Supplemental Material: Unsupervised Moving Object Segmentation with Atmospheric Turbulence

Dehao Qin¹, Ripon Kumar Saha², Woojeh Chung², Suren Jayasuriya², Jinwei Ye³, and Nianyi Li¹

 ¹ Clemson University, Clemson SC 29634, USA {dehaoq,nianyil}@clemson.edu
² Arizona State University, Tempe AZ 85287, USA {rsaha8,wchung25,sjayasur}@asu.edu
³ George Mason University, Fairfax VA 22030, USA jinweiye@gmu.edu

In this supplementary document, we present additional details on our algorithm, datasets, and results. We also encourage readers to visit our website for reference: https://turb-research.github.io/segment_with_turb.

1 Threshold δ_{seed} in Region-growing Scheme

Recall that in our region-growing scheme, we use the following criteria to determine whether or not to include a pixel in the growth:

$$|M_t(\mathbf{p}_{new}) - M_t(\mathbf{p}_{seed})| < \delta_{seed},\tag{1}$$

where M_t is the motion featrue map; \mathbf{p}_{new} is the pixel under consideration; \mathbf{p}_{seed} is the seed pixel that we grow from; and δ_{seed} is the threshold for stopping the growth.

Although we mentioned $\delta_{seed} = 0.2 \times M_t(\mathbf{p}_{seed})$ in Section 3.2 of the main paper, this threshold is in fact dependent on turbulence strength and needs to be adjusted for extreme cases. $\delta_{seed} = 0.2 \times M_t(\mathbf{p}_{seed})$ is used for normal turbulence strength. For stronger turbulence that causes severe distortions, we prefer larger δ_{seed} and increase the multiplier from 0.2 to 0.3. For scenes with weak turbulence, we decrease the multiplier to 0.1. In our experiments, we use $\delta_{seed} = 0.2 \times M_t(\mathbf{p}_{seed})$ for "normal turbulence" scenes; and $\delta_{seed} = 0.3 \times M_t(\mathbf{p}_{seed})$ for "severe turbulence" scenes.

2 Dataset Details

We have tested our methods on two datasets: a real long-range video dataset (referred to as "dynamic object segmentation in turbulence" dataset or DOST) and a synthetic dataset simulated by introducing the turbulence effect to an existing segmentation dataset.

2 Qin et al.



Fig. 1: Example scenes from our real dataset (DOST). We show frames from two types of scenes, categorized based on turbulence strength: normal turbulence and severe turbulence.

Real Dataset (DOST): DOST has 38 high-definition videos with a resolution 1920×1080 , captured using a Nikon Coolpix P1000 camera with telelens. Scenes recorded in the videos are in the range of 50 meters to 1 kilometer from the camera. Our scenes contain a variety of subjects, including people, cars, airplanes, bikes, etc., and various everyday movements and interactions, such as walking, running, driving, etc.

For each video, we save individual frames as images in PNG format. In DOST, the number of frames per video ranges from 24 to 56. Fig. 1 shows example frames from different videos under varied turbulence strengths, *i.e.*, "normal" and "severe". We manually annotate per-frame masks for moving objects in each video using the latest online segmentation tools, *i.e.*, Computer Vision Annotation Tool (CVAT) [1]. Our masks are binary with 1 indicating moving objects and 0 indicating static background. It is important to note that our dataset is the first turbulent video dataset with motion segmentation masks. Furthermore, DOST is not only designed for benchmarking the motion segmentation tasks, but also can potentially benefit other tasks, including turbulence restoration, object detection and tracking etc.

Synthetic Dataset: We also generated a set of synthetic turbulence videos for more comprehensive evaluations. We take videos and ground truth segmentation masks from the DAVIS 2016 dataset [5] and use a physics-based turbulence simulator P2S [4] to add turbulence to DAVIS video frames. By controlling the ratio between the telescope aperture diameter (D) and the atmospheric coherence diameter (r_0) used in the simulator, our synthetic set provides a comprehensive range of imaging conditions and turbulence intensities, ensuring a robust assessment of the object segmentation performance across different environments. Fig. 2 shows examples of simulated turbulent video frames of two different strengths.



Fig. 2: Sample frames from our synthetic dataset simulated with a physics-based turbulence simulator, showing two different turbulence strengths.

3 Additional DOST Results

In this section, we present additional qualitative comparisons using videos from our DOST dataset. Based on the analysis detailed in our paper, TMO [2] emerges as the most effective method for segmenting moving objects in videos affected by atmospheric turbulence. Therefore, here, we mainly show comparison results with TMO.

Fig. 3 presents segmentation results of consecutive frames from two videos in the "normal turbulence" category. We can see that our segmentation results have better accuracy and robustness than TMO. In "Video 1", our method accurately segments the walking person, whereas TMO's segmentation includes the static board in the front. This is because their algorithm uses more appearance cues 4 Qin et al.



Fig. 3: Additional visual comparison results on DOST dataset (normal turbulence). In video 2, although TMO can segment the moving person, its mask is not tight to the object (e.g., the arms).



Fig. 4: Additional visual comparison results on DOST dataset (severe turbulence).

and is less dependent on motion. In "Video 2", although TMO can segment the moving woman, its mask is not tight to the object. In contrast, our method well discerns details of the moving object, such as the person's arms.

Fig. 4 shows segmentation results of consecutive frames from two videos in the "severe turbulence" category. We can see that TMO's performance downgrades significantly in both examples. It even fails to generate segmentation masks for "Video 4". In contrast, our method still outputs accurate segmentation masks under severe turbulence distortions. This demonstrates the robustness of our method.

Moreover, Fig. 5 presents additional results from SAM [3]. SAM, which focuses on segmenting all semantic objects, solves a different problem compared to our approach that targets only moving objects. SAM requires user annotation or prompting to initialize the algorithm, which contrasts with our automated segmentation of moving objects. While SAM demonstrates significant challenges in segmenting whole objects under strong turbulence, it performs fairly accurately in semantic segmentation under weak to medium turbulence. These qualitative comparisons highlight the limitations of SAM in the presence of strong turbulence. However, it remains future work to investigate the potential of SAM for semantic segmentation in turbulent environments.



Fig. 5: Qualitative and quantitative comparisons against SAM. Considering that SAM can generate multiple masks, we select the mask with the maximum IoU to represent SAM's performance.

4 Additional Synthetic Results

Fig. 6 shows visual comparison results for synthetic videos with two different turbulence strengths (turb strength: $D/r_0 = 2.5$ and 5.0). Note that this example is very challenging as the moving car has similar color to the background. Although the accuracy of both TMO and our method downgrades in the severe case, our segmentation has been better consistency than TMO.

5 Additional Ablation Results

Here we show qualitative results for ablation on our method's variants. Recall that we test on three variants: A only employs the region-growing algorithm (with Refine-Net excluded); B uses both region-growing and Refine-Net but excludes the grouping loss for refinement; and C is implemented as our full approach. The segmentation results on a car scene are shown in Fig. 7.

6 Qin et al.



Fig. 6: Visual comparison results on synthetic data with different turbulence strengths.

The segmentation results directly obtained from the seeded region growing may be inconsistent and incomplete (see model A results). Introducing the Refine-Net enhances the mask consistency across all frames (see model B results). Finally, adding the grouping function in network optimization further improves the spatial consistency of masks (see model C results).

The main paper showed ablation studies for region growing and the refinement network. In Table 1, we also evaluated the effectiveness of our optical flow stabilization and geometric consistency check.

Table 1: Ablation studies

Metric	[c]w/o optical	flow stab.	& geo.	check	[c]	w/o geo.	check	[c]Ours
\mathcal{J}		0.354				0.685		0.703



Fig. 7: Qualitative comparison results on variants of our method (C is our full model).

References

- 1. Computer vision annotation tool (2024), https://www.cvat.ai/ 2
- Cho, S., Lee, M., Lee, S., Park, C., Kim, D., Lee, S.: Treating motion as option to reduce motion dependency in unsupervised video object segmentation. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE (Jan 2023). https://doi.org/10.1109/wacv56688.2023.00511, http://dx.doi. org/10.1109/WACV56688.2023.00511 3
- 3. Kirillov, A., et al.: Segment anything (2023) 5
- Mao, Z., Chimitt, N., Chan, S.H.: Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14759–14768 (2021) 3
- Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 724–732 (2016). https://doi.org/10.1109/CVPR.2016.85 3