Benchmarking Object Detectors under Real-World Distribution Shifts in Satellite Imagery

Sara A. Al-Emadi[1,2] Yin Yang[2] Ferda Ofli[1]
[1] Qatar Computing Research Institute, HBKU & [2] College of Science and Engineering, HBKU
{salemadi, yyang, fofli}@hbku.edu.qa

Abstract

Object detectors have achieved remarkable performance in many applications; however, these deep learning models are typically designed under the i.i.d. assumption, meaning they are trained and evaluated on data sampled from the same (source) distribution. In real-world deployment, however, target distributions often differ from source data, leading to substantial performance degradation. Domain Generalisation (DG) seeks to bridge this gap by enabling models to generalise to Out-Of-Distribution (OOD) data without access to target distributions during training, enhancing robustness to unseen conditions. In this work, we examine the generalisability and robustness of state-of-the-art object detectors under real-world distribution shifts, focusing particularly on spatial domain shifts. Despite the need, a standardised benchmark dataset specifically designed for assessing object detection under realistic DG scenarios is currently lacking. To address this, we introduce Real-World Distribution Shifts (RWDS), a suite of three novel DG benchmarking datasets that focus on humanitarian and climate change applications. These datasets enable the investigation of domain shifts across (i) climate zones and (ii) various disasters and geographic regions. To our knowledge, these are the first DG benchmarking datasets tailored for object detection in real-world, high-impact contexts. We aim for these datasets to serve as valuable resources for evaluating the robustness and generalisation of future object detection models. Our dataset and code are available at https://github.com/saraalemadi/RWDS.

1. Introduction

Deep learning has achieved remarkable success in various applications, including flood mapping [32, 51], medical diagnostics [4, 35], and self-driving cars [19, 60]. However, these machine learning models are typically developed under the i.i.d. assumption, where they are trained and evaluated on data samples drawn from the same *source* distribution. Consequently, when deployed in real-world environments with differing *target* distributions, these models experience significant performance degradation, hindering their large-scale deployment. This phenomenon is known as distribution or domain shift. In this paper, we focus a specific type of domain shift, referred to as spatial domain shift (i.e., covariate shift) on a global scale, which is driven by visual variations in land cover and built structures influenced by factors such as natural landscapes, climate zones, architectural styles, economic and financial development, social and cultural attributes, and human settlement patterns.

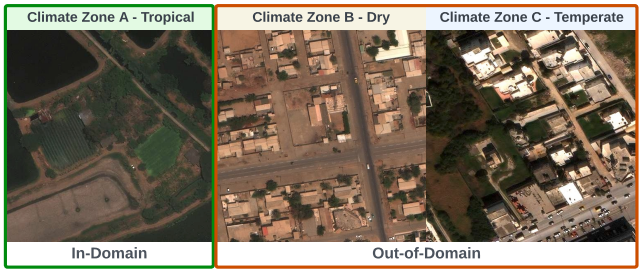


Figure 1. Example images from different climate zones

In satellite imagery-based object detection, spatial domain shift poses a significant challenge, especially when environmental conditions vary unpredictably, as demonstrated in Figure 1. To examine this, we use the Köppen climate classification system [61] and consider a scenario where an object detector is trained on images from a *tropical* climate zone but evaluated on *Out-of-Distribution (OOD)* target domains, specifically *dry* and *temperate* climate zones. Since these target domains exhibit distinct visual characteristics compared to the source domain, a performance drop is expected, highlighting the impact of spatial domain shift when applying object detection models across diverse climatic contexts.

Several studies have sought to mitigate the issue of domain shift through data augmentation [59], transfer learning [40], and domain adaptation, in which the model has access to unlabelled samples from the target distribution during training [43, 65]. However, in real-world applications, models often encounter distributions that cannot be fore-

seen before deployment. To tackle this challenge, recent research has focused on domain shift under this constraint, a problem known as Domain Generalisation (DG).

DG datasets are essential for assessing a model's ability to generalise to unseen target distributions. In image classification, a substantial body of research have been devoted to curating DG datasets for broad use cases, such as PACS [29] and DomainNet [44], as well as datasets introducing synthetic domain shifts, like RotatedMNIST [15], or real-world distribution shifts, as seen in WILDS [26]. However, the study of domain shift in object detection remains relatively underexplored. To bridge this gap, Mao *et al*. introduced COCO-O [41], a DG benchmark for object detection with six domains including Sketch, Weather, Cartoon, Painting, Tattoo, and Handmade, to evaluate both in-domain (ID) and OOD performance. While the domain shifts in COCO-O are evident, such as the differences between sketches and paintings, further investigation of the practical motivation for training a model on sketches and testing it on paintings could provide valuable insights. To our knowledge, a DG benchmark does not currently exist for evaluating the behaviours of object detectors on OOD test data in a common, real-world application setting.

Motivated by this, we introduce *Real-World Distribution Shifts* (*RWDS*), a suite of three realistic DG benchmarking datasets, namely, RWDS-CZ, RWDS-FR and RWDS-HE, which focus on humanitarian and climate change applications and investigate domain shifts across (i) climate zones and (ii) different disasters and geographic regions, respectively. Moreover, we benchmark and analyse the performance of several state-of-the-art (SOTA) object detection algorithms on RWDS under two setups: single-source, where an object detector is trained on only one source domain, and multi-source, where training incorporates multiple source domains. We then evaluate these models on the unseen target domains to provide comprehensive insights into their generalisation performance. We trained around 100 object detector models and conducted over 200 experiments. Our contributions are summarised as follows:

- We propose RWDS, a suite of novel, realistic and challenging DG datasets designed to evaluate spatial domain shifts in real-world object detection tasks.
- We provide the community with in-depth benchmarking analyses on the performance of the SOTA object detectors on RWDS datasets.
- We analyse the impact of single-source versus multi-source training in DG for spatial domain shifts in satellite imagery, concluding that multi-source training enhances generalisability of object detectors.

The rest of the paper is organised as follows: Section 2 reviews literature on object detection and robustness benchmarks. Section 3 introduces the RWDS datasets with details on data cleaning and preprocessing. Section 4 describes the evaluation metrics, experimental setups, and selected object detectors. Section 5 presents the results and provides a comprehensive analysis and Section 6 concludes the paper.

2. Related Work

2.1. Object Detection

Early attempts in deep-learning-based object detection used a set of bounding boxes and masked regions as input to the CNN architecture to incorporate shape information into the classification process to perform object localisation [12, 16, 54, 55]. Later on, end-to-end techniques were proposed based on shared computation of convolutions for simultaneous detection and localization of the objects [9, 17, 22, 24, 37, 46, 48, 52]. These methods can be generally divided into two categories: one-stage detectors [11, 13, 28, 37, 45, 46, 50, 56, 57, 70] and two-stage detectors [5, 16, 17, 23, 24, 33, 47, 53]. More recently, transformer-based object detection models have proved more efficient and accurate, thanks to their ability to not require anchor boxes and non-maximum suppression procedure [6, 67, 71]. Besides, with the advances in foundation models (large vision models or vision-language models), open-set and open-world object detection has become popular [31, 36, 62]. Following these trends, remote sensing community has also integrated deep learning-based object detection models into their research [1, 10, 21, 30, 38, 39, 66]. However, accurate object detection from satellite imagery *at scale* remains a challenging task.

2.2. Robustness Benchmarks

Various benchmark studies have been developed to assess the robustness of object detection models under distribution shifts. For instance, COCO-C [42] evaluates model performance by applying synthetic corruptions, such as JPEG compression and Gaussian noise, to the COCO test set. Similarly, OOD-CV [68] and its extended version, OOD-CV-v2 [69], include OOD examples across 10 object categories from PASCAL VOC and ImageNet, spanning variations in pose, shape, texture, context, and weather conditions. These datasets enable benchmarking across multiple tasks like image classification, object detection, and 3D pose estimation. COCO-O [41] introduces natural distribution shifts in COCO-based datasets, spanning six domains—weather, painting, handmade, cartoon, tattoo, and sketch. Their study has shown that there is a significant performance gap of 55.7% between ID and OOD performance, highlighting the domain generalisation challenges under such shifts. However, despite their contributions, these datasets still lack the complexity of real-world distribution shifts. More realistic benchmarks include those reflecting environmental changes in autonomous driving [25] and object variations in aerial imagery [63], which better capture

the dynamic and unpredictable conditions faced in practical applications. However, they remain limited in scope, as they do not comprehensively account for geographic and temporal variability, environmental and weather conditions, occlusion, clutter, and object appearance changes within a unified framework. In contrast, our RWDS datasets aim to bridge this gap by providing a diverse and realistic evaluation setting that encapsulates these real-world domain shifts more holistically.

3. Our RWDS Datasets

3.1. RWDS Across Climate Zones

While there are increasing efforts to mitigate and reduce the negative and potential impact of climate change on the global ecosystem including natural resources, weather and the natural landscapes [49], there is a need to develop robust models to support computer vision tasks under these circumstances, more particularly, object detection task. In order to investigate their robustness and generalisability across different climate zones, we propose **RWDS across Climate Zones (RWDS-CZ)** dataset where we focus on Köppen's climate zone classification [3, 61]. Given the scarcity of global satellite imagery that covers all climate zones, we use the raw satellite imagery from the xView dataset [27], an open-source object detection dataset featuring high-resolution (0.3m) images captured at a global scale across 60 object classes. For this study, we focus on three distinct climate zones: Zone A (CZ A)—tropical or equatorial, Zone B (CZ B)—arid or dry, and Zone C (CZ C)—warm/mild temperate. These serve as our distinct domains for studying spatial domain shifts in satellite-based object detection.

To create the domains, we first map the geo-coordinates of each image to its respective climate zone and proceed with splitting the overall dataset into domains. However, this results in a mismatch between the classes available across the domains. To resolve this, we retain only those classes that appear in all climate zones. Additionally, we set a threshold of 30 samples per class to ensure sufficient data for training. Any class with fewer than 30 samples is excluded from all domains. This process yields a total of 16 classes. To maintain consistent distribution of object instances across the training, validation, and test sets within each domain, we follow the procedure outlined in Algorithm 1. This process is repeated for each domain, resulting in the final RWDS-CZ dataset. Table 1 summarises the dataset statistics, while Figure 2 shows the distribution of training samples across classes in all domains.

To visualize the domain shift in RWDS-CZ, we extract image embeddings using RemoteCLIP [34] and project them into 2D using t-SNE [58]. Figure 3-A showcases the shift between images from CZ A and CZ B.

Algorithm 1 Dataset Split Procedure

Input: Set of images $\mathcal{I}$, class labels $\mathcal{C}$
Initialise: Training set $\mathcal{T} \leftarrow \emptyset$, Validation set $\mathcal{V} \leftarrow \emptyset$, Testing set $\mathcal{S} \leftarrow \emptyset$
function ALLOCIMAGES($\mathcal{I}, \mathcal{N}$)
  for each class $c \in \mathcal{C}$ **do**
    $I^* \leftarrow \underset{I \in \mathcal{I}}{\operatorname{argmax}}\, \text{count}(I, c)$ ▷ Select image with most instances of c
    $\mathcal{N} \leftarrow \mathcal{N} \cup \{I^*\}$ ▷ Append to designated set
    $\mathcal{I} \leftarrow \mathcal{I} \setminus \{I^*\}$ ▷ Remove allocated image
  end for
  return $(\mathcal{N}, \mathcal{I})$ ▷ Return final dataset splits
end function
while $\mathcal{I} \neq \emptyset$ **do**
  for $i = 1$ to 3 **do** ▷ Repeat 3 times for training set
    $\mathcal{T}, \mathcal{I} \leftarrow$ ALLOCIMAGES($\mathcal{I}, \mathcal{T}$) ▷ Update training set
  end for
  $\mathcal{V}, \mathcal{I} \leftarrow$ ALLOCIMAGES($\mathcal{I}, \mathcal{V}$) ▷ Update validation set
  $\mathcal{S}, \mathcal{I} \leftarrow$ ALLOCIMAGES($\mathcal{I}, \mathcal{S}$) ▷ Update test set
end while

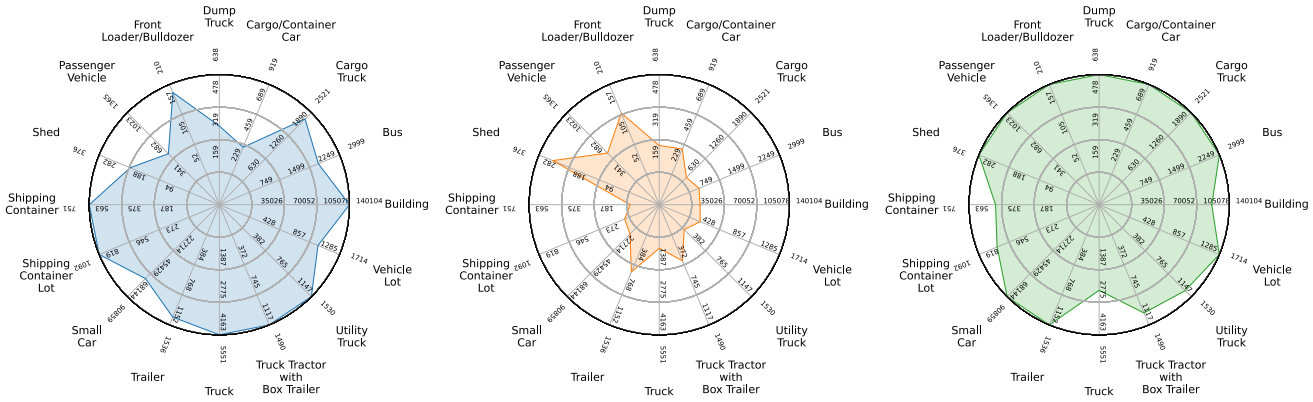
Split	CZ A	CZ B	CZ C
Training	117,265	43,272	124,717
Validation	58,997	13,423	47,362
Test	56,954	24,745	60,310

Table 1. RWDS-CZ overall object instances per partition

3.2. RWDS in Disaster Damage Assessment

A notable consequence of climate change is the increasing frequency and severity of natural disasters such as hurricanes, storms, floods, wildfires, earthquakes, tsunamis, etc. Damage assessment is essential during and after disasters to support aid delivery, guide building reconstruction efforts, and provide governments and humanitarian agencies with an estimate of the disaster's impact. Generally, large amount of satellite imagery is captured around the disaster-hit locations. However, given the sheer volume of data, the cleaning, preprocessing and re-training of object detectors for each disaster on the spot is time consuming and might not be feasible due to lack of annotations, highlighting the crucial need for robust models that can generalise well to unseen distributions beyond those they were trained on. Hence, we investigate the robustness of SOTA object detectors in this application under two different scenarios.

In the first use-case, we examine the shift of the same disaster type across distant geographic regions with different socioeconomical characteristics. Specifically, we define domains in terms of collection of events that caused *floods* across the United States and India, respectively. We refer to this use-case as **RWDS across Flooded Regions (RWDS-FR)**. Whereas, in the second use-case, we focus on understanding the shift in the behaviour of these models



(a) Climate Zone A (b) Climate Zone B (c) Climate Zone C

Figure 2. Class-wise distribution of training data for each domain as well as the overall data distribution across domains in RWDS-CZ

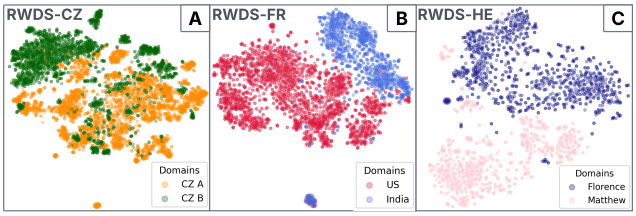


Figure 3. Embedding space representations of the RWDS datasets

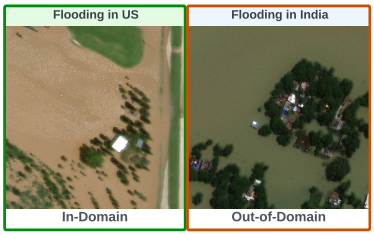


Figure 4. Comparison of flood scenes between the US and India

across different disaster events of the same type, namely, *hurricanes*, in North America. We refer to this use-case as **RWDS across Hurricane Events (RWDS-HE)**. Similar to the discussion related to the scarcity of open-source satellite imagery, we utilise the raw satellite images released in the xDB building damage assessment dataset [20] for both RWDS-FR and RWDS-HE.

3.2.1 RWDS Across Flooded Regions (RWDS-FR)

We start by creating the metadata for the raw images. xDB dataset provides disaster event, damage type, and polygons of buildings for segmentation application. However, given that we are interested in object detection, we convert polygons of buildings into bounding boxes. Furthermore, similar to Section 3.1, we map the latitude and longitude coordinates of the polygons to find the corresponding location of each object instance in terms of country, region, continent, etc. We then extract the flooded objects in India and US. Figure 4 shows example images illustrating the shift between the domains, with a close-up visualization of image embeddings in Figure 3-B. Unlike the original data, where the instances are categorised into four classes, namely, no damage, minor damage, major damage and destroyed, when extracting the flooded instances in the US and India, we observe a class imbalance between those classes. Therefore, we transform the task into a binary categorisation, leaving us with two classes, namely, damaged (D) and no damage (ND). We then follow the same logic as that discussed in Section 3.1 to create the training, validation and testing splits. This yields the RWDS-FR dataset. Table 2 represents the resulting domain and class distributions per split.

	India		US	
Split	D	ND	D	ND
Training	5,023	14,841	10,680	20,055
Validation	2,532	8,320	5,470	9,834
Test	2,802	8,064	5,452	10,034

Table 2. RWDS-FR object instances per partition.

3.2.2 RWDS Across Hurricane Events (RWDS-HE)

In contrast to RWDS-FR dataset, where domains are defined by diverse geographic regions, this dataset focuses on hurricane events across North America, which are geographically closer in proximity. As a result, the dataset consists of four hurricane events as domains: Florence, Michael, Harvey, and Matthew, as shown in Figure 5. Figure 3-C presents the image embeddings for hurricanes Florence and Matthew as an example. We adhere to the preprocessing, metadata creation, binary class categorisation, and data splitting procedures outlined in Section 3.2.1. This

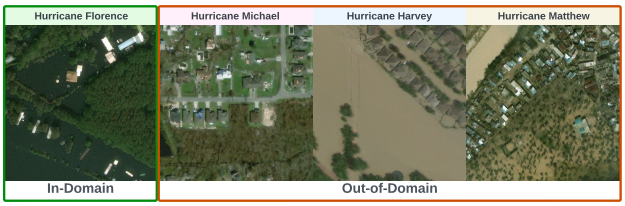


Figure 5. Comparison of hurricane scenes from different events

	Florence		Michael		Harvey		Matthew	
Split	D	ND	D	ND	D	ND	D	ND
Training	1,102	4,196	6,132	11,347	9,270	9,223	8,919	1,938
Validation	578	2,112	3,075	5,455	4,670	4,594	4,910	1,042
Test	582	2,158	3,229	5,890	4,796	4,821	4,743	1,078

Table 3. RWDS-HE object instances per partition.

defines the RWDS-HE dataset. Table 3 presents the final per-class and per-split distribution for each domain.

4. Experiments

4.1. Single-Source and Multi-Source Setup

We investigate DG in two setups. The first involves training an object detector on a single ID source training set, then assessing its performance on both the held-out OOD target domains and the ID test set. This setup reflects scenarios with a limited diversity of data distributions. Whereas, in the second setup, we incorporate training an object detector on a collection of source domains, mirroring real-world scenarios, where data from a variety of distributions may be available. For quantitative comparison, we evaluate the trained object detector on each OOD target domain separately, as well as on the average performance across ID domains.[1]

4.2. DG Evaluation Metrics

Methods for evaluating DG models remain an active and open area of investigation. Researchers have, however, adapted existing approaches to assess the performance of deep learning models on OOD datasets for classification tasks. Among these, the *leave-one-domain-out* evaluation strategy [18] is widely regarded. In this setup, one domain is excluded from training, enabling it to serve as an independent test domain to rigorously evaluate model performance without any additional tuning. Inspired by this, we adapt this evaluation technique for object detection under the single- and multi-source setups.

We assess the performance of the object detectors using the standard mean Average Precision (mAP) metric which is commonly used in object detection applications [72]. More specifically, we use the MS-COCO AP metric, which is calculated as the average over multiple IoU thresholds ranging from 0.50 to 0.95 with a stepsize of 0.05.[2]

[1]The single- and multi-source setups are formally defined in Supplementary A.

Object Detector	Backbone
Faster R-CNN [47]	ResNeXt-101-64x4d
Mask R-CNN [24]	ResNeXt-101-64x4d and FPN
TOOD [13]	ResNeXt-101-64x4d, DCNv2 and FPN
DINO (5scale) [67]	Swin-L
Grounding DINO [36]	BERT and Swin-B
GLIP (L) [31]	BERT and Swin-L

Table 4. Object detectors and their backbone architectures.

Performance Drop (PD). A metric frequently used in the DG community for assessing the generalisability of classification tasks is the *Performance Drop*, which quantifies the percentage of performance degradation observed in the model when subjected to OOD data from a target domain. Drawing inspiration from this approach, we apply it in the context of DG for object detection. This is formulated as follows:

$$PD = 100 \times \frac{mAP_{ID} - mAP_{OOD}}{mAP_{ID}} \quad (1)$$

where mAP_{ID} and mAP_{OOD} represent the average mAP of the combination of detectors tested on a specific domain's ID and OOD test sets, respectively.

Harmonic Mean (H). To compare the ID and OOD performance of object detectors based on their mAP, we adopt the widely recognised *Harmonic Mean* as another evaluation metric. This choice is motivated by its use in recent generalised open-set zero-shot learning studies [8, 14, 64] to compute a joint score reflecting model performance across in-domain and out-of-domain test sets. This is defined as:

$$H = \frac{2 \times mAP_{OOD} \times mAP_{ID}}{mAP_{OOD} + mAP_{ID}} \quad (2)$$

4.3. Object Detectors and Hyperparameters

We conduct all the experiments using the MMDetection toolbox [7]. We selected object detectors across classical (Faster R-CNN [47], Mask R-CNN [24]), recent (DINO [67], TOOD [13]), and foundation model-based approaches (Grounding DINO [36], GLIP [31]). Table 4 presents the top-performing backbone architecture selected for each detector, as evaluated on standard object detection datasets by MMDetection.

To train the object detectors, we perform preprocessing to unify the image sizes of the raw images. We start by cropping all the images into 512×512 tiles with an overlapping ratio of 0.2 using SAHI [1] while preserving the original resolution of the images. To ensure a fair comparison

[2]mAP_{50} & mAP_{75} results are included in Supplementary B.

Methods	Target											
	CZ A				CZ B				CZ C			
	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
Faster R-CNN	7.2	3.9	47	5.0	7.5	6.0	20	6.7	7.7	3.4	56	4.7
Mask R-CNN	7.3	3.7	49	4.9	7.7	5.8	25	6.6	7.8	3.5	55	4.8
TOOD	7.8	4.0	49	5.2	7.8	6.1	22	6.8	8.2	4.0	52	5.3
DINO	11.0	5.6	49	7.4	9.6	8.0	17	8.7	11.0	5.6	49	7.4
Grounding DINO	**12.9**	**7.5**	42	**9.5**	**10.8**	**10.0**	**7**	**10.4**	**13.1**	**7.1**	46	**9.2**
GLIP	9.8	6.3	**36**	7.6	8.8	8.2	**7**	8.5	9.2	5.4	**41**	6.8

Table 5. Single-source DG analysis of SOTA detectors on RWDS-CZ.

Methods	Target											
	CZ A				CZ B				CZ C			
	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
Faster R-CNN	7.7	4.9	36	6.0	8.2	7.1	13	7.6	7.7	4.1	47	5.4
Mask R-CNN	7.5	4.7	37	5.8	8.1	6.9	15	7.5	7.9	4.3	46	5.6
TOOD	8.2	5.0	39	6.2	8.7	7.0	19	7.7	8.3	4.8	42	6.1
DINO	11.6	7.2	38	8.9	11.5	9.6	16	10.4	11.8	7.0	40	8.8
Grounding DINO	**13.1**	**8.8**	33	**10.5**	**12.5**	**11.0**	12	**11.7**	**13.1**	**9.3**	**29**	**10.9**
GLIP	10.6	8.0	**24**	9.1	9.8	9.2	**6**	9.5	9.8	6.8	31	8.0

Table 6. Multi-source DG analysis of SOTA detectors on RWDS-CZ.

of model performances and to mimic real-world conditions where hyperparameter optimisation may be impractical, we use the default hyperparameters specified for each model.

5. Results and Analyses

In the single-source experiment, we evaluate OOD performance by calculating the average performance across models tested on OOD domains, while the ID performance is assessed on the test set of the ID source domain. In contrast, for the multi-source setup, we calculate ID performance as the average performance of all object detectors trained on source domains, while OOD performance is evaluated on the test set of the left-out OOD target domain.

5.1. RWDS across Climate Zones (RWDS-CZ)

5.1.1 Single-Source DG Experiment

In the single-source setup, we train all six detectors on the three climate zones, namely, CZ A, CZ B and CZ C, individually. This results in a total of 18 trained object detectors. We then proceed to evaluate the performance of the trained detectors on the different ID and target OOD test sets, yielding a total of an additional 54 inference experiments. Table 5 summarises the performance of the detectors on each climate zone.[3]

When comparing the performance on OOD climate zones to the ID test sets, it can be observed from Table 5 that all object detectors exhibit a significant performance drop of above 35% for CZ A, 7% for CZ B and 40% for CZ C, highlighting the challenges posed by domain shift across different climate zones and the limitations of current models in handling OOD data efficiently.

While GLIP experiences the lowest drop between ID and OOD performance for all the climate zones, Grounding DINO achieves the highest overall tradeoff, balancing both ID and OOD performance most effectively. Moreover, highlighted by H-scores, among the SOTA object detectors evaluated, Grounding DINO outperforms other detectors, both in terms of ID and OOD performance. A plausible explanation to this observation could be that Grounding DINO was designed to generalise to unseen classes in an open-set setting and such capabilities not only boost the performance in an open-set setting but also under a DG setting. The qualitative performance across domains are analysed in Supplementary B.1.2.

5.1.2 Multi-Source DG Experiment

Similar to the single-source experiment, we evaluate the performance of the trained detectors on the different ID and OOD test sets, yielding a total of an additional 54 experiments. Table 6 summarises the performance of each detector across the various combinations of the climate zones.[4]

When comparing the performance of the object detectors trained on a single-source to those trained under the

[3]The detectors' cross-domain results on RWDS-CZ under single-source setup are in Supplementary B.1.1.

[4]The detectors' cross-domain results on RWDS-CZ under multi-source setup are in Supplementary B.1.3.

Methods	Target							
	India				US			
	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
Faster R-CNN	4.5	1.3	71	2.0	25.5	1.8	93	3.4
Mask R-CNN	4.3	1.2	72	1.9	25.9	2.0	92	3.7
TOOD	5.1	1.6	69	2.4	27.6	2.4	91	4.4
DINO	**7.0**	2.2	69	3.3	30.8	4.3	86	7.5
Grounding DINO	6.7	**3.3**	**51**	**4.4**	**31.3**	10.8	65	16.1
GLIP	6.7	**3.3**	**51**	**4.4**	30.7	**11.9**	**61**	**17.2**

Table 7. Single-source DG analysis of SOTA detectors on RWDS-FR.

Methods	Target															
	Florence				Michael				Harvey				Matthew			
	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
Faster R-CNN	34.5	8.6	75	13.8	18.6	6.5	65	9.7	25.1	3.7	85	6.4	1.5	0.3	78	0.5
Mask R-CNN	34.0	8.3	76	13.3	19.1	6.9	64	10.1	25.6	3.7	86	6.4	1.7	0.4	78	0.6
TOOD	35.7	10.4	71	16.1	21.0	7.1	66	10.6	27.5	4.4	84	7.5	2.4	0.5	78	0.9
DINO	36.5	12.0	67	18.0	20.6	7.6	63	11.1	**31.4**	4.9	84	8.5	2.5	0.8	69	1.2
Grounding DINO	39.3	17.4	56	24.2	**24.2**	9.3	62	13.4	31.0	**7.7**	**75**	**12.4**	3.3	1.2	65	1.7
GLIP	**40.8**	**19.0**	**53**	**25.9**	23.9	**10.2**	**57**	**14.3**	29.2	7.0	76	11.3	**3.7**	**1.3**	**64**	**2.0**

Table 8. Single-source DG analysis of SOTA detectors on RWDS-HE.

multi-source setup, it can be observed that training on multiple source domains enhances not only the object detector's performance on the ID test set but also provides a more substantial boost to the OOD performance.

Furthermore, consistent with the findings of the single-source experiment, all object detectors evaluated under the multi-source setting experience a notable performance drop when tested on OOD target test set compared to the ID test set. Moreover, across the object detectors benchmarked, Grounding DINO demonstrates the highest performance on the ID and OOD test sets, as evidenced by its H-score.

5.2. RWDS across Flooded Regions (RWDS-FR)

Since RWDS-FR includes only India and the US as domains, this dataset inherently fits within a single-source setting. We train a total of 12 object detectors and conduct a total of 24 experiments to evaluate their ID and OOD performances, respectively. Table 7 summarises the ID and OOD performance of all the object detectors.[5]

A notable decline in performance can be observed across both India and the US on the OOD test sets with drops exceeding 52% and 62%, respectively, indicating a significant impact of the domain-specific variations between the two regions. Grounding DINO and GLIP exhibit comparable performance, both outperforming other object detectors in terms of ID and OOD performance. Their superior OOD performance highlights their slight robustness in handling domain shifts, making them more effective than the other object detection models when the domain shift is defined in terms of disparate geographic regions. We present qualitative performance analyses across domains in Supplementary B.2.2.

5.3. RWDS across Hurricane Events (RWDS-HE)

5.3.1 Single-Source DG Experiment

A significant decline in performance can be observed, in Table 8, between hurricane events, suggesting that variations in the nature and characteristics of these events contribute to discrepancies in detection precision.[6] This drop highlights the challenge of generalising disaster object detection across different types of extreme weather events due to domain-specific variations, underscoring the need for further research on developing more robust and generalisable detectors to account for such discrepancies.

Consistent with the RWDS-FR experiment, both Grounding DINO and GLIP achieve the highest performance, with GLIP performing slightly better and excelling in OOD detection compared to the other object detectors. It is also evident that the object detectors have the weakest ID performance when evaluated on the test set of Hurricane Matthew, which suggests that the underlying data is difficult and might suffer from factors such as label noise and class imbalance, as discussed in Section 3.2.2, influencing the performance of the models negatively. Qualitative domain performance analyses are provided in Supplementary B.2.2.

[5]The detectors' cross-domain results on RWDS-FR are in Supplementary B.2.1.

[6]The detectors' cross-domain results on RWDS-HE under single-source setup are in Supplementary B.3.1.

Methods	Target															
	Florence				Michael				Harvey				Matthew			
	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
Faster R-CNN	32.8	12.7	61	18.3	19.0	8.9	53	12.1	25.0	5.2	79	8.6	1.7	0.4	76	0.6
Mask R-CNN	33.6	13.3	60	19.1	19.3	9.1	53	12.4	25.8	5.4	79	8.9	1.6	0.7	56	1.0
TOOD	34.2	14.0	59	19.9	19.7	9.6	51	12.9	27.2	5.3	81	8.9	2.2	0.5	77	0.8
DINO	37.3	17.0	54	23.3	21.4	10.6	50	14.2	31.3	7.7	75	12.4	2.8	0.8	71	1.2
Grounding DINO	39.6	28.2	29	32.9	**24.3**	**12.8**	**47**	**16.8**	**32.2**	**9.4**	**71**	**14.5**	3.1	**1.5**	**52**	**2.0**
GLIP	**40.8**	**30.7**	**25**	**35.0**	**24.3**	11.4	53	15.5	30.9	7.8	75	12.5	**3.2**	1.1	66	1.6

Table 9. Multi-source DG analysis of SOTA detectors on RWDS-HE.

5.3.2 Multi-Source DG Experiment

In our experimental results illustrated in Table 9, we observe a slight performance drop in the ID performance of the object detectors when comparing the multi-source setup to the single-source.[7] This decline in ID performance suggests that training on multiple source domains introduces additional complexity, which can marginally reduce the model's ability to generalise effectively to the ID domain. Therefore, the inclusion of multiple domains likely causes the model to adapt to a broader set of domain features, which may dilute its focus on optimising the performance specifically on the ID test set. While the multi-source setup generally enhances OOD performance, this comes at the cost of slightly decreased ID accuracy.

However, a notable improvement in OOD performance is observed across nearly all models and domains when compared to the single-source setup. This improvement indicates that the model benefits significantly from exposure to a diverse set of source domains, enhancing its ability to generalise to unseen, OOD domain. Furthermore, consistent with the previous experiments, Grounding DINO and GLIP had a comparably high ID and OOD performances in comparison to the other evaluated object detectors.

5.4. Error Analysis of Object Detectors

We analyse detection errors using the TIDE [2] toolbox, as shown in Figures 6. Evaluating on the OOD data of RWDS-CZ as a use-case, where the model is trained on CZ A and tested on CZ B, we find that classification errors are the main factor for performance drop, followed by background errors for both GLIP and Grounding DINO, and missed classifications for Faster R-CNN. Moreover, Faster R-CNN has the highest classification and missed groundtruth errors, which aligns with the findings in Section 5.1, where it consistently had the weakest performance. While GLIP makes less errors than Faster R-CNN, its higher rate of background errors explains its weaker performance compared to Grounding DINO. A similar trend appears when evaluating the performance of the model trained of CZ B and evaluated on its ID test set, where Faster R-CNN and Grounding DINO exhibit higher classification errors, while GLIP has the highest background error rate.

[7] The detectors' cross-domain results on RWDS-HE under multi-source setup are in Supplementary B.3.3.

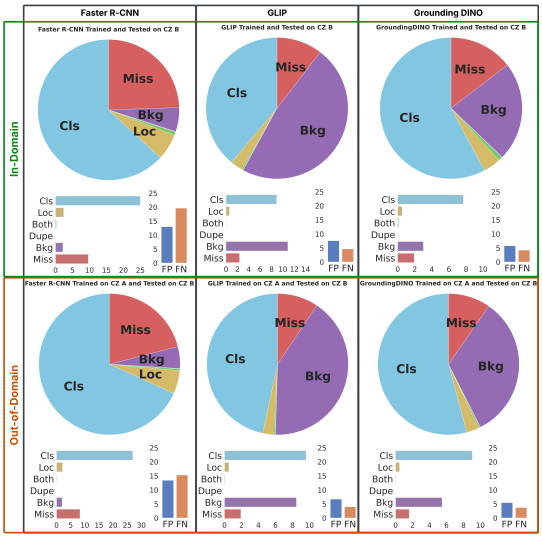


Figure 6. Object detection errors of detectors trained on CZ A and evaluated on ID (top-row) and OOD (bottom-row) data of CZ B.

6. Conclusion

Object detectors typically perform well under the assumption that training and evaluation data come from the same distribution. However, real-world target distributions often differ, causing performance drops due to the distribution shift. DG aims to address this by enabling models to generalise to OOD data without access to target distributions during training. This study examines the generalisability of SOTA object detectors under spatial domain shifts in real world applications and introduces three novel DG benchmark datasets focused on humanitarian and climate change applications. Supported by our findings under single-source and multi-source setups, these datasets, covering domain shifts across climate zones, regions and disaster events, are the first to assess object detection in high-impact real-world contexts and aim to provide valuable resources for evaluating future models' robustness and generalisation.

References

[1] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. *2022 IEEE International Conference on Image Processing (ICIP)*, pages 966–970, 2022. 2, 5

[2] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *ECCV*, 2020. 8

[3] Tyson Brown. Köppen Climate Classification System - National Geographic, 2024. 3

[4] Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. 1

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019. 2

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[8] Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European conference on computer vision*, pages 572–588. Springer, 2020. 5

[9] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3992–4000, Boston, MA, USA, 2015. IEEE. 2

[10] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7778–7796, 2021. 2

[11] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 2

[12] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2155–2162, Columbus, OH, USA, 2014. IEEE. 2

[13] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. TOOD: Task-aligned One-stage Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499, Montreal, QC, Canada, 2021. IEEE. 2, 5

[14] Yanwei Fu, Xiaomei Wang, Hanze Dong, Yu-Gang Jiang, Meng Wang, Xiangyang Xue, and Leonid Sigal. Vocabulary-informed zero-shot and open-set learning. *IEEE transactions on pattern analysis and machine intelligence*, 42(12):3136–3152, 2019. 5

[15] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 2

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, 2014. IEEE. 2

[17] Ross B Girshick. Fast R-CNN. In *International Conference on Computer Vision*, pages 1440–1448, Boston, MA, USA, 2015. IEEE. 2

[18] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *International Conference on Learning Representations*, 2021. 5

[19] Abhishek Gupta, Alagan Anpalagan, Ling Guan, and Ahmed Shaharyar Khwaja. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10:100057, 2021. 1

[20] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xBD: A Dataset for Assessing Building Damage from Satellite Imagery, 2019. arXiv:1911.09296 [cs]. 4

[21] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, Online, 2021. IEEE. 2

[22] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, pages 297–312, Zurich, Switzerland, 2014. Springer. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 2

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international*

conference on computer vision, pages 2961–2969, Venice, Italy, 2017. IEEE. 2, 5

[25] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *IEEE Conference on Robotics and Automation*, 2017. 2

[26] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 2

[27] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xView: Objects in Context in Overhead Imagery, 2018. arXiv:1802.07856 [cs]. 3

[28] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 2

[30] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 2

[31] 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, 2022. arXiv:2112.03857. 2, 5

[32] Lin Lin, Chaoqing Tang, Qiuhua Liang, Zening Wu, Xinling Wang, and Shan Zhao. Rapid urban flood risk mapping for data-scarce environments using social sensing and region-stable deep neural network. *Journal of Hydrology*, 617:128758, 2023. 1

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[34] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 3

[35] Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 475–485. Springer, 2020. 1

[36] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, 2024. arXiv:2303.05499. 2, 5

[37] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, pages 21–37, Amsterdam, The Netherlands, 2016. Springer. 2

[38] Wenchao Liu, Long Ma, Jue Wang, et al. Detection of multiclass objects in optical remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 16(5):791–795, 2018. 2

[39] Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, 2017. 2

[40] Yuchi Ma, Shuo Chen, Stefano Ermon, and David B. Lobell. Transfer learning in environmental remote sensing. *Remote Sensing of Environment*, 301:113924, 2024. 1

[41] Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. COCO-O: A Benchmark for Object Detectors under Natural Distribution Shifts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6316–6327, Paris, France, 2023. IEEE. 2

[42] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 2

[43] Daifeng Peng, Haiyan Guan, Yufu Zang, and Lorenzo Bruzzone. Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022. 1

[44] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 2

[45] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, Las Vegas, NV, USA, 2016. IEEE. 2

[47] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 2, 5

[48] Shaoqing Ren, Kaiming He, Ross Girshick, Xiangyu Zhang, and Jian Sun. Object detection networks on convolutional feature maps. *arXiv:1504.06066 (v2)*, 2016. 2

[49] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling Climate Change with Machine Learning. *ACM Computing Surveys*, 55(2):1–96, 2023. 3

[50] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 2

[51] Rizwan Sadiq, Zainab Akhtar, Muhammad Imran, and Ferda Ofli. Integrating remote sensing and social sensing for flood mapping. *Remote Sensing Applications: Society and Environment*, 25:100697, 2022. 1

[52] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR 2014)*, Banff, AB, Canada, 2014. CBLS. 2

[53] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 2

[54] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. In *Advances in Neural Information Processing Systems 26*, pages 2553–2561, Lake Tahoe, NV, USA, 2013. Curran Associates, Inc. 2

[55] Christian Szegedy, Scott E. Reed, Dumitru Erhan, Dragomir Anguelov, and Sergey Ioffe. Scalable, high-quality object detection. *arXiv:1405.0312 (v3)*, 2015. 2

[56] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 2

[57] Z Tian, C Shen, H Chen, and T He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, 2019. 2

[58] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 3

[59] Riccardo Volpi and Vittorio Murino. Addressing Model Vulnerability to Distributional Shifts Over Image Transformation Sets. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7979–7988, Seoul, Korea (South), 2019. IEEE. 1

[60] Li-Hua Wen and Kang-Hyun Jo. Deep learning-based perception systems for autonomous driving: A comprehensive survey. *Neurocomputing*, 2022. 1

[61] Wikipedia. Köppen climate classification, 2024. Page Version ID: 1256836310. 1, 3

[62] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*, pages 207–224. Springer, 2024. 2

[63] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, Salt Lake City, UT, USA, 2018. IEEE. 2

[64] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591, 2017. 5

[65] Tao Xu, Xian Sun, Wenhui Diao, Liangjin Zhao, Kun Fu, and Hongqi Wang. Fada: Feature aligned domain adaptive object detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 1

[66] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8232–8241, Long Beach, CA, USA, 2019. IEEE. 2

[67] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection, 2022. arXiv:2203.03605 [cs]. 2, 5

[68] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *European conference on computer vision*, pages 163–180. Springer, 2022. 2

[69] Bingchen Zhao, Jiahao Wang, Wufei Ma, Artur Jesslen, Siwei Yang, Shaozuo Yu, Oliver Zendel, Christian Theobalt, Alan Yuille, and Adam Kortylewski. Ood-cv-v2: An extended benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[70] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 850–859, 2019. 2

[71] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 2

[72] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object Detection in 20 Years: A Survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. Conference Name: Proceedings of the IEEE. 5

Benchmarking Object Detectors under Real-World Distribution Shifts in Satellite Imagery

Supplementary Material

A. Mathematical Formulation of DG Setups

Let X and Y be the input and target spaces, with domain D having joint distribution P_{XY} on $X \times Y$. DG aims to learn a model $f : X \rightarrow Y$ from source data that minimises error on both source (ID) and target (OOD) test data.

Single-source domain generalisation. We assume that there is only one source domain, D_s, where s represents a unique source available during the training phase. Therefore, the training set, D_{train}, is defined as follows:

$$D_{train} = D_s = \{(x_i, y_i)\}_{i=1}^{M} \tag{3}$$

where x_i, y_i being the i^{th} sample and label pairs from the source domain and M indicating the total number of training samples. Furthermore, D_s is associated with a joint distribution P_{XY}^s.

Multi-source domain generalisation. We consider a training scenario with access to N distinct yet related source domains, denoted as D_s for $s \in \{1, ..., N\}$. Accordingly, the training set is defined as D_{train}:

$$D_{train} = \bigcup_{s=1}^{N} D_s \tag{4}$$

$$D_s = \{(x_i^s, y_i^s)\}_{i=1}^{M_s}$$

here, x_i^s represents the i^{th} sample with label y_i^s, and M_s denotes the total number of training samples in domain D_s. Each source domain D_s is characterised by a joint distribution P_{XY}^s. While the distributions across source domains may be related, they are not equivalent, i.e., $P_{XY}^s \neq P_{XY}^{s'}$ for $s \neq s'$, where $s, s' \in \{1, ..., N\}$.

The target domain. We define the OOD target domain(s) as D_t, where t represents a target domain distinct from the source domains ($t \neq s$). The target domain follows a joint distribution P_{XY}^t that differs from all source distributions, i.e., $P_{XY}^t \neq P_{XY}^s$, $\forall s \in \{1, ..., N\}$. Accordingly, the test set is defined as:

$$D_{test} = \{D_t | t \in \{1, ..., K\}\} \tag{5}$$

$$D_t = \{(x_i^t, y_i^t)\}_{i=1}^{M_t}$$

where K denotes the total number of target domains, x_j^t is the j^{th} sample with label y_j^t, and M_t signifies the total number of test samples from the target domain D_t.

B. Detailed Benchmarking Results

In this section, we provide a detailed breakdown of the experiments conducted in our paper, focusing on the individual domains. We begin by analysing the domain shift experienced by the selected object detectors, as discussed in Section 4.3, across different climate zones using RWDS-CZ in Section B.1 under both single- and multi-source setups. Similarly, we discuss the findings related to the generalisation capabilities of the object detectors across flooded regions using RWDS-FR under the single-source setup in Section B.2. Additionally, Section B.3 presents an examination of the impact of domain shift on the performance of object detectors across various hurricane events in RWDS-HE, also under both single- and multi-source setups.

The Upper Bound Experiments. To establish a baseline for evaluating model performance under the best-case scenario—where the i.i.d. assumption holds and the model is only tested on samples from the same underlying distribution seen during training, we present the upper-bound (**UB**) experimental results for RWDS-CZ, RWDS-FR, and RWDS-HE. More specifically, these experiments represent the oracle setup, in which an object detector is trained on the training set from all domains, including the target domain, and evaluated on the test set of each domain respectively.

B.1. Further Analyses of RWDS-CZ Experiments

In Section 5.1, we investigated the performance of the selected SOTA object detectors on RWDS-CZ, showing that there exist a shift in the underlying distribution of data gathered from different climate zone. To gain a better intuition on the relationship between these domains, if any, and the influence of domain shift across the different climate zones, we provide a fine-grained analyses of domain shift under the single- (Section B.1.1) and multi-source setups (Section B.1.3)along with a qualitative assessment (Section B.1.2) .

B.1.1 Single-Source DG Experiment

Table S1 presents the results of the single-source experiment using mAPs over different IoU regions, namely, mAP_{50}, mAP_{75} and $\text{mAP}_{50:95}$. The overall trends discussed in Section 5.1 remain consistent across evaluations using mAP50, mAP75, and $\text{mAP}_{50:95}$.

Furthermore, Table S2 illustrates the performance of object detectors on UB, which is underlined in the table, CZ A, CZ B and CZ C for each of the six object detectors under the

Metric	Methods	Target											
		CZ A				CZ B				CZ C			
		mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
mAP_{50}	Faster R-CNN	16.7	9.3	45	11.9	15.7	13.0	17	14.2	17.4	8.6	51	11.5
	Mask R-CNN	16.9	9.2	46	11.9	16.3	12.6	23	14.2	17.3	8.8	49	11.7
	TOOD	17.1	9.3	46	12.0	15.4	12.5	19	13.8	17.3	9.4	46	12.2
	DINO	25.2	13.8	45	17.8	19.2	16.6	14	17.8	24.3	14.0	43	17.7
	Grounding DINO	**27.9**	**17.0**	39	**21.1**	**21.5**	**20.1**	**7**	**20.8**	**28.1**	**16.7**	41	**20.9**
	GLIP	20.7	13.4	**35**	16.3	17.1	15.8	8	16.4	19.2	12.1	**37**	14.8
mAP_{75}	Faster R-CNN	5.2	2.5	52	3.4	5.9	4.5	24	5.1	5.3	2.0	63	2.9
	Mask R-CNN	4.9	2.3	53	3.1	5.9	4.1	31	4.8	5.7	2.0	66	2.9
	TOOD	5.8	2.6	55	3.6	6.8	5.1	26	5.8	6.8	2.7	61	3.8
	DINO	8.2	4.0	52	5.3	9.0	6.9	23	7.8	9.2	4.0	57	5.5
	Grounding DINO	**11.0**	**6.1**	45	**7.8**	**10.2**	**9.1**	11	**9.6**	**11.5**	**5.4**	53	**7.3**
	GLIP	8.0	5.0	**38**	6.1	8.0	7.4	**7**	7.7	7.5	4.1	**46**	5.3
$mAP_{50:95}$	Faster R-CNN	7.2	3.9	47	5.0	7.5	6.0	20	6.7	7.7	3.4	56	4.7
	Mask R-CNN	7.3	3.7	49	4.9	7.7	5.8	25	6.6	7.8	3.5	55	4.8
	TOOD	7.8	4.0	49	5.2	7.8	6.1	22	6.8	8.2	4.0	52	5.3
	DINO	11.0	5.6	49	7.4	9.6	8.0	17	8.7	11.0	5.6	49	7.4
	Grounding DINO	**12.9**	**7.5**	42	**9.5**	**10.8**	**10.0**	**7**	**10.4**	**13.1**	**7.1**	46	**9.2**
	GLIP	9.8	6.3	**36**	7.6	8.8	8.2	**7**	8.5	9.2	5.4	**41**	6.8

Table S1. Single-source DG analysis of SOTA detectors on RWDS-CZ where ID/OOD denotes the mAP scores over different IoUs.

single-source setup. The diagonal, indicated in bold, highlights their ID performance. Aligned with the observations made in Section 5.1.1, there is always a performance drop when testing the OOD test sets. Furthermore, the performances on the UB is always higher than not only the OOD but also the ID. A plausible explanation is that the models benefits from being exposed to a more diverse data distributions during training which makes them more robust in comparison to training on a single source domain.

B.1.2 Qualitative DG Performance Comparison

In order to gain insights on the quality and behaviour of the object detector among the different domains, we select the highest performing object detector, Grounding DINO, and sample the output predictions under the single-source setup. A set of these are illustrated in Figure S1, where the diagonal images, highlighted in purple, indicate the performance on the ID test samples. It is important to note that we selected the samples with few number of bounding boxes for visualisation purposes.

- **CZ A**: When tested on CZ A, aligned with the results found in Table S2, one can observe that the best performing model in comparison to the ground truth is the one trained on the ID training set. Whereas, the object detectors trained on CZ B and CZ C miss detecting a building.
- **CZ B**: When evaluated on CZ B, in-line with the results found in Table S2, the best performing model in comparison to the ground truth is the one trained on the ID training set. Whereas, the object detectors trained on CZ B and CZ C miss detecting a number of buildings.
- **CZ C**: When tested on CZ C, similar to the findings mentioned in the previous points and the results presented in Table S2, it can be observed that the best performing model in comparison to the ground truth is the one trained on the ID training set. Whereas, the object detectors trained on CZ A and CZ B miss detecting a number

Methods	Source	Target		
		CZ A	CZ B	CZ C
Faster R-CNN	UB	8.1	9.0	7.5
	CZ A	**7.2**	5.3	3.5
	CZ B	3.6	**7.5**	3.3
	CZ C	4.1	6.7	**7.7**
Mask R-CNN	UB	7.8	9.0	7.6
	CZ A	**7.3**	5.0	3.6
	CZ B	3.6	**7.7**	3.4
	CZ C	3.8	6.5	**7.8**
TOOD	UB	8.2	9.1	8.4
	CZ A	**7.8**	5.0	4.2
	CZ B	3.8	**7.8**	3.7
	CZ C	4.1	7.1	**8.2**
DINO	UB	12.2	12.5	11.8
	CZ A	**11.0**	7.5	5.9
	CZ B	5.1	**9.6**	5.3
	CZ C	6.1	8.4	**11.0**
Grounding DINO	UB	13.5	13.8	12.9
	CZ A	**12.9**	8.6	7.2
	CZ B	6.9	**10.8**	6.9
	CZ C	8.1	11.4	**13.1**
GLIP	UB	11.1	10.7	10.1
	CZ A	**9.8**	7.2	5.7
	CZ B	5.8	**8.8**	5.1
	CZ C	6.7	9.2	**9.2**

Table S2. $mAP_{50:95}$ results on RWDS-CZ for the single-source setup.

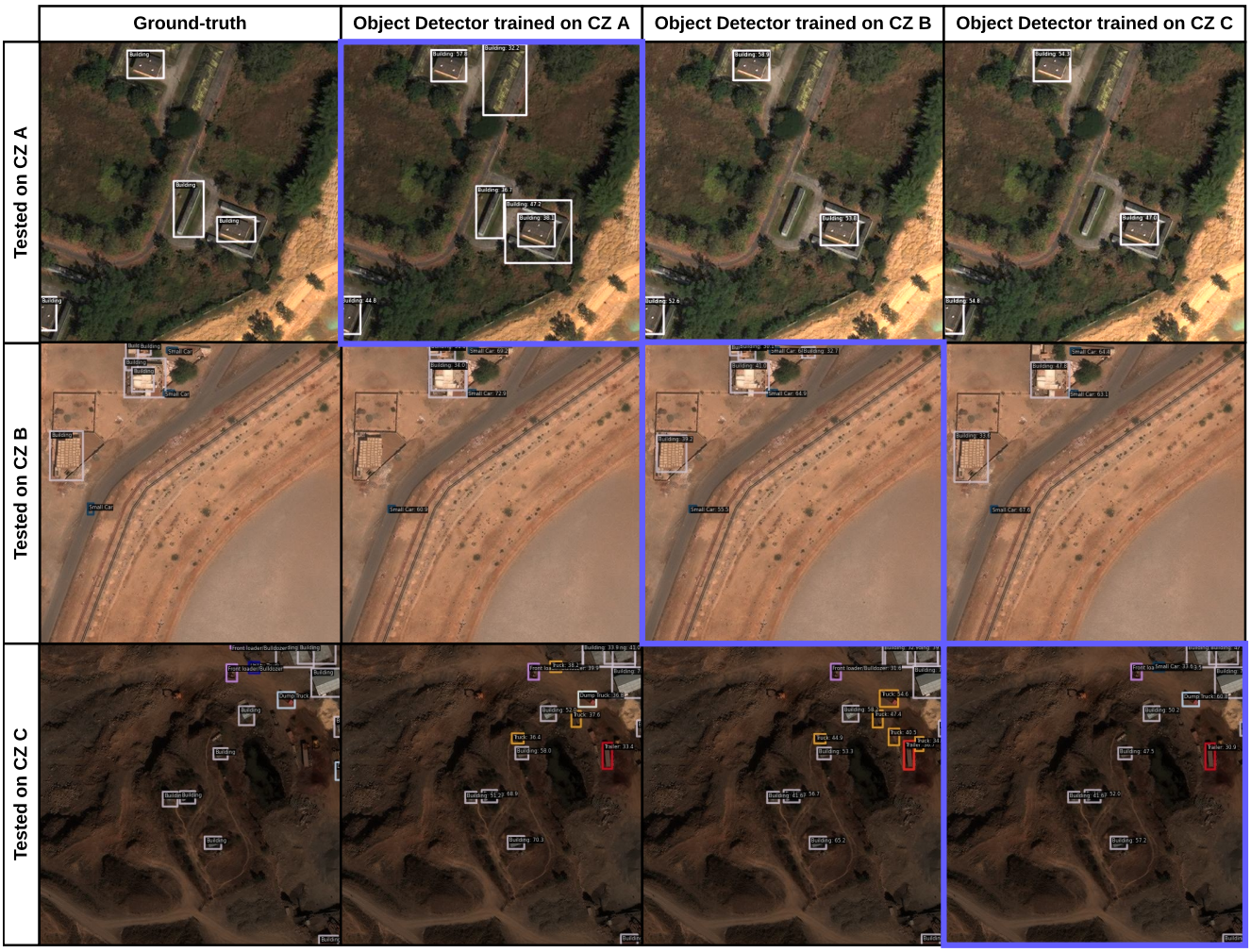


Figure S1. Qualitative DG performance comparison of Grounding DINO among different climate zones, where the diagonal images highlighted in purple indicate the performance on the ID test sample.

of buildings and have higher rates of false negatives and false positives.

B.1.3 Multi-Source DG Experiments

Table S3 presents the results of the multi-source experiment using mAPs over different IoU regions, namely, mAP_{50}, mAP_{75} and $mAP_{50:95}$, where general trends presented in Section 5.1 are consistently observed across evaluations utilising mAP_{50}, mAP_{75}, and $mAP_{50:95}$.

Moreover, Table S4 presents the performance of the six object detectors under the multi-source setup, where an object detector is trained on a collection of source domains and tested on the individual ID test sets in addition to the left out target domain's test set. The diagonal, indicated in bold, highlights their OOD performance.

Unlike the observations made in the single-source setup where the model trained on the UB always had the highest performance, it can be observed from Table S4 that this is not always the case. For example, when trained on the collection of source domains excluding CZ A, Faster R-CNN, GLIP and Grounding DINO achieve an outstanding performance on the ID test set of CZ C in comparison to the UB. This suggests that eliminating CZ A from training actually improves the ID performance of the models on CZ C. A possible explanation to this phenomena is that the distribution of CZ A is quite different than that of CZ B and CZ C. A similar pattern is observed for multiple other combination of domains and methods, as shown in Table S4.

B.2. Further Analyses of RWDS-FR Experiments

In Section 5.2, we evaluated the performance of the selected object detectors on RWDS-FR, highlighting the existence of distribution shifts in data originating from different flooded regions. To better understand the potential

Metric	Methods	Target											
		CZ A				CZ B				CZ C			
		mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
mAP_{50}	Faster R-CNN	17.5	11.6	34	13.9	17.0	14.6	14	15.7	17.5	10.2	42	12.9
	Mask R-CNN	17.2	11.0	36	13.4	16.7	14.5	13	15.5	17.9	10.6	41	13.3
	TOOD	17.9	11.3	37	13.8	16.6	14.3	14	15.3	17.5	11	37	13.5
	DINO	25.9	17.2	34	20.7	22.2	18.9	15	20.4	25.8	16.6	36	20.2
	Grounding DINO	**28.9**	**19.8**	31	**23.5**	**23.9**	**21.7**	9	**22.7**	**27.7**	**21.1**	**24**	**24.0**
	GLIP	22.6	16.7	**26**	19.2	18.4	17.4	**5**	17.9	20.3	14.6	28	17.0
mAP_{75}	Faster R-CNN	5.4	3.4	37	4.2	6.6	5.4	18	5.9	5.5	2.3	58	3.2
	Mask R-CNN	5.2	3.2	38	3.9	6.6	5.4	18	5.9	5.8	2.6	55	3.6
	TOOD	6.4	3.6	43	4.6	7.7	5.8	25	6.6	6.8	3.4	50	4.5
	DINO	9.2	5.4	41	6.8	10.8	8.8	18	9.7	9.9	5.5	44	7.1
	Grounding DINO	**10.9**	**6.9**	36	**8.4**	**12.1**	**10.4**	14	**11.2**	**11.9**	**7.5**	**37**	**9.2**
	GLIP	8.6	6.7	**22**	7.5	9.2	8.3	**10**	8.7	8.3	5.2	**37**	6.4
$mAP_{50:95}$	Faster R-CNN	7.7	4.9	36	6.0	8.2	7.1	13	7.6	7.7	4.1	47	5.4
	Mask R-CNN	7.5	4.7	37	5.8	8.1	6.9	15	7.5	7.9	4.3	46	5.6
	TOOD	8.2	5.0	39	6.2	8.7	7.0	19	7.7	8.3	4.8	42	6.1
	DINO	11.6	7.2	38	8.9	11.5	9.6	16	10.4	11.8	7.0	40	8.8
	Grounding DINO	**13.1**	**8.8**	33	**10.5**	**12.5**	**11.0**	12	**11.7**	**13.1**	**9.3**	**29**	**10.9**
	GLIP	10.6	8.0	**24**	9.1	9.8	9.2	**6**	9.5	9.8	6.8	31	8.0

Table S3. Multi-source DG analysis of SOTA detectors on RWDS-CZ where ID/OOD denotes the mAP scores over different IoUs.

Methods	Source	Target		
		CZ A	CZ B	CZ C
Faster R-CNN	UB	8.1	9.0	7.5
	Unseen CZ A	**4.9**	9.1	7.7
	Unseen CZ B	7.6	**7.1**	7.7
	Unseen CZ C	7.7	7.3	**4.1**
Mask R-CNN	UB	7.8	9.0	7.6
	Unseen CZ A	**4.7**	8.8	8.0
	Unseen CZ B	7.5	**6.9**	7.8
	Unseen CZ C	7.4	7.4	**4.3**
TOOD	UB	8.2	9.1	8.4
	Unseen CZ A	**5.0**	9.2	8.3
	Unseen CZ B	8.3	**7.0**	8.3
	Unseen CZ C	8.0	8.1	**4.8**
DINO	UB	12.2	12.5	11.8
	Unseen CZ A	**7.2**	11.8	11.3
	Unseen CZ B	12.1	**9.6**	12.2
	Unseen CZ C	11.0	11.1	**7.0**
Grounding DINO	UB	13.5	13.8	12.9
	Unseen CZ A	**8.8**	12.9	13.3
	Unseen CZ B	12.8	**11.0**	12.8
	Unseen CZ C	13.4	12.1	**9.3**
GLIP	UB	11.1	10.7	10.1
	Unseen CZ A	**8.0**	10.2	10.3
	Unseen CZ B	10.2	**9.2**	9.3
	Unseen CZ C	10.9	9.4	**6.8**

Table S4. $mAP_{50:95}$ results on RWDS-CZ for the multi-source setup.

relationships between these domains and the effects of domain shifts across various flooded regions, we provide a detailed analyses of the domain shift under the single-source setup (Section B.2.1)accompanied by a qualitative assessment and discussion (Section B.2.2) below.

B.2.1 Single-Source DG Experiment

As mentioned in Section 5.2, RWDS-FR inherently falls under the single-source setup given that it consist of two domains. Table S5 presents the results of the single-source experiment using mAPs over different IoU regions, namely, mAP_{50}, mAP_{75} and $mAP_{50:95}$. The patterns outlined in Section 5.2 are observed across evaluations using mAP50, mAP75 (with minor variations), and $mAP_{50:95}$.

Furthermore, Table S6 showcases the breakdown of each object detector's performance on the ID and OOD test sets. The bolded diagonal indicates their in-domain performance. While the model trained on India maintains its performance, when comparing the single-source performance versus the UB, the model trained on the US performs slightly better than the UB when evaluated on the ID test set. A plausible explanation for such a behaviour is that the training set of India is naturally difficult and its distribution is further away in the latent space from that of the US, thus hurting the model's ID performance when combined during the training phase. Moreover, aligned with the observations made in Section 5.2, the OOD performance of the model trained on India on the US test set is notably low, highlighting the existence of a significant domain shift between the two domains.

B.2.2 Qualitative DG Performance Comparison

Figure S2 illustrates the performance on the ID and OOD test sets of the best performing object detector, Grounding DINO, where the diagonal samples highlighted in purple indicate the performance on the ID test sample. It is

Metric	Methods	Target							
		India				US			
		mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
mAP_{50}	Faster R-CNN	14.7	3.8	74	6.0	56.8	4.6	92	8.5
	Mask R-CNN	14.8	3.7	75	5.9	56.7	4.9	91	9.0
	TOOD	17.2	5.2	70	8.0	59.2	6.0	90	10.9
	DINO	**24.0**	8.8	63	12.9	64.6	13.4	79	22.2
	Grounding DINO	23.3	**12.5**	**46**	**16.3**	**67.7**	**31.3**	54	**42.8**
	GLIP	20.5	11.0	**46**	14.3	64.0	31.0	**52**	41.8
mAP_{75}	Faster R-CNN	1.7	0.8	53	1.1	19.6	1.1	94	2.1
	Mask R-CNN	1.5	0.5	67	0.8	20.2	1.3	94	2.4
	TOOD	1.5	0.8	**47**	1.0	22.1	1.5	93	2.8
	DINO	**2.9**	0.6	79	1.0	**26.7**	1.9	93	3.5
	Grounding DINO	2.6	1.0	62	1.4	25.8	5.1	80	8.5
	GLIP	**2.9**	**1.1**	62	**1.6**	25.4	**6.6**	**74**	**10.5**
$mAP_{50:95}$	Faster R-CNN	4.5	1.3	71	2.0	25.5	1.8	93	3.4
	Mask R-CNN	4.3	1.2	72	1.9	25.9	2.0	92	3.7
	TOOD	5.1	1.6	69	2.4	27.6	2.4	91	4.4
	DINO	**7.0**	2.2	69	3.3	30.8	4.3	86	7.5
	Grounding DINO	6.7	**3.3**	**51**	**4.4**	**31.3**	10.8	65	16.1
	GLIP	6.7	**3.3**	**51**	**4.4**	30.7	**11.9**	**61**	**17.2**

Table S5. Single-source DG analysis of SOTA detectors on RWDS-FR where ID/OOD denotes mAP scores over different IoUs.

Methods	Source	Target	
		India	United States
Faster R-CNN	UB	4.5	25.2
	India	**4.5**	1.8
	United States	1.3	**25.5**
Mask R-CNN	UB	4.4	25.8
	India	**4.3**	2.0
	United States	1.2	**25.9**
TOOD	UB	5.1	27.4
	India	**5.1**	2.4
	United States	1.6	**27.6**
DINO	UB	7.0	30.7
	India	**7.0**	4.3
	United States	2.2	**30.8**
Grounding DINO	UB	6.9	31.2
	India	**6.7**	10.8
	United States	3.3	**31.3**
GLIP	UB	6.5	30.8
	India	**6.7**	11.9
	United States	3.3	**30.7**

Table S6. $mAP_{50:95}$ results on RWDS-FR for the single-source setup.

worth noting that samples with a limited number of bounding boxes were deliberately chosen to facilitate visualization and enhance clarity in explanation.

It is evident, from Table S2, that the model trained on India and tested on the ID test set misses a number of bounding boxes. Similarly, the model trained on the US and tested on the OOD test set from India, not only misses a number of bounding boxes, but also consists of false positive detections. However, when evaluated on the test set from the US, its ID performance is closer to the ground-truth. Furthermore, a drop in OOD performance of the model trained on India is observed on when evaluated on OOD US test set, where the model fails in detecting a number of bounding boxes. These observations are aligned with the results previously reported in Table S2.

B.3. Further Analyses of RWDS-HE Experiments

In Section 5.3, we analysed the performance of the selected SOTA object detectors on RWDS-HE, emphasising the presence of a distribution shift in data collected from different hurricane events. To gain deeper insights into the potential relationships between these domains and the impact of domain shifts across various hurricane events, we present fine-grained analyses of the object detectors' performance under the single- (Section B.3.1 and multi-source (Section B.3.3) setups,

B.3.1 Single-Source DG Experiment

Table S7 presents the results of the single-source experiment using mAPs over different IoU regions, namely, mAP_{50}, mAP_{75} and $mAP_{50:95}$. The general trends presented in Section 5.3 are consistently observed, with minor variations, across evaluations utilising mAP_{50}, mAP_{75}, and $mAP_{50:95}$.

Furthermore, Table S8 outlines the performance of object detectors on UB, Hurricanes Florence, Michael, Harvey and Matthew under the single-source setup. The bolded diagonal indicates their ID performance. In line with the findings in Section 5.3.1, all object detectors experience performance degradation when tested on OOD test sets across all

Figure S2. Qualitative DG performance comparison of Grounding DINO among different flooded regions, namely, India and the US, where the diagonal images highlighted in purple indicate the performance on the ID test sample.

		Target															
		Florence				Michael				Harvey				Matthew			
Metric	Methods	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
mAP_{50}	Faster R-CNN	64.5	19.3	70	29.7	42.7	17.2	60	24.5	56.9	9.5	83	16.3	5.5	1.2	79	1.9
	Mask-CRNN	63.3	18.6	71	28.7	42.9	17.7	59	25.1	57.1	9.5	83	16.2	6.7	1.2	83	2.0
	TOOD	64.3	23.2	64	34.1	45.5	18.2	60	26.0	59.9	11.4	81	19.1	7.9	2.1	73	3.3
	DINO	66.5	25.9	61	37.2	46.5	19.3	59	27.2	65.8	13.0	80	21.8	9.4	2.8	71	4.3
	Grounding DINO	**70.6**	36.3	49	47.9	**52.8**	23.8	55	32.8	**67.9**	**20.1**	**70**	**31.0**	**12.5**	**4.5**	64	**6.6**
	GLIP	70.4	**37.2**	**47**	**48.7**	50.8	**24.6**	**52**	**33.1**	63.3	17.5	72	27.5	11.1	**4.5**	**59**	6.4
mAP_{75}	Faster R-CNN	33.2	6.5	80	10.9	13.8	3.6	74	5.7	19.1	2.1	89	3.8	0.3	0.1	67	0.2
	Mask-CRNN	33.7	6.1	82	10.3	14.2	4.1	71	6.4	19.7	2.1	89	3.8	0.4	0.1	75	0.2
	TOOD	35.8	7.8	78	12.8	17.0	4.4	74	7.0	22.3	2.7	88	4.8	1.0	0.1	87	0.2
	DINO	37.4	9.6	74	15.2	16.3	4.8	71	7.4	**27.3**	3.0	89	5.4	0.7	0.3	**62**	0.4
	Grounding DINO	40.4	14.5	64	21.3	19.6	5.4	72	8.5	25.2	**4.7**	81	**7.9**	1.1	0.3	70	0.5
	GLIP	**42.8**	**17.2**	**60**	**24.6**	**19.8**	**6.7**	**66**	**10.0**	22.9	4.5	**80**	7.6	**1.9**	**0.4**	77	**0.7**
$mAP_{50:95}$	Faster R-CNN	34.5	8.6	75	13.8	18.6	6.5	65	9.7	25.1	3.7	85	6.4	1.5	0.3	78	0.5
	Mask-CRNN	34.0	8.3	76	13.3	19.1	6.9	64	10.1	25.6	3.7	86	6.4	1.7	0.4	78	0.6
	TOOD	35.7	10.4	71	16.1	21.0	7.1	66	10.6	27.5	4.4	84	7.5	2.4	0.5	78	0.9
	DINO	36.5	12.0	67	18.0	20.6	7.6	63	11.1	**31.4**	4.9	84	8.5	2.5	0.8	69	1.2
	Grounding DINO	39.3	17.4	56	24.2	**24.2**	9.3	62	13.4	31.0	**7.7**	**75**	**12.4**	3.3	1.2	65	1.7
	GLIP	**40.8**	**19.0**	**53**	**25.9**	23.9	**10.2**	**57**	**14.3**	29.2	7.0	76	11.3	**3.7**	**1.3**	**64**	**2.0**

Table S7. Single-source DG analysis of SOTA detectors on RWDS-HE where ID/OOD denotes the mAP scores over different IoUs.

domains.

Generally, UB outperforms the other models on the test sets, which is an expected behaviour. However, it can be observed that for rare cases such as when a model, more specifically either of Faster R-CNN, Mask R-CNN or TOOD, is trained on Florence and evaluated on the ID test set, its performance is better than that of the UB. One possible interpretation is that the diversity provided by other distributions hurt the ID performance on Florence compared to training on Florence exclusively.

Furthermore, the results in Table S8 clearly show that the model trained on Hurricane Matthew exhibits the weakest performance on both ID and OOD test sets. A likely explanation for this poor performance is that the underlying dataset is challenging and may contain label noise or class imbalance due to the limited number of instances in the raw

Methods	Source	Target			
		Florence	Michael	Harvey	Matthew
Faster R-CNN	UB	33.2	19.3	25.1	1.9
	Florence	**34.5**	8.4	5.2	0.4
	Michael	8.7	**18.6**	4.8	0.2
	Harvey	14.4	6.9	**25.1**	0.4
	Matthew	2.8	4.3	1.1	**1.5**
Mask R-CNN	UB	32.8	19.3	25.5	1.7
	Florence	**34.1**	9.2	4.7	0.4
	Michael	8.1	**19.1**	4.8	0.3
	Harvey	13.5	7.0	**25.6**	0.4
	Matthew	3.3	4.4	1.5	**1.7**
TOOD	UB	34.1	19.8	27.7	2.4
	Florence	**35.7**	9.1	5.6	0.5
	Michael	11.4	**21.0**	6.2	0.7
	Harvey	16.1	7.3	**27.5**	0.4
	Matthew	3.7	5.0	1.3	**2.4**
DINO	UB	37.7	22.2	32.0	2.8
	Florence	**36.5**	9.6	6.7	1.0
	Michael	11.6	**20.6**	6.3	0.7
	Harvey	19.5	7.8	**31.4**	0.6
	Matthew	4.8	5.4	1.8	**2.5**
Grounding DINO	UB	40.4	24.7	32.2	3.0
	Florence	**39.3**	10.5	8.1	1.1
	Michael	18.2	**24.2**	10.2	1.4
	Harvey	23.7	9.4	**31.0**	1.0
	Matthew	10.4	7.9	4.9	**3.3**
GLIP	UB	41.0	24.2	31.1	3.2
	Florence	**40.8**	10.6	7.8	0.9
	Michael	17.9	**23.9**	9.0	1.4
	Harvey	26.3	10.3	**29.2**	1.7
	Matthew	12.8	9.6	4.3	**3.7**

Table S8. $mAP_{50:95}$ results on RWDS-HE for the single-source setup.

dataset. These factors, which are independent of domain shift, represent an open area of research and fall outside the scope of this paper.

B.3.2 Qualitative DG Performance Comparison

Figure S3 illustrates the performance of the best-performing object detector, Grounding DINO, on both ID and OOD test sets. The diagonal samples, highlighted in purple, represent the performance on the ID test samples. Notably, samples with a small number of bounding boxes were intentionally selected to aid visualization and facilitate for clarity in the explanation.

It can be observed that the ID performance across each domain closely matches the ground truth, consistent with the earlier findings from Table S8. However, in certain cases, such as when analysing the ID performance of the model trained on Hurricane Matthew, the model fails to detect several bounding boxes or makes incorrect detections.

Moreover, when examining the OOD performance, the models appear to make similar mistakes during detection. For instance, when testing on the Florence test set, the object detector trained on Florence performs exceptionally well, in contrast to the detectors trained on Michael, where the false negatives are notably higher or in even a worse case, Matthew, where the model fails to generalise effectively.

B.3.3 Multi-Source DG Experiments

Table S9 presents the results of the multi-source experiment using mAPs over different IoU regions, namely, mAP_{50}, mAP_{75} and $mAP_{50:95}$. The results across those three regions exhibit a similar trend to the one reported in Section 5.3.

Moreover, Table S10 presents the performance of the object detectors under the multi-source setup, where each detector is trained on a combination of source domains and tested on both the individual ID test sets and the excluded target domain test set. The diagonal, indicated in bold, highlights their OOD performance.

Similar to the observations in the previous section, we can see the domain shift experienced by the object detectors through the performance decline between ID and OOD test sets. Additionally, it is evident that in RWDS-HE, training on multiple domains helps improve the generalisation of the object detectors to OOD test sets, although this may slightly affect the average ID performance due to this trade-off. This is particularly noticeable when examining the OOD performance on Florence for GLIP and Grounding DINO.

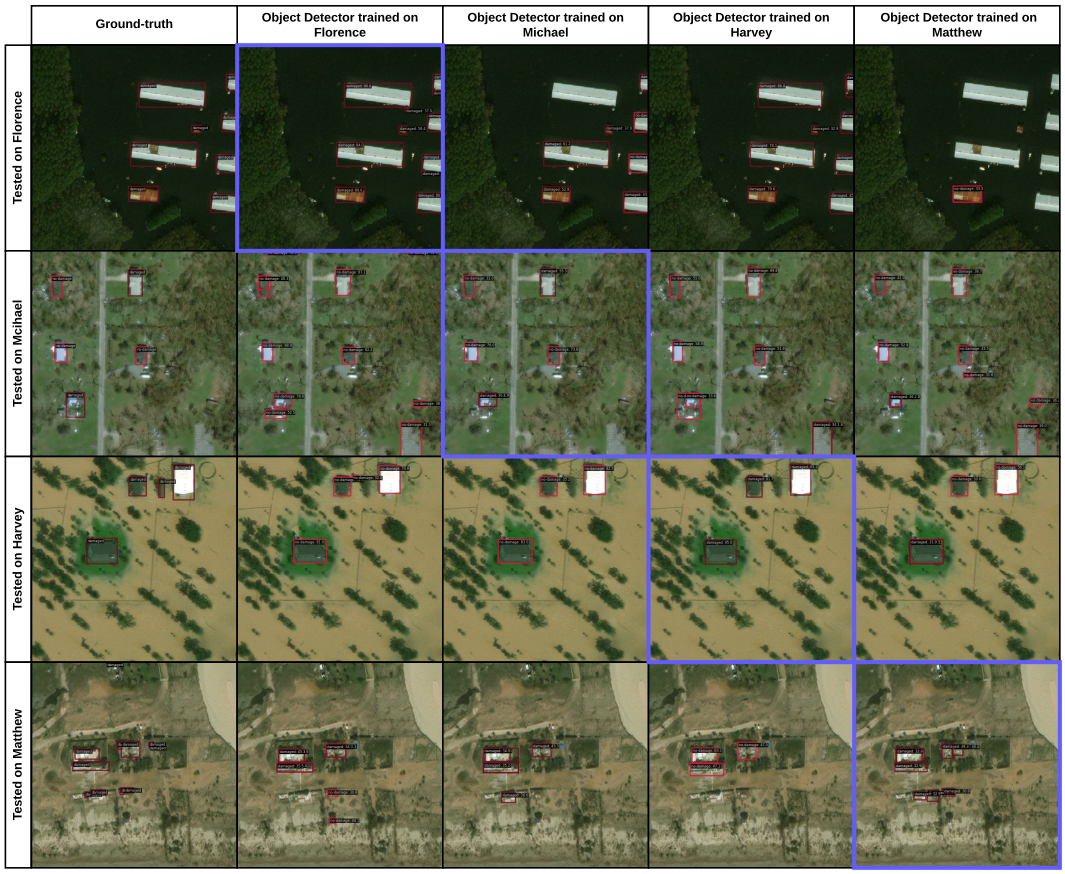


Figure S3. Qualitative DG performance comparison of Grounding DINO among different hurricane events, namely, hurricanes Florence, Michael, Harvey and Matthew, where the diagonal images highlighted in purple indicate the performance on the ID test sample.

		Target															
		Florence				Michael				Harvey				Matthew			
Metric	Methods	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑	mAP_{ID}	mAP_{OOD}	PD ↓	H ↑
		IoU: 0.50															
mAP_{50}	Faster R-CNN	63.1	27.5	56	38.3	43.3	20.8	52	28.1	56.6	13.2	77	21.4	6.0	1.4	77	2.3
	Mask-CRNN	63.8	28.2	56	39.1	43.9	21.0	52	28.4	57.8	14.0	76	22.5	5.8	2.1	64	3.1
	TOOD	62.7	29.7	53	40.3	43.5	22.6	48	29.8	59.6	13.6	77	22.1	7.3	1.9	74	3.0
	DINO	68.2	35.2	48	46.4	47.3	24.7	48	32.5	65.9	18.7	72	29.1	10.6	3.2	70	4.9
	Grounding DINO	70.8	53.4	25	60.9	**52.7**	**29.0**	**45**	**37.4**	**69.1**	**23.2**	**66**	**34.7**	**11.4**	**5.6**	**51**	**7.5**
	GLIP	**71.0**	**55.8**	**21**	**62.5**	51.0	25.2	51	33.7	65.6	18.4	72	28.7	10.2	3.9	62	5.6
mAP_{75}	Faster R-CNN	31.4	9.9	68	15.1	14.5	6.4	56	8.9	18.6	3.2	83	5.5	0.5	0.1	80	0.2
	Mask-CRNN	32.2	11.4	65	16.8	14.6	6.6	55	9.1	19.6	3.1	84	5.4	0.4	0.2	**45**	0.3
	TOOD	33.3	11.6	65	17.2	15.7	6.9	56	9.6	21.7	3.1	86	5.4	0.6	0.2	68	0.3
	DINO	37.1	14.6	61	21.0	17.3	7.8	55	10.8	**27.5**	5.1	81	8.6	1.0	0.3	71	0.5
	Grounding DINO	40.4	28.3	30	33.3	19.8	**9.6**	**51**	**12.9**	26.9	**6.3**	**77**	**10.2**	1.2	**0.5**	58	**0.7**
	GLIP	**42.8**	**30.7**	**28**	**35.8**	**20.5**	9.0	56	12.5	25.3	5.4	79	8.9	**1.4**	0.4	71	0.6
$mAP_{50:95}$	Faster R-CNN	32.8	12.7	61	18.3	19.0	8.9	53	12.1	25.0	5.2	79	8.6	1.7	0.4	76	0.6
	Mask-CRNN	33.6	13.3	60	19.1	19.3	9.1	53	12.4	25.8	5.4	79	8.9	1.6	0.7	56	1.0
	TOOD	34.2	14.0	59	19.9	19.7	9.6	51	12.9	27.2	5.3	81	8.9	2.2	0.5	77	0.8
	DINO	37.3	17.0	54	23.3	21.4	10.6	50	14.2	31.3	7.7	75	12.4	2.8	0.8	71	1.2
	Grounding DINO	39.6	28.2	29	32.9	**24.3**	**12.8**	**47**	**16.8**	**32.2**	**9.4**	**71**	**14.5**	3.1	**1.5**	**52**	**2.0**
	GLIP	**40.8**	**30.7**	**25**	**35.0**	**24.3**	11.4	53	15.5	30.9	7.8	75	12.5	**3.2**	1.1	66	1.6

Table S9. Multi-source DG analysis of SOTA detectors on RWDS-HE where ID/OOD denotes the mAP scores over different IoUs.

Methods	Source	Target			
		Florence	Michael	Harvey	Matthew
Faster R-CNN	UB	33.2	19.3	25.1	1.9
	Un. Florence	**12.7**	18.9	25.1	1.9
	Un. Michael	32.4	**8.9**	25.0	1.5
	Un. Harvey	32.9	19.0	**5.2**	1.7
	Un. Matthew	33.1	19.2	24.8	**0.4**
Mask R-CNN	UB	32.8	19.3	25.5	1.7
	Un. Florence	**13.3**	19.2	25.4	1.6
	Un. Michael	33.7	**9.1**	25.9	1.5
	Un. Harvey	33.5	19.3	**5.4**	1.7
	Un. Matthew	33.6	19.4	26.1	**0.7**
TOOD	UB	34.1	19.8	27.7	2.4
	Un. Florence	**14.0**	19.7	27.2	2.2
	Un. Michael	34.3	**9.6**	27.2	2.2
	Un. Harvey	34.6	19.6	**5.3**	2.1
	Un. Matthew	33.6	19.9	27.2	**0.5**
DINO	UB	37.7	22.2	32.0	2.8
	Un. Florence	**17.0**	21.5	31.4	2.7
	Un. Michael	37.6	**10.6**	31.4	2.9
	Un. Harvey	37.2	21.2	**7.7**	2.7
	Un. Matthew	37.0	21.5	31.1	**0.8**
Grounding DINO	UB	40.4	24.7	32.2	3.0
	Un. Florence	**28.2**	24.2	32.2	3.0
	Un. Michael	38.4	**12.8**	32.0	2.8
	Un. Harvey	40.3	24.3	**9.4**	3.5
	Un. Matthew	40.1	24.3	32.3	**1.5**
GLIP	UB	41.0	24.2	31.1	3.2
	Un. Florence	**30.7**	24.0	30.8	3.5
	Un. Michael	40.1	**11.4**	30.7	3.1
	Un. Harvey	41.3	24.5	**7.8**	3.1
	Un. Matthew	40.9	24.3	31.3	**1.1**

* Un. means Unseen

Table S10. $mAP_{50:95}$ results on RWDS-HE for the multi-source setup.