physical exposure parameters.

Covariance of the integrated response. For completeness, define $H_{t,T} = T^{-1}\int_t^{t+T} H_s\,ds$, $H_{\mathrm{LE}} = \mathbb{E}[H_t]$, $\delta H_t = H_t - H_{\mathrm{LE}}$, and $C_H(\tau) = \mathbb{E}[\delta H_{t+\tau}\overline{\delta H_t}]$, suppressing the frequency argument. The fluctuation of the exposure average satisfies

$$\begin{aligned}\mathbb{E}\big[|H_{t,T} - H_{\mathrm{LE}}|^2\big] &= \frac{1}{T^2}\int_t^{t+T}\int_t^{t+T} \mathbb{E}[\delta H_s\overline{\delta H_u}]\,ds\,du \\ &= \frac{1}{T^2}\int_{-T}^{T}(T-|\tau|)C_H(\tau)\,d\tau \\ &= \frac{2}{T^2}\int_0^T (T-\tau)\,\mathrm{Re}\,C_H(\tau)\,d\tau,\end{aligned} \tag{A.7}$$

where stationarity gives $\mathbb{E}[\delta H_s\overline{\delta H_u}] = C_H(s-u)$ and $C_H(-\tau) = \overline{C_H(\tau)}$. If $\int_0^\infty |C_H(\tau)|\,d\tau$ is finite, the right-hand side is bounded above by $2T^{-1}\int_0^\infty |C_H(\tau)|\,d\tau$, which tends to zero. Since $\mathbb{E}[H_{t,T}] = H_{\mathrm{LE}}$, temporal integration reduces fluctuations around the ensemble response under these assumptions. In training, finite video averages serve as scene-content proxies as defined in Section 3.2.

B Training and Evaluation Details

B.1 Training and inference

Both stages use FocalNet (Cui et al., 2023), with 8 residual blocks in Stage I and 20 in Stage II. Each is trained for 50 epochs with Adam, batch size four, 256×256 crops, geometric augmentation, and cosine learning-rate decay from 10^{-4} to 10^{-6}. The Charbonnier loss in Eq. (8) is $\mathcal{L}_{\mathrm{Char}}(U,V) = M^{-1}\sum_{j=1}^{M}\sqrt{(U_j - V_j)^2 + 10^{-6}}$, where M counts scalar entries. Supervision uses full, half, and quarter resolution; the additional loss weights are $\lambda_{\mathrm{con}} = 1.6$ and $\lambda_{\mathrm{fft}} = 0.1$.

Stage I uses 800 RLR-AT-Static training scenes (Xu et al., 2024), disjoint from the 65 test scenes. Each provides 100 short-exposure inputs and one shared 50-frame temporal-average proxy. Global translation relative to a middle frame is corrected before averaging, yielding 80,000 frame–proxy pairs.

For Stage II, we generate 88,000 blurred/clean pairs from Places365 (Zhou et al., 2018). The two blur strengths are $\gamma_j = ba_j$, with $b \sim \mathcal{U}(3, 100)$, independent $a_j \sim \mathcal{U}(0.5, 1.5)$, and shared $\beta \sim \mathcal{U}(5/3, 2)$. The approximate inverse-gamma transform $\Gamma(x) = x^{2.2}$ maps each encoded RGB channel, normalized to $[0, 1]$, to an approximately linear intensity domain for filtering. A smoothed, bicubically upsampled random map with 128-pixel grid spacing blends the two branches. The result is clipped to $[0, 1]$, mapped back through $\Gamma^{-1}(x) = x^{1/2.2}$, and quantized to 8 bits without added noise.

For multi-frame inference, Algorithm 1 uses grayscale global registration at width 512. A corner-track/RANSAC gate bypasses refinement above target/window translation thresholds of 8/16 pixels or rotation thresholds of $0.25^\circ/0.35^\circ$. Accepted windows use Euclidean ECC and a median reference, followed by Farnebäck flow (three levels, 35-pixel window, five iterations). Flow is smoothed with $\sigma = 6$ pixels and clipped to two pixels; reference correction uses $\sigma = 1$ and $\alpha = 1$. The operator is the identity when $K = 1$.

B.2 Evaluation and measurements

The released weights for AT-Net (Yasarla & Patel, 2021) and TurbNet (Mao et al., 2022) were unavailable to us, so we retrain these single-frame methods on TMT-Static (Zhang et al., 2024b)

using their official training code and compare them only in Table 1. TMT (Zhang et al., 2024b), DATUM (Zhang et al., 2024a), and MambaTM (Zhang et al., 2025) use their released weights and official inference code in both the one-frame and twelve-frame settings of Tables 1 and 2. Turb-Seg-Res (Saha et al., 2024) and CDSP (Xu et al., 2024) require multiple input frames; we use their released weights and official inference code only in the twelve-frame setting of Table 2. Evaluation uses 100 frames per scene on RLR-AT-Static, RLR-AT-Dynamic, and UoB-Turb, and the central 12 on CLEAR, at 512×512. TurbText provides 100 sequences of 100 frames with five annotated words per frame. Fixed CRNN, DAN, and ASTER recognizers use the supplied crops; accuracy is exact word match after lowercasing and trimming whitespace.

Figure 1(b) uses CPBD (Narvekar & Karam, 2011) for blur and nTII for fluctuating geometry. Lower CPBD indicates stronger apparent blur; lower nTII indicates greater temporal stability (Appendix B.3). Here "tilt" is an image-based proxy, not a calibrated optical tilt angle.

For Figure 5, we select isolated edges from held-out RLR-AT-Static scenes. We use linearized luminance to sample each edge-spread function over ± 22 pixels, then Fourier-transform its windowed derivative. Equal-scene 10th, 50th, and 90th percentiles are reported over 0–0.5 cycles/pixel.

B.3 Temporal stability metric

Temporal stability is evaluated from frame-to-frame changes in a bounded structure representation. For a sequence $V = \{V_t\}_{t=1}^{T}$, each image is decoded as 8-bit grayscale and converted to floating point while retaining the intensity range $[0, 255]$. Let Ω denote the image grid and $\mathcal{G}_2$ Gaussian filtering with a standard deviation of two pixels. We form the high-pass response and its normalized structure map as

$$H_t = V_t - \mathcal{G}_2(V_t), \qquad S_t(p) = \frac{H_t(p)}{|H_t(p)| + \frac{1}{|\Omega|}\sum_{q\in\Omega}|H_t(q)| + 10^{-3}}. \tag{B.1}$$

The normalization bounds the structure response and reduces its dependence on high-frequency amplitude. For temporal stride d and $T > 2d$, the Temporal Instability Index is the mean absolute second-order temporal difference:

$$\mathrm{TII}_d(V) = \frac{1}{(T-2d)|\Omega|}\sum_{t=1}^{T-2d}\sum_{p\in\Omega}\left|S_{t+2d}(p) - 2S_{t+d}(p) + S_t(p)\right|. \tag{B.2}$$

For each scene, we normalize the restored sequence $\widehat{X}$ by its corresponding turbulent input Y with matching frame indices:

$$\mathrm{nTII}_d = \frac{\mathrm{TII}_d(\widehat{X})}{\max(\mathrm{TII}_d(Y), 10^{-6})}, \qquad \mathrm{nTSS}_d = 100\exp(-\mathrm{nTII}_d). \tag{B.3}$$

Lower nTII and higher nTSS indicate greater temporal stability relative to the input. Table 1 uses $d = 1$, while Table 2 uses $d = 12$. Figure 7 compares fixed strides $d \in \{1, 12, 24\}$ and adaptive stride $d = K \in \{1, 6, 12, 24, 48\}$, applied consistently to each restored sequence and its input reference. The exponential mapping is applied per scene before aggregation; reported nTII and nTSS are the arithmetic means of their respective scene-level values. The single-to-multi-frame comparison in Section 4.2 uses $d = 12$ for both settings, yielding single-frame nTII/nTSS of $0.831/43.822$; percentage changes are computed from unrounded means.

C Additional Experiments and Results

C.1 Paired Synthetic Evaluation

We further evaluate the two-stage architecture when paired synthetic data are available. For this experiment, Stage I is trained on TMT-Static short-exposure/temporal-average pairs, and Stage II on Stage-I-output/clean pairs from the training split. AT-Net and TurbNet are retrained on the same dataset; all three methods receive a single turbulent frame at evaluation.

Table 5: Single-frame restoration on 1,518 paired TMT-Static evaluation images. Mean RGB PSNR (dB) and SSIM are reported.

Method	PSNR↑	SSIM↑
AT-Net	23.973	0.701
TurbNet	25.659	0.753
Ours	**26.096**	**0.761**

Our full output achieves the highest PSNR and SSIM among these three methods (Table 5). In Figure 8, Stage I produces a geometrically more regular representation that retains residual blur; Stage II restores the train's overhead wires, aircraft lettering and engine rim, and building facade lines. Together, the full-reference gains and the transition from intermediate to final output support the effectiveness of decoupling geometric stabilization from detail restoration.

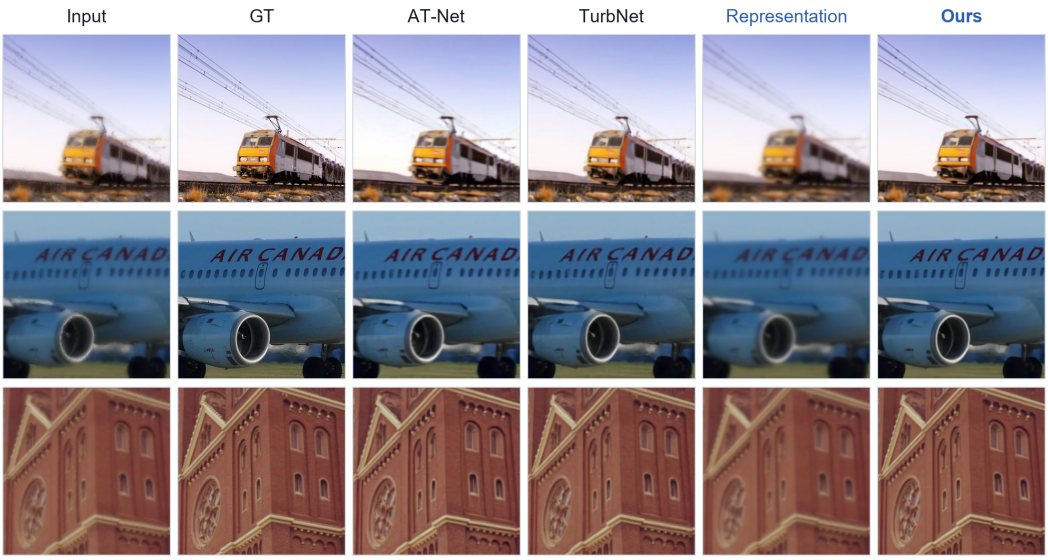

Figure 8: Qualitative TMT-Static comparisons on three outdoor scenes: a train, an aircraft, and a building (top to bottom). Each row uses aligned 320×320 crops. Representation and Ours denote the outputs of Stages I and II, respectively.

C.2 Geometric Stabilization in the Learned Representation

We compare turbulent inputs, single-frame Stage-I outputs (Representation$_1$), and their 12-frame refined outputs (Representation$_{12}$) on 65 held-out RLR-AT-Static scenes, using 100 target frames per scene. Both representations are evaluated before Stage II. Following Appendix B.3, we measure nTII and nTSS at strides $d = 1$ and $d = 12$ and average the scores over scenes.

Table 6: Temporal stability before detail restoration. Input normalization gives nTII $= 1$ and nTSS $= 100/e$ for the input sequence.

	$d = 1$		$d = 12$	
Sequence	nTII↓	nTSS↑	nTII↓	nTSS↑
Input	1.000	36.788	1.000	36.788
Representation$_1$	0.784	46.233	0.755	47.322
Representation$_{12}$	**0.349**	**71.347**	**0.539**	**58.790**

Table 6 shows that Stage I reduces nTII by 21.6% and 24.5% at the two strides, respectively. Twelve-frame refinement further reduces nTII by 55.5% and 28.6% relative to Representation$_1$, with improvements in all 65 scenes at both strides. Figure 9 distinguishes the roles of these two steps:

Stage I produces straighter building lines and more regular sign boundaries, while multi-frame refinement further suppresses fluctuations in their space–time traces. The gains across temporal scales show that the learned intermediate supports both single-frame geometric stabilization and further temporal refinement before detail recovery.

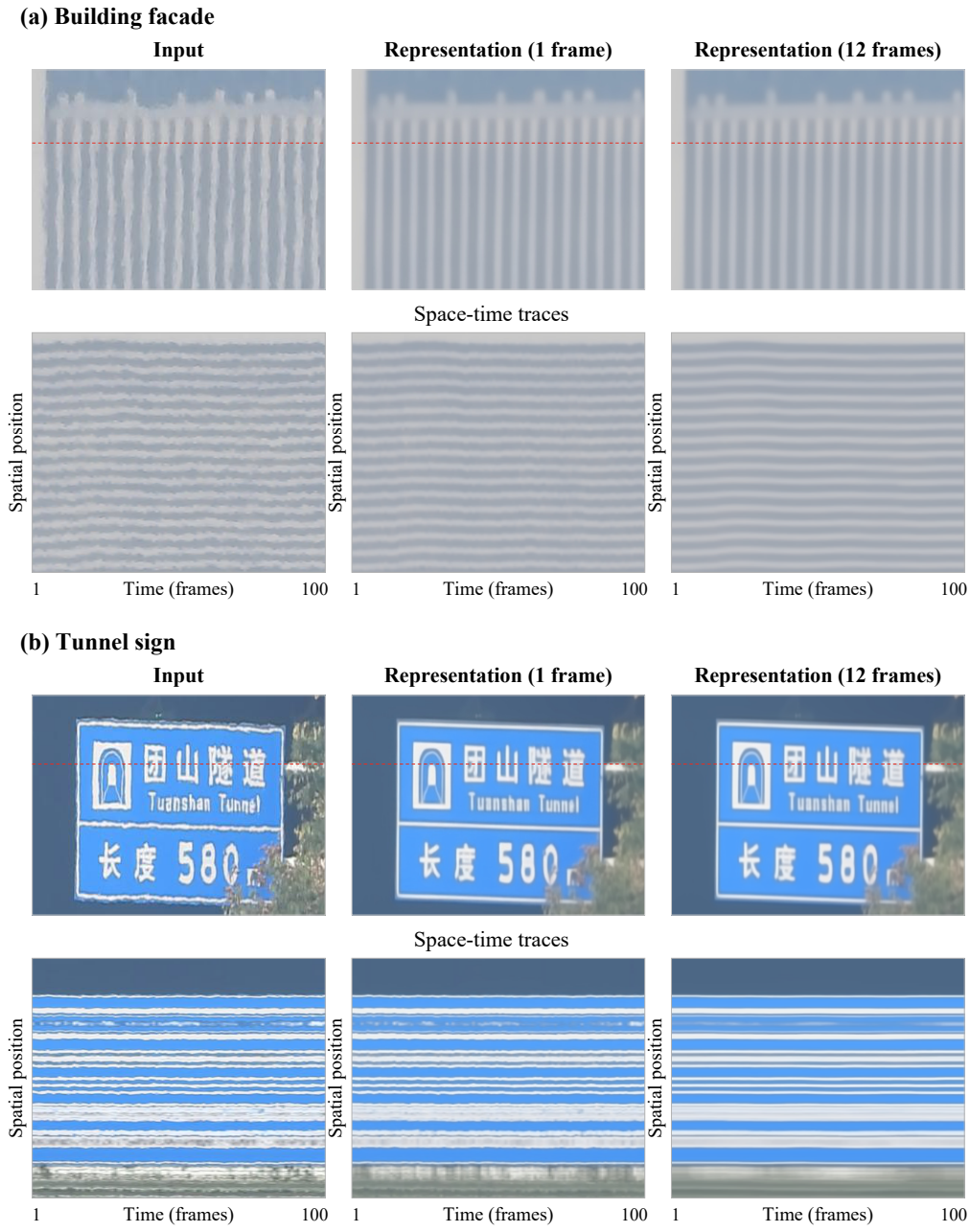

Figure 9: Input, single-frame representation, and 12-frame refined representation for two held-out scenes. Each row uses the same target frame and crop. Red dashed lines mark identical spatial sampling locations. The traces below show these samples across 100 frames, with time increasing from left to right. Stage I stabilizes spatial structures, and multi-frame refinement further reduces temporal fluctuations.

C.3 Stage-II Capacity and Representation Reuse

To examine whether the learned representation supports detail restorers of different capacities, we keep Stage I unchanged and replace only the Stage-II FocalNet. The small (S), medium (M), and large (L) variants have 8, 20, and 32 residual blocks, respectively; M is the model used in the main paper. Table 7 reports their results on RLR-AT-Static with one or twelve input frames.

Table 7: Perceptual quality of Stage-II capacity variants on RLR-AT-Static. S, M, and L denote 8, 20, and 32 residual blocks in Stage II.

Inputs	Model	LIQE↑	MUSIQ↑	NIMA↑	MANIQA↑	Q-Align↑	ARNIQA↑	HyperIQA↑
One frame	Ours-S	2.357	66.336	4.806	0.413	2.660	0.594	0.634
	Ours-M	2.539	67.356	4.859	0.427	2.672	0.604	0.656
	Ours-L	**2.647**	**67.869**	**4.870**	**0.436**	**2.679**	**0.611**	**0.664**
Twelve frames	Ours-S	2.344	66.395	4.835	0.419	2.669	0.594	0.634
	Ours-M	2.537	67.571	4.889	0.435	2.684	0.604	0.657
	Ours-L	**2.653**	**67.874**	**4.892**	**0.445**	**2.686**	**0.608**	**0.664**

All seven reported perceptual metrics improve from S to M to L in both the single-frame and twelve-frame settings. The consistent gains indicate that increasing Stage-II capacity improves the perceptual quality of the final restoration. Since the same Stage-I representation is used by all three restorers, these results also support its role as a reusable interface for subsequent detail recovery. Figure 10 shows progressively clearer window frames, roof lines, and sign lettering while preserving the stabilized structures of the shared representation.

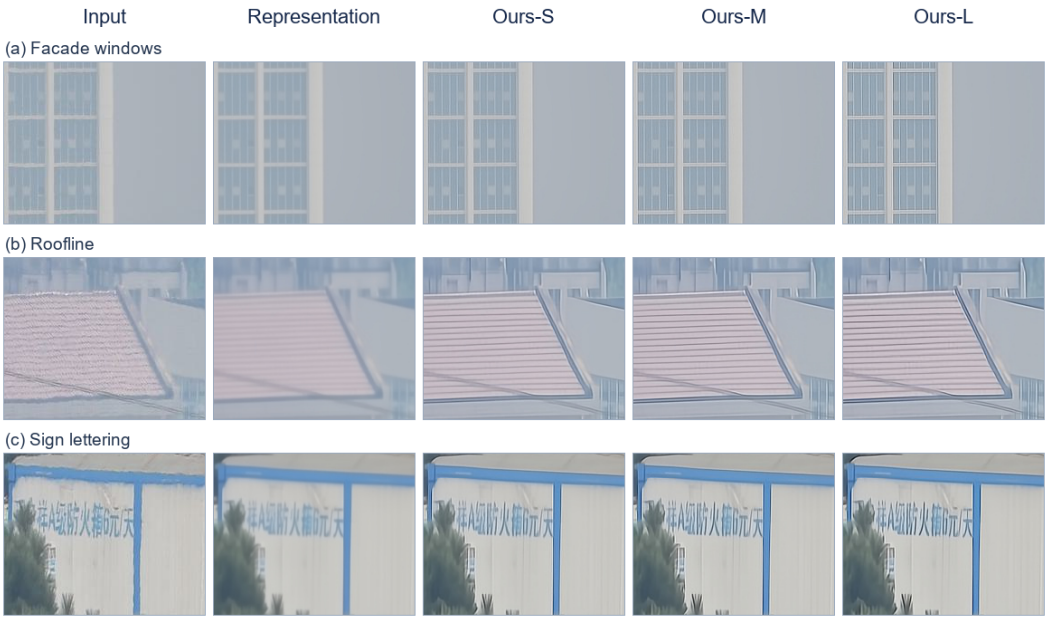

Figure 10: Stage-II capacity on three held-out RLR-AT-Static scenes. Each row shows the same single input frame and matched crop, its Stage-I representation, and restorations from Stage-II models with 8 (S), 20 (M), or 32 (L) residual blocks. Rows show building windows, a roofline, and sign lettering from top to bottom.

C.4 Effect of Stage II on Real-World Restoration

We compare the Stage-I representation $\widehat{Z}_t$ with its Stage-II restoration $\widehat{X}_t = G_\phi(\widehat{Z}_t)$ under the single-frame evaluation protocol of Table 1. Table 8 reports seven perceptual-quality metrics on four real-world benchmarks.

Table 8: Stage-I representations and Stage-II restorations in the single-frame setting. Restored scores correspond to Ours in Table 1; higher is better for every metric.

		Perceptual quality						
Dataset	Output	LIQE↑	MUSIQ↑	NIMA↑	MANIQA↑	Q-Align↑	ARNIQA↑	HyperIQA↑
RLR-AT-Static	Representation	1.310	46.497	4.194	0.297	2.255	0.463	0.387
	Restored	**2.539**	**67.356**	**4.859**	**0.427**	**2.672**	**0.604**	**0.656**
RLR-AT-Dynamic	Representation	1.267	47.206	4.069	0.259	2.493	0.465	0.380
	Restored	**2.127**	**66.743**	**4.722**	**0.369**	**2.771**	**0.605**	**0.631**
UoB-Turb	Representation	1.156	28.763	3.894	0.221	2.000	0.401	0.287
	Restored	**2.356**	**55.577**	**4.633**	**0.325**	**2.910**	**0.585**	**0.466**
CLEAR	Representation	1.368	33.605	4.134	0.189	1.959	0.428	0.286
	Restored	**1.671**	**48.515**	**4.666**	**0.242**	**2.403**	**0.515**	**0.409**

Stage II improves all seven reported metrics on every benchmark. The perceptual gain is especially pronounced on UoB-Turb, where MUSIQ rises from 28.763 to 55.577; substantial gains also appear on the other three benchmarks. This consistent improvement shows that the learned representation provides an effective input for Stage-II detail restoration and the effective of Stage II across diverse real-world scenes.

D Model Size and Inference Efficiency

Table 9 compares model size and inference time at 512×512 resolution on an NVIDIA RTX 4090. Each time is the average number of milliseconds required to produce one output frame in the indicated input-frame setting; dashes denote settings without reported results.

Table 9: Model parameters and average inference time per output frame on an NVIDIA RTX 4090 with 512×512 images.

Method	Parameters (M)	One frame (ms)	Twelve frames (ms)
AT-Net	7.01	3,004	—
TurbNet	26.60	135	—
TMT	22.95	7869	1,315
DATUM	5.78	151	39.7
MambaTM	7.24	37.4	14.7
Turb-Seg-Res	31.38	—	776
CDSP	50.88	—	17,800
Ours	**22.62**	**67.2**	**300.7**