Evaluation of Spatio-Temporal Small Object Detection in Real-World Adverse Weather Conditions

Michel van Lier Martin van Leeuwen Bastian van Manen Leo Kampmeijer
and Nicolas Boehrer
TNO Intelligent Imaging, The Netherlands
{michel.vanlier, martin.vanleeuwen, bastian.vanmanen, leo.kampmeijer, nicolas.boehrer}@tno.nl

Abstract

Deep learning-based object detection methods, such as YOLO, are promising for surveillance applications. However, detecting small objects in large-scale scenes with cluttered backgrounds and adverse weather remains challenging. Recent advancements leverage spatio-temporal information to enhance small object detection, yet the impact of (temporal) adverse weather conditions on such methods remains largely unexplored due to the lack of comprehensive evaluation datasets. This paper evaluates the performance of spatio-temporal YOLOv8 (TYOLOv8) for detecting small objects in real-world adverse weather conditions, comparing it to spatial YOLOv8 and the 3FN moving object detection method. Additionally, we propose haze augmentation to improve object detection performance in challenging hazy weather. Due to the lack of suitable datasets for evaluation, this paper introduces a novel real-world video dataset for small object detection, referred to as Nano-VID-weather, with an average object size of 16.4^2 *pixels, consisting of a Tiny Objects subset and three challenging weather subsets: Wind, Rain and Haze. Our findings reveal that TYOLOv8 is resilient to real-world adversarial weather conditions, like wind, rain, and haze. Notably, on average TYOLOv8 outperformed both 3FN and YOLOv8 with* $+0.21mAP$ *across all our subsets. These results demonstrate that TYOLOv8 can enhance surveillance capabilities for small object detection under real-world adverse weather conditions.*

(a) Tiny Objects subset.

(b) Rain subset.

(c) Haze subset.

Figure 1. Impression of challenges in Nano-VID-weather dataset including, detected objects from our spatio-temporal TYOLOv8.

1. Introduction

Object detection is a critical component in modern surveillance systems, with deep learning CNN-based object detection methods such as the YOLO-family [34], [44], [72], [35] and transformer-based methods like DETR-family [10], [89], [8], [88], showing remarkable success across a range of surveillance applications. However, typical object detection model are not optimized for small object detection. Large-area monitoring with a single camera or long-range object detection for early warnings can not be performed reliably with these methods.

In small object detection, the reduced performance of spatial object detectors, such as YOLO and DETR, is often attributed to limited pixel representation and a lack of

distinctive features of small objects. Spatio-temporal object detection methods address this challenge by leveraging both spatial and temporal information, processing multiple frames simultaneously [16, 24, 71, 73, 86].

Although these methods show promising results, it remains unclear whether their reliance on temporal information makes them more susceptible to the challenges posed by real-world (temporal) adverse weather conditions. Conditions such as wind, rain, and haze can obscure objects, distort their appearance, and create apparent movement, which may lead to higher rates of false positives or missed object detections. However, the effect of such conditions on spatio-temporal small object detectors remains unexplored in the literature.

To address this gap, a dedicated video dataset that includes challenging weather conditions and carefully annotated small (moving) objects with sizes below 20 by 20 pixels is needed. As such a dataset is unavailable, we contribute a novel evaluation dataset *Nano-VID-weather*, derived from real-world surveillance video footage. It features small objects, including pedestrians, cyclists, vehicles, and animals, with an average object size ranging from 11.98 by 11.98 up to 19.27 by 19.27 pixels. The Nano-VID-weather includes three distinct weather conditions: wind, rain, and haze, as well as a tiny objects baseline with clear weather. This enables a comprehensive evaluation of detector performance across various environmental scenarios on small objects.

Our second contribution is the evaluation of the state-of-the-art spatio-temporal object detector TYOLOv8 [71] on our proposed Nano-VID-weather dataset. We compare its object detection performance with that of YOLOv8 [35] and 3FN, a traditional frame-differencing method for moving object detection.

While the 3FN and spatial YOLOv8 methods exhibit highly unreliable performance across all adverse weather conditions, TYOLOv8 consistently demonstrates robust performance, even without weather-specific training data. This highlights the model's strong resilience to challenging real-world weather conditions and suggests that similar spatio-temporal detectors are likely to exhibit comparable robustness.

2. Related Works

2.1. Small object detection

Small object detection based on visual cues is challenging due to several factors, such as cluttered, high-frequency and noisy backgrounds, limited pixel representation of small objects, and the scarcity of distinctive features. As such, most traditional small object detection approaches rely on motion features rather than appearance features to detect small objects. The motion features are usually constructed from absolute differences between two or more frames. Subsequently, morphological operations can be applied to these difference maps to localize objects. While frame-differencing techniques apply these maps directly, background subtraction aims to improve precision in noisy environments by detecting backgrounds and employing these maps to attenuate detected motion at these locations [54, 70]. This paper will compare against a frame-differencing technique known as 3FN. This technique uses three frames to establish a noise map and a frame difference map. By relating frame differences to the detected noise, it becomes possible to reduce false positives, which are common for frame-differencing approaches.

With the advent of deep learning, numerous architectures for object detection have been introduced, such as R-CNN [26], YOLO [36, 61], DETR [9], GLIP [47], and Grounding DINO [49]. While these architectures are typically intended for larger objects (for example 20 by 20 pixels or larger), they can also be employed for smaller objects. However, various challenges specific to small object detection may limit their effectiveness [11]. For instance, they do not consider motion as a feature, and they may use IOU-based loss functions or fail to address the foreground-to-background imbalance.

Recently, some methods have been proposed to address some of these issues. For example, Centernet [18] resolves the issue with the foreground-background imbalance and IOU-loss functions by predicting the centres of bounding boxes instead of using anchor boxes. With this adaptation, strong object detection performance can be achieved for small and densely packed objects. Alternatively, TYOLOv8 has been introduced as an adaptation of the YOLOv8 model to further enhance the model's capability for detecting small moving objects [71]. Following the approach in [16], this method exchanges the RGB channels of the model's input with grey-scale data from three separate input frames to provide temporal context. Additionally, it incorporates various augmentation techniques to improve training for small objects. With these adaptations, combining the ease of use and efficiency of the YOLOv8 network with strong small object detection capabilities becomes possible. Due to these practical considerations, we will evaluate TYOLOv8 and YOLOv8 in this work and leave Centernet as future work.

2.1.1 Object detection in adverse weather

Most object detection models are optimized for general conditions, often assuming high-quality images captured in clear weather conditions. However, their object detection performance significantly degrades in adverse weather conditions such as rain, haze, or fog, which reduces visibility. Various methods have been proposed to address this, which can generally be categorized into three main strategies. The

first category of approaches involves directly training object detection models on images affected by real-world or simulated adverse weather. While this can improve object detection performance under challenging conditions, it requires generating datasets of diverse real-world adverse weather scenarios or highly realistic weather simulations, which can be both time-consuming and resource-intensive.

A second category of approaches adopts a two-stage process: first, applying image restoration algorithms, such as dehazing or deraining, followed by processing the restored images using object detection models, as demonstrated in methods like DCP [28], AOD-Net [42], GridDehazeNet [50], MSBDN [17], AECR-Net [76], DehazeFormer [66], and DH-YOLO [82]. Although these methods can improve image clarity, they often yield sub-optimal results and fail to outperform the first category. This is because restored images may lack essential latent features critical for accurate object detection [14], particularly for small objects, while the two-stage process suffers from sub-optimal optimization.

The third category focuses on jointly training object detection models with weather-specific image enhancement techniques or leveraging domain adaptation to bridge the gap between clear training images and those captured in adverse weather conditions. Examples of these approaches include D-YOLO [14], FriendNet [21], BAD-Net [45], IA-Det [46], and R-YOLO [74]. While these approaches generally outperform those in the second category, their object detection performance varies when compared to models trained directly on adverse weather images, as demonstrated by [14], [21], [45], [45], [46] and [74]. Since these methods cannot provide a reliable solution for our application, we explored spatio-temporal object detection with TYOLOv8.

2.2. Small objects datasets

Evaluation datasets provide a standardized basis for comparing different solutions. The PASCAL VOC 2007 [19], 2012 [20], and MS COCO [48] datasets have become widely adopted standards for evaluating object detectors. While these datasets are diverse and cover multiple object classes in various contexts, they have notable limitations. Specifically, they lack sufficient representation of small objects (see Table 1), lack video sequences, and do not include annotated weather conditions.

Several specialized datasets have been introduced to address the gap in evaluating the performance of small object detection. Notable examples include CityPerson [83], WiderFace [79], TinyPerson [81], XS-VID [4], ShipRSImageNet [84], Airbus Ship Detection [31], COWC [56], CARPK [29], EDAI, and DOTA [77]. These datasets feature significantly smaller object sizes. However, despite their emphasis on small objects, these datasets lack video sequences required for spatio-temporal object detection evaluation.

While MSCOCO serves as a standard benchmark for spatial object detection, ImageNetVID [63] has emerged as a widely used dataset for video object detection. It includes numerous object classes across various environments, but focuses mainly on larger objects. Video datasets, such as UA-DETRAC [75], Marine Obstacle Detection Dataset V1.0 [41], V2.0 [5], VisDrone [87], MOT15 [13], MOT17 [68], Okutama-Action [2], MOR-UAV [52], DAC-SDC [78], DroneSURF [37], and SODA-A, SODA-D [12], SeaDronesSee [38], UAV123 [55], Stanford Drone Dataset [62], include smaller objects compared to ImageNetVID, but do not sufficiently emphasize small objects. Additionally, these datasets are not captured from a stationary surveillance perspective, which limits their suitability for evaluating TYOLOv8.

VIRAT [57] and Nano-VID [71] do include a diverse set of stationary surveillance perspective video footage with small objects. However, VIRAT [57] leaves some of the smallest objects not annotated, leading to label noise. Therefore, Nano-VID is selected as training data for our experiments since it includes more complete annotations. Due to the absence of annotated weather, an additional evaluation dataset is required.

2.3. Adverse weather datasets

Numerous datasets include explicit annotations for weather conditions, such as fog or haze RADIATE [65], Foggy Driving [64], or rain in datasets like RainDS [60], WCity [85], Raindrop [59], Raindrop Clarity [33], Boxy [3], A*3D [58], aiMotive [53]. Some datasets, such as MUAD [23], RaidaR [32], MUSES [6], Waymo OD [67], nuScenes, Kitti-c [39], JAAD [40], BKD100K [80], RTTS [69] Boreas [7], and SODA10M [27], offer a combination of multiple weather conditions. Additionally, some datasets are created by simulating weather. For instance Multifog KITTI [51], FoggyCityscapes [64], RainCityscapes [30], add simulated weather to clean weather datasets like KITTI [25] or Cityscapes [15]. These datasets primarily target automotive applications, presenting scenarios and perspectives distinct from those of surveillance. They often place less emphasis on small objects and do not include stationary video footage. To address this gap, we developed our own real-world dataset with small objects called *Nano-VID-weather*, that has been captured from a stationary surveillance perspective and scenario. Nano-VID-weather includes annotated video footage under challenging adverse weather conditions, offering a unique and valuable resource for this domain and will be open sourced.

3. Methods

3.1. Dataset

Unlike the datasets discussed in Sections 2.2 and 2.3, Nano-VID-weather introduces annotated stationary video data recorded from a surveillance perspective over an extended period in a surveillance scenario. ano-VID-weather specifically targets small objects and incorporates corresponding meteorological data. Several factors contribute to the increased complexity of ano-VID-weather for small object detection. First, objects often move against backgrounds such as grass and sandy dunes, limiting the contrast between them and their surroundings. Additionally, natural vegetation frequently causes partial occlusions, while varying weather conditions introduce further challenges, including reduced visibility, altered brightness and contrast, obscured or distorted object appearances, and apparent background movement. ano-VID-weather is divided into four subsets to evaluate object detection performance under specific weather conditions. The Tiny Objects subset focuses on the smallest annotated objects under clear weather conditions. The remaining subsets include weather-specific scenarios: wind, rain, and hazy weather. Figure 3 illustrates the detailed distribution of absolute object sizes across these subsets. In contrast, Table 1 compares the average absolute and relative object sizes to those in other popular datasets. This paper defines the absolute size of an object as $\sqrt{h \cdot w}$, where h and w represent the height and width of the object, respectively. On the other hand, the relative size expresses the object's size as a proportion of the image size defined as $\sqrt{\frac{h \cdot w}{H \cdot W}}$ where H and W represent the height and width of the image. Furthermore, persons, vehicles, cyclists, birds, dogs, and horses are annotated in each frame. Since this study focuses on object detection rather than classification, all annotated objects are assigned to the same class.

3.1.1 Data collection

The three weather subsets and part of the Tiny Objects subset are collected in a typical real-world surveillance scenario from static cameras mounted at 30 meters elevation that have recorded video footage daily from March to December in 2023, covering a large area and massive foreground-background imbalance, as shown in Figure 1. This extended recording period introduces significant diversity in the dataset, encompassing variations in light intensity, background movement, changes in vegetation, contrast, and occlusions. The videos were captured using Teledyne Adimec TMX-55 cameras [1], with a native resolution of 4K (4096×2176 pixels) at 15 frames per second. To account for small enough objects, the frames were downscaled using bi-cubic interpolation to resolutions of 2K and 1K. This reduced the average pixel count per object while preserving sufficient small relative object size to mimic the large foreground-background imbalance.

Meteorological data, including rainfall (mm/h), wind speed (m/s), and temperature (°C), were obtained from a nearby weather station during the video recordings. This information was used to categorize videos into weather-specific subsets (Tiny Object, Rain, Wind, Haze) through an initial automated pre-selection process, followed by manual verification to finalize the categorization.

3.1.2 Tiny Objects subset

The Tiny Objects subset is a set of videos used to evaluate the object detection performance of spatial and spatio-temporal small object detection models in different surveillance scenes and as a baseline with "clear weather conditions". This subset differentiates itself in two main ways from the three adverse weather subsets. First, the videos selected for this subset does not contain severe rain, wind, haze or other disturbing weather phenomena. Instead, it focuses on more common weather conditions, such as sunny and lightly clouded. Second, the subset contains much more variation in scenes, brightness and contrast, viewing angles, object sizes and vegetation since additional proprietary datasets are added to evaluate the generalization capability of the small object detection models. 60.85% of the annotated objects originate from videos from the grassy dune scene as described in Section 3.1, and 39.15% are annotations extracted from two proprietary datasets. These two datasets contain images taken at ground level, looking at an open field with low bush vegetation and trees in the background. The Tiny Objects subset contains 2506 annotated objects in 1046 frames with an average absolute object size of 11.98 pixels. See Table 1.

3.2. Adverse weather subsets

The adverse weather subsets contain videos of the grassy dune scene (3.1) in either wind, rain or haze conditions.

Wind subset: The Wind subset was created to evaluate the object detection performance of spatial and spatio-temporal small object detection models in wind and contains wind speeds of over 35 km/h. This subset captures a scenario characterized by significant environmental disturbances, such as extensive moving vegetation due to high winds. The large amount of background movement is expected to be a challenge for spatio-temporal detectors. Figure 2a provides an impression of the Wind subset, illustrating the background motion through a frame-difference overlay. The Wind subset contains 366 annotated objects in 99 frames. The average absolute object size is 15.24 pixels. See Table 1.

Rain subset: The Rain subset was created to evaluate the object detection performance of spatial and spatio-

temporal small object detection models in rain, and features rainfall rates of approximately 5 mm/h. The falling raindrops in the subset cause additional movement and noise in the background while reducing the scene's visibility, brightness and contrast. However, the wind speed for this sequence was only measured at 17 km/h, so the precipitation mainly generated additional background noise. Figure 1b shows an impression, while Figure 2b highlights the additional background movement caused by the rain. The subset contains 235 annotated objects in 79 frames, with an average absolute object size of 19.20 pixels. See Table 1.

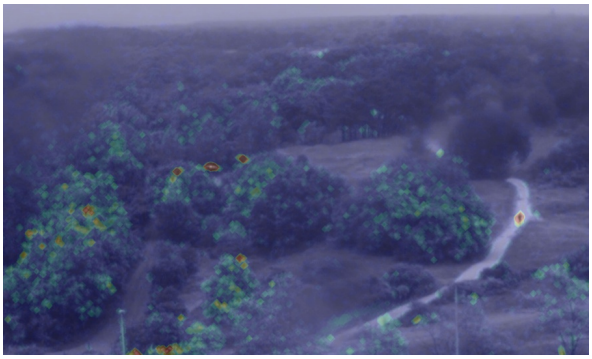
Haze subset: The Haze subset was created to evaluate the object detection performance of spatial and spatio-temporal small object detection models in moderate to heavy haze conditions. The degree of haze ranges from low overhanging haze providing partial coverage to full opaque coverage of the entire dune scene. Using the setup described in Section 3.1, 245 frames have been annotated from multiple videos during heavy haze in November and December 2023, resulting in 867 annotated objects. Figure 1c shows an impression. The average absolute object size is 19.27 pixels. See Table 1.

Datasets	Object size	
	Absolute (↓)	**Relative** (↓)
Tiny Objects	$\mathbf{11.98}_{7.37}$	$\mathbf{0.004}_{0.002}$
Wind	$15.24_{7.81}$	$0.007_{0.003}$
Rain	$19.20_{14.5}$	$0.009_{0.006}$
Haze	$19.27_{8.3}$	$0.009_{0.004}$
TinyPerson [81]	$18.0_{17.3}$	$0.012_{0.010}$
Visdrone [87]	$45.37_{44.90}$	$0.030_{0.028}$
BKD100K [80]	$51.00_{63.96}$	$0.053_{0,067}$
Virat-Ground [57]	$59.9_{45.47}$	$0.052_{0,033}$
KITTI [25]	$72.3_{62.23}$	$0.106_{0.091}$
CityPerson [83]	$79.8_{0.67}$	$0.055_{0.046}$
MSCOCO [48]	99.5_{108}	$0.109_{0.203}$

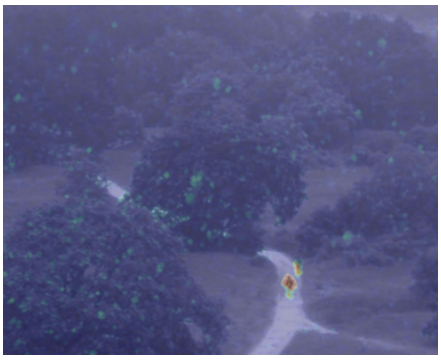
Table 1. Object sizes from Nano-VID-weather subsets compared to popular datasets. The subscripts represent standard deviations (STD) for the corresponding values.

3.3. Haze Augmentation

We propose haze augmentation to enhance the object detection performance and robustness in hazy weather. As outlined in Section 2.1.1, incorporating hazy weather images into the training process is expected to improve object detection performance solely on clear-weather images. Since a real-world training dataset meeting our conditions is absent, as discussed in Section 2.3 and capturing and annotating real-world images in all weather conditions is time-intensive, our approach uses simulated haze to overcome



(a) Cropped still of a video in our Wind subset, with a heatmap overlay, indicating the amount of motion due to wind.



(b) Cropped still of a video in our Rain subset, with a heatmap overlay, indicating the amount of motion due to rain.

Figure 2. An impression of motion in our Wind and Rain subsets using a motion overlay shows amplified movement. Both moving trees and falling raindrops might be expected to affect the object detection performance of a spatio-temporal models.

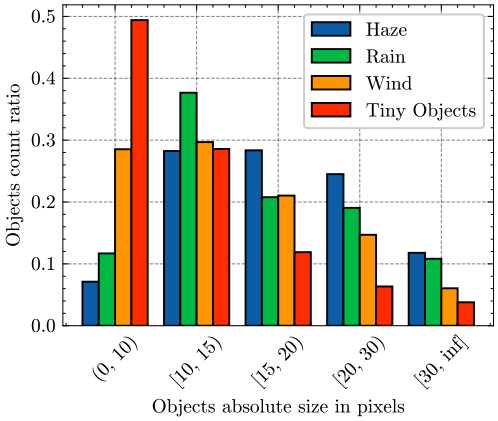


Figure 3. Distribution of object's sizes in Nano-VID-weather.

these limitations. Our haze augmentation is based on a simplified atmospheric scattering model [43], which assumes that the same amount of light that is scattered away from the camera is scattered towards the camera from all other locations. This can be expressed as follows:

$$I(x,y) = J(x,y) \cdot t(x,y) + A(1 - t(x,y)) \quad (1)$$

Where $I(x,y)$ represents the pixel's final intensity (color) at position x, y. $J(x,y)$ is the intensity (color) of the pixel in the original image scene without atmospheric effects (referred to as "non-hazy image"), to which we want to apply haze augmentation. A is a coefficient representing the global atmospheric light. The global atmospheric light A is randomly sampled within the $[150, 255]$ range to simulate various intensities of airlight. $t(x,y)$ is the transmission in the direction of pixel x, y and represents how much of the original light reaches the camera without being scattered by the haze, which decreases with increasing distances. $t(x,y)$ can be expressed as follows:

$$t(x,y) = e^{-d(x,y)\beta} \quad (2)$$

Where $d(x,y)$ is the distance to the observed scene of $J(x,y)$ for pixel x, y, and β is the scattering coefficient of the atmosphere. The scattering coefficient β is generated using the diamond-square algorithm [22] to simulate the real-world uneven, low-hanging haze. This algorithm mimics the natural variability of haze, with the "wibble decay" (WD) parameter randomly selected within the range $[1.6, 1.8]$ to control the size and density of haze clouds. Lower WD values produce smaller, thinner clouds, while higher values generate larger, denser clouds. The haze intensity is further varied by scaling β randomly within the range $[2, 8]$. Figure 1c illustrates real-world haze from Nano-VID-weather Haze subset, while Figure 4 shows examples of our simulated haze augmentation. We estimate a single depth map for the entire dataset based on known camera parameters, intrinsic properties, vertical FOV, and camera height. Assuming a flat-earth model, we estimate the scene distance for each horizontal pixel row y as follows:

$$d(y) = \frac{h_{\text{cam}}}{\cos\left(a + (IFOV \cdot y)\right)} \quad (3)$$

$$d(y) = \begin{cases} d_{\text{horizon}} & \text{if } d(y) > d_{\text{horizon}} \\ & \text{or } (a + \text{IFOV} \cdot y) < 0, \\ d(y) & \text{otherwise.} \end{cases} \quad (4)$$

Where $d(y)$ represents the depth at pixel row y, h_{cam} is the camera's height above the ground (in meters), a is the base angle, $IFOV$ is the instantaneous FOV per pixel. d_{horizon} is the distance to the horizon, calculated as follows:

$$d_{\text{horizon}} = \sqrt{2 \cdot h_{\text{cam}} \cdot 6.4 \cdot 10^6} \quad (5)$$

$6.4 \cdot 10^6$ approximates the radius of the earth in meters. The depth map is constructed so that distant objects have the same depth, creating a realistic haze effect similar to natural scenes. Varying haze conditions are simulated by randomly reducing d_{horizon}.

(a) $A = 220, \beta = 2, WD = 2$ (b) $A = 160, \beta = 1, WD = 0.5$

Figure 4. Impression of our haze augmention applied to an clean image from training set.

4. Experiments & Results

The following section outlines the setup for the evaluation method, metrics, and model training configuration. Followed by the experiments conducted to evaluate TYOLOv8 and the proposed haze augmentation on Nano-VID-weather, along with the corresponding results and discussion. For comparison, we include the frame-differencing method, 3FN, and the spatial approach (using a single frame) with YOLOv8 as baselines for these experiments. These object detectors were selected to facilitate a comparative evaluation of spatial and spatio-temporal approaches.

4.1. Evaluation & Metrics

Common metrics for evaluating object detectors include recall, precision, and mean Average Precision (mAP), typically using an Intersection over Union (IoU) threshold of at least 0.5 [20, 48]. We use various modifications to these metrics to better assess small object detection performance, as proposed in earlier work [71]. This includes scaling up bounding boxes such that each dimension of the bounding box dimension is at least 15 pixels and reducing the IOU threshold to 0.25. Furthermore, multiple object detections on the same annotation and multiple annotations using a single object detection are both considered correct. This approach enables us to measure the ability of the detectors to localize a target rather than its ability to make a perfectly fitting bounding box.

4.2. Model configuration

Our experiments use the pre-trained spatio-temporal TYOLOv8 model from [71]. Additionally, we use a spatial YOLOv8 model, which has been trained using the same hyperparameters and techniques as proposed in the same paper but takes a single RGB frame as input instead of multiple grey-scale frames. The Nano-VID dataset [71] was used as a training dataset for these models. Nano-VID focuses on typical weather, such as sunny and lightly clouded weather, and excludes severe rain, wind, or haze. A one-second interval between input frames is used for the spatio-temporal methods 3FN and TYOLOv8.

Haze augmentation: For the haze augmentation experiment, the YOLOv8 and TYOLOv8 models are fine-tuned for 20 epochs on Nano-VID, using the training techniques

proposed in [71] but includes our haze augmentations as proposed in Section 3.3. The haze augmentation is applied to non-hazy image samples from Nano-VID with a probability of 50%. The models are trained and evaluated three times with a random seed to account for random differences between training runs. The PR (precision-recall) curves for the haze experiments are based on the result from the first run, while the reported mAP scores are averaged between runs.

4.3. Results

The YOLOv8, TYOLOv8 and 3FN object detectors are evaluated on our proposed Tiny Objects, Wind, Rain and Haze subsets, to assess the impact of different (temporal) weather conditions on their small object performance.

The object detection performance is summarized in Table 2, providing the mAP@.25 score per object size range and per weather condition. Depending on the subset, TYOLOv8 achieves mAP@.25 scores between 0.74 and 0.81 on all object sizes, which outperforms YOLOv8 with an mAP@.25 score of 0.21 on average on Nano-VID-weather. Depending on the subset, YOLOv8 model yields mAP@.25 scores between 0.41 and 0.61 on all object sizes. While Nano-VID-weather is not as extensive as some established benchmarks, the substantial differences in relative performance across models enables us to draw robust conclusions.

4.3.1 Adverse weather conditions

The object detector's PR curves, trained without hazy weather and without haze augmentation, are presented in Figure 5. Here, we observe curves with similar characteristics for the TYOLOv8 model across all subsets. The optimal F1 score is around a recall of 0.75 and a precision of 0.9 for each weather condition and subset. As expected, the TYOLOv8 model achieves a strong mAP@.25 score of 0.78, and outperforms YOLOv8 with an mAP@.25 score of 0.22 on the Tiny Objects subset. It can exploit motion features without experiencing too much hinder from background noise. Additionally, TYOLOv8 also significantly outperforms the other approaches, on subsets that feature adverse weather conditions such as wind, rain on average with an mAP@.25 score of 0.22. These conditions considerably increase the background noise and reduce the visibility, as shown in Figure 1. Although the TYOLOv8 model was not trained on adverse weather data, it effectively minimizes false positives, distinguishing viable targets from background elements such as moving vegetation or rain. This ability is reflected in the model's consistent mAP scores and PR curves across adverse weather conditions.

For the YOLOv8 object detection performance, more substantial differences can be observed between subsets and

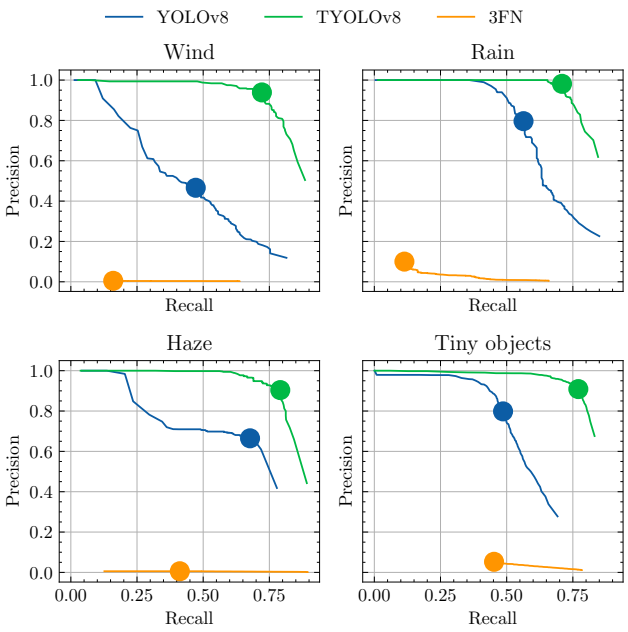


Figure 5. PR curve results on Nano-VID-weather, featuring YOLOv8 (blue), TYOLOv8 (green) and the 3FN (orange). Similar to Figure 6, the dot represents the optimal F1 score.

weather conditions. However, despite the lack of temporal information, the YOLOv8 model can still detect a portion of the annotated objects. However, without motion, the model depends on the limited visual features of small objects, which may share many similarities to features found in the background. Nano-VID-weather contain a low foreground-to-background ratio, leading to low precision. While the amount of background features that mimic the foreground may be affected by the weather, it is also influenced by many other parameters. For instance, the Wind subset appears to include more challenging background clutter, leading to increased false alarms for the YOLOv8 model. As such, deducing a relationship between YOLOv8 object detection performance and weather has not been possible.

The 3FN method achieves the lowest precision and thus fails to yield competitive mAP scores, and are therefore not reported. Even in good conditions, balancing sensitivity to objects with rejection of background noise in a large-scale background is challenging. Although most traditional approaches are intended to be paired with tracking to reduce false positives, this is undesirable for applications constrained by latency.

4.3.2 Haze augmentation

The results in Figure 6 show the PR curves for YOLOv8 models trained without and with our proposed haze augmentation. As expected, the PR curve improves when

Subsets	mAP@25 per absolute object size range (↑)					
	(0, 10)	**[10, 15)**	**[15, 20)**	**[20, 30)**	**[30, inf]**	**All**
	Y/TY	Y/TY	Y/TY	Y/TY	Y/TY	Y/TY
Tiny Objects	0.65/**0.81**	0.58/**0.82**	0.53/**0.73**	0.47/**0.76**	0.46/**0.60**	0.56/**0.78**
Wind	0.43/**0.79**	0.49/**0.88**	0.52/**0.62**	0.41/**0.78**	0.52/**0.59**	0.41/**0.74**
Rain	0.64/**0.79**	0.56/**0.79**	0.50/**0.81**	0.44/**0.80**	0.59/**0.60**	0.67/**0.78**
Haze	0.42/**0.66**	0.59/**0.78**	0.60/**0.84**	0.83/**0.89**	0.65/**0.81**	0.59/**0.81**
Haze + haze aug	0.39/**0.77**	0.67/**0.79**	0.65/**0.87**	0.80/0.80	0.69/**0.74**	0.64/**0.80**

Table 2. YOLOv8 (Y) / TYOLOv8 (TY) object detection performance in mAP on Nano-VID-weather subsets for five object size ranges. **Bold** represents the best results. **Haze + haze aug** represents the results were our proposed haze augmentation was applied during training.

haze augmentation is incorporated during training, exhibiting higher recall and precision.

The results in Table 2 reveal that the most significant improvement of the YOLOv8 models occurs in the object size range [10, 20), where haze often compromises visibility. However, there is no improvement for the smallest objects in range (0, 10), likely due to insufficient spatial information for object detection. In contrast, augmentation does not contribute significantly to object detection performance for larger objects in range [20-inf), where the haze is less dense or absent.

Similarly, Figure 6 displays the PR curves for TYOLOv8 models trained without and with our haze augmentation. While the overall object detection performance is not increasing for the TYOLOv8 model, Table 2 shows an improvement in the object range (0, 20), where dense haze typically reduces visibility. This suggests that haze augmentation can help the TYOLOv8 model better detect small objects by leveraging spatio-temporal information in dense haze.

When comparing the PR curves in (Figure 6) of YOLOv8 with the TYOLOv8, as well as their respective mAP@0.25 scores (Table 2), TYOLOv8 consistently achieve higher recall across all precision levels. It outperforms both YOLOv8 models in hazy weather (trained without and with our haze augmentation) on average with 0.19 mAP@0.25. This highlights the TYOLOv8 model's robustness in detecting objects under hazy conditions, even when haze augmentation is not incorporated during training. In contrast, the YOLOv8 model shows improvement when haze augmentation is incorporated during training, which underscores the need for training data to adapt its appearance-based features.

This advantage of the TYOLOv8 model is particularly beneficial since acquiring real-world hazy training data can be expensive and time-consuming. By effectively utilizing motion as an additional feature, TYOLOv8 enhances its ability to detect small objects even in challenging real-world hazy weather, offering a robust solution for object detection in diverse and demanding real-world scenarios.

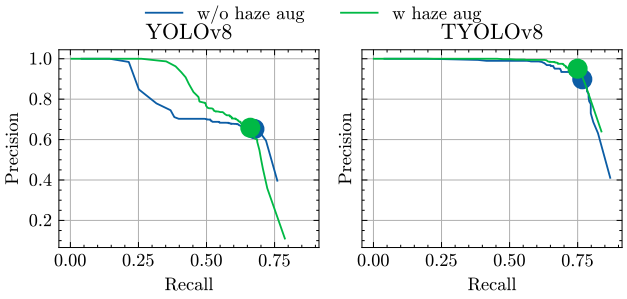


Figure 6. PR curve results on Haze subset featuring YOLOv8 (left) and TYOLOv8 (right) trained with our haze augmentation (green) and without haze augmentation (blue).

5. Conclusion

This study investigated the effect of real-world adverse weather conditions such as wind, rain and haze on small object detection performance of spatial YOLOv8 and spatio-temporal TYOLOv8 and 3FN moving object detectors. Despite that adverse weather conditions reduce visibility and introduce spatio-temporal background noise, TYOLOv8 consistently outperforms its spatial-only counterparts across all object size ranges and weather conditions, achieving this without the need incorporating weather data during the training phase. These results demonstrate that TYOLOv8 is capable of combining spatio-temporal features in such a way that it can discriminate between foreground movement and background movement. While we find that the object detection performance of YOLOv8 can be improved substantially for hazy weather by applying realistic haze augmentations, the object detection performance for TYOLOv8 remains significantly higher. Thus, the spatio-temporal detector delivers the highest quality object detection's across all evaluated scenarios, making it highly suitable for a wide range of real-world surveillance applications, without the need for incorporating real-world or simulated weather during training, which can be both costly and time-consuming.

References

[1] Adimec. Adimec tmx-55. https://www.adimec.com/cameras/global-security/tmx-cmos-cameras/tmx-55/. Visited on 2024-11-11. 4

[2] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. *CoRR*, abs/1706.03038, 2017. 3

[3] Karsten Behrendt. Boxy vehicle detection in large images. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 840–846, 2019. 3

[4] XS-VID benchmark. Xs-vid: An extra small object video detection dataset. Visited on 2024-07-11. 3

[5] Borja Bovcon, Rok Mandeljc, Janez Perš, and Matej Kristan. Stereo obstacle detection for unmanned surface vehicles by imu-assisted semantic segmentation. *Robotics and Autonomous Systems*, 2018. 3

[6] Tim Brödermann, David Bruggemann, Christos Sakaridis, Kevin Ta, Odysseas Liagouris, Jason Corkill, and Luc Van Gool. Muses: The multi-sensor semantic perception dataset for driving under uncertainty, 2024. 3

[7] Keenan Burnett, David J. Yoon, Yuchen Wu, Andrew Zou Li, Haowei Zhang, Shichen Lu, Jingxing Qian, Wei-Kang Tseng, Andrew Lambert, Keith Y. K. Leung, Angela P. Schoellig, and Timothy D. Barfoot. Boreas: A multi-season autonomous driving dataset, 2023. 3

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020. 1

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229, 2020. 2

[10] Qiang Chen, Jian Wang, Chuchu Han, Shan Zhang, Zexian Li, Xiaokang Chen, Jiahui Chen, Xiaodi Wang, Shuming Han, Gang Zhang, Haocheng Feng, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Group DETR v2: Strong Object Detector with Encoder-Decoder Pretraining, Nov. 2022. arXiv:2211.03594 [cs] version: 1. 1

[11] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[12] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13467–13488, Nov. 2023. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 3

[13] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. *CoRR*, abs/1904.04989, 2019. 3

[14] Zihan Chu. D-yolo a robust framework for object detection in adverse weather conditions, 2024. 3

[15] 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. *CoRR*, abs/1604.01685, 2016. 3

[16] Christof W. Corsel, Michel van Lier, Leo Kampmeijer, Nicolas Boehrer, and Erwin M. Bakker. Exploiting temporal context for tiny object detection. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1–11, 2023. 2

[17] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2154–2164, 2020. 3

[18] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. 2019. 2

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 3

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012. 3, 6

[21] Yihua Fan, Yongzhen Wang, Mingqiang Wei, Fu Lee Wang, and Haoran Xie. Friendnet: Detection-friendly dehazing network, 2024. 3

[22] Alain Fournier, Don Fussell, and Loren Carpenter. *Computer rendering of stochastic models*, page 189–202. Association for Computing Machinery, New York, NY, USA, 1998. 6

[23] Gianni Franchi, Xuanlong Yu, Andrei Bursuc, Angel Tena, Rémi Kazmierczak, Séverine Dubuisson, Emanuel Aldea, and David Filliat. MUAD: Multiple Uncertainties for Autonomous Driving, a benchmark for multiple uncertainty types and tasks, Oct. 2022. arXiv:2203.01437 [cs] version: 2. 3

[24] Masato Fujitake and Akihiro Sugimoto. Video Sparse Transformer With Attention-Guided Memory for Video Object Detection. *IEEE Access*, 10:65886–65900, 2022. Conference Name: IEEE Access. 2

[25] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32:1231 – 1237, 2013. 3, 5

[26] R. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2

[27] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. SODA10M: towards large-scale object detection benchmark for autonomous driving. *CoRR*, abs/2106.11118, 2021. 3

[28] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1956–1963, 2009. 3

[29] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network. *CoRR*, abs/1707.05972, 2017. 3

[30] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-Attentional Features for Single-Image Rain Removal. pages 8022–8031, 2019. 3

[31] Martin inversion, Jeff Faudi. Airbus ship detection challenge, 2018. 3

[32] Jiongchao Jin, Arezou Fatemi, Wallace Lira, Fenggen Yu, Biao Leng, Rui Ma, Ali Mahdavi-Amiri, and Hao Zhang. RaidaR: A Rich Annotated Image Dataset of Rainy Street Scenes, Oct. 2021. arXiv:2104.04606 [cs]. 3

[33] Yeying Jin, Xin Li, Jiadong Wang, Yan Zhang, and Malu Zhang. Raindrop Clarity: A Dual-Focused Dataset for Day and Night Raindrop Removal, July 2024. arXiv:2407.16957 [cs] version: 1. 3

[34] Glenn Jocher. YOLOv5 by Ultralytics, May 2020. 1

[35] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, Jan. 2023. 1, 2

[36] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO-v8 by Ultralytics. Technical report, 2023. 2

[37] Isha Kalra, Maneet Singh, Shruti Nagpal, Richa Singh, Mayank Vatsa, and P. B. Sujit. Dronesurf: Benchmark dataset for drone-based face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7, 2019. 3

[38] Benjamin Kiefer, Matej Kristan, Janez Perš, Lojze Žust, Fabio Poiesi, Fabio Augusto de Alcantara Andrade, Alexandre Bernardino, Matthew Dawkins, Jenni Raitoharju, Yitong Quan, et al. 1st workshop on maritime computer vision (macvi) 2023: Challenge results. *arXiv preprint arXiv:2211.13508*, 2022. 3

[39] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions, 2023. 3

[40] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. Joint attention in autonomous driving (JAAD). *CoRR*, abs/1609.04741, 2016. 3

[41] Matej Kristan, Vildana Sulic, Stanislav Kovacic, and Janez Pers. Fast image-based obstacle detection from unmanned surface vehicles. *CoRR*, abs/1503.01918, 2015. 3

[42] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4780–4788, 2017. 3

[43] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4780–4788, 2017. 5

[44] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, Yiduo Li, Bo Zhang, Yufei Liang, Linyuan Zhou, Xiaoming Xu, Xiangxiang Chu, Xiaoming Wei, and Xiaolin Wei. Yolov6: A single-stage object detection framework for industrial applications, 2022. 1

[45] Chengyang Li, Heng Zhou, Yang Liu, Caidong Yang, Yongqiang Xie, Zhongbo Li, and Liping Zhu. Detection-friendly dehazing: Object detection in real-world hazy scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8284–8295, 2023. 3

[46] Hao Li, Kailong Yu, Junhui Qiu, Zheng Wang, and Yang Yang. IA-Det: Iterative Attention-Based Robust Object Detection in Adverse Traffic Scenes. *IEEE Transactions on Instrumentation and Measurement*, 73:1–14, 2024. 3

[47] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 2

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, Feb. 2015. arXiv:1405.0312 [cs]. 3, 5, 6

[49] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection, 2023. 2

[50] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing, Aug. 2019. arXiv:1908.03245 [cs, eess]. 3

[51] Nguyen Anh Minh Mai, Pierre Duthon, Louahdi Khoudour, Alain Crouzil, and Sergio A. Velastin. 3d object detection with sls-fusion network in foggy weather conditions. *Sensors*, 21(20), 2021. 3

[52] Murari Mandal, Lav Kush Kumar, and Santosh Kumar Vipparthi. MOR-UAV: A benchmark dataset and baselines for moving object recognition in UAV videos. *CoRR*, abs/2008.01699, 2020. 3

[53] Tamás Matuszka, Iván Barton, Ádám Butykai, Péter Hajas, Dávid Kiss, Domonkos Kovács, Sándor Kunsági-Máté, Péter Lengyel, Gábor Németh, Levente Pető, Dezső Ribli, Dávid Szeghy, Szabolcs Vajna, and Bálint Varga. aiMotive Dataset: A Multimodal Dataset for Robust Autonomous Driving with Long-Range Perception, Sept. 2023. arXiv:2211.09445 [cs] version: 3. 3

[54] Joshua Migdal and W. Eric L. Grimson. Background subtraction using markov thresholds. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, volume 2, pages 58–65, 2005. 2

[55] Matthias Mueller, Neil Smith, and Bernard Ghanem. *A Benchmark and Simulator for UAV Tracking*, volume 9905. arXiv, Oct. 2016. Pages: 461. 3

[56] T. Nathan Mundhenk, Goran Konjevod, Wesam A. Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. *CoRR*, abs/1609.04453, 2016. 3

[57] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160, Colorado Springs, CO, USA, June 2011. IEEE. 3, 5

[58] Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A*3d dataset: Towards autonomous driving in challenging environments. *CoRR*, abs/1909.07541, 2019. 3

[59] Rui Qian, Robby T. Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive Generative Adversarial Network for Raindrop Removal from a Single Image, May 2018. arXiv:1711.10098 [cs] version: 4. 3

[60] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing Raindrops and Rain Streaks in One Go. pages 9147–9156, 2021. 3

[61] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 2

[62] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 549–565, Cham, 2016. Springer International Publishing. 3

[63] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, Jan. 2015. arXiv:1409.0575 [cs]. 3

[64] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic Foggy Scene Understanding with Synthetic Data. *International Journal of Computer Vision*, 126(9):973–992, Sept. 2018. arXiv:1708.07819 [cs]. 3

[65] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. RADIATE: A Radar Dataset for Automotive Perception in Bad Weather, Apr. 2021. arXiv:2010.09076 [cs] version: 3. 3

[66] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision Transformers for Single Image Dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. arXiv:2204.03883 [cs]. 3

[67] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[68] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *CoRR*, abs/1810.11780, 2018. 3

[69] test. rtts dataset. https://universe.roboflow.com/test-mdnu9/rtts, jun 2022. Visited on 2024-10-01. 3

[70] Gopal Thapa, Kalpana Sharma, and Mrinal Kanti Ghose. Moving object detection and segmentation using frame differencing and summing technique. *International Journal of Computer Applications*, 102(7):20–25, 2014. 2

[71] Michel van Lier, Martin C. van Leeuwen, Bastian van Manen, and Leo Kampmeijer. Real-time small object detection on embedded hardware for 360-degree vision. In Judith Dijk and Jose Luis Sanchez-Lopez, editors, *Autonomous Systems for Security and Defence*, volume 13207, page 132070E. International Society for Optics and Photonics, SPIE, 2024. 2, 3, 6, 7

[72] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, July 2022. arXiv:2207.02696 [cs]. 1

[73] Han Wang, Jun Tang, Xiaodong Liu, Shanyan Guan, Rong Xie, and Li Song. PTSEFormer: Progressive Temporal-Spatial Enhanced TransFormer Towards Video Object Detection, Sept. 2022. arXiv:2209.02242 [cs] version: 1. 2

[74] Lucai Wang, Hongda Qin, Xuanyu Zhou, Xiao Lu, and Fengting Zhang. R-yolo: A robust object detector in adverse weather. *IEEE Transactions on Instrumentation and Measurement*, 72:1–11, 2023. 3

[75] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking, Jan. 2020. arXiv:1511.04136 [cs]. 3

[76] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive Learning for Compact Single Image Dehazing, Apr. 2021. arXiv:2104.09367 [cs]. 3

[77] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge J. Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. *CoRR*, abs/1711.10398, 2017. 3

[78] Xiaowei Xu, Xinyi Zhang, Bei Yu, Xiaobo Sharon Hu, Christopher Rowen, Jingtong Hu, and Yiyu Shi. DAC-SDC low power object detection challenge for UAV applications. *CoRR*, abs/1809.00110, 2018. 3

[79] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A Face Detection Benchmark, Nov. 2015. 3

[80] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, Seattle, WA, USA, June 2020. IEEE. 3, 5

[81] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale Match for Tiny Person Detection, Dec. 2019. arXiv:1912.10664 [cs]. 3, 5

[82] Kaiwen Zhang, Xuefeng Yan, Yongzhen Wang, and Junchen Qi. Adaptive Dehazing YOLO for Object Detection. In *Artificial Neural Networks and Machine Learning – ICANN 2023: 32nd International Conference on Artificial Neural Networks, Heraklion, Crete, Greece, September 26–29, 2023, Proceedings, Part VII*, pages 14–27, Berlin, Heidelberg, Sept. 2023. Springer-Verlag. 3

[83] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. CityPersons: A Diverse Dataset for Pedestrian Detection, Feb. 2017. arXiv:1702.05693 [cs]. 3, 5

[84] Zhengning Zhang, Lin Zhang, Yue Wang, Pengming Feng, and Ran He. Shiprsimagenet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8458–8472, 2021. 3

[85] Xian Zhong, Shidong Tu, Xianzheng Ma, Kui Jiang, Wenxin Huang, and Zheng Wang. Rainy WCity: A Real Rainfall Dataset with Diverse Conditions for Semantic Driving Scene Understanding. volume 2, pages 1743–1749, July 2022. ISSN: 1045-0823. 3

[86] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: End-to-end video object detection with spatial-temporal transformers. *CoRR*, abs/2201.05047, 2022. 2

[87] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and Tracking Meet Drones Challenge, Oct. 2021. arXiv:2001.06303 [cs]. 3, 5

[88] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. *CoRR*, abs/2010.04159, 2020. 1

[89] Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with Collaborative Hybrid Assignments Training, Aug. 2023. arXiv:2211.12860 [cs] version: 5. 1