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.

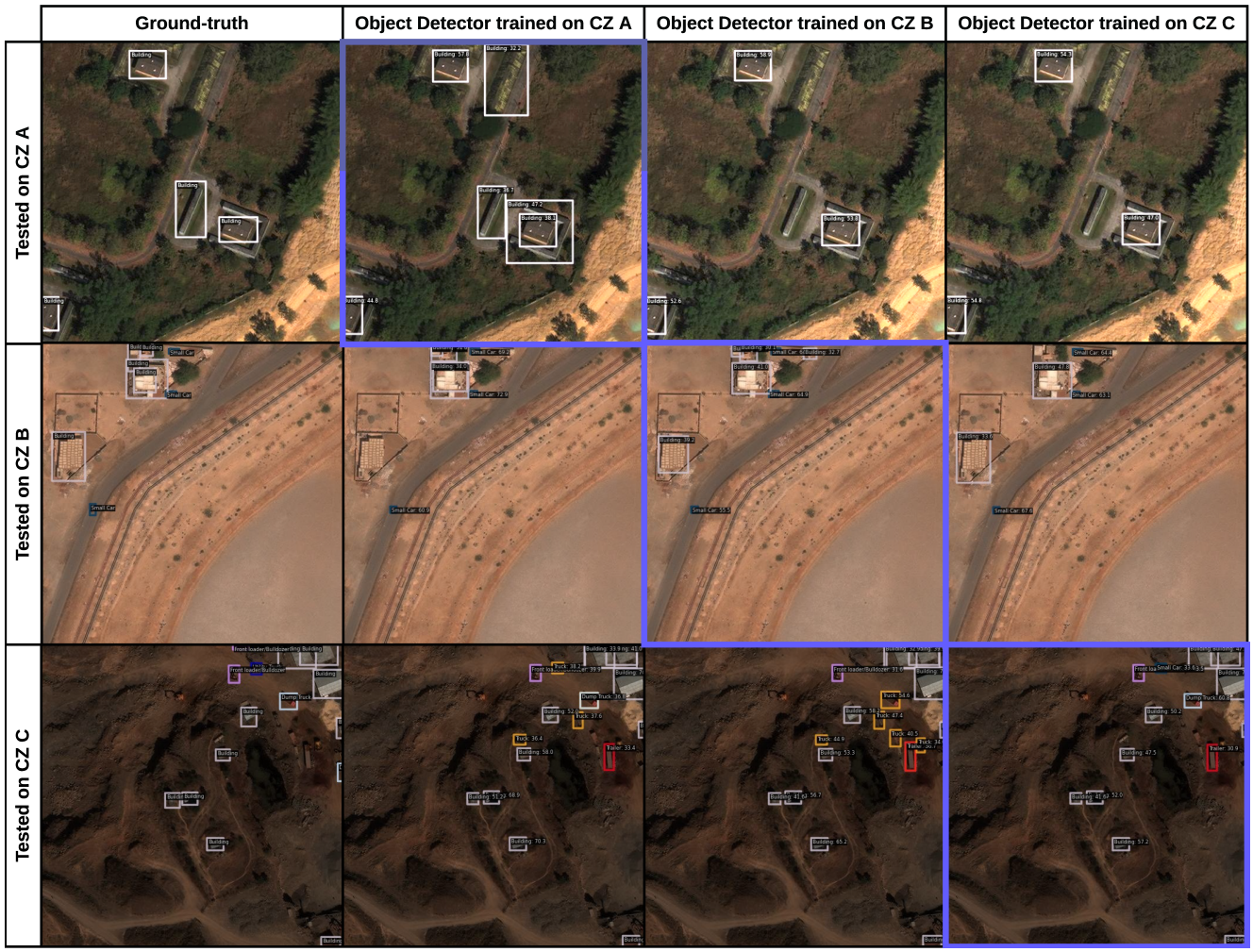

Figure S1. Qualitative DG performance comparison of Grounding DINO among different climate zones, where the diagonal images highlighted in purple indicate the performance on the ID test sample.

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 alongside a qualitative assessment (Section B.3.2) below.

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 perfor-

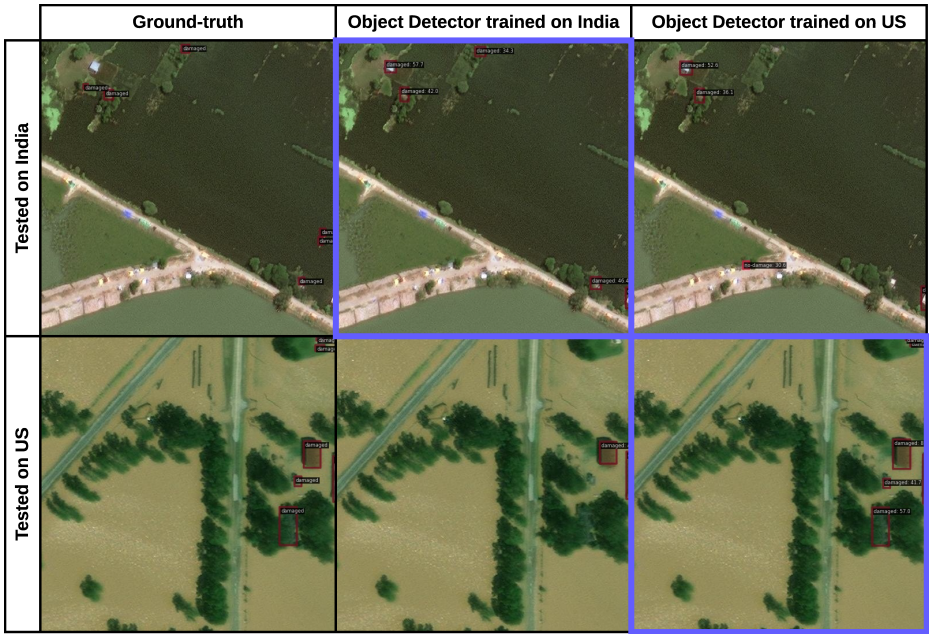

Figure S2. Qualitative DG performance comparison of Grounding DINO among different flooded regions, namely, India and the US, where the diagonal images highlighted in purple indicate the performance on the ID test sample.

		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.

mance degradation when tested on OOD test sets across all 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

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.

imbalance due to the limited number of instances in the raw 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.

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