Unsupervised Moving Object Segmentation with Atmospheric Turbulence

Dehao Qin[1(✉)], Ripon Kumar Saha[2], Woojeh Chung[2], Suren Jayasuriya[2], Jinwei Ye[3], and Nianyi Li[1]

[1] Clemson University, Clemson, USA
dehaoq@g.clemson.edu
[2] Arizona State University, Tempe, USA
[3] George Mason University, Virginia, USA
https://turb-research.github.io/segment_with_turb

Abstract. Moving object segmentation in the presence of atmospheric turbulence is highly challenging due to turbulence-induced irregular and time-varying distortions. In this paper, we present an unsupervised approach for segmenting moving objects in videos downgraded by atmospheric turbulence. Our key approach is a detect-then-grow scheme: we first identify a small set of moving object pixels with high confidence, then gradually grow a foreground mask from those seeds to segment all moving objects. This method leverages rigid geometric consistency among video frames to disentangle different types of motions, and then uses the Sampson distance to initialize the seedling pixels. After growing per-frame foreground masks, we use spatial grouping loss and temporal consistency loss to further refine the masks in order to ensure their spatio-temporal consistency. Our method is unsupervised and does not require training on labeled data. For validation, we collect and release the first real-captured long-range turbulent video dataset with ground truth masks for moving objects. Results show that our method achieves good accuracy in segmenting moving objects and is robust for long-range videos with various turbulence strengths.

Keywords: Unsupervised learning · Object segmentation · Turbulence

1 Introduction

Moving object segmentation is critical for motion understanding with an important role in numerous vision applications such as security surveillance [24,35], remote sensing [39,66], and environmental monitoring [5,18]. Although tremendous success has been achieved in motion segmentation and analysis [6,26], the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72658-3_2.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15064, pp. 18–37, 2025.
https://doi.org/10.1007/978-3-031-72658-3_2

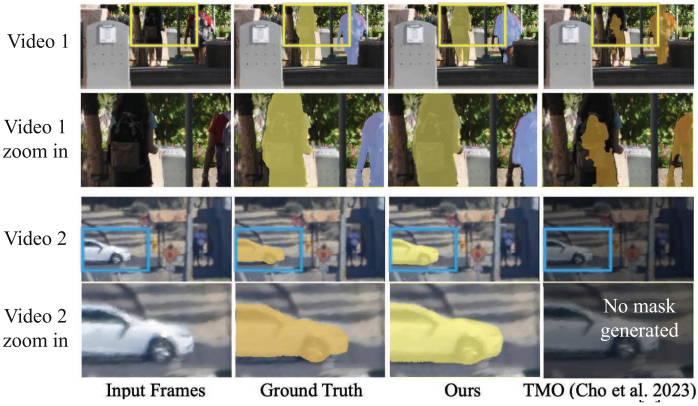


Fig. 1. Our method robustly segments moving objects under various turbulence strengths, while state-of-the-art methods may fail under strong turbulence (2nd video).

problem becomes highly challenging for long-range videos captured with ultra-telephoto lenses (e.g., focal length $>$ 800 mm). These videos often suffer from perturbations caused by atmospheric turbulence which can geometrically shift or warp pixels in images. When mixed with rigid motions in a dynamic scene, they break down the underlying assumption of most motion analysis algorithms: the intensity structures of local regions are constant under motion. Further, when averaged over time, the turbulent perturbation also yields blurriness in images, which blurs out moving object edges and makes it challenging to maintain the spatio-temporal consistency of segmentation masks.

Most motion segmentation algorithms solely consider static backgrounds and assume rigid body movement. Under these premises, learning-based approaches [6,26,53], either supervised or unsupervised, have achieved remarkable success in motion segmentation. However, these algorithms' performance significantly downgrades when applied to videos with turbulence effects (see failure examples in Fig. 6). Supervised methods [19,28,53], even trained on extensive labeled data, cannot generalize well on turbulent videos. Unsupervised methods [6,26], on the other hand, typically rely on optical flow, which becomes inaccurate when rigid motion is perturbed by turbulence. Furthermore, imaging at long distance makes videos highly susceptible to camera shake and motion due to the limited field of view when zooming, further complicating the segmentation task.

In this paper, we present an unsupervised approach for segmenting moving objects in long-range videos affected by air turbulence. The unsupervised nature of our approach is highly desirable, since real-captured turbulent video datasets are scarce and difficult to acquire. Our method directly takes in a turbulent video and outputs per-frame masks that segment all moving objects without the need for data supervision. The overall pipeline of our approach is illustrated in Fig. 2.

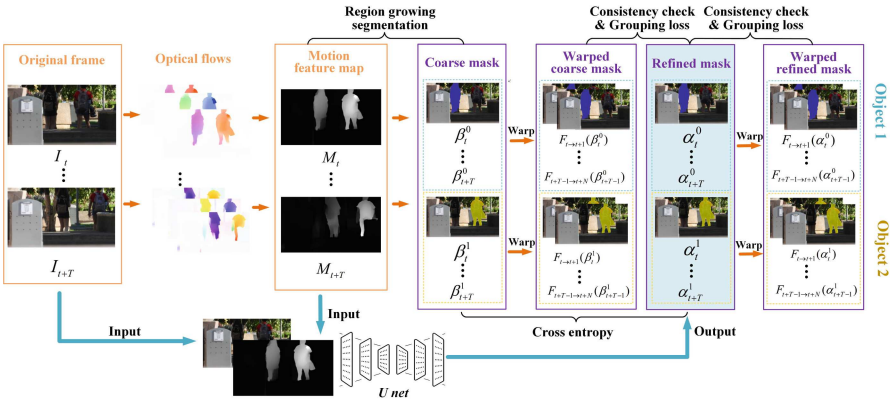


Fig. 2. Overall pipeline of our unsupervised motion segmentation method. We first generate motion feature maps by applying a geometry-based consistency check on optical flows. We then adopt a region-growing scheme to generate coarse segmentation masks. Finally, we refine the masks using cross entropy-based consistency losses to enforce their spatio-temporal consistency.

Our method starts with calculating bidirectional optical flow. To disentangle actual object motion from turbulent motion, we use a novel epipolar geometry-based consistency check to generate motion feature maps that only preserve object motions. We then adopt a region-growing scheme that generates per-object motion segmentation masks from a small set of seed pixels. Finally, we develop a U-Net [42] trained by our proposed bidirectional consistency losses and a pixel grouping function to improve the spatio-temporal consistency of estimated motion segmentation masks.

For evaluation, we collect and release a turbulent video dataset captured with an ultra-telephoto lens called Dynamic Object Segmentation in Turbulence (DOST), and manually annotate ground-truth masks for moving objects, which, to our knowledge, is the *first* dataset in this application domain. We benchmark our approach, as well as other state-of-the-art segmentation algorithms, on this real dataset. Our method is able to handle multiple objects in a dynamic scene and is robust to videos captured with various turbulence strengths (see Fig. 1). More details about DOST datasest can be found in https://turb-research.github.io/DOST/. Our key contributions include:

- A rigid geometry-based and consistency enhanced framework for motion disentanglement in long-range videos.
- Region-growing scheme for generating robust spatio-temporal consistent masks with tight object boundaries.
- A refinement pipeline with novel training losses, which improve the spatio-temporal consistency for segmenting dynamic objects in videos.
- The *first* real-captured long-range turbulent video dataset with ground-truth motion segmentation masks.

2 Related Work

Unsupervised Motion Segmentation. Early approaches tackle this problem using handcrafted features such as objectness [63], saliency [52], and motion trajectory [34]. Recent learning-based unsupervised methods largely rely on either optical flow or feature alignment for locating moving objects. Common optical flow models are RAFT [49], PWC-Net [48], and FLow-Net [8,14]. MP-Net [50] uses optical flow as the only cue for motion segmentation. Many other methods combine optical flow with other information, such as appearance features [21,36,57], long-term temporal information [41,64], and boundary similarity [25,68], to guide motion segmentation. Our method also uses optical flow but segregates different types of motion using a geometry-based consistency check to overcome turbulence-induced errors.

Another class of methods [10,27,58] rely more on appearance features and reduce their dependency on optical flow, so to achieve robust performance regardless of motion. The recent TMO [6] achieves state-of-the-art performance on object segmentation by prioritizing feature alignment and fusion. However, all these methods face challenges when the input video is recorded under air turbulence. Further, the literature on supervised motion segmentation (e.g., [22,29,61]) is not effective due to the lack of large, labelled turbulent video datasets.

Geometric Constraints for Motion Analysis. Geometric constraints are extensively studied for understanding spatial information, such as depth estimation, pose estimation, camera calibration [40,62,70], and 3D reconstruction [47,69]. They can also be useful in understanding motion. Valgaerts et al. [51] propose a variational model to estimate the optical flow, along with the fundamental matrix. Wedel et al. [54] use the fundamental matrix as a weak constraint to guide optical flow estimation. Yamaguchi et al. [56] convert the problem of optical flow estimation into a 1D search problem by using precomputed fundamental matrices with small motion assumptions. Wulff et al. [55] use semantic information to separate dynamic objects from static backgrounds and apply strong geometric constraints to the static backgrounds. Zhong et al. [67] integrate global geometric constraints into network learning for unsupervised optical flow estimation. More recently, Ye et al. [59] use the Sampson error, which measures the consistency of epipolar geometry, to model a loss function for layer decomposition in videos.

Inspired by these works, our method adopts a geometry-based consistency check to separate rigid motion from other types of motion (i.e., turbulent motion and camera motion). Similar to [59], we use the Sampson distance to measure the geometric consistency between neighboring video frames.

Turbulent Image/Video Restoration and Segmentation. Analyzing images or videos affected by air turbulence has been a challenging problem in computer vision due to distortions and blur caused by turbulence. Most techniques have focused on turbulent image and video restoration. Early physics-based approaches [12,20,38] investigate the physical modeling of turbulence (e.g., the Kolmogorov model [20]), and then invert the model to restore clear images.

A popular class of methods [3,9] uses "lucky patches" to reduce turbulent artifacts, although they typically assume static scenes. Motion cues, such as optical flow, are also used for turbulent image restoration [33,46,71]. Notably, Mao et al. [31] adopt an optical-flow guided lucky patch technique to restore images of dynamic scenes. More recently, neural networks have been used to restore turbulent images or videos. Mao et al. [32] introduce a physics-inspired transformer model for restoration, and Zhang et al. [65] improve this thread of work to demonstrate state-of-the video restoration. Li et al. [23] propose an unsupervised network to mitigate the turbulence effect using deformable grids. Jiang et al. [16] extend this deformable grid model to handle more realistic turbulence effects.

In contrast, object segmentation with atmospheric turbulence is relatively understudied. Cui and Zhang [7] propose a supervised network for semantic segmentation with turbulence. They generate a physically-realistic simulated training dataset, but the method cannot handle scene motion and suffers from real-world domain generalization. Saha et al. [45] use simple optical flow-based segmentation in their turbulent video restoration pipeline. Unlike these existing methods, our approach aims to segment moving objects without restoring or enhancing the turbulent video. In this way, our method is able to generate segmentation masks that are consistent with the actual turbulent video.

3 Unsupervised Moving Object Segmentation

Algorithm Overview. Given an input turbulent video: $\{I_t|t = 1, 2, \ldots, T\}$ (where T is the total number of frames, and I_t represents a frame in the video), we first calculate its bidirectional optical flow: $\mathcal{O}_t = \{F_{t\to t\pm i}|i = 1, \ldots, B\}$ (where B is the maximum number of frames used for calculation; $F_{t\to t+i}$ is the forward flow, and $F_{t\to t-i}$ is the backward flow). We then perform an epipolar geometry-based consistency check to disentangle rigid object motion from turbulence-induced motions and camera motions (Sect. 3.1). We output per-frame motion feature maps: $\{M_t|t = 1, 2, \ldots, T\}$ to characterize candidate motion regions. Next, we leverage a detect-then-grow strategy, named "region growing", to generate motion segmentation masks: $\{\beta_t^m|t = 1, 2, \ldots, T\}$ for every moving object m, from a small set of seedling pixels selected from $\{M_t\}_{t=1}^{T}$ (Sect. 3.2). Finally, we further refine the masks by using a U-Net trained with our proposed bidirectional spatial-temporal consistency losses and pixel grouping loss (Sect. 3.3). The final output is a set of per-frame, per-object binary masks: $\{\alpha_t^m|t = 1, 2, \ldots, T\}$ segmenting each moving object in a dynamic scene.

3.1 Epipolar Geometry-based Motion Disentanglement

We first tackle the problem of motion disentanglement, which is a major challenge posed by turbulence perturbation in rigid motion analysis. Our key idea is to check on the rigid geometric consistency among video frames: pixels on moving objects do not obey the geometric consistency constraint posed by the

image formation model. Specifcially, we use the Sampson distance, which measures geometric consistency with a given epipolar geometry, to improve the spatial-temporal consistency among video frames. We first average the optical flow between adjacent frames to stabilize the direct estimations $\{\mathcal{O}_t\}$, since they are susceptible to turbulence perturbation. We then calculate the Sampson distance using fundamental matrices estimated from the averaged optical flow. Next, we merge the Sampson distance maps as the motion feature maps $\{M_t|t = 1, 2, \ldots, T\}$. We use the motion feature map values as indicators of how likely a pixel has rigid motion (the higher the value, the higher the likelihood). Our epipolar geometry-based motion disentanglement pipeline is shown in Fig. 3.

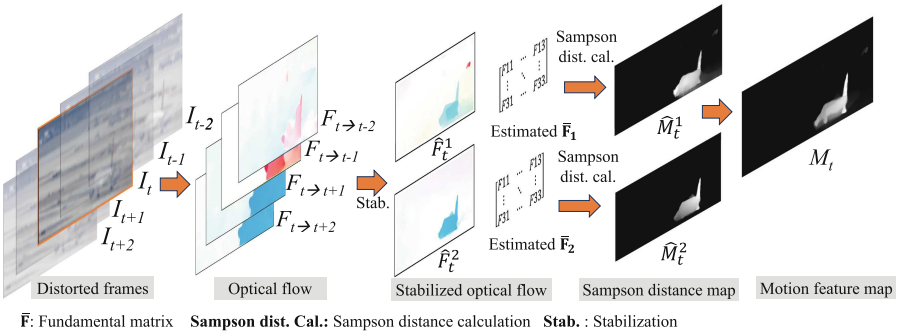


Fig. 3. Pipeline of epipolar geometry-based motion disentanglement. Since the raw optical flows are downgraded by turbulence, we apply a geometry-based consistency check to generate motion feature maps that only preserve object motion.

Optical Flow Stabilization. Since atmospheric turbulence causes erratic pixel shifts in video frames [31], estimating optical flow is prone to error. We address this by first stabilizing the optical flow estimations: we assume consistent object motion during a short period of time, and then average the optical flow within a small time step to reduce the error caused by turbulence perturbation without losing features of the actual rigid motion. Specifically, given a sequence of bi-directional optical flow $\mathcal{O}_t = \{F_{t\to t\pm i}\}_{i=0}^{B}$, we calculate a sequence of per-frame stabilized flow $\hat{\mathcal{O}}_t = \{\hat{F}_t^j | j = 1, 2, ..., A\}$ (where A is the total number of stabilized flows for each frame) by averaging the original sequence within a short interval:

$$\hat{F}_t^j = \frac{1}{|\mathcal{K}^j|} \sum_{i \in \mathcal{K}^j} \frac{F_{t\to t+i}}{i}, \tag{1}$$

where $\mathcal{K}^j$ is the temporal interval used for calculating $\hat{F}_t^j$, namely the subset of $\{x \mid x \in \mathbb{Z}, -B \leq x \leq B\}$

Geometric Consistency Check. Our fundamental assumption is that pixels on moving objects have larger geometric consistency errors compared with static background, when mapping a frame to its neighboring time frame using

foundamental matrix. The Sampson distance [13] measures rigid geometric consistency by calculating the distance between a frame pair in video, constrained by epipolar geometry. Since moving objects do not obey the epipolar geometry that assumes a static scene, their correspondences will have a large Sampson distance. Although the turbulent perturbation also breaks down the epipolar geometry, the errors that they introduce are much smaller and more random, and thus easily eliminated through averaging.

Given a stabilized optical flow $\hat{F}_t^j$, we calculate its Sampson distance map $\hat{M}_t^j$ as:

$$\hat{M}_t^j(\mathbf{p}_1, \mathbf{p}_2) = \frac{(\mathbf{p}_2^\mathsf{T}\bar{\mathbf{F}}\mathbf{p}_1)^2}{(\bar{\mathbf{F}}\mathbf{p}_1)_1^2 + (\bar{\mathbf{F}}\mathbf{p}_1)_2^2 + (\bar{\mathbf{F}}^\mathsf{T}\mathbf{p}_2)_1^2 + (\bar{\mathbf{F}}^\mathsf{T}\mathbf{p}_2)_2^2}, \tag{2}$$

where $\mathbf{p}_1$ and $\mathbf{p}_2$ are the homogeneous coordinates of a pair of corresponding points in two neighboring frames. We determine the correspondence using the stabilized optical flow: $\mathbf{p}_2 = \mathbf{p}_1 + \hat{F}_t^j(\mathbf{p}_1)$. $\bar{\mathbf{F}}$ is the fundamental matrix between the two frames estimated by Least Median of Squares (LMedS) regression [44].

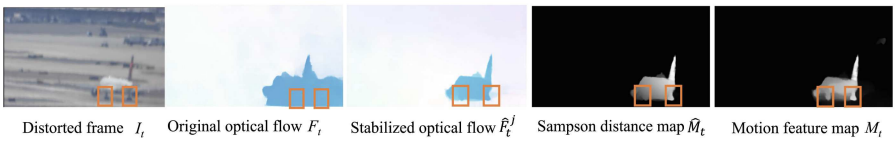


Fig. 4. Step-by-step intermediate results for motion feature map estimation.

We further average all available Sampson distance maps $\{\hat{M}_t^j | j = 1, 2, ..., A\}$ for a frame I_t to obtain the per-frame motion feature map: $M_t = \frac{1}{A}\sum_{j=1}^{A}\hat{M}_t^j$. This map indicates how likely a pixel is to belong to a moving object. Figure 4 compares the intermediate results when generating the motion feature map. Note that the original optical flow map is corrupted by turbulence and cannot resolve the airplane. After our stabilization and consistency check, the final motion feature map preserves the airplane's shape including the highlighted wheels.

3.2 Region Growing-Based Segmentation

Next, based on the motion feature maps, we adopt a "region growing" scheme to generate segmentation masks for moving objects (see Fig. 5). Note that while motion feature maps effectively characterize object motions, they tend to be non-binary and exhibit fuzziness at the object boundaries.

Initial Seed Selection. We first select a small set of seedling pixels that have high confidence of being on a moving object. This selection is based on the motion feature map which encodes how likely a pixel is to be in motion. Note that since we apply the sliding window on the motion map M_t, the size and appearance of the object would not affect the seed selection. Specifically, our sliding windows

$\{W_k\}_{k=1}^{K}$ are of size $D \times D$, where D is adaptively chosen based on the input resolution, to scan through M_t to determine the initial seed of each moving object k. A seed is detected when pixels within the search window are of similar large values, indicating large Sampson distances in this area. Specifically, for a search window W_k, we consider it has found a seed when it satisfies two criteria: (1) its average value $\bar{M}_t(W_k)$ is greater than a threshold δ_1; and (2) its variance $\sigma^2(M_t(W_k))$ is less than a small threshold δ_2. Each selected seedling region is assigned an integer mask ID to uniquely identify multiple moving objects.

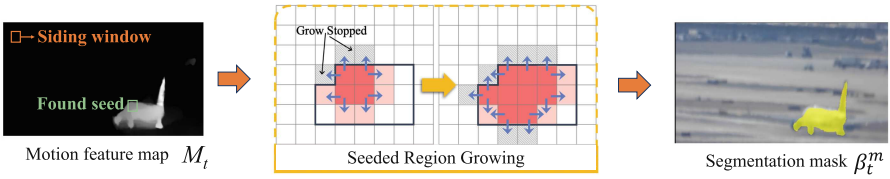


Fig. 5. Pipeline of region growing-based segmentation. We select seeds on the motion feature map using a sliding window. We then grow the seeds to full segmentation masks.

Seeded Region Growing. We then grow full segmentation masks from the initialized seedling pixels. We gradually expand each seeded region outwards to the nearest neighbors of boundary pixels using this criteria for pixel inclusion:

$$|M_t(\mathbf{p}_{new}) - M_t(\mathbf{p}_{seed})| < \delta_{seed}, \tag{3}$$

where M_t is the motion feature map; $\mathbf{p}_{new}$ is the pixel under consideration; $\mathbf{p}_{seed}$ is the seed pixel that we grow from; and $\delta_{seed} = 0.2 \times M_t(\mathbf{p}_{seed})$ is the threshold for stopping the growth. Note that this threshold value depends on turbulence strength and needs to be adjusted for extreme cases. For stronger turbulence, we prefer larger δ_{seed} and increase the multiplier from 0.2 to 0.3. For weak turbulence, we decrease the multiplier to 0.1. The object's border tends to get blurred for videos with severe turbulence. Meanwhile, a larger threshold makes the region grow harder, resulting in more reliable results.

When growing from multiple seeds, we skip the pixels already examined so that different object masks are non-overlapping. We thus obtain a set of per-frame segmentation masks $\{\beta_t^m\}_{t=1}^{T}$.

Mask ID Unification. When a scene has multiple moving objects (say K), our region-growing algorithm will generate K segmentation masks with IDs ranging from 1 to K for each frame. Since the region-growing module is applied to each frame independently, the mask ID among different frames may be inconsistent with respect to objects. There is a need to unify the mask IDs across frames, so that the same mask ID always maps to the same object. We propose a K-means-based filtering technique to do so.

We first represent each mask region by its centroid: $\mathbf{c}_t^m = \text{mean}(\mathbf{p}_t^m)$ (where $\mathbf{p}_t^m$ is the coordinates of all foreground pixels in the mask). We take the mask

centroids for all frames and all objects, and optimize K K-means cluster centroids μ^m:

$$\underset{\mu^m}{\operatorname{argmin}} \sum_{m=1}^{K} \sum_{t=1}^{T} ||\mathbf{c}_t^m - \mu^m||^2, \tag{4}$$

where K is the total number of objects, and T is the total number of frames.

We then re-ID the masks for each frame by comparing their centroids to the K-means cluster centroids. The K-means cluster IDs act as a global reference for all frames. Each mask is assigned the ID m of a K-means cluster that it is closest to: $m^* = \underset{m}{\operatorname{argmin}} ||\mathbf{c}_t^m - \mu^m||$. After this re-assignment, the mask IDs for all frames are consistent with each ID uniquely identifying a moving object in the scene. This allows our method to handle multiple moving objects in a scene.

3.3 Spatio-temporal Refinement

Finally, we further refine the masks to improve their spatio-temporal consistency. We develop a Refine-Net Φ_θ with a U-Net [43] backbone to refine the masks generated by region growing.

Parameter Initialization. We concatenate the video frame I_t with its motion feature map M_t. The concatenated tensor is fed as input to the Refine-Net Φ_θ: $\alpha_t^m = \Phi_\theta(I_t, M_t)$ (here α_t^m is the output of Φ_θ, which is a refined mask). We use the following loss function for initializing the parameters of Φ_θ:

$$\mathcal{L}_{ini} = \gamma_1 \mathcal{L}_1 + \gamma_2 \sum_g \mathcal{L}_2^g + \gamma_3 \sum_g \mathcal{L}_3^g, \tag{5}$$

where γ_1, γ_2, and γ_3 are balancing weights for each loss term. We run 20–30 epochs for initialization. Below, we describe our loss terms in detail.

$\mathcal{L}_1$ is a pixel-wise cross-entropy loss that enforces consistency between the refined output mask α_t^m and coarse input mask β_t^m. It is calculated as:

$$\mathcal{L}_1 = \frac{1}{\Omega} \sum_{\mathbf{p} \in \Omega} (-\alpha_t^m(\mathbf{p}) \log \alpha_t^m(\mathbf{p}) + \beta_t^m(\mathbf{p}) \log \beta_t^m(\mathbf{p})), \tag{6}$$

where $\mathbf{p}$ is the pixel coordinates, and Ω represents the spatial domain.

$\mathcal{L}_2^g$ is a bidirectional consistency loss that enforces flow consistency between α_{t+g}^m and the optical flow-warped input mask: $\hat{\beta}_t^m = F_{t \to t+g}(\beta_t^m)$. We also use the cross-entropy for comparison, and $\mathcal{L}_2^g$ is written as:

$$\mathcal{L}_2^g = \frac{1}{\Omega} \sum_{\mathbf{p} \in \Omega} (-\alpha_{t+g}^m(\mathbf{p}) \log \alpha_{t+g}^m(\mathbf{p}) + \hat{\beta}_t^m(\mathbf{p}) \log \hat{\beta}_t^m(\mathbf{p})). \tag{7}$$

$\mathcal{L}_3^g$ is another bidirectional consistency loss that enforces flow consistency between α_{t+g}^m and the optical flow-warped version of itself: $\hat{\alpha}_t^m = F_{t \to t+g}(\alpha_t^m)$. $\mathcal{L}_3^g$ is written as:

$$\mathcal{L}_3^g = \frac{1}{\Omega} \sum_{\mathbf{p} \in \Omega} (-\alpha_{t+g}^m(\mathbf{p}) \log \alpha_{t+g}^m(\mathbf{p}) + \hat{\alpha}_t^m(\mathbf{p}) \log \hat{\alpha}_t^m(\mathbf{p})). \tag{8}$$

After initialization, the output mask α_t^m is aligned with the input mask β_t^m and has improved temporal consistency with the two bidirectional losses.

Iterative Refinement Using Grouping Function. To further improve the mask quality and consistency, we adopt an iterative refinement constrained by a grouping function. The refinement runs for 10 epochs. We update the input reference mask β_t^m using a K-means-based grouping function for every 3 epochs. Specifically, we concatenate each pixel's mask values $\{\beta_t^m(\mathbf{p})\}_{t=1}^T$, with its coordinates $\mathbf{p} = (x, y)$, to form a new tensor $\{T_t^m(\mathbf{p})\}_{t=1}^T$ that combines both the motion and spatial information. The combined tensor $T_t^m(\mathbf{p})$ of all pixels is used to optimize two K-means cluster centroids θ^1, θ^2: one for foreground cluster (θ^1) and the other for background (θ^2). The centroids are optimized using the following function:

$$\underset{\theta^1,\theta^2}{\operatorname{argmin}} \sum_{i=1}^{2} \sum_{\mathbf{p}\in\Omega}^{\Omega} ||T_t^i - \theta_t^i||^2, \tag{9}$$

where Ω is the domain of all pixels. We then re-assign the values of β_t^m: if $T_t^m(\mathbf{p})$ is closer to θ^1, we assign the mask value as 1; otherwise, we consider the pixel as background and assign the mask value as 0.

The loss function we use for network optimization is the same as the initialization step, but in this refinement step, β_t^m is updated using the K-means-based grouping every 3 epochs. Without this grouping-based refinement, the output mask tends to have gaps or other spatial inconsistent artifacts.

Table 1. Comparison of existing datasets on turbulent images or videos.

Dataset	Source	Format	Purpose	Num. of video	Availability
Turb Pascal VOC [7]	Synthetic	Image	Segmentation	✗	✗
Turb ADE20K [7]	Synthetic	Image	Segmentation	✗	✗
TSRW-GAN [17]	Real	Video	Restoration	27*	✓
OTIS [11]	Real	Video	Restoration	5	✓
BVI-CLEAR [2]	Real	Video	Restoration	3	✓
DOST (ours)	Real	Video	Segmentation	38	✓

* The actual number of videos in TSRW-GAN is greater than 27, but many of the videos have large overlaps. The number of unique (or non-overlapped) videos is 27.

4 Experiments

4.1 DOST Dataset

We capture a long-range turbulent video dataset, which we call *Dynamic Object Segmentation in Turbulence (DOST)*, to evaluate our method. DOST consists of 38 videos, all collected outdoors in hot weather using long focal length settings. All videos contain instances of moving objects, such as vehicles, aircraft, and

pedestrians. We manually annotate the video to provide per-frame ground truth masks for segmenting moving objects. Specifically, we use a Nikon Coolpix P1000 to capture the videos. The camera has an adjustable focal length of up to 539mm (125× optical zoom), which is equivalent to 3000mm focal length in 35mm sensor format. We record videos with a resolution of 1920×1080. In total, our dataset has 38 videos with 1719 frames. We annotate moving objects in each video frame using the Computer Vision Annotation Tool (CVAT) [1]. Our dataset is the first of its kind to provide a ground truth moving object segmentation mask in the context of long-range turbulent video. DOST is designed for motion segmentation, but can be used for other tasks (e.g., turbulent video restoration).

Table 1 compares DOST with existing datasets designed for turbulent image or video processing. Since real images/videos with air turbulence are very difficult to acquire, real datasets are scarce, and existing ones are usually small in size. Cui and Zhang [7] synthesize large turbulent image datasets using standard image datasets (Pascal VOC 2012 [7] and ADE20K [7]) and a turbulence simulator. However, there is a domain gap between simulated data and real data. Further, their datasets only contain single images and cannot be used for studying motion. Other real turbulent video datasets [2,11,17] are all designed for the restoration task. Although some [11,17] have bounding box annotations, none provide object-tight segmentation masks.

Table 2. Quantitative comparisons with state-of-the-art unsupervised methods on DOST w.r.t. various turbulence strengths.

Model	$\mathcal{J}$			$\mathcal{F}$			$\mathcal{G}$
	Normal turb.	Severe turb.	Overall	Normal turb.	Severe turb.	Overall	Overall
TMO [6]	0.643	0.235	0.439	0.757	0.315	0.536	0.487
DSprites [60]	0.427	0.101	0.264	0.772	0.203	0.374	0.319
DS-net [26]	0.361	0.191	0.276	0.422	0.232	0.327	0.302
Ours	**0.851** ↑	**0.557** ↑	**0.703** ↑	**0.812** ↑	**0.634** ↑	**0.723** ↑	**0.713** ↑

4.2 Implementation Details

We implemented our network using PyTorch on a supercomputing node equipped with an NVIDIA GTX A100 GPU. The input frames are resized to a lower resolution of 240×432 for faster optical flow calculation and network training. We use RAFT [49] for optical flow estimation with a maximum frame interval of 4. The stabilized optical flow and Sampson distance maps are subsequently calculated based on the RAFT output. The region-growing algorithm's stopping threshold δ_{seed} (see Eq. 3) is dependent on the turbulence strength, and we need to adjust this parameter for varied strength turbulent videos (see more details in the supplementary material). For the bidirectional consistency losses, we compare with four neighboring time frames, $g \in \{-2, -1, 1, 2\}$ in Eqs. 7–8.

Evaluation Metrics. Given a ground-truth mask, we evaluate the accuracy of the estimated segmentation mask using two standard metrics [4]: (1) Jaccard's

Index ($\mathcal{J}$) that calculates the intersection over the union of two sets (also known as intersection-over-union or IoU measure); and (2) F1-Score ($\mathcal{F}$) that calculates the harmonic mean of precision and recall (also known as dice coefficient). We also calculate the average of $\mathcal{J}$ and $\mathcal{F}$, and denote this overall metric as $\mathcal{G}$.

4.3 Comparison with State-of-the-Art Methods

We compare our method against recent state-of-the-art *unsupervised* methods for motion segmentation, including TMO [6], Deformable Sprites [60], and DS-Net [26]. TMO [6] achieves high accuracy in object segmentation regardless of motion. It therefore has certain advantages in handling turbulent videos, although it is not specifically designed for this purpose. Deformable Sprites [60] integrates appearance features with optical flow, and further enforces consistency with optical flow-guided grouping loss and warping loss. DS-Net [26] uses multi-scale spatial and temporal features for segmentation and is able to achieve good performance when the input is noisy. For all three methods, we use the code implementations provided by authors and train their networks using settings described in papers. For methods that need optical flow, we use RAFT to estimate optical flow between consecutive frames.

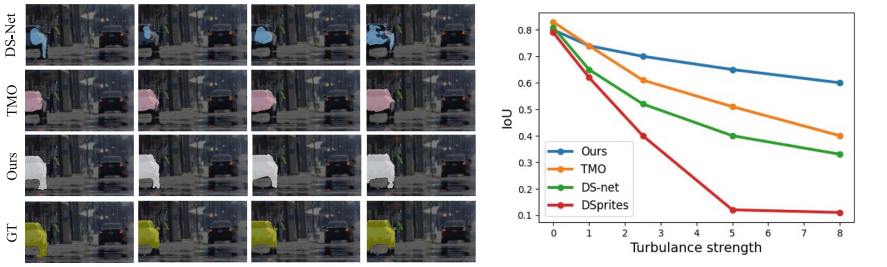
Quantitative Comparisons. We show quantitative comparison results in Table 2. All methods are evaluated on our DOST dataset. We organize videos into two sets, "normal turb." and "severe turb.", according to their exhibited turbulence strength. Our method significantly outperforms these state-of-the-art on motion segmentation accuracy under various turbulence strengths. In normal cases, some can still achieve decent performance, whereas our method scores much higher in all metrics. Compared to TMO, whose overall score is the highest among the three state-of-the-art, our accuracy is increasing by 60.1% in $\mathcal{J}$ and 34.9% in $\mathcal{F}$. In severe cases, the performance of all state-of-the-art significantly downgrades, with all $\mathcal{J}$ values lower than 0.25 and $\mathcal{F}$ lower than 0.35. In contrast, our method is relatively robust to strong turbulence.

We also experiment on a larger synthetic dataset to evaluate our robustness with respect to turbulence strength. We use a physics-based turbulence simulator [30] and the DAVIS 2016 dataset [37] to synthesize videos with various strengths of turbulence. We test the three state-of-the-art methods on the synthetic data as well. $\mathcal{J}$ score (or IoU) plots with respect to turbulence strength are shown in Fig. 7b. The results demonstrate that our method is robust to various turbulence strengths. Note that when there is no turbulence in the scene (strength = 0), TMO has a slightly higher $\mathcal{J}$ score, as it incorporates more visual cues for segmentation, whereas our method focuses on turbulence artifacts.

Qualitative Comparisons. We show visual comparisons with state-of-the-art in Fig. 6. We can see that the optical estimations are largely affected by turbulence, especially in severe cases. Our method is able to generate segmentation masks that are tight to the object and works well for multiple moving objects. State-of-the-art methods face challenges when the turbulence strength is strong, or the moving object is too small. Their segmentation masks are incomplete in many cases. In the airplane scenes, DS-Net and DSprites fail to detect moving

Fig. 6. Qualitative comparisons with state-of-the-art methods on DOST w.r.t. various turbulence strengths. Here we also show the raw optical flows ("Flow"), our motion feature maps ("M_t"), and our ground truth masks ("GT").

objects. Notably, our method achieves the highest robustness to turbulent distortions, and camera shakes, as shown in Fig. 7. We analyzed the effects of various camera motions, including complex multi-directional shake, on segmentation results in video sequences, as illustrated in Fig. 7a. Quantitatively, our method achieved an average IoU score of 0.712 on videos featured by camera shaking



(a) Comparisons w.r.t. camera shake. **(b)** Comparisons w.r.t. turbulence strength.

Fig. 7. Example results of handling videos (a) with camera motion and (b) impacted by different turbulence strength. (a) Our results achieves robust performance on different time frames (2nd to 4th columns) in videos suffering from significant camera shake. (b) We can also tell that our method achieves the best robustness across different turbulence strengths, even when it is very strong.

Variants	$\mathcal{L}_2^g + \mathcal{L}_3^g$	$\mathcal{L}_{GR}$	$\mathcal{J}$
A	✗	✗	0.663
B	✔	✗	0.690
C	✔	✔	**0.703**

(a) Ablation studies on our key components.

Metrics \ Interval	1	2	3	4	5
IoU	0.55	0.67	0.72	0.73	0.73
Dice	0.52	0.63	0.69	0.70	0.70

(b) Ablation study on optical flow interval.

Fig. 8. Ablation studies. (a) Variants of our network. *A*: Simply using region-growing; *B*: Full pipeline without grouping loss; *C*: Our full pipeline. (b) The interval $i = 1...5$ is the maximum temporal gap we used to stablize optical flow in Eq. 1.

in our dataset, compared with TMO/0.305, DS-Net/0.267, and DSprites/0.235, respectively.

We notice that supervised segmentation methods can hardly generalize to DOST, since these methods were trained solely using turbulence-free data. For instance, the foundation model for segmentation, i.e., SAM [19], which has been trained on 11 million images, and over 1B masks, still fails at segmenting whole objects under strong turbulence, as shown in Fig. 6 (last row). Additionally, SAM does not interpret motion , and SAM requires a user to click or prompt the algorithm, whereas we only segment moving objects without any need for user input. More comparison results can be found in supplemental material.

4.4 Ablation Studies

We perform ablation studies to evaluate individual components of our method. All experiments are performed on DOST. We test on three variants of our method: 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. Their $\mathcal{J}$ score comparison results are shown in Fig. 8a. Each component is clearly effective, and our full model achieves the best performance. We also evaluate the influence of the optical flow interval (i.e., the number of temporal frames used for optical flow calculation). Accuracy scores with respect to the interval length are shown in Fig. 8b. We can see that the score achieves the plateau when the interval is greater than 4. Therefore, we set the interval to 4.

We also evaluated the effectiveness of our optical flow stabilization and geometric consistency check. Without the optical flow stabilization step, the IoU immediately drops to 0.354; without the geometric consistency check step, the IoU is 0.685, compared with IoU of 0.703 in our full pipeline.

5 Conclusions

In summary, we present an unsupervised approach for segmenting moving objects in videos affected by air turbulence. Our method uses a geometry-based consistency check to disambiguate motions and a region-growing scheme to generate tight segmentation masks. The masks are further refined with spatio-temporal

consistency losses. Our method significantly outperforms the existing state-of-the-art in terms of accuracy and robustness when handling turbulent videos. We also contribute the first long-range turbulent video dataset designed for motion segmentation. Nevertheless, due to the unsupervised nature of our method, the current version can only achieve a latency of 0.95 FPS. Our future work will focus on optimizing the approach to reduce this latency by embedding results from foundations models such as SAM. Our method also has limited performance in separating overlapping moving objects in the videos. To address this, we plan to integrate additional visual cues, such as appearance and saliency.

Acknowledgements. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022–21102100003 and NSF IIS-2232298/2232299/2232300. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotation therein. We also wish to thank ASU Research Computing for providing GPU resources [15] to support this research.

References

1. Computer vision annotation tool (2024). https://www.cvat.ai/
2. Anantrasirichai, N., Achim, A., Kingsbury, N.G., Bull, D.R.: Atmospheric turbulence mitigation using complex wavelet-based fusion. IEEE Trans. Image Process. **22**(6), 2398–2408 (2013)
3. Aubailly, M., Vorontsov, M.A., Carhart, G.W., Valley, M.T.: Video enhancement through automated lucky region fusion from a stream of atmospherically distorted images. In: Frontiers in Optics 2009/Laser Science XXV/Fall 2009 OSA Optics & Photonics Technical Digest. p. CThC3. Optica Publishing Group (2009). https://doi.org/10.1364/COSI.2009.CThC3
4. Chen, L., Wu, Y., Stegmaier, J., Merhof, D.: Sortedap: rethinking evaluation metrics for instance segmentation. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3925–3931. IEEE Computer Society, Los Alamitos, CA, USA (2023). https://doi.org/10.1109/ICCVW60793.2023.00424
5. Chen, X., et al.: Moving object segmentation in 3D lidar data: a learning-based approach exploiting sequential data. IEEE Rob. Autom. Lett. **6**, 6529–6536 (2021). https://doi.org/10.1109/LRA.2021.3093567
6. 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: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5140–5149 (2023)
7. Cui, L., Zhang, Y.: Accurate semantic segmentation in turbulence media. IEEE Access **7**, 166749–166761 (2019)
8. Dosovitskiy, A., et al.: Flownet: learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2758–2766 (2015)
9. Fried, D.L.: Probability of getting a lucky short-exposure image through turbulence. JOSA **68**(12), 1651–1658 (1978)

10. Garg, S., Goel, V.: Mask selection and propagation for unsupervised video object segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1680–1690 (2021)
11. Gilles, J., Ferrante, N.B.: Open turbulent image set (OTIS). Pattern Recogn. Lett. **86**, 38–41 (2017)
12. Gutierrez, D., Seron, F.J., Munoz, A., Anson, O.: Simulation of atmospheric phenomena. Comput. Graph. **30**(6), 994–1010 (2006)
13. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, USA (2003)
14. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2462–2470 (2017)
15. Jennewein, D.M., et al.: The sol supercomputer at Arizona state university. In: Practice and Experience in Advanced Research Computing. pp. 296–301. PEARC '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3569951.3597573
16. Jiang, W., Boominathan, V., Veeraraghavan, A.: Nert: implicit neural representations for unsupervised atmospheric turbulence mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4235–4242 (2023)
17. Jin, D., et al.: Neutralizing the impact of atmospheric turbulence on complex scene imaging via deep learning. Nat. Mach. Intell. **3**(10), 876–884 (2021)
18. Johnson, B.A., Ma, L.: Image segmentation and object-based image analysis for environmental monitoring: recent areas of interest, researchers' views on the future priorities. Remote Sens. **12**(11) (2020). https://doi.org/10.3390/rs12111772, https://www.mdpi.com/2072-4292/12/11/1772
19. Kirillov, A., et al.: Segment anything. arXiv **2304**, 02643 (2023)
20. Kolmogorov, A.N.: Dissipation of energy in locally isotropic turbulence. Akademiia Nauk SSSR Doklady **32**, 16 (1941)
21. Lee, M., Cho, S., Lee, S., Park, C., Lee, S.: Unsupervised video object segmentation via prototype memory network. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5924–5934 (2023)
22. Li, H., Chen, G., Li, G., Yu, Y.: Motion guided attention for video salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 7274–7283 (2019)
23. Li, N., Thapa, S., Whyte, C., Reed, A., Jayasuriya, S., Ye, J.: Unsupervised non-rigid image distortion removal via grid deformation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2502–2512 (2021). https://doi.org/10.1109/ICCV48922.2021.00252
24. Ling, Q., Yan, J., Li, F., Zhang, Y.: A background modeling and foreground segmentation approach based on the feedback of moving objects in traffic surveillance systems. Neurocomputing **133**, 32–45 (2014). https://doi.org/10.1016/j.neucom.2013.11.034, https://www.sciencedirect.com/science/article/pii/S0925231214000654
25. Liu, D., Yu, D., Wang, C., Zhou, P.: F2net: learning to focus on the foreground for unsupervised video object segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2109–2117 (2021)
26. Liu, J., Wang, J., Wang, W., Su, Y.: Ds-net: Dynamic spatiotemporal network for video salient object detection. Digit. Signal Process.

130, 103700 (2022). https://doi.org/10.1016/j.dsp.2022.103700, https://www.sciencedirect.com/science/article/pii/S1051200422003177
27. Lu, X., Wang, W., Danelljan, M., Zhou, T., Shen, J., Van Gool, L.: Video object segmentation with episodic graph memory networks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, pp. 661–679. Springer (2020). https://doi.org/10.1007/978-3-030-58580-8_39
28. Lu, X., Wang, W., Shen, J., Crandall, D.J., Van Gool, L.: Segmenting objects from relational visual data. IEEE Trans. Pattern Anal. Mach. Intell. **44**(11), 7885–7897 (2022). https://doi.org/10.1109/TPAMI.2021.3115815
29. Mahadevan, S., Athar, A., Ošep, A., Hennen, S., Leal-Taixé, L., Leibe, B.: Making a case for 3D convolutions for object segmentation in videos. arXiv preprint arXiv:2008.11516 (2020)
30. Mao, Z., Chimitt, N., Chan, S.H.: Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14739–14748. IEEE Computer Society, Los Alamitos, CA, USA (2021). https://doi.org/10.1109/ICCV48922.2021.01449, https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01449
31. Mao, Z., Chimitt, N., Chan, S.H.: Image reconstruction of static and dynamic scenes through anisoplanatic turbulence. IEEE Trans. Comput. Imaging **6**, 1415–1428 (2020). https://doi.org/10.1109/TCI.2020.3029401
32. Mao, Z., Jaiswal, A., Wang, Z., Chan, S.H.: Single frame atmospheric turbulence mitigation: a benchmark study and a new physics-inspired transformer model. In: European Conference on Computer Vision, pp. 430–446. Springer (2022). https://doi.org/10.1007/978-3-031-19800-7_25
33. Nieuwenhuizen, R., Dijk, J.S.K.D.T.M.F.L.R.I.I.T.P.O.L.M.O.: Dynamic turbulence mitigation for long-range imaging in the presence of large moving objects. In: EURASIP J Image Video Process, pp. 1–8 (2019). https://doi.org/10.1186/s13640-018-0380-9
34. Ochs, P., Brox, T.: Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In: 2011 International Conference on Computer Vision, pp. 1583–1590 (2011). https://doi.org/10.1109/ICCV.2011.6126418
35. Osorio, R., López, I., Peña, M., Lomas, V., Lefranc, G., Savage, J.: Surveillance system mobile object using segmentation algorithms. IEEE Lat. Am. Trans. **13**, 2441–2446 (2015). https://doi.org/10.1109/TLA.2015.7273810
36. Pei, G., Shen, F., Yao, Y., Xie, G.S., Tang, Z., Tang, J.: Hierarchical feature alignment network for unsupervised video object segmentation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV, pp. 596–613. Springer (2022). https://doi.org/10.1007/978-3-031-19830-4_34
37. 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: Computer Vision and Pattern Recognition (2016)
38. Potvin, G., Forand, J., Dion, D.: A parametric model for simulating turbulence effects on imaging systems. DRDC Valcartier TR **2006**, 787 (2007)
39. Rai, M., Al-Saad, M., Darweesh, M., Al-Mansoori, S., Al-Ahmad, H., Mansoor, W.: Moving objects segmentation in infrared scene videos. In: 2021 4th International Conference on Signal Processing and Information Security (ICSPIS), pp. 17–20 (2021). https://doi.org/10.1109/icspis53734.2021.9652436

40. Ranjan, A., et al.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12232–12241. IEEE Computer Society, Los Alamitos, CA, USA (2019). https://doi.org/10.1109/CVPR.2019.01252, https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.01252
41. Ren, S., Liu, W., Liu, Y., Chen, H., Han, G., He, S.: Reciprocal transformations for unsupervised video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15455–15464 (2021)
42. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
43. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, pp. 234–241. Springer International Publishing, Cham (2015)
44. Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**(388), 871–880 (1984). http://www.jstor.org/stable/2288718
45. Saha, R.K., Qin, D., Li, N., Ye, J., Jayasuriya, S.: 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 (2024)
46. Shimizu, M., Yoshimura, S., Tanaka, M., Okutomi, M.: Super-resolution from image sequence under influence of hot-air optical turbulence. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008). https://doi.org/10.1109/CVPR.2008.4587525
47. Shin, D., Ren, Z., Sudderth, E., Fowlkes, C.: 3D scene reconstruction with multi-layer depth and epipolar transformers. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2172–2182. IEEE Computer Society, Los Alamitos, CA, USA (2019)
48. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-net: CNNS for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8934–8943 (2018)
49. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision, pp. 402–419. Springer (2020). https://doi.org/10.1007/978-3-030-58536-5_24
50. Tokmakov, P., Alahari, K., Schmid, C.: Learning motion patterns in videos. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 531–539. IEEE Computer Society, Los Alamitos, CA, USA (2017). https://doi.org/10.1109/CVPR.2017.64, https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.64
51. Valgaerts, L., Bruhn, A., Weickert, J.: A variational model for the joint recovery of the fundamental matrix and the optical flow. In: Rigoll, G. (ed.) Pattern Recogn., pp. 314–324. Springer, Berlin Heidelberg, Berlin, Heidelberg (2008)
52. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3395–3402 (2015). https://doi.org/10.1109/CVPR.2015.7298961
53. Wang, W., Zhou, T., Porikli, F., Crandall, D.J., Gool, L.: A survey on deep learning technique for video segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 7099–7122 (2021). https://doi.org/10.1109/TPAMI.2022.3225573
54. Wedel, A., Cremers, D., Pock, T., Bischof, H.: Structure- and motion-adaptive regularization for high accuracy optic flow. In: 2009 IEEE 12th International Con-

ference on Computer Vision, pp. 1663–1668 (2009). https://doi.org/10.1109/ICCV.2009.5459375
55. Wulff, J., Sevilla-Lara, L., Black, M.J.: Optical flow in mostly rigid scenes. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6911–6920 (2017). https://doi.org/10.1109/CVPR.2017.731
56. Yamaguchi, K., McAllester, D., Urtasun, R.: Robust monocular epipolar flow estimation. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1862–1869 (2013). https://doi.org/10.1109/CVPR.2013.243
57. Yang, S., Zhang, L., Qi, J., Lu, H., Wang, S., Zhang, X.: Learning motion-appearance co-attention for zero-shot video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1564–1573 (2021)
58. Yang, Z., Wang, Q., Bertinetto, L., Hu, W., Bai, S., Torr, P.H.: Anchor diffusion for unsupervised video object segmentation. In: Proceedings of the IEEE/CVF International Conference On Computer Vision, pp. 931–940 (2019)
59. Ye, V., Li, Z., Tucker, R., Kanazawa, A., Snavely, N.: Deformable sprites for unsupervised video decomposition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2647–2656 (2022). https://doi.org/10.1109/CVPR52688.2022.00268
60. Ye, V., Li, Z., Tucker, R., Kanazawa, A., Snavely, N.: Deformable sprites for unsupervised video decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2657–2666 (2022)
61. Yin, Y., Xu, D., Wang, X., Zhang, L.: Agunet: annotation-guided u-net for fast one-shot video object segmentation. Pattern Recogn. **110**, 107580 (2021)
62. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1983–1992. IEEE Computer Society, Los Alamitos, CA, USA (2018). https://doi.org/10.1109/CVPR.2018.00212, https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00212
63. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 628–635 (2013). https://doi.org/10.1109/CVPR.2013.87
64. Zhang, K., Zhao, Z., Liu, D., Liu, Q., Liu, B.: Deep transport network for unsupervised video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8781–8790 (2021)
65. Zhang, X., Mao, Z., Chimitt, N., Chan, S.H.: Imaging through the atmosphere using turbulence mitigation transformer. IEEE Trans. Comput. Imaging **10**, 115–128 (2024). https://doi.org/10.1109/tci.2024.3354421, http://dx.doi.org/10.1109/TCI.2024.3354421
66. Zheng, Z., Zhong, Y., Wang, J., Ma, A.: Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4095–4104 (2020). https://doi.org/10.1109/CVPR42600.2020.00415
67. Zhong, Y., Ji, P., Wang, J., Dai, Y., Li, H.: Unsupervised deep epipolar flow for stationary or dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
68. Zhou, Y., Xu, X., Shen, F., Zhu, X., Shen, H.T.: Flow-edge guided unsupervised video object segmentation. IEEE Trans. Circuits Syst. Video Technol. **32**(12), 8116–8127 (2021)

69. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3D reconstruction. In: CVPR (2023)
70. Zou, Y., Luo, Z., Huang, J.B.: DF-Net: unsupervised Joint Learning Of Depth And Flow Using Cross-task Consistency. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 38–55. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01228-1_3, http://dx.doi.org/10.1007/978-3-030-01228-1_3
71. Çaliskan, T., Arica, N.: Atmospheric turbulence mitigation using optical flow. In: 2014 22nd International Conference on Pattern Recognition, pp. 883–888 (2014). https://doi.org/10.1109/ICPR.2014.162