Into the Fog: Evaluating Robustness of Multiple Object Tracking

Nadezda Kirillova
nadezda.kirillova@tugraz.at

M. Jehanzeb Mirza
mirza@tugraz.at

Horst Bischof
horst.bischof@tugraz.at

Horst Possegger
possegger@tugraz.at

Institute of Computer Graphics and Vision
Graz University of Technology
Graz, Austria

Abstract

State-of-the-art Multiple Object Tracking (MOT) approaches have shown remarkable performance when trained and evaluated on current benchmarks. However, these benchmarks primarily consist of clear weather scenarios, overlooking adverse atmospheric conditions such as fog, haze, smoke and dust. As a result, the robustness of trackers against these challenging conditions remains underexplored. To address this gap, we introduce physic-based volumetric fog simulation method for arbitrary MOT datasets, utilizing frame-by-frame monocular depth estimation and a fog formation optical model. We enhance our simulation by rendering both homogeneous and heterogeneous fog and propose to use the dark channel prior method to estimate atmospheric light, showing promising results even in night and indoor scenes. We present the leading benchmark MOTChallenge (third release) augmented with fog (smoke for indoor scenes) of various intensities and conduct a comprehensive evaluation of MOT methods, revealing their limitations under fog and fog-like challenges.

1 Introduction

Multiple object tracking (MOT) plays a core role in providing situation awareness for intelligent video surveillance systems, smart city technologies and autonomous driving. Driven by advancements in computer vision and deep learning, MOT methods have achieved remarkable results when trained and evaluated on current benchmarks [2, 9, 13, 20, 51]. However, these benchmarks primarily consist of clear weather scenarios, overlooking performance degradation under adverse atmospheric conditions, including fog and fog-like phenomena such as haze, dust, and smoke. To the best of our knowledge, no comprehensive work is available that extensively analyzes the robustness of MOT methods to such challenges. Moreover, only a limited number of appropriate datasets exist [18, 43, 58], which are, however, still insufficient and do not cover the full range of fog intensity.

Figure 1: Visualization of MOT performance degradation with increasing fog intensity (from left to right), demonstrated on fog-augmented video sequences with different scenarios. Top row: a daytime front-view static camera scene. Bottom row: a heavily illuminated night crowd surveillance scene. (Best viewed on screen.)

Collecting and annotating new data is time-consuming and labor-intensive, especially under adverse atmospheric conditions. Moreover, data protection laws in many countries may raise privacy concerns due to potential collection of sensitive data. To overcome these challenges, we utilize existing clear MOT datasets and generate photorealistic fog (smoke for indoor scenes) of various intensity levels. The approach of simulating adverse conditions in real-world clear images has been successfully demonstrated in various computer vision tasks such as image classification [8, 24, 25, 28], object detection [43, 59] and semantic segmentation [50]. However, its application to the MOT domain has remained unexplored.

Our work presents a novel extension of volumetric fog rendering to the MOT task, which, unlike the previously mentioned tasks, operates on videos rather than single images. In comparison with previous methods [43, 50, 58], which apply volumetric fog rendering, we do not rely solely on autonomous driving (AD) datasets, where accurate depth information is provided by range sensors, such as LiDAR, stereo camera or ToF. We introduce a novel twist by leveraging monocular depth estimation. This adaptation allows us to extend the feasibility of volumetric fog simulation beyond the scope of AD, opening up new possibilities across diverse scenarios. Additionally, we enhance our simulation by rendering of both homogeneous and heterogeneous fog effects to achieve more diversity in the representation of this complex natural phenomenon. Unlike previous studies [43, 50, 58], our approach is not restricted to images where the sky is visible. This flexibility stems from leveraging the dark channel prior method [22] for atmospheric light estimation and subsequent fog color determination, yielding promising results even in night and indoor scenes. Finally, we augment with fog (smoke for indoor scenes) of various intensities the leading tracking benchmark MOTChallenge (third release: MOT17 dataset) [13], which represents the combination of datasets from different domains and known for its diversity and wide coverage of various real-world scenarios. We conduct a comprehensive evaluation of different state-of-the-art (SOTA) MOT methods, quantitatively analyzing the impact of fog and fog-like phenomena on MOT robustness, thereby revealing their limitations, as illustrated in Fig. 1.

2   Related Work

Relevant work to our research involves studies on MOT robustness, availability of appropriate benchmarks, semi-synthetic simulations in real-world images and methods for extracting depth information from a 2D scene.

2.1 MOT Robustness

MOT is a complex computer vision task that includes various steps, such as object detection, motion estimation, instance re-identification (re-ID) and data association. MOT deals with video sequences rather than separate images, incorporating the temporal information. Frame-by-frame, it processes visual data from the surrounding environment, detects and tracks targets over time, forming their trajectories. Abundant SOTA methods [2, 3, 6, 63, 69, 70, 71] suggest different approaches to address various MOT challenges such as occlusion, viewpoint variation, pose modification, appearance changes, scale diversity, crowded scenes, lighting conditions, real-time processing, and model complexity. However, research on MOT robustness in adverse atmospheric conditions is still lacking. Some early works [22, 30, 54, 55] propose general dehazing techniques to address issues caused by fog or haze by restoring clear scenes. However, image restoration approaches face limitations in model bias and generalization ability, loss of image details or textures, and misinterpretation of scene ambiguity. The problem of developing MOT methods capable to handle adverse atmospheric conditions, including fog, remains relevant and underexplored.

2.2 Adverse Weather Datasets

Availability of datasets with adverse weather conditions is crucial for improving MOT model robustness. While several works release such datasets (see Tab. 1), especially in the AD domain, they still face limitations and need more diversity. The real-world BDD100K [67] dataset, recorded from car dashcams, includes video footage of adverse weather conditions but suffers from issues, such as camera shaking, poor image quality, and does not contain foggy videos. The DAWN [29] dataset, composed from web sources, offers diverse images of fog, rain, snow and sand, but lacks video sequences necessary for tracking purposes. Among datasets for autonomous driving (AD), such as KITTI [20], NuScenes [9], Waymo Open [51], Cityscapes [11], Oxford RobotCar [38], CADC [45] and DENSE [7], only the last one has foggy images. While all of them provide range data sufficient for fog rendering, only the first three offer MOT annotations, and they all are constrained to specific camera positions typical for the AD domain. Fully synthetic FRIDA [56] and FRIDA2 [57] datasets, which provide foggy scenes, are also captured from the driver's perspective. They are designed to address dehazing problems and do not contain videos, which makes them unsuitable for MOT tasks. CARLA [15] is a constantly growing open-source simulator for AD research that offers a realistic synthetic environment for testing and developing AD models across various scenarios, including foggy and rainy conditions. However, the fully synthetic Omni-MOT dataset [53], which was derived from CARLA and created for tracking purposes, does not include foggy scenarios. The MOTSynth [18] dataset, generated from the Grand Theft Auto V (GTA-V) computer game, aims to address tracking problems in various scenarios but lacks variation in fog intensity and scene diversity. Moreover, MOTSynth is fully synthetic, that only widens the gap with reality compared to our semi-synthetic approach.

The MOTChallenge [13] benchmark is a leading platform for evaluating MOT methods and video scene understanding. Its third release, the MOT17 dataset, covers a wide range of data diversity across different domains, including scenes with moving and static cameras, captured indoors and outdoors from different camera positions, as well as scenarios from surveillance and autonomous driving, and spanning from day to night. In this work, we demonstrate our method for photorealistic fog simulation using the MOTChallenge benchmark, as it provides the most general combination of different MOT dataset types.

2.3   Semi-synthetic Simulations in Real-world Images

Efforts to create semi-synthetic images based on real ones have been undertaken across various domains and shown promising results in improving the robustness of deep learning models. In the image classification task, some works [8, 24, 25] propose common image corruption techniques for IMAGENET [14] and CIFAR [32], paving the way to enhance model generalization. However, they involve non-realistic planar fog simulations, diverging from real-world scenarios. To overcome these limitations, Kar *et al.* [28] propose a monocular depth estimation approach to integrate scene geometry for more realistic simulations, demonstrating its effectiveness. However, these corruptions are constrained to homogeneous volumetric fog (smoke) in indoor single images captured from a front-view camera position.

Sakaridis *et al.* [50] explore the fog impact on semantic segmentation performance in the AD domain. They generate volumetric fog in real-world single images and collect a small foggy dataset to validate the approach. However, their study is limited to the Cityscapes [11] dataset and relies on provided stereo images along with corresponding disparity maps. Moreover, their fog simulation is only feasible in images containing the sky and is restricted by homogeneity. Similarly, some other AD works with available range data [43, 59] encounter such limitations, focusing on fog simulation in KITTI [20] and NuScenes [9] datasets with the aim to enhance the object detection task. Despite the promising results of employing synthetic fog in these contexts, there are currently no physics-based volumetric fog simulations available to address MOT requirements with arbitrary scenarios and camera viewpoints.

2.4   Monocular Depth Estimation

Photorealistic volumetric fog simulation requires 3D scene information and relies on image depth maps. Accurate monocular depth estimation (MDE) remains an ongoing research challenge due to the inherent uncertainty in reconstructing the 3D structure of a scene from a single 2D image. MDE faces various challenges, including scale ambiguity and generalization ability across diverse domains. Moreover, processing videos on a per-frame basis results in temporal inconsistency for dynamic scenes, leading to fog flickering over time [21, 45]. Recent MDE methods [6, 46, 47, 55] use vision transformers [60], which, compared to convolutional neural networks, exhibit higher accuracy. One line of these works [6, 55] focuses on *metric depth* estimation directly from a model. However, they are predominantly fine-tuned on specific narrow datasets, resulting in overfitting and encountering generalization limitations. To facilitate effective training across multiple datasets and enhance generalization across different domains, *relative depth* estimation approaches [53, 46, 47] are employed. Such methods predict pixel-wise disparity without providing metric values and allow us to utilize MDE on arbitrary MOT sequences, where fine-tuning is not possible.

3   Volumetric Fog Rendering

Depending on various factors such as humidity, temperature and air movement, natural fog varies in density and thickness, significantly affecting the visibility of distant objects [39, 41, 42]. Accurately depicting such a complex phenomenon is a challenging task that must consider many factors. This section describes our framework for physic-based volumetric fog rendering applicable to an arbitrary MOT dataset.

	Dataset	Year	Weather	Task	Camera View	3D Info	Augmentation
Real-world data	KITTI [■]	2014	clear	D2D, MOT2D, DE, LD, OF, SG, SLAM, ST	AD	depth map, range data	fog [■], rain [■]
	MOTChallenge [■, ■]	2015-20	clear	MOT2D, VSU	diverse	-	-
	Cityscapes [■]	2016	clear	D3D, SG	driver	depth map from stereo	fog [■, ■], rain [■]
	BDD100K [■]	2017	clear, rain, snow, fog*	D2D, MOT2D, LD, MOTS, SG	dashcam	-	-
	Oxford RobotCar [■]	2017	clear, rain, snow	SLAM	AD	range data	-
	NuScenes [■]	2019	clear, rain	D3D, MOT3D, SLAM, MP	AD	range data	fog [■], rain [■]
	Waymo Open [■]	2019	clear, rain	D2D, MOT2D, D3D, MOT3D, DA, HPE, SG, SLAM, MP	AD	range data	-
	CADC [■]	2020	snow (variety)	D3D, SLAM	AD	range data	-
	DAWN [■]	2020	fog, rain, sand, snow	D2D	web	-	-
	DENSE [■]	2020	clear, rain, snow, fog chamber	D2D, D3D, DA, DE, MSF	AD	depth map, range data	-
Synthetic data	FRIDA [■] FRIDA2 [■]	2010 2012	clear, fog	dehazing	driver	depth map, calibration	
	CARLA [■]	2017	clear, fog, rain	driving simulator	virtual AD multi-cam.	depth map, range data	
	Omni-MOT [■]	2020	clear, rain	MOT2D	CARLA	velocity, calibration	
	MOTSynth [■]	2021	clear, fog, rain snow, thunder	MOT2D, MOTS, VSU	GTA-V	depth map	

Table 1: Overview of datasets providing either adverse weather, its simulation, or MOT annotations, with the following task abbreviations: *D2D* (*D3D*): 2D (3D) object detection; *DA*: domain adaptation; *DE*: depth estimation; *HPE*: human pose estimation; *LD*: lane detection; *MOT2D* (*MOT3D*): 2D (3D) MOT; *MOTS*: MOT and segmentation; *MP*: motion prediction; *MSF*: multi-sensor fusion; *OF*: optical flow; *SG*: segmentation; *SLAM*: simultaneous localization and mapping; *ST*: stereo evaluation; *VSU*: video scene understanding.

*only one foggy sequence is available but was incorrectly annotated (is actually *clear*).

3.1 Depth Maps

Since we aim to render fog in monocular videos without any range data provided, we employ MDE method on a per-frame basis. We utilize the top-performing MiDaS 3.0 [■] method with a powerful vision transformer [■] backbone DPT BEiT$_{512}$-Large [■, ■]. This model has robust generalization capabilities across diverse domains due to its zero-shot cross-dataset transfer ability and two mixing training strategies. MiDaS achieves strong results in single-image *relative inverse depth* $\mathbf{d}(\mathbf{x})$ estimation (at pixel $\mathbf{x} = (x, y)$), with accuracy up to scale s and shift t. If 3D ground truth reference points of the captured scene are available, the actual *metric depth* $\mathbf{D}(\mathbf{x})$ can be obtained. (Details are provided in the supplementary material.) This allows us to align fog intensity simulations with meteorological *visibility* measured in meters. We denote fog severity levels ranging from 1 to $k = 4$, where we apply visibility thresholds of less than $100\ m$, $50\ m$, $20\ m$ and $10\ m$ for outdoor scenes, and $20\ m$, $10\ m$, $5\ m$ and $3\ m$ for indoor scenes, respectively. Without 3D reference points, approximate metric depth conversion still enables effective fog simulation. However, the fog's severity may not precisely correspond to the actual visibility.

3.2 Fog Formation Optical Model

To understand the physics of vision in the atmosphere, we draw on the work of Narasimhan and Nayar [■], which consolidates fundamental research in this field, including contribu-

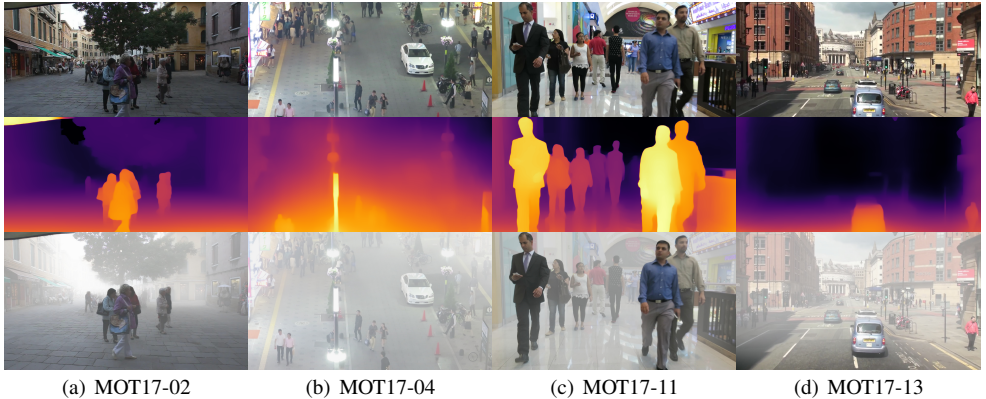(a) MOT17-02   (b) MOT17-04   (c) MOT17-11   (d) MOT17-13

Figure 2: Illustration of fog (smoke) rendering in scenes from different domains: (a) daytime static frontal view; (b) nighttime surveillance view of a heavily illuminated, crowded square; (c) indoor crowd scene captured by a forward-moving camera; (d) daytime road view from a shaking, moving double-decker bus. From top to bottom: clear images, depth maps, foggy images.

tions from Middleton [41] and McCartney [39]. We also utilize Koschmieder's [31] mathematical model of haze and fog formation:

$$\mathbf{I}(\mathbf{x}) = \mathbf{I_0}(\mathbf{x}) \cdot \mathbf{T}(\mathbf{x}) + L_\infty(1 - \mathbf{T}(\mathbf{x})), \qquad \mathbf{T}(\mathbf{x}) = e^{-\beta \mathbf{D}(\mathbf{x})}, \qquad \beta = \frac{-ln(0.05)}{V}, \quad (1)$$

where $\mathbf{I}(\mathbf{x})$ is the foggy image, $\mathbf{I_0}(\mathbf{x})$ is the clear image, $\mathbf{T}(\mathbf{x})$ is the *transmission map*, which defines the portion of object radiance reaching the camera after scattering in the atmosphere, and L_∞ is the *atmospheric light at the horizon*. The *attenuation coefficient* β is inversely proportional to the *visibility V* and regulates the fog intensity. The human eye contrast threshold 0.05 is proposed by Horvath [26]. This model is commonly used in computer vision, computer graphics and dehazing approaches [21, 22, 28, 34, 50, 54, 56]. Examples of the β calculation can be found in the supplementary material.

3.3 Atmospheric Light

Natural fog appearance varies significantly depending on lighting conditions, time of day, atmospheric pollutants, scene composition, and camera settings. These variations lead to a wide range of fog hues and tones observed in photographs. In our fog formation model (see, Eq. (1)), the *atmospheric light at the horizon* L_∞ regulates the fog color. If the sky is visible, we estimate it by averaging intensities of "far away" pixels on the estimated depth map. This aligns the fog color with the sky, ensuring consistency with the overall scene illumination. However, in non-sky scenarios, such as surveillance settings where cameras point downward or in indoor scenes, the light at the horizon is not directly accessible. Moreover, highly illuminated night scenes pose significant difficulties.

Traditional methods, such as defining the image's brightest pixels as atmospheric light [54] or using the image's average intensity [28], can result in fog that is either too bright or unrealistically dark, especially in surveillance and indoor scenes (see the supplementary material). This occurs because the brightest pixels often correspond to objects like white cars or white clothing, while the average illumination in non-sky images tends to be low.

To address these issues, we employ the *dark channel prior* concept introduced by He *et al.* [22], which provides a novel prior for image pixel intensity distribution based on statistical analysis of numerous non-sky haze-free images. The dark channel map $\mathbf{I}^{\text{dark}}(\mathbf{x})$ of an image $\mathbf{I}(\mathbf{x})$ can be calculated from:

$$\mathbf{I}^{\text{dark}}(\mathbf{x}) = \min_{\mathbf{y} \in \Omega(\mathbf{x})} \left(\min_{c \in \{R,G,B\}} \mathbf{I}^c(\mathbf{y}) \right), \tag{2}$$

where $\Omega(\mathbf{x})$ is a local patch centered at pixel $\mathbf{x}$ and $\mathbf{I}^c(\mathbf{y})$ is the intensity across three color channel c for each pixel $\mathbf{y}$ of the patch. Constructing the dark channel by sliding a small window over the image and pooling each patch's brightness by its minimum value across the three color channels, we generate a new intensity distribution, the top 10% brightest pixels of which describe the atmospheric light. This method filters out excessively bright values and effectively thresholds dark pixels, resulting in a stable atmospheric light estimation and a more realistic fog appearance. Detailed comparisons and visualizations of fog appearance using different methods across various scenarios are provided in the supplementary material.

3.4 Heterogeneous Effect

In nature, air turbulence can cause random variations in fog intensity across different spatial regions, leading to the formation of irregular cloud clusters in the atmosphere. To better capture this diversity and the dynamic nature of fog, we simulate heterogeneous fog by incorporating a *turbulence texture* $\tau(\mathbf{x})$. Following [21, 58], we generate it by combining N levels of pseudo-random gradient *Perlin noise* $\mathbf{P_n}(\mathbf{x})$ [44, 51] with various frequencies and amplitudes. We then redefine the transmission map $\mathbf{T}(\mathbf{x})$ in Eq. (1) by blending the depth map $\mathbf{D}(\mathbf{x})$ with the turbulence texture $\tau(\mathbf{x})$:

$$\mathbf{T}(\mathbf{x}) = e^{-\beta\,\tau(\mathbf{x})\cdot\mathbf{D}(\mathbf{x})}, \quad \text{where} \quad \tau(\mathbf{x}) = \sum_{n=1}^{N} \frac{\mathbf{P_n}(\mathbf{x})}{2^n}. \tag{3}$$

To maintain visual consistency and avoid flickering effects throughout a video, we use the same turbulence texture for an entire sequence. Since the videos in MOT datasets are typically short (only a few seconds), we can reasonably neglect the gradual movement of fog over time, which is usually slow.

Fig. 2 demonstrates the results achieved with our fog simulation framework. For more detailed explanations and further visualizations, please refer to the supplementary material.

3.5 Validating Fog Appearance

To assess the perceptual quality of fog simulations and validate our approach, we conducted a user study where participants compared real foggy photographs, simulations from other methods, and our approach. Viewers also rated pairs of augmented videos – with homogeneous and heterogeneous fog. We recruited 48 participants with varying levels of experience in computer graphics and image processing. Given the subjectivity of human visual perception, we employed the Mean Opinion Score to compare and evaluate all methods. Our approach, when evaluated on images, was judged to be 12% more realistic for heterogeneous fog and 29% more realistic for homogeneous fog compared to the state-of-the-art methods (see the supplementary materials). Remarkably, heterogeneous fog simulation in videos was perceived as more realistic than homogeneous, contrary to the findings in image

augmentation. The dynamic nature of videos, with their capture of motion and changes over time, provides a more immersive experience, helping viewers engage more deeply with the content. This makes heterogeneous fog in videos more realistic, unlike to the uniform and regularized homogeneous fog typically seen in photographs. This finding highlights the importance of heterogeneous fog rendering in videos, which, to our knowledge, has not been done before and actually makes our method more suitable for MOT tasks. Comprehensive validation results, including both quantitative and qualitative evaluations of fog appearance, are available in the supplementary materials.

4 MOT in Foggy Environments

Understanding the impact of fog on MOT performance and its limitations is crucial for further developing MOT robustness. In this section, we provide a comprehensive evaluation of existing MOT methods on the MOT17 dataset augmented with various types of fog.

4.1 Multiple Object Trackers

We select five MOT methods representing different tracking concepts and paradigms to analyze their robustness to foggy conditions: ByteTrack [70], Tracktor++ [3], CenterTrack [71], FairMOT [69] and TransCenter [64]. These trackers show strong performance and SOTA results on established benchmarks consisting of clear atmospheric conditions.

ByteTrack [70] follows the classical tracking-by-detection paradigm, similar to Deep-Sort [63], and treats detection and tracking as separate steps. It employs a Kalman filter [27] for motion prediction and utilizes appearance features for re-ID and data association. To enhance DeepSort, the authors introduce two key improvements: Firstly, they handle all detections, regardless of their confidence scores, removing detector's thresholds and delegating detection filtering to the tracking stage. This enables the tracker to discover valuable detections, mitigating potential detector failures. Secondly, they propose an improved association method BYTE, to effectively handle object linking. Our work adopts the official ByteTrack implementation using the anchor-based YOLOX [19] detector.

Tracktor++ [3] aims to challenge the prevailing tracking-by-detection paradigm by introducing a new concept of employing a detector for motion prediction. By applying a regression head, the detector learns to propagate bounding boxes to the next frame, reducing the need for traditional motion prediction models, such as the Kalman filter, and facilitating the data association step. To address low-frame-rate videos, it applies the constant velocity assumption and utilizes re-ID with a Siamese neural network to handle long-term occlusion. Tracktor++ demonstrates strong performance regarding false negatives (FN) and identity preservation (IDF1) compared to other trackers. Our research employs the official Tracktor++ implementation, which is built on the Faster R-CNN [48] detector with Feature Pyramid Network [36] and ResNet-101 [23] backbone.

CenterTrack [71], like Tracktor++, follows the idea of jointly learning detection and displacement over time, eliminating separate motion prediction and data association models, but adopts a point-based object representation, tracking their centers. Built on the Center-Net [17] detector, CenterTrack propagates a heatmap of object keypoints across adjacent frames and predicts temporal point offsets using optical flow techniques.

FairMOT [69] belongs to a class of trackers that jointly learn features for object detection and instance re-ID. Finding a feature balance for such two distinct tasks is challenging. Detection treats objects within a class as identical, focusing to distinguish between different

classes, while re-ID needs to differentiate between instances of the same class. Simultaneous learning leads to a model bias toward object detection. FairMOT addresses this by employing two parallel heads on the ResNet34 [23] feature extractor, using the DLA [66] multi-layer feature fusion model. Simular to CenterTrack, it is based on the CenterNet detector, utilizing center heatmaps and tracking objects as points. Using a traditional Kalman filter [27] and appearance features for re-ID, FairMOT is capable to handle long-range associations and effectively address occlusion scenarios.

TransCenter [64] is a recent transformer-based MOT approach that utilizes object center heatmap representations. It employs powerful attention-based encoder and decoder, and integrates a Query Learning Network to generate tracking queries for producing object displacement vectors over time. The TransCenter's decoder jointly learns object detection and its displacements, resulting in higher performance. Built on the improved Pyramid Vision Transformer (PVT *v2*) [62] encoder, it overcomes efficiency issues present in other transformer-based trackers such as TrackFormer [40] and TransTrack[52], caused by the use of DETR [10]. Differently to the DETR encoder, PVT is a completely convolution-free transformer backbone, which significantly reduces transformer sequence lengths and therefore computational cost. Moreover, unlike ViT [16], PVT is enhanced by a feature pyramid and copes well with the dense pixel-level structure, making it ideal for center heatmap predictions and yielding superior results compared to other trackers.

4.2 Robustness Evaluations

We augment multi-domain video scenarios of the MOTChallenge benchmark (third release: MOT17 dataset) with fog or smoke (for indoor scenes) of varying intensity levels, denoted as *Fog 1*, *Fog 2*, *Fog 3*, and *Fog 4* to categorize decreasing visibility as described in Sec. 3.1. We then assess MOT robustness to fog using HOTA [37], CLEAR-MOT [4] and IDF1 [49] metrics. Each metric evaluates overall MOT performance but emphasizes different aspects. HOTA provides a balanced score for both detection and association, while traditional MOTA tends to prioritize detection accuracy over association. IDF1 focuses on the identity assignments, i.e. association. Since the MOTChallenge benchmark primarily focuses on pedestrian tracking, our analysis is reported for this particular object class. We present the quantitative evaluation results in Tab. 2.

Despite demonstrating strong performance under clear atmospheric conditions, all trackers exhibit a significant drop in foggy scenarios. Considering homogeneous fog, ByteTrack shows the highest robustness with slow performance degradation from *Clear* to *Fog 4* level, showing $-0.4\%, -5\%, -26\%, -65\%$ HOTA drops and $-0.7\%, -5\%, -38\%, -85\%$ MOTA drops. These results may be attributed to its strategy of handling even uncertain detections regardless of their confidence score, allowing the tracker to perform well even in low visibility conditions, thereby correcting potential detector failures. Despite having a slightly larger percentage decrease compared to the classical ByteTrack, the transformer-based TransCenter achieves the best tracking performance until *Fog 2*. FairMOT exhibits better robustness than CenterTrack, despite having a similar detector, possibly due to its re-ID branch and separate motion prediction and association models. Although Tracktor++ and CenterTrack utilize the same tracking paradigm, the robustness of CenterTrack is better compared to Tracktor++, possibly because of its point-based object representation instead of an anchor-based one. Tracktor++ shows the weakest performance with significant drops already at the beginning, degrading by $-10\%, -38\%, -71\%, -86\%$ (HOTA) and $-15\%, -49\%, -87\%, -96\%$

Method		Setup	HOTA↑(%)	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	IDsw↓	Setup	HOTA↑(%)	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	IDsw↓
	ByteTrack	Clear	78.53	76.14	84.78	78.95	389	Clear	78.53	76.14	84.78	78.95	389
		Fog 1	78.20	75.64	84.64	77.70	363	Fog 1	78.42	76.06	84.41	78.40	371
		Fog 2	74.74	72.65	84.72	74.78	324	Fog 2	76.68	74.29	84.73	76.45	333
		Fog 3	**58.15**	**47.25**	84.14	**57.21**	205	Fog 3	69.35	61.98	84.77	69.69	261
		Fog 4	27.54	11.41	83.04	19.16	79	Fog 4	**38.02**	**19.14**	**85.11**	**31.06**	127
	Tracktor++	Clear	65.64	61.77	**88.81**	67.82	**217**	Clear	65.64	61.77	**88.81**	67.82	**217**
		Fog 1	56.19	52.28	86.44	55.43	287	Fog 1	59.37	55.89	86.86	59.45	281
		Fog 2	40.66	31.69	84.08	36.96	270	Fog 2	48.38	43.35	84.76	46.64	248
		Fog 3	18.79	8.19	79.76	12.05	33	Fog 3	26.34	13.57	82.47	19.68	58
		Fog 4	9.24	2.45	74.51	3.88	4	Fog 4	10.10	2.39	76.54	3.74	16
	CenterTrack	Clear	60.39	64.74	81.10	53.91	1849	Clear	60.39	64.74	81.10	53.91	1849
		Fog 1	58.20	60.78	80.97	52.03	1826	Fog 1	59.32	62.35	80.98	53.04	1799
		Fog 2	48.33	43.89	81.61	43.18	1736	Fog 2	51.46	50.6	80.76	45.47	1960
		Fog 3	29.39	17.57	**84.23**	23.01	548	Fog 3	37.34	26.27	83.82	30.78	888
		Fog 4	11.51	3.68	82.99	5.68	187	Fog 4	17.21	9.33	83.44	11.01	414
	FairMOT	Clear	75.40	72.96	83.25	78.36	299	Clear	75.40	72.96	83.25	78.36	299
		Fog 1	73.38	69.13	83.40	75.35	264	Fog 1	75.09	70.26	83.43	77.69	233
		Fog 2	66.45	59.94	83.37	67.85	212	Fog 2	69.10	63.69	83.46	70.67	231
		Fog 3	39.18	24.88	81.65	35.42	148	Fog 3	53.18	44.06	82.15	53.00	205
		Fog 4	13.05	4.10	**83.71**	7.43	18	Fog 4	24.99	11.65	82.03	19.11	40
	TransCenter	Clear	**84.66**	**85.53**	87.19	**83.86**	280	Clear	**84.66**	**85.53**	87.19	**83.86**	280
		Fog 1	**83.09**	**80.50**	**86.90**	**82.87**	274	Fog 1	**83.89**	**81.93**	**86.97**	**83.21**	276
		Fog 2	**78.68**	**74.62**	**85.28**	**78.42**	257	Fog 2	**80.47**	**77.32**	**86.14**	**80.31**	258
		Fog 3	50.87	34.26	82.32	46.61	157	Fog 3	**67.90**	**57.42**	83.44	**66.93**	223
		Fog 4	26.66	6.26	77.96	12.42	120	Fog 4	35.24	16.32	81.34	27.01	137

(Left block: Homogeneous Fog; Right block: Heterogeneous Fog)

Table 2: MOT evaluation on the augmented MOT17 dataset with homogeneous (left) and heterogeneous (right) fog. Increasing fog intensity denoted from *Fog 1* to *Fog 4*. The best scores across all trackers in the same fog level are highlighted in bold.

(MOTA), respectively. Detailed degradation scores for all trackers can be found in the supplementary material.

MOTA scores degrade worse than HOTA, indicating the crucial role of detection for robust tracking. MOTP scores, which evaluate the localization precision, consistently yield high results unaffected by fog. The IDsw metric, indicating the number of identity switches during tracking, correlates with the total number of detected instances and decreases accordingly across all trackers due to the increased number of missed tracks.

MOT approaches perform slightly better under heterogeneous fog, likely due to transparent regions between clouds of thicker fog. A detailed analysis is provided in the supplementary material. Nevertheless, while some trackers perform better in reduced visibility than others, a huge gap persists compared to their clear performance. Fig. 1 already illustrated that despite the human eye's ability to track objects in extremely severe fog, current MOT methods fail. A key takeaway is the importance of selecting the right tracking architecture to cope with fog. As observed, a successful strategy involves mitigating detector failures by processing even uncertain detections while improving an association model. Choosing a pixel-wise object center heatmap representation over anchor-based also yields better results. Incorporating the prediction of temporal object displacement into detectors instead of using a separate model is still under-explored in terms of MOT robustness, while transition from convolution-based architectures to vision transformers drastically enhances MOT.

5 Conclusion

By introducing a photorealistic fog augmentation method for arbitrary MOT datasets, we bridge the data gap for further enhancement of MOT robustness. Our evaluations shed light on the limitations of various MOT paradigms and architectures when faced with foggy conditions and set the foundation for developing MOT methods capable to handle adverse atmospheric conditions. Our augmentation method can be applied on any tracking dataset, enabling our community to investigate and improve the robustness of MOT approaches.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. In *ICLR*, 2022.

[2] Bashar, Samia Islam, Kashifa K. Hussain, Bakhtiar Hasan, Ashikur Rahman, and Hasanul Kabir. Multiple Object Tracking in Recent Times: A Literature Review. *arXiv preprint: 2209.04796*, 2022.

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *ICCV*, 2019.

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT metrics. *EURASIP JIVP*, 2008.

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In *ICIP*, 2016.

[6] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. *arXiv preprint: 2302.12288*, 2023.

[7] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather. In *CVPR*, 2020.

[8] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 2020.

[9] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. DETR: End-to-End Object Detection With Transformers. In *ECCV*, 2020.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 2016.

[12] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint:2003.09003*, 2020.

[13] Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. *IJCV*, 2021.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[15] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *Conference on Robot Learning (CoRL)*, 2017.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

[17] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. CenterNet: Keypoint Triplets for Object Detection. In *ICCV*, 2019.

[18] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri1, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara1, Laura Leal-Taixé, and Rita Cucchiara. MOT-Synth: How Can Synthetic Data Help Pedestrian Detection and Tracking? In *ICCV*, 2021.

[19] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO Series in 2021. *arXiv preprint: 2107.08430*, 2021.

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012.

[21] Fan Guo, Jin Tang, and Xiaoming Xiao. Foggy Scene Rendering Based on Transmission Map Estimation. *International Journal of Computer Games Technology*, 2014.

[22] Kaiming He, Jian Sun, and Xiaoou Tang. Single Image Haze Removal Using Dark Channel Prior. *IEEE TPAMI*, 2011.

[23] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2015.

[24] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *ICLR*, 2019.

[25] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *ICLR*, 2020.

[26] Helmuth Horvath. On the applicability of the koschmieder visibility formula. *Atmospheric Environment*, 1971.

[27] Rudolph Emil Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Fluids Engineering*, 1960.

[28] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3D Common Corruptions and Data Augmentation. In *CVPR*, 2022. Oral.

[29] Mourad A. Kenk and Mahmoud Hassaballah. DAWN: Vehicle Detection in Adverse Weather Nature, 2020. https://data.mendeley.com/datasets/766ygrbt8y/3.

[30] Louis Kratz Ko Nishino and Stephen Lombardi. Bayesian Defogging. *IJCV*, 2011.

[31] Harald Koschmieder. *Theorie der horizontalen Sichtweite*. Beiträge zur Physik der freien Atmosphäre. Keim & Nemnich, 1924.

[32] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. https://www.cs.toronto.edu/ kriz/cifar.html.

[33] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *CVPR*, 2019.

[34] Zhongping Lee and Shaling Shang. Visibility: How Applicable is the Century-Old Koschmieder Model? *Journal of the Atmospheric Sciences*, 2016.

[35] Zhaoshuo Li, Wei Ye, Dilin Wang, Francis X. Creighton, Russell H. Taylor, Ganesh Venkatesh, and Mathias Unberath. Temporally Consistent Online Depth Estimation in Dynamic Scenes. In *CVPR*, 2023.

[36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *CVPR*, 2017.

[37] Jonathon Luiten, Aljoša Ošep, Simone Calderara1, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A Higher Order Metric for Evaluating Multi-object Tracking. *IJCV*, 2020.

[38] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *IJRR*, 2017.

[39] Earl J McCartney. *Optics of the atmosphere: Scattering by molecules and particles*. Wiley series in pure and applied optics, 1976.

[40] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. TrackFormer: Multi-object Tracking with Transformers. In *CVPR*, 2022.

[41] Knowles W. E. Middleton. *Visibility in Meteorology*. American Meteorological Society, 1951.

[42] Srinivasa G. Narasimhan and Shree K. Nayar. Vision and the Atmosphere. *IJCV*, 2002.

[43] Dipkumar Patel. Nuscenes fog augmented samples for bad weather pre. https://www.kaggle.com/datasets/dipkumar/nuscenes-fog-augmented-samples-for-bad-weather-pre.

[44] Ken Perlin. An image synthesizer. In *ACM SIGGRAPH*, 1985.

[45] Matthew Pitropov, Danson Garcia, Jason Rebello, Michael Smart, Carlos Wang, Krzysztof Czarnecki, and Steven Waslander. Canadian Adverse Driving Conditions dataset. *IJRR*, 2020.

[46] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *ICCV*, 2021.

[47] 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 TPAMI*, 2022.

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015.

[49] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Dataset for Multi-Target, Multi-Camera Tracking. In *ECCVW*, 2016.

[50] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic Foggy Scene Understanding with Synthetic Data. *IJCV*, 2018.

[51] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *CVPR*, 2020. https://waymo.com/open.

[52] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. TransTrack: Multiple Object Tracking with Transformer. *arXiv preprint: 2012.15460*, 2020.

[53] ShiJie Sun, Naveed Akhtar, XiangYu Song, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Simultaneous Detection and Tracking with Motion Modelling for Multiple Object Tracking. In *ECCV*, 2020.

[54] Robby T. Tan. Visibility in bad weather from a single image. In *CVPR*, 2008.

[55] Jean-Philippe Tarel and Nicolas Hautière. Fast visibility restoration from a single color or gray level image. In *ICCV*, 2009.

[56] Jean-Philippe Tarel, Nicolas Hautière, Aurélien Cord, Dominique Gruyer, and Houssam Halmaoui. Improved visibility of road scene images under heterogeneous fog. In *Intelligent Vehicles Symposium (IVS)*, 2010.

[57] Jean-Philippe Tarel, Nicholas Hautiere, Laurent Caraffa, Aurélien Cord, Houssam Halmaoui, and Dominique Gruyer. Vision Enhancement in Homogeneous and Heterogeneous Fog. *IEEE Intelligent Transportation Systems Magazine*, 2012. http://perso.lcpc.fr/tarel.jean-philippe/bdd/frida.html.

[58] Maxime Tremblay, Shirsendu Sukanta Halder, Raoul de Charette, and Jean-François Lalonde. Weather, rainy and foggy, augmented datasets. https://team.inria.fr/rits/computer-vision/weather-augment.

[59] Maxime Tremblay, Shirsendu Sukanta Halder, Raoul de Charette, and Jean-François Lalonde. Rain Rendering for Evaluating and Improving Robustness to Bad Weather. *IJCV*, 2021.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[61] Pierre Vigier. Perlin noise with numpy. https://github.com/pvigier/perlin-numpy?tab=readme-ov-file.

[62] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with Pyramid Vision Transformer. *Computational Visual Media*, 2022.

[63] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple Online and Realtime Tracking with a Deep Association Metric. In *ICIP*, 2017.

[64] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. TransCenter: Transformers With Dense Representations for Multiple-Object Tracking. *IEEE TPAMI*, 2023.

[65] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *CVPR*, 2024.

[66] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep Layer Aggregation. In *CVPR*, 2018.

[67] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *CVPR*, 2020.

[68] Dorota Zdrojewska. Real time rendering of heterogenous fog based on the graphics hardware acceleration. In *CESCG*, 2004.

[69] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. *IJCV*, 2021.

[70] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In *ECCV*, 2022.

[71] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020.