

ConVRT: Consistent Video Restoration Through Turbulence with Test-time Optimization of Neural Video Representations

Haoming Cai^{*1}, Jingxi Chen^{*1}, Brandon Y. Feng², Weiyun Jiang³,
Mingyang Xie¹, Kevin Zhang¹, Ashok Veeraghavan³, Christopher Metzler^{†1}
¹University of Maryland, College Park ²Massachusetts Institute of Technology ³Rice University

<https://convrt-2024.github.io/>

Abstract

Atmospheric turbulence presents a significant challenge in long-range imaging. Current restoration algorithms often struggle with temporal inconsistency, as well as limited generalization ability across varying turbulence levels and scene content different than the training data. To tackle these issues, we introduce a self-supervised method, **Consistent Video Restoration through Turbulence (ConVRT)** a test-time optimization method featuring a neural video representation designed to enhance temporal consistency in restoration. A key innovation of ConVRT is the integration of a pretrained vision-language model (CLIP) for semantic-oriented supervision, which steers the restoration towards sharp, photorealistic images in the CLIP latent space. We further develop a principled selection strategy of text prompts, based on their statistical correlation with a perceptual metric. ConVRT’s test-time optimization allows it to adapt to a wide range of real-world turbulence conditions, effectively leveraging the insights gained from pretrained models on simulated data. ConVRT offers a comprehensive and effective solution for mitigating real-world turbulence in dynamic videos.

1. Introduction

Atmospheric turbulence often occurs in aerial photography and astronomical observations and significantly degrades imaging quality, leading to blurred, warped, or otherwise distorted imagery. Effective turbulence mitigation is not only crucial for enhancing the clarity and reliability of visual information, but also plays a pivotal role in various applications ranging from remote sensing and security surveillance to scientific research and environmental monitoring.

While turbulence mitigation on static scenes has seen re-

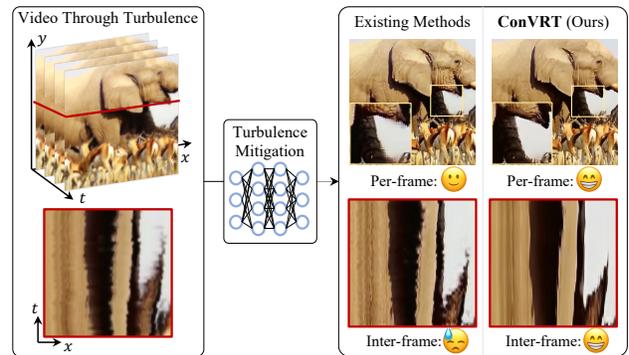


Figure 1. Existing turbulence mitigation methods produce good per-frame results but often fail to maintain consistency across frames, which is vital for downstream tasks. Our work introduces ConVRT, which effectively removes turbulence while preserving temporal consistency in the restored video.

markable advancements, largely fueled by the availability of extensive datasets, improved turbulence simulators, and the development of more capable machine learning algorithms, the domain of dynamic video restoration under turbulence conditions lags behind. This lag can be attributed to several unique challenges inherent to video processing.

A primary obstacle in video-based restoration is maintaining high temporal consistency. Unlike static scenes, video observations of dynamic scenes comprise a sequence of frames where each frame is not only expected to be clear, but also consistent with other frames in terms of quality and continuity. This requirement for temporal coherence adds a layer of complexity to the restoration process. Furthermore, turbulence distortions vary not only spatially across a single frame but also temporally across the sequence of frames, making the restoration process significantly more intricate.

The prevailing approaches to video turbulence mitigation typically involve either the iterative application of single-image restoration methods to each frame or the use of

^{*}Equal Contribution

[†]Corresponding author.

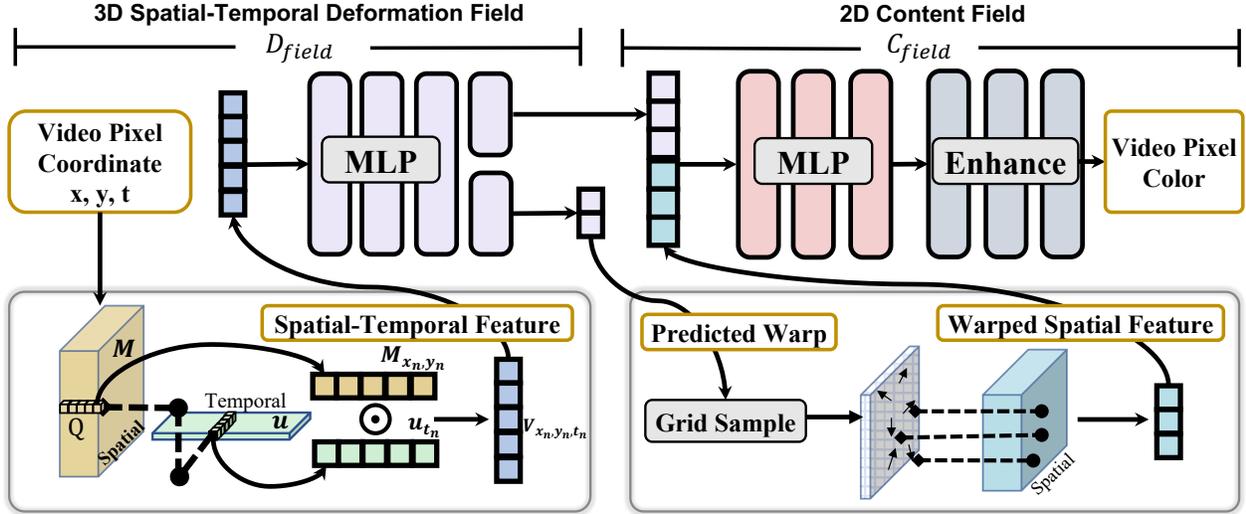


Figure 2. Illustration of the proposed method : ConVRT decomposes a video into two fields: 3D Spatial-Temporal Deformation Field D_{field} and 2D Content Field C_{field} . Low-rank decomposition in D_{field} reduces parameters and preserves spatial-temporal details. The predicted warp from D_{field} shapes the spatial features in C_{field} , affecting the RGB frame output.

video-based restoration networks trained on simulated data. However, these strategies are not without significant drawbacks. When single-image methods are applied independently to each frame, they often fail to maintain inter-frame continuity, resulting in jittery or inconsistent visual outputs. Furthermore, the reliance on simulators for the development and testing of these methods introduces a notable accuracy gap. While simulators are useful for dataset generation, they may not accurately replicate the complex and dynamic nature of real-world atmospheric turbulence. As such, methods with good performance on simulated data do not always perform effectively in real-world scenarios.

These challenges highlight an urgent need for more robust, adaptable, and specialized approaches in video-based turbulence mitigation. A pivotal question arises: Is it possible to develop a method that not only leverages the valuable insights gained from pre-trained methods using simulated data, but also adapts to the constantly changing conditions of real-world turbulence during test time?

This paper presents, **Consistent Video Restoration through Turbulence (ConVRT)**, a strategy which combines the adaptability required for real-world application with the foundational strengths of simulation-based training. Moving away from the conventional reliance on complex deep learning models or intricate turbulence simulators for inverse rendering, ConVRT innovatively combines recent advances in machine learning with neural signal representations to address video-based turbulence mitigation.

At the heart of ConVRT lies a neural video representation, which is composed of a 2D content field and a 3D spatial-temporal deformation field. This dual-field repre-

sentation allows for a more nuanced and accurate restoration of video content distorted by atmospheric turbulence. We employ a test-time optimization framework to train this video representation, effectively modeling both the dynamic scene and the turbulence distortions. After optimization, the dynamic scene can decouple from the turbulence distortion, resulting in a sharply restored video. To further refine the reconstruction, ConVRT incorporates semantic-oriented supervision using priors from the Contrastive Language-Image Pre-Training (CLIP) [41] model. We propose a novel strategy that selects prompts based on the statistical correlation between CLIP and the LPIPS [55] metric, guiding the restoration process such that the output is more closely aligned with the ideal prompt within the projected embedding space of CLIP. Additionally, a key focus of ConVRT is on improving temporal consistency, achieved through the careful design of the neural representation of the deformation field. Crucially, our test-time optimization framework circumvents the typical generalization issues of deep-learning-based methods while retaining the flexibility to incorporate pretrained knowledge from existing models.

ConVRT addresses the key challenges of restoring video distorted by turbulence. The major contributions include:

- An innovative test-time optimization framework for turbulence mitigation, improving per-frame restoration fidelity and inter-frame temporal coherence.
- An efficient neural representation of videos tailored specifically for turbulence mitigation, including a pair of content field and deformation field.
- A novel, semantic-oriented enhancement module using the pretrained CLIP model, including developing

a principled strategy for prompt selection based on the statistical correlation between CLIP and LPIPS.

- A comprehensive evaluation against existing methods, showing it outperforming existing methods on visual quality and coherence of the restored video content.

2. Related Work

Implicit neural representations. Our work leverages a coordinate-based implicit neural representation (INRs), which has been commonly adopted to model 2D images or 3D videos as multi-layer perceptions (MLPs). INRs take 2D pixel coordinates (x, y) , or 3D pixel coordinates with temporal encoding, (x, y, t) and output the corresponding pixel values. These INRs demonstrate exceptional performance when fitting images [14–16, 35–37, 48], videos [3, 8, 9, 46], and 3D shapes [18, 40, 46]. Not only they are able to represent these 2D or 3D signals, but they also show strong priors for solving inverse problems, such as image super resolution [10], video inpainting [9], phase retrieval [49, 57], and reducing optical aberration [5, 20, 30].

Neural video representation. Our work aligns closely with the evolving field of neural video representation [19, 29, 39, 50]. While there are existing approaches [25, 28, 39, 54] that seek to represent a video into decomposed layers, these primarily focus on clean videos and are not applicable to videos with severe degradation turbulence. Our work extends the application of neural video representation to scenarios heavily impacted by atmospheric turbulence. This extension is not trivial, as it involves addressing the unique challenges posed by the dynamic and unpredictable nature of turbulence, which are not considered in conventional video representations.

Atmospheric turbulence mitigation. Attempts to mitigate atmospheric turbulence [21, 38] have applied optical flow [6, 32], B-spline grid [45], and diffeomorphism [22] to unwarped each distorted image and then fuse and combine these registered distorted images into a clean and sharp image. The fusion is usually modeled as patch-wise stitching [32] or blind deconvolution [2]. Recent development of high-performance GPUs and fast turbulence simulators leads to new progress in turbulence mitigation [11, 12, 17, 23, 33, 34, 56]. However, previous efforts tend to overlook the importance of temporal consistency on the reconstructed video. Our method, ConVRT, is specifically designed to restore temporal consistency with on test-time optimization of a neural video representation.

3. Method

This section describes the proposed design of a neural video representation tailored to turbulence mitigation.

3.1. General Pipeline

The framework of our method, ConVRT, is presented in Figure 2. During training, TurbNet’s output [34] serves as the sole supervision signal for our ConVRT. ConVRT is designed to adapt to future advancements in turbulence mitigation algorithms. As for the pipeline design, ConVRT employs a dual-field approach: a 3D Spatial-Temporal Deformation Field D_{field} for adapting to temporal variations, and a 2D Content Field C_{field} for canonical 2D content. D_{field} generates spatial-temporal features (x, y, t) , which are transformed into hidden features and a predicted warp. This warp guides C_{field} to produce warped spatial features, which are then concatenated with hidden features for the MLP in C_{field} , creating RGB frames. An enhancement module is applied to finalize the restoration, enhancing visual quality.

3.2. 3D Spatial-Temporal Deformation Field

We construct a spatial-temporal feature space $V \in \mathbb{R}^{Q \times M \times N \times T}$, storing Q -channel feature vectors $V_{x_n, y_n, t_n} = \{V_q\}_{q=1}^Q$ at each location. The spatial dimensions x_n and y_n correspond to the video frame’s height and width. A compact MLP transforms these feature vectors into predicted warp and hidden spatial-temporal features at (x_n, y_n, t_n) .

Adopting a low-rank-decomposed representation similar to TensorRF [7], V is modeled using a 1D vector u for temporal variation and a full-rank matrix M for spatial variations in x and y . The Q -channel feature vectors in u and M are dynamically updated during optimization.

To extract V_{x_n, y_n, t_n} at t_n , we project (x_n, y_n) onto M and t_n onto u , resulting in M_{x_n, y_n} and u_{t_n} . The final feature vector is the Hadamard product:

$$V_{x_n, y_n, t_n} = M_{x_n, y_n} \odot u_{t_n}, \quad (1)$$

where \odot is the Hadamard product. This efficiently approximates a 3D feature as a tensor product of a 2D matrix and 1D vector, reducing parameters while capturing spatial-temporal details. In the subsequent MLP, Layer normalization is enabled to promote stable training.

3.3. 2D Content Field

Within the 2D Content Field C_{field} , we obtain spatial feature vectors M_{x_n, y_n} with the spatial coordinates (x_n, y_n) . These vectors used to warped the grid points used to sample from D_{field} , resulting in warped spatial features. Subsequently, each warped vector is concatenated with hidden spatial-temporal features derived from D_{field} . These concatenated features are then fed into the MLP and the subsequent enhancement module to produce the final RGB frame.

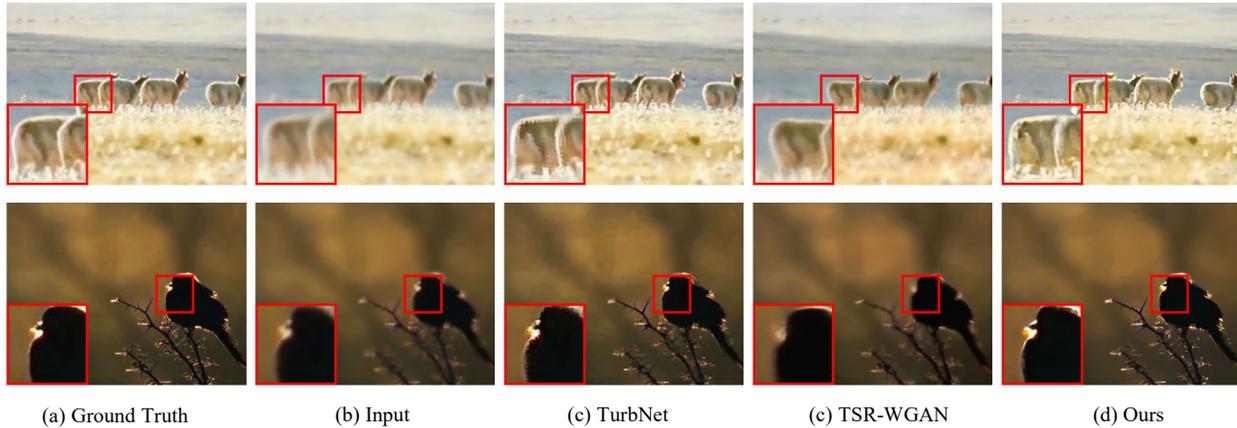


Figure 3. Comparing single-frame restoration in synthetic turbulence videos with TurbNet[34] and TSR-WGAN[24]: Our results are the sharpest and most accurate.

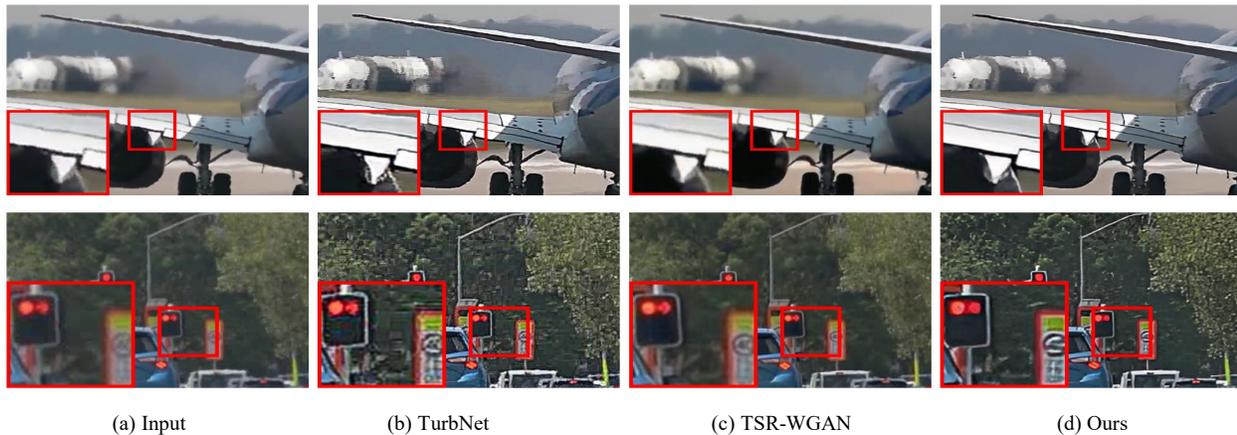


Figure 4. Comparison of single-frame restoration quality on real-world turbulence videos: TurbNet[34] results are with distorted boundaries and artifacts; TSR-WGAN[24] results do not really remove the turbulence blur as shown on zoom-in views.

3.4. Semantic Enhancement

In our approach, the output from TurbNet [34] serves as a solitary supervisory signal, which may inadvertently introduce artifacts into our results if relied upon exclusively. To enhance the quality of the visual output and imbue it with semantic depth, we utilize text-driven models, specifically CLIP, known for their proficiency in aligning visual content with textual descriptions. The Semantic Enhancement step is crucial for transcending mere fidelity, aiming instead for semantically enriched and contextually nuanced outputs.

Nevertheless, the arbitrary selection of text prompts could yield suboptimal results. To address this, we present a principled prompt-selection methodology, which employs statistical correlation between the Learned Perceptual Image Patch Similarity (LPIPS) and CLIP scores to select the

most appropriate prompts. As demonstrated in Figure 8, we conduct a comparative analysis of various degraded images using the same sequence of frames. We observe the relationship between LPIPS scores and CLIP losses, alongside a correlation study illustrated by the Kendall Rank Correlation Coefficient (KRCC) and Spearman’s Rank Correlation Coefficient (SRCC) in the right figure of Figure 8.

The selection process yields “a degraded image with noise and turbulence distortion” as the negative text prompt and “a clean and sharp natural image” as the positive text prompt, which have shown the highest relevance in our tests. These prompts, meticulously chosen, are then integrated into our training as a new loss term:

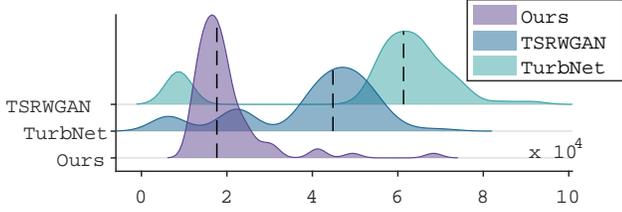


Figure 5. Histogram of TV for optical flow between frames in a video sequence with real-world turbulence. ConVRT obtains more defined optical flow, leading to lower total variation.

$$\mathcal{L}_{text} = - \left(\frac{\langle Enc_i(I), Enc_t(T_{pos}) \rangle}{\|Enc_i(I)\| \|Enc_t(T_{pos})\|} - \frac{\langle Enc_i(I), Enc_t(T_{neg}) \rangle}{\|Enc_i(I)\| \|Enc_t(T_{neg})\|} \right). \quad (2)$$

$Enc_i(I)$ is the feature vector extracted from the predicted image I using CLIP’s image encoder. $Enc_t(T_{pos})$ is the feature vector obtained from the positive prompt T_{pos} using CLIP’s text encoder, and $Enc_t(T_{neg})$ is derived similarly from the negative prompt T_{neg} . With this term, we guide the learning process toward the direction of positive prompt semantics and away from the negative prompt semantics.

3.5. Training Objectives

Temporal Consistency Regularization. To ensure temporal stability across video frames, we employ a disparity estimation network (MiDas [42]) and calculate pixel-wise disparities as weight for the predicted warp (one of D_{field} ’s output) to maintain spatial consistency over time. The loss is defined as:

$$\mathcal{L}_{temp} = (1 - \text{Disparity}(I)) \cdot \|\text{Predicted Warp}\|_1 \quad (3)$$

where $\text{Disparity}(I)$ measures the pixel-level disparity, and $\|\text{Predicted Warp}\|_1$ enforces sparsity in the grid changes. The design of \mathcal{L}_{temp} minimizes the L1 norm of the predicted warp, conditioned by $1 - \text{Disparity}(I)$, to prioritize consistency in far regions based on the depth information. This focused approach on temporal consistency significantly reduces the propagation of turbulence-induced distortions, ensuring a smooth transition between frames.

Similarity Loss. The Similarity Loss Term is given by:

$$\mathcal{L}_{sim} = \lambda_{mse} \mathcal{L}_{mse} + \lambda_{ssim} \mathcal{L}_{ssim} + \lambda_{lpipe} \mathcal{L}_{lpipe} \quad (4)$$

where λ_{mse} , λ_{ssim} , and λ_{lpipe} are weights for each term. This loss term assesses the fidelity of the predicted output compared to TurbNet output, incorporating Mean Squared Error (MSE), Structural Similarity Index Measure

Table 1. Evaluation metrics tested on video with synthetic turbulence. \uparrow : higher is better, \downarrow : lower is better.

Method	PSNR _{Img} \uparrow	SSIM \uparrow	LPIPS \downarrow	E _{warp} \downarrow	PSNR _{x-t} \uparrow
TSRWGAN [24]	23.58	0.739	0.230	0.0026	23.77
TurbNet [34]	23.44	0.732	0.228	0.0057	23.54
TurbNet+Real-ESRGAN [51]	22.48	0.713	0.213	0.0074	22.67
ConVRT (Ours)	24.90	0.787	0.189	0.0014	25.73

(SSIM) [52], and Learned Perceptual Image Patch Similarity (LPIPS). This multifaceted approach ensures a comprehensive evaluation of reconstruction quality.

Overall Loss The overall loss combines the similarity loss with temporal consistency and semantic enhancement:

$$\mathcal{L}_{total} = \mathcal{L}_{sim} + \lambda_{temp} \mathcal{L}_{temp} + \lambda_{text} \mathcal{L}_{text}. \quad (5)$$

4. Experiments

In this section, we provide the experimental details and results which validate the performance improvement enabled by our method. Additional experimental results are provided in the supplementary file.

4.1. Datasets and Training Details

We adopt subsets of the standard datasets, BVI-CLEAR dataset [1] and the TSR-WGAN dataset [24], for fair comparison since baselines are trained among them. These datasets feature real-world turbulence and clean videos. To generate synthetic turbulence videos, we use the P2S atmospheric turbulence simulator [33] on clean real-world videos, producing our synthetic distorted video sequences. We employ turbulence parameters $D/r_0 = 2$ and $corr = 1$. Overall, this subset includes 8 sequences of real-world and synthetic videos with turbulence. We train the ConVRT model for 6000 iterations with a learning rate of 2×10^{-3} . The Adam optimizer [26] is employed. The enhancement module is Real-ESRGAN [51], which remains fixed during the training of ConVRT.

4.2. Evaluation Strategy

Two state-of-the-art methods of Turbulence Mitigation are used for fair comparison : TurbNet and TSRW-GAN. TMT [56], the only video-based turbulence mitigation method, is skipped because the well-trained weight is inaccessible. To evaluate the consistent removal of turbulence in video, we use four metrics for qualitative evaluation and two interframe-related methods for qualitative evaluation.

Per-frame Quality and Temporal Consistency. We use PSNR and SSIM to measure the per-frame quality of reconstruction. LPIPS is used to measure the perceptual quality.

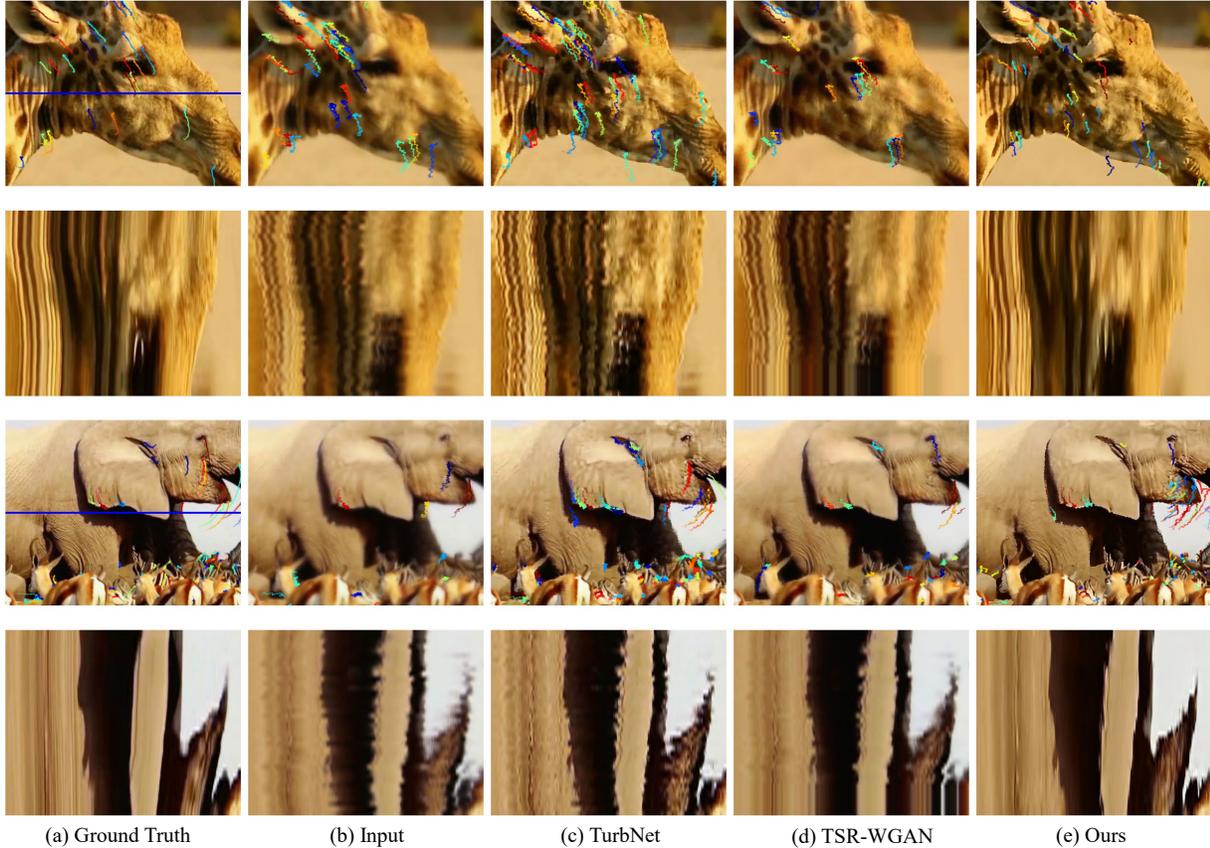


Figure 6. A comparison of synthetic turbulence video consistency: Two scenes are shown, each with two rows. The first row shows the KLT tracking trajectories, and the second shows an $x-t$ slice with a blue line indicating the slice’s position. The “zig-zag” patterns in TurbNet[34] and TSR-WGAN[24] trajectories show temporal inconsistency, and TSR-WGAN produces fewer trajectories, like the blurry input. In contrast, our method produces a smooth and reasonable number of trajectories. The $x-t$ slices from TurbNet and TSR-WGAN are non-smooth, whereas ours are smooth.

Following [27], we use the average warp error to measure the temporal consistency for the restored video. For the warp error between two consecutive frames, it can be defined as following:

$$E_{\text{warp}}(V_t, V_{t+1}) = \frac{1}{\sum_{i=1}^N M_t^{(i)}} \sum_{i=1}^N M_t^{(i)} \left\| V_t^{(i)} - \hat{V}_{t+1}^{(i)} \right\|_2^2, \quad (6)$$

where $\hat{V}_{t+1}^{(i)}$ is the warped frame by optical flow at time $t+1$ and $M_t^{(i)} \in \{0, 1\}$ is the occlusion mask estimated by the methods proposed in [43]. The average warp error is:

$$E_{\text{warp}}(V) = \frac{1}{T-1} \sum_{t=1}^{T-1} E_{\text{warp}}(V_t, V_{t+1}) \quad (7)$$

which is the average of consecutive warp errors across the entire video sequence.

KLT Trajectories. We use the KLT tracker [31] to track the feature points, and then plot their trajectories as shown in Figure 7. KLT tracking is directly based on the image gradient information such that the common issues in turbulence restoration, i.e., blurriness, artifacts, temporal inconsistency, will be reflected in the tracked trajectories.

$x-t$ Slice. We plot $x-t$ slices to show the motion of a row of pixels as shown in Figure 7. If the video restoration is temporally inconsistent, the $x-t$ slice plot will show the non-smooth shape for curves.

4.3. Comparison on Synthetic Turbulence Videos

Our method achieves high temporal consistency while maintaining fidelity. As observed in Figure 3, ConVRT surpasses other approaches in both turbulence removal and texture detail restoration. Figure 6 demonstrates that the video dynamics generated by our method closely resemble those in the corresponding ground-truth videos (first column).

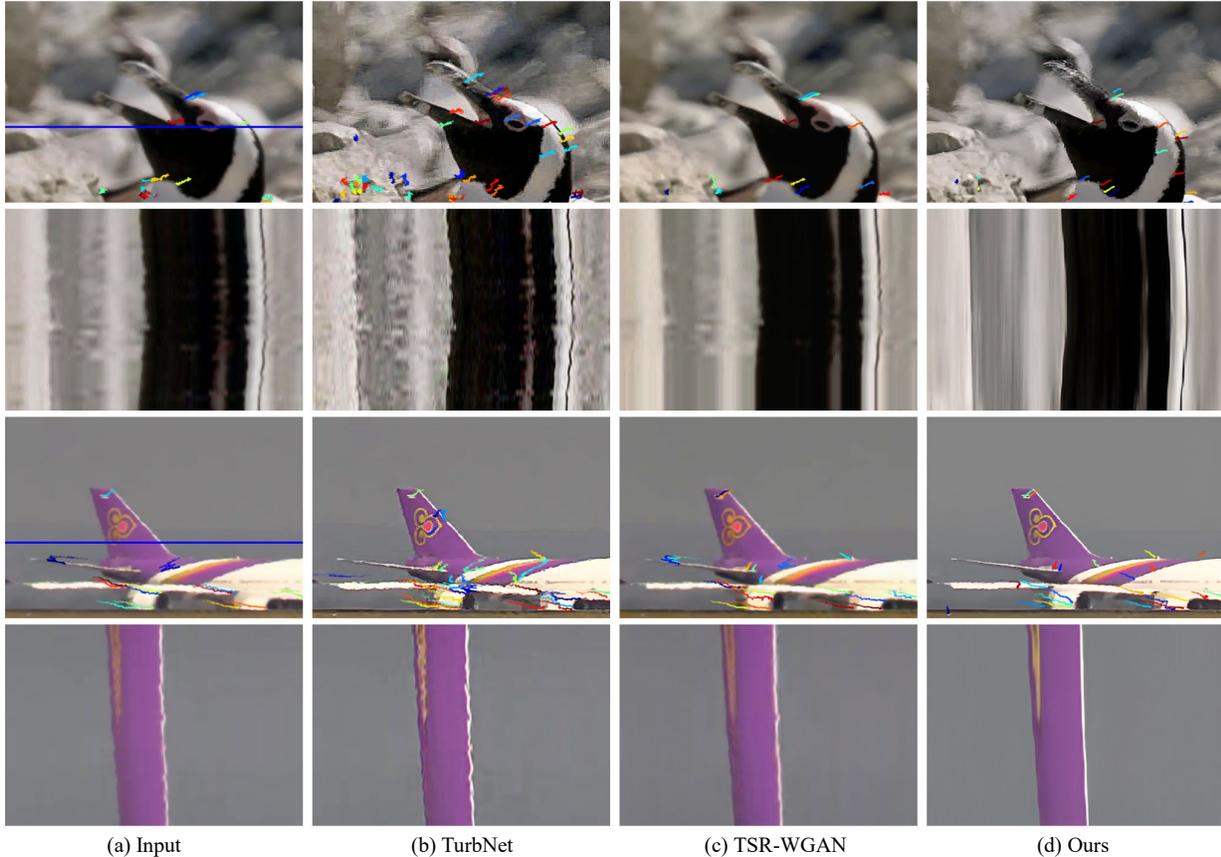


Figure 7. A comparison of temporal consistency in real-world turbulence videos with TurbNet[34] and TSR-WGAN[24]: Demonstrated in two scenes, each scene is represented by KLT tracking trajectories and an $x-t$ slice, indicated by a blue line. Our method shows the best temporal consistency in these real-world scenarios. Notably, in the first penguin scene with a stationary background, only our method accurately reflects stationary tracking trajectories (single-dot trajectories on the rock).

Table 2. Evaluation Text Prompts for Turbulence Mitigation Learning Guidance

Text Index	Positive Prompts	Negative Prompt
1	“a sharp image”	“a blur image”
2	“a sharp image”	“a image with blur and turbulence distortion”
3	“a clean and sharp natural image”	“a degraded image with noise and turbulence distortion”
4	“a clean and sharp natural image”	“a degraded image with mosaic and turbulence distortion”
5	“a clean and sharp natural image”	“a low-resolution image with mosaic and turbulence distortion”
6	“a clean and sharp natural image with table and alarm clock and books”	“a low-resolution image with mosaic and turbulence distortion”

ConVRT uniquely captures the scene dynamics, effectively smoothing out atmospheric turbulence, a distinction not seen in other methods for turbulence mitigation. The KLT trajectories further substantiate this temporal consistency. In contrast, TurbNet and TSR-WGAN produce “zig-zag” tracking trajectories, indicative of temporally inconsistent video restoration. Notably, the KLT tracker generates only a few trajectories for TSR-WGAN’s restoration and frame with turbulence (second column). This scarcity of trajectories might be attributed to this GAN-based method’s gener-

ation of inconsistent content and failure to adequately de-blur the scene across the video. Table 1 presents these fidelity and temporal consistency metrics. With the best per-frame quality, ConVRT outperforms all baseline methods in the temporal consistency metrics E_{warp} and $PSNR(x-t)$.

4.4. Comparison on Real-world Turbulence Videos

Real-world atmospheric turbulence presents a significant and challenging domain gap compared to simulations. However, our method achieves outstanding fidelity and tem-

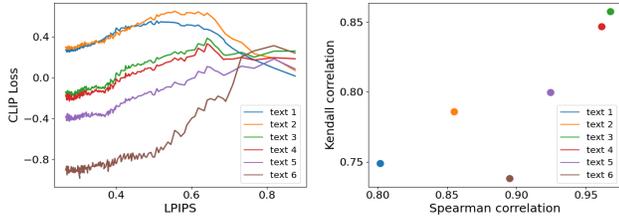


Figure 8. CLIP text prompt selection. Left: LPIPS against CLIP loss during optimization. Right: correlation scores between CLIP loss and LPIPS sequences for each text prompt. Text 3 is the final choice due to its highest correlation scores.

poral consistency in various video sequences distorted by real atmospheric turbulence, as demonstrated in the following figures. Figure 4 showcases the comparison of fidelity, where ConVRT restores outlines with sharper edge.

As in Figure 7, while the original turbulence and baseline methods result in “zig-zag” tracking trajectories by KLT, ConVRT achieves smoother trajectories. This indicates that, unlike other methods, ConVRT consistently removes turbulence throughout the video. In addition, the x-t slice in Figure 7 reveals that ConVRT smoothens row pixel motion more effectively, further enhancing single-frame turbulence removal (also shown in Figure 4).

Figure 5 presents a histogram of Total Variation (TV) for optical flow between frames in a video with real-world turbulence. ConVRT attains lower TV , indicating more accurate optical flow, particularly on static backgrounds. Incorrect restoration of these backgrounds often leads to high TV due to chaotic optical flow.

4.5. CLIP Text Prompt Selection

Since we use the CLIP model to semantically guide the video restoration process, we investigate the effectiveness of different text prompts for guiding the turbulence removal and get better quality restoration. As shown in Figure 8, we use the ground-truth information to learn the turbulence removal process on a single image, we also simultaneously record the LPIPS learning curve and clip loss for different prompts from the Table 2, then calculating the correlation between LPIPS sequence and clip loss sequences. The best text prompt is the text prompt that has loss with highest correlation score with LPIPS sequence. In this way we can quantitatively select the most correlated clip text prompt to guide the learning process of turbulence removal purpose.

5. Discussion

Difference to existing neural video representations?

ConVRT does not merely represents the observed video, but actually restores the content before distorted by turbulence. ConVRT accounts for the nuanced relationship between the

temporal and spatial distortions of turbulence and adapting the neural representation to effectively model and counteract these distortions. Conventional neural video representations do not account for such complex, dynamic distortions.

Difference between CLIP guidance to perceptual loss?

Perceptual loss is less effective for turbulence mitigation as they typically require a reference clean image, which is often unavailable in turbulence-distorted scenarios. Our CLIP-guided module, on the other hand, bypasses this limitation by leveraging the pretrained CLIP model’s ability to understand and interpret complex image content without needing a direct clean image reference, and yet allows for the integration of human priors through text prompts.

Significance of improving the temporal consistency?

Our method achieves smoother KLT tracking trajectories and maintains stationary backgrounds, which is crucial for downstream tasks like SLAM [4, 13, 47], NeRF [36], pose estimation [44], and object segmentation and tracking [53]. Improving temporal consistency in turbulence-mitigated videos will bring significant benefits to these tasks.

Why not use a simulator?

Existing turbulence simulators may not accurately model distortions on backgrounds or distant objects. This is likely because 2D-based simulators (e.g., [33]) do not consider the depth effect on pixel displacement. Consequently, models based on these simulators struggle with far or background objects, as in Fig 7, where baseline results show vibrating tracked points on stationary rocks. Simulating realistic turbulence for 3D scenes and objects are beyond the capacity of existing simulators.

6. Conclusion

This paper presents ConVRT, a novel approach combining neural video representation with semantic supervision, enhanced by the pretrained CLIP model. ConVRT significantly improves temporal coherence and visual quality in video restoration, outperforming existing methods. ConVRT not only enhances video restoration quality under severe atmospheric turbulence, but also enables application scenarios like long-range object tracking and scene reconstruction. Overall, ConVRT represents a major step forward in long-range imaging, merging machine learning advancements to address key challenges and opening new avenues in computer vision and optical imaging.

Acknowledgements

This work was supported in part by the AFOSR Young Investigator Program Award no. FA9550-22-1-0208 and ONR award no. N000142312752.

References

- [1] Nantheera Anantrasirichai. Atmospheric turbulence removal with complex-valued convolutional neural network. *Pattern Recognition Letters*, 171:69–75, 2023. 5
- [2] Nantheera Anantrasirichai, Alin Achim, and David Bull. Atmospheric turbulence mitigation for sequences with moving objects using recursive image fusion. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 2895–2899. IEEE, 2018. 3
- [3] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O’Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16610–16620, 2023. 3
- [4] Josep Aulinas, Yvan Petillot, Joaquim Salvi, and Xavier Lladó. The slam problem: a survey. *Artificial Intelligence Research and Development*, pages 363–371, 2008. 8
- [5] Emrah Bostan, Reinhard Heckel, Michael Chen, Michael Kellman, and Laura Waller. Deep phase decoder: self-calibrating phase microscopy with an untrained deep neural network. *Optica*, 7(6):559–562, 2020. 3
- [6] Tufan Caliskan and Nafiz Arica. Atmospheric turbulence mitigation using optical flow. In *2014 22nd International Conference on Pattern Recognition*, pages 883–888. Ieee, 2014. 3
- [7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3
- [8] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. *Advances in Neural Information Processing Systems*, 34:21557–21568, 2021. 3
- [9] Hao Chen, Matthew Gwilliam, Ser-Nam Lim, and Abhinav Shrivastava. Hnerv: A hybrid neural representation for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10270–10279, 2023. 3
- [10] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 3
- [11] Nicholas Chimitt and Stanley H Chan. Anisoplanatic optical turbulence simulation for near-continuous profiles without wave propagation. *arXiv preprint arXiv:2305.09036*, 2023. 3
- [12] Nicholas Chimitt, Xingguang Zhang, Yiheng Chi, and Stanley H Chan. Scattering and gathering for spatially varying blurs. *arXiv preprint arXiv:2303.05687*, 2023. 3
- [13] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007. 8
- [14] Brandon Yushan Feng and Amitabh Varshney. Signet: Efficient neural representation for light fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14224–14233, 2021. 3
- [15] Brandon Yushan Feng and Amitabh Varshney. Neural subspaces for light fields. *IEEE Transactions on Visualization and Computer Graphics*, 2022.
- [16] Brandon Yushan Feng, Susmija Jabbireddy, and Amitabh Varshney. Viinter: View interpolation with implicit neural representations of images. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3
- [17] Brandon Y Feng, Mingyang Xie, and Christopher A Metzler. Turbugan: An adversarial learning approach to spatially-varying multiframe blind deconvolution with applications to imaging through turbulence. *IEEE Journal on Selected Areas in Information Theory*, 3(3):543–556, 2022. 3
- [18] Brandon Y Feng, Yinda Zhang, Danhang Tang, Ruofei Du, and Amitabh Varshney. Prif: Primary ray-based implicit function. In *European Conference on Computer Vision*, pages 138–155. Springer, 2022. 3
- [19] Brandon Y Feng, Hadi Alzayer, Michael Rubinstein, William T Freeman, and Jia-Bin Huang. 3d motion magnification: Visualizing subtle motions from time-varying radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9837–9846, 2023. 3
- [20] Brandon Y Feng, Haiyun Guo, Mingyang Xie, Vivek Boominathan, Manoj K Sharma, Ashok Veeraraghavan, and Christopher A Metzler. Neuws: Neural wavefront shaping for guidestar-free imaging through static and dynamic scattering media. *Science Advances*, 9(26):eadg4671, 2023. 3
- [21] David L Fried. Probability of getting a lucky short-exposure image through turbulence. *JOSA*, 68(12):1651–1658, 1978. 3
- [22] Jérôme Gilles, Tristan Dagober, and Carlo De Franchis. Atmospheric turbulence restoration by diffeomorphic image registration and blind deconvolution. In *Advanced Concepts for Intelligent Vision Systems: 10th International Conference, ACIVS 2008, Juan-les-Pins, France, October 20-24, 2008. Proceedings 10*, pages 400–409. Springer, 2008. 3
- [23] Weiyun Jiang, Vivek Boominathan, and Ashok Veeraraghavan. Nert: Implicit neural representations for unsupervised atmospheric turbulence mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4235–4242, 2023. 3
- [24] 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. 4, 5, 6, 7
- [25] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 3
- [26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [27] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 6

- [28] Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14317–14326, 2023. 3
- [29] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023. 3
- [30] Esther YH Lin, Zhecheng Wang, Rebecca Lin, Daniel Miao, Florian Kainz, Jiawen Chen, Xuaner Cecilia Zhang, David B Lindell, and Kiriakos N Kutulakos. Learning lens blur fields. *arXiv preprint arXiv:2310.11535*, 2023. 3
- [31] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 6
- [32] Zhiyuan Mao, Nicholas Chimitt, and Stanley H Chan. Image reconstruction of static and dynamic scenes through anisoplanatic turbulence. *IEEE Transactions on Computational Imaging*, 6:1415–1428, 2020. 3
- [33] Zhiyuan Mao, Nicholas Chimitt, and Stanley H Chan. Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14759–14768, 2021. 3, 5, 8
- [34] 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*, pages 430–446. Springer, 2022. 3, 4, 5, 6, 7
- [35] Ishit Mehta, Michaël Gharbi, Connelly Barnes, Eli Shechtman, Ravi Ramamoorthi, and Manmohan Chandraker. Modulated periodic activations for generalizable local functional representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14214–14223, 2021. 3
- [36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 8
- [37] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3
- [38] Robert J Noll. Zernike polynomials and atmospheric turbulence. *JOsA*, 66(3):207–211, 1976. 3
- [39] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926*, 2023. 3
- [40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [42] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 5
- [43] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings 38*, pages 26–36. Springer, 2016. 6
- [44] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 8
- [45] Masao Shimizu, Shin Yoshimura, Masayuki Tanaka, and Masatoshi Okutomi. Super-resolution from image sequence under influence of hot-air optical turbulence. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 3
- [46] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 3
- [47] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: A survey from 2010 to 2016. *IPSPJ Transactions on Computer Vision and Applications*, 9(1):1–11, 2017. 8
- [48] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 3
- [49] Hao Wang and Lei Tian. Local conditional neural fields for versatile and generalizable large-scale reconstructions in computational imaging. *arXiv preprint arXiv:2307.06207*, 2023. 3
- [50] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. *arXiv preprint arXiv:2306.05422*, 2023. 3
- [51] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 5
- [52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

- [53] Rui Yao, Guosheng Lin, Shixiong Xia, Jiaqi Zhao, and Yong Zhou. Video object segmentation and tracking: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4):1–47, 2020. [8](#)
- [54] Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, and Noah Snavely. Deformable sprites for unsupervised video decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2666, 2022. [3](#)
- [55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. [2](#)
- [56] Xingguang Zhang, Zhiyuan Mao, Nicholas Chimitt, and Stanley H Chan. Imaging through the atmosphere using turbulence mitigation transformer. *arXiv preprint arXiv:2207.06465*, 2022. [3](#), [5](#)
- [57] Haowen Zhou, Brandon Y Feng, Haiyun Guo, Mingshu Liang, Christopher A Metzler, Changhui Yang, et al. Fpm-inr: Fourier ptychographic microscopy image stack reconstruction using implicit neural representations. *arXiv preprint arXiv:2310.18529*, 2023. [3](#)