Supplementary material for "Open-vocabulary vs. Closed-set: Best Practice for Few-shot Object Detection Considering Text Describability"

Anonymous Author(s)
Affiliation
Address
email

A Access to Our Code

All the code used in our experiments is available at: https://github.com/rsCPSyEu/ovd_cod

B Detailed Design of Our Benchmark Test

Our benchmark test is designed by repurposing the ODinW (Object Detection in the Wild) datasets [1]. The 35 included datasets are partitioned into three splits based on their CLIP zero-shot performance, as explained in Sec. 3.3.

Table 4 shows the list of datasets included in each split and the CLIP classification accuracy for each dataset. See also Fig. 2 in the main paper. The datasets highlighted in red indicate the 13 datasets that existing OVD studies employ for evaluation. Most of these belong to S1 and S2, suggesting that their class names are relatively easy to describe in text. S3 consists only of datasets with classes that are difficult to describe.

Table 4: A list of 35 Datasets in ODinW [1] with corresponding CLIP zero-shot accuracy. Red color shows 13 datasets that are typically used in previous methods [2].

S1(#12)		S2(#12)		S3(#11)	
Dataset	Acc.	Dataset	Acc.	Dataset	Acc.
CottontailRabbits	100.0	ShellfishOpenImages	45.3	HardHatWorkers	6.2
Packages	100.0	AerialMaritimeDrone(large)	38.2	AmericanSignLanguageLetters	5.7
Raccoon	100.0	AerialMaritimeDrone(tiled)	38.2	UnoCards	5.5
NorthAmericaMushrooms	93.3	SelfDrivingCar	33.7	BoggleBoards	4.5
Pothole	88.9	Plantdoc	26.2	WebsiteScreenshots	2.8
OxfordPets(breed)	87.8	BrackishUnderwater	24.7	ThermalDogsAndPeople	1.4
VehiclesOpenImages	71.7	Pistols	24.6	BCCD	0.9
MountainDewCommercial	59.7	MaskWearing	15.5	Dice	0.2
PASCAL VOC	57.6	EgoHands(generic)	10.0	OpenPoetryVision	0.1
OxfordPets(species)	56.6	ChessPiecesPieces	8.6	PKLot640	0.0
WildfireSmoke	52.7	ThermalCheetah	6.5	EgoHands(specific)	0.0
Aquarium	49.6	DroneControl	6.3		

In our experiments, we evaluate the object detection performance of multiple methods in a few-shot setting. For each dataset listed in Table 4, we randomly select few-shot training samples from the dataset's training split; the original validation/testing split is used for validation/testing. We repeat

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Table 5: Detailed configurations of the few-shot training per a dataset. The numbers separated by a slash represent the total number of images (#Img.) and the total number of bounding boxes (#Ann.) used for the few-shot training, respectively. Configurations for $K = 1$ and 3 are shown.

Dataset	Classes	$K = 1$ (#Img. / #Ann.)					$K = 3$ (#Img. / #Ann.)				
		seed=0	seed=1	seed=2	seed=3	seed=4	seed=0	seed=1	seed=2	seed=3	seed=4
CottontailRabbits	1	1/1	1/1	1/1	1/1	1/1	3/3	3/3	3/3	3/3	3/3
Packages	1	1/2	1/1	1/2	1/2	1/2	3/6	3/4	3/4	3/4	3/5
Raccoon	1	1/1	1/1	1/1	1/1	1/1	3/4	3/3	3/3	3/3	3/3
NorthAmericaMushrooms	2	2/8	2/2	2/4	2/3	2/2	6/14	6/7	6/9	6/15	6/7
Pothole	1	1/1	1/3	1/1	1/2	1/3	3/4	3/24	3/5	3/6	3/9
OxfordPets(breed)	37	37/37	37/37	37/37	37/37	37/37	111/111	111/111	111/111	111/111	111/111
VehiclesOpenImages	5	5/14	5/18	5/7	5/17	5/6	15/32	15/46	15/25	15/28	15/28
MountainDewCommercial	1	1/39	1/49	1/51	1/51	1/24	3/90	3/110	3/98	3/90	3/78
PASCAL VOC	20	17/35	17/42	18/41	19/54	16/40	50/126	51/143	51/134	55/156	51/132
OxfordPets(species)	2	2/2	2/2	2/2	2/2	2/2	6/6	6/6	6/6	6/6	6/6
WildfireSmoke	1	1/1	1/1	1/1	1/1	1/1	3/3	3/3	3/3	3/3	3/3
Aquarium	7	6/50	6/53	6/49	6/26	6/15	17/151	19/100	18/107	18/97	17/90
ShellfishOpenImages	3	3/7	3/3	3/3	3/5	3/6	9/16	9/15	9/10	9/19	9/18
AerialMaritimeDrone(large)	5	3/96	4/53	2/48	2/28	3/36	6/141	10/139	7/110	5/76	7/109
AerialMaritimeDrone(tiled)	5	4/10	3/8	3/11	5/23	4/13	12/42	12/40	10/39	13/48	12/42
SelfDrivingCar	11	6/50	8/88	7/64	9/98	9/74	20/194	25/261	22/190	24/227	24/222
Plantdoc	30	30/123	30/142	30/123	30/137	30/158	86/323	86/417	87/344	87/421	86/381
BrackishUnderwater	6	6/18	5/12	4/16	5/16	5/22	18/42	15/41	13/38	14/36	15/42
Pistols	1	1/1	1/1	1/1	1/1	1/1	3/3	3/3	3/3	3/3	3/3
MaskWearing	2	1/20	2/3	2/24	2/5	2/11	5/48	5/17	4/45	5/18	6/34
EgoHands(generic)	1	1/4	1/3	1/3	1/2	1/4	3/10	3/8	3/11	3/8	3/10
ChessPiecesPieces	13	3/52	5/73	3/31	2/30	3/34	8/131	10/135	9/128	6/106	8/130
ThermalCheetah	2	2/6	2/7	2/5	2/6	2/4	6/17	6/16	6/17	6/17	6/14
DroneControl	8	8/10	8/10	8/10	8/12	8/10	23/27	23/27	23/27	23/29	23/27
HardHatWorkers	3	3/11	3/16	3/14	3/10	2/17	9/39	8/39	9/38	8/29	7/35
AmericanSignLanguageLetters	26	26/26	26/26	26/26	26/26	26/26	78/78	78/78	78/78	78/78	78/78
UnoCards	15	9/27	8/24	11/33	8/24	10/30	23/69	23/69	21/63	23/69	21/63
BoggleBoards	36	12/301	14/373	14/324	16/396	15/378	30/787	32/859	36/903	32/801	36/943
WebsiteScreenshots	8	2/91	4/166	4/129	3/110	3/98	7/327	10/356	11/446	11/569	10/385
ThermalDogsAndPeople	2	2/3	2/3	1/2	2/3	2/2	6/8	5/7	4/8	6/7	6/7
BCCD	3	2/34	1/16	3/28	2/29	2/22	4/62	4/56	6/73	5/78	4/56
Dice	6	5/8	4/6	6/12	3/45	5/8	14/23	13/58	14/25	13/59	12/64
OpenPoetryVision	43	24/85	26/75	21/82	28/104	23/79	62/208	66/212	53/182	67/229	60/202
PKLot640	2	1/40	1/100	1/100	1/28	1/100	4/280	3/240	3/168	3/228	4/268
EgoHands(specific)	4	2/6	2/6	1/4	2/7	2/6	5/17	4/14	3/12	5/17	5/15

this sampling, followed by training and evaluating each method five times, and report the average mean AP across all object classes.

We consider a K-shot setting where $K = 1$, 3, 5, and 10. Following previous studies, K indicates the number of images per object class, not the number of bounding boxes. Depending on the dataset, a single image may contain multiple object instances annotated with different bounding boxes, so the number of bounding boxes used for training varies.

Table 5 and 6 provide detailed configurations. For each dataset listed in a row, the column "classes" indicates the number of object classes. Columns from "seed=0" to "seed=4" represent individual trials in the five random samplings, each reporting the number of images and bounding boxes, separated by '/,' used for training.

C More Details of Experimental Settings

This section provides the comprehensive configurations of our experiments reported in the paper.

Table 6: Detailed configurations of the few-shot training per a dataset. The numbers separated by a slash represent the total number of images (#Img.) and the total number of bounding boxes (#Ann.) used for the few-shot training, respectively. Configurations for $K = 5$ and 10 are shown.

Dataset	Classes	$K = 5$ (#Img / #Ann)					$K = 10$ (#Img / #Ann)				
		seed=0	seed=1	seed=2	seed=3	seed=4	seed=0	seed=1	seed=2	seed=3	seed=4
CottontailRabbits	1	5/5	5/5	5/5	5/5	5/5	10/10	10/11	10/10	10/10	10/11
Packages	1	5/9	5/7	5/7	5/8	5/7	10/18	10/16	10/16	10/18	10/15
Raccoon	1	5/6	5/5	5/5	5/5	5/5	10/12	10/10	10/10	10/11	10/13
NorthAmericaMushrooms	2	10/24	10/17	10/16	10/19	10/12	20/38	20/32	20/27	20/30	20/25
Pothole	1	5/15	5/31	5/10	5/8	5/12	10/29	10/46	10/22	10/29	10/26
OxfordPets(breed)	37	185/186	185/185	185/185	185/185	185/185	370/371	370/370	370/370	370/371	370/371
VehiclesOpenImages	5	25/49	25/63	25/44	25/48	25/45	50/91	50/119	50/97	50/81	50/85
MountainDewCommercial	1	5/126	5/144	5/149	5/137	5/178	10/264	10/324	10/276	10/267	10/323
PASCAL VOC	20	84/259	85/223	85/262	88/286	84/249	169/618	166/496	172/535	171/501	168/490
OxfordPets(species)	2	10/10	10/10	10/10	10/10	10/10	20/20	20/20	20/20	20/20	20/20
WildfireSmoke	1	5/5	5/5	5/5	5/5	5/5	10/10	10/10	10/10	10/10	10/10
Aquarium	7	31/198	32/187	29/175	32/200	29/208	60/353	65/421	60/497	64/359	61/408
ShellfishOpenImages	3	15/28	15/25	15/24	15/28	15/42	30/59	30/60	30/56	30/56	30/75
AerialMaritimeDrone(large)	5	10/221	15/290	11/167	10/155	11/236	24/402	27/486	25/394	22/297	23/424
AerialMaritimeDrone(tiled)	5	17/53	20/68	18/65	20/71	19/64	37/122	40/128	38/136	37/132	41/124
SelfDrivingCar	11	32/336	39/399	37/363	35/315	40/409	65/667	73/698	72/701	74/660	70/716
Plantdoc	30	139/519	140/629	141/576	139/601	139/655	273/981	272/1108	274/1116	274/1119	274/1124
BrackishUnderwater	6	29/70	25/64	23/60	25/59	26/84	52/145	47/135	48/116	51/148	50/137
Pistols	1	5/8	5/5	5/5	5/5	5/5	10/18	10/11	10/10	10/10	10/10
MaskWearing	2	8/54	8/43	7/63	9/40	9/48	16/101	15/69	15/123	15/59	16/116
EgoHands(generic)	1	5/16	5/15	5/16	5/13	5/16	10/31	10/31	10/32	10/30	10/34
ChessPiecesPieces	13	14/194	14/230	13/235	10/191	11/190	25/315	27/401	25/403	24/358	20/366
ThermalCheetah	2	10/25	10/32	10/32	10/30	10/28	18/50	18/48	18/50	17/49	18/48
DroneControl	8	37/42	37/41	37/41	37/43	37/42	72/78	72/77	72/76	71/78	72/77
HardHatWorkers	3	15/75	14/67	14/59	13/47	13/71	30/156	28/157	29/140	28/144	28/155
AmericanSignLanguageLetters	26	130/130	130/130	130/130	130/130	130/130	260/260	260/260	260/260	260/260	260/260
UnoCards	15	38/114	36/108	35/105	36/108	34/102	69/207	65/195	69/207	69/207	64/192
BoggleBoards	36	49/1280	51/1383	52/1328	50/1298	57/1448	77/1935	80/2085	80/2003	76/1910	86/2141
WebsiteScreenshots	8	13/610	16/723	19/725	18/734	17/827	29/1418	31/1259	37/1565	34/1468	32/1412
ThermalDogsAndPeople	2	9/12	9/12	8/14	9/10	9/12	17/26	16/23	16/25	18/22	17/23
BCCD	3	8/107	8/107	8/100	7/110	7/84	17/251	15/194	14/180	14/209	15/197
Dice	6	23/68	20/83	22/38	20/102	20/79	46/106	40/155	46/73	40/138	37/149
OpenPoetryVision	43	93/305	105/330	94/307	108/359	91/300	172/555	184/563	186/572	180/576	177/564
PKLot640	2	7/364	5/440	5/248	7/424	7/424	15/972	12/1008	12/684	12/804	14/800
EgoHands(specific)	4	9/30	6/20	7/23	10/33	9/28	18/61	13/45	13/43	18/59	15/48

C.1 Training

C.1.1 Object Detection Pre-training

To pre-train the models with the Object365 dataset, we utilized the AdamW optimizer [3], setting the batch size to 64 across 8 V100 GPUs. The training duration was 30 epochs for DyHead, GLIP(A), and Faster RCNN, while F-ViT was trained for 20 epochs. Notably, the original F-ViT training protocol [6] employed only 3 epochs, leveraging its robust pre-training via CLIPSelf self-supervised learning.

The initial learning rate was set at 1.0×10^{-4}. The weight decay parameters were set at 0.1 for F-ViT and 0.05 for the other models. The learning rate was reduced by a factor of 10 at 67% and 89% of the total iterations. For data augmentation, we applied a standard random horizontal flip with a 0.5 probability and implemented multi-scale training. Input images were resized such that their shorter side is sampled from [480,560,640,720,800].

C.1.2 Finetuning

The finetuning process retains the same training configurations as pre-training but includes several modifications. For all finetuning datasets (i.e., ODinW), the batch size is set to 4 on 4 V100 GPUs. We specifically increase the initial learning rate to 1.0×10^{-3} for DyHead with the TFA [5] approach. In line with the GLIP [2] implementation, we adjust learning rates during finetuning, based on the detection performance assessed on each dataset's validation data. Specifically, we employ a Pytorch ReduceLROnPlateau scheduler with a patience of 3 and a factor of 0.1 to decrease the learning rate

Table 7: Detection performance for DyHead. All methods are finetuned with Full-FT approach.

	K=1					K=3					K=5					K=10				
Seed	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
CottontailRabbits	63.8	44.7	51.6	58.3	60.4	62.4	54.7	57.6	56.4	54.2	70.7	66.1	54.9	62.5	62.1	67.5	66.7	53.3	66.0	65.5
Packages	44.5	61.5	30.6	51.1	68.6	43.5	62.3	56.3	45.9	73.6	63.1	60.1	55.2	39.1	62.8	48.4	60.7	55.4	59.1	63.9
Raccoon	41.6	33.8	32.1	34.8	32.6	30.8	35.5	43.2	42.7	34.9	33.6	27.7	50.9	43.3	50.4	44.5	46.2	54.4	45.8	27.4
NorthAmericaMushrooms	77.7	68.0	86.2	66.0	68.9	81.5	39.6	86.8	85.4	71.5	84.2	66.9	89.3	78.5	72.1	90.7	69.1	84.7	84.4	85.8
Pothole	3.3	18.8	12.8	9.8	12.2	19.8	18.8	22.4	18.8	18.5	26.4	20.3	24.8	18.7	14.5	28.6	19.1	30.3	27.4	27.4
OxfordPets(breed)	14.7	14.6	16.8	16.9	11.9	33.0	33.6	33.4	33.2	35.2	45.4	44.8	46.6	44.5	45.3	59.9	62.2	61.5	59.8	60.3
VehiclesOpenImages	29.5	36.3	33.0	36.2	26.9	45.7	47.6	47.7	45.5	47.5	51.1	50.9	52.5	52.3	51.5	53.8	57.4	53.0	54.3	55.0
MountainDewCommercial	13.2	4.3	9.2	8.9	23.1	16.6	21.4	13.3	21.0	21.3	18.2	14.7	11.3	19.8	14.7	20.0	17.1	17.1	21.1	12.3
PASCAL VOC	22.4	24.4	23.4	21.0	21.3	38.0	38.4	40.8	38.3	39.8	41.6	43.9	44.4	39.0	41.2	46.1	45.5	48.2	45.4	46.4
OxfordPets(species)	19.4	16.8	33.9	11.0	26.3	39.4	50.3	48.5	50.7	45.0	52.3	49.6	47.0	58.6	57.8	63.7	52.2	58.5	62.9	57.8
WildfireSmoke	0.1	0.6	3.3	3.5	0.0	8.6	19.7	21.5	16.1	8.9	21.4	14.6	1.8	9.1	1.1	18.1	23.0	29.4	26.2	22.5
Aquarium	21.0	17.5	14.9	18.3	13.4	31.3	22.8	23.6	28.0	31.6	37.6	26.3	28.5	36.8	35.1	40.3	39.2	43.2	42.2	40.4
ShellfishOpenImages	5.9	6.6	9.9	6.9	11.2	17.2	13.6	13.1	16.3	16.7	25.7	23.4	21.1	26.6	18.3	27.8	33.3	25.5	32.8	26.0
AerialMaritimeDrone(large)	13.4	16.0	13.0	16.6	17.5	22.9	24.8	25.9	16.9	23.8	21.7	28.0	30.3	22.6	21.4	26.3	28.3	31.8	30.3	30.7
AerialMaritimeDrone(tiled)	6.4	14.8	10.4	18.3	17.1	23.7	24.1	29.5	22.3	23.8	24.8	31.9	26.9	26.1	25.9	26.6	35.2	34.1	30.8	29.1
SelfDrivingCar	9.6	8.6	11.6	9.1	10.4	12.6	13.7	14.6	15.4	15.4	15.0	16.5	17.2	18.7	19.8	19.8	20.3	20.9	21.6	21.4
Plantdoc	6.0	5.2	11.4	8.2	10.7	19.2	14.9	20.2	16.3	21.7	20.9	20.4	23.7	20.4	25.7	30.5	26.3	31.9	28.6	27.3
BrackishUnderwater	8.0	4.1	8.3	8.1	14.7	16.0	18.1	15.1	21.7	21.6	25.2	23.8	20.7	28.9	29.4	32.9	29.2	30.6	32.9	35.6
Pistols	23.5	24.4	29.3	9.7	0.3	33.2	27.9	42.7	32.4	33.9	18.8	20.3	32.6	29.5	39.1	15.8	26.7	39.3	31.4	43.5
MaskWearing	20.8	13.1	32.1	29.8	27.5	39.1	34.6	40.6	51.4	47.1	43.9	45.2	44.2	42.0	42.9	43.4	42.9	52.0	48.0	55.1
EgoHands(generic)	39.3	42.6	26.8	44.8	45.9	53.0	43.6	51.0	55.0	53.9	47.6	59.8	43.6	52.6	57.6	62.1	62.5	60.8	62.7	62.3
ChessPiecesPieces	61.5	74.5	59.3	53.6	53.6	74.1	72.0	74.4	74.5	75.2	73.8	75.7	75.3	77.6	77.0	78.0	78.5	77.1	77.4	78.5
ThermalCheetah	31.4	35.2	26.2	37.9	32.9	48.2	59.4	52.0	51.3	53.9	57.3	57.7	45.5	55.2	45.3	54.9	42.9	52.4	52.4	55.0
DroneControl	28.2	26.7	26.9	25.4	30.5	30.2	36.2	40.8	40.5	38.8	41.8	40.5	42.7	45.6	45.3	53.1	51.6	51.8	48.0	54.5
HardHatWorkers	36.4	34.6	26.8	20.4	18.9	38.5	37.8	36.5	35.3	37.1	38.1	39.1	38.7	37.7	38.9	39.1	40.8	40.9	39.2	40.5
AmericanSignLanguageLetters	38.7	27.7	36.4	38.3	32.2	64.6	51.6	49.7	56.0	55.3	63.1	57.9	62.1	59.1	68.1	68.5	66.6	68.0	69.2	70.5
UnoCards	43.3	34.0	41.8	31.5	35.8	69.3	68.6	61.7	67.6	60.2	75.4	75.6	75.2	76.8	74.7	81.4	81.2	79.8	79.8	80.7
BoggleBoards	45.7	51.4	43.8	49.7	49.0	67.7	68.0	63.5	63.6	63.7	70.9	70.1	68.7	68.4	70.3	73.9	73.0	70.9	70.9	72.5
WebsiteScreenshots	5.2	10.6	6.7	4.2	5.8	10.2	10.5	9.9	7.8	13.1	12.1	11.6	12.6	7.3	13.6	14.5	13.1	14.3	10.3	15.7
ThermalDogsAndPeople	21.0	36.9	29.7	34.6	25.8	63.6	52.4	62.3	55.9	49.2	59.3	56.0	65.5	56.9	54.6	57.1	67.1	69.7	56.1	60.4
BCCD	51.5	14.1	44.0	51.6	37.3	52.1	53.4	50.9	53.7	51.3	55.5	57.3	52.6	54.5	53.5	59.5	54.6	52.9	56.1	56.4
Dice	4.0	6.3	7.8	16.9	3.9	9.1	14.4	13.0	23.3	19.0	25.2	17.5	14.2	27.6	25.3	30.7	31.2	18.9	42.3	22.6
OpenPoetryVision	2.9	2.9	3.8	3.3	3.1	7.9	8.9	9.7	9.5	8.2	11.7	13.1	15.2	14.9	12.8	16.2	27.2	26.5	27.5	29.3
PKLot640	9.4	15.1	15.7	5.1	15.4	38.0	42.3	41.0	37.9	41.9	45.3	25.4	29.9	28.8	44.8	58.4	31.7	59.4	55.2	53.0
EgoHands(specific)	13.7	15.2	17.5	13.7	20.0	28.5	33.0	24.9	26.9	31.6	34.5	21.1	30.1	34.7	37.4	42.4	25.1	40.6	32.9	44.6

when no improvement is observed in the validation dataset. Furthermore, we terminate the fine-tuning process if there is no improvement in validation for 8 consecutive epochs.

D More Experimental Results

This section reports all the detection performance results for reference, including those omitted in the main paper due to space constraints.

D.1 Results of Full-FT

Table 7 through 11 show all fintuning results across different number of few-shot samples ($K = [1, 3, 5, 10]$) and different random seeds (seed$= [0, 1, 2, 3, 4]$) for the compared methods, including DyHead, GLIP(A), GLIP, Faster RCNN, and F-ViT, respectively. Each model is finetuned using Full-FT approach.

D.2 Results of Finetuning Approaches

In Sec. 4.3.2, we evalute TFA (Two-stage Fine-tuning Approach) [5] and FSCE (Few-Shot object detection via Contrastive proposals Encoding) [4] as finetuning approaches. Table 12 to 15 show the results of TFA with four main compared methods (DyHead, GLIP(A), Faster RCNN, and F-ViT, sequentially). Additionally, Table 16 and 17 show the results of FSCE. These fine-tuning approaches are evaluated only in the $K = 3$ setting.

D.3 Results of Different Pre-training Data

Table 18 presents all results associated with the experiments on pre-training data in Sec 4.3.3 in the main paper.

Table 8: Detection performance for GLIP(A). All methods are finetuned with Full-FT approach.

	K=1					K=3					K=5					K=10				
Seed	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
CottontailRabbits	65.0	61.4	66.0	53.3	67.0	65.5	62.2	62.0	65.1	69.1	62.6	66.4	64.3	64.1	67.0	63.6	70.8	57.7	67.9	68.5
Packages	58.6	42.8	56.9	61.2	75.2	57.9	62.6	70.3	63.0	75.6	59.1	72.2	72.8	62.0	70.2	64.3	56.9	74.3	45.6	54.3
Raccoon	45.5	53.0	55.9	58.0	50.4	56.5	55.3	55.4	54.3	48.4	59.6	55.2	56.9	55.1	53.6	44.0	58.7	54.3	59.1	55.8
NorthAmericaMushrooms	74.8	56.5	70.4	70.0	62.1	79.9	66.1	72.4	84.9	71.2	85.4	68.2	86.5	84.4	65.9	86.3	70.7	79.4	84.4	68.0
Pothole	10.3	8.2	6.6	19.2	19.1	23.7	18.0	15.9	23.3	16.9	27.2	24.3	24.5	25.5	25.5	32.3	32.2	32.9	32.8	30.8
OxfordPets(breed)	11.8	13.3	10.0	12.5	12.0	30.3	31.3	32.7	29.9	29.9	43.7	43.8	46.1	42.6	45.0	60.4	56.2	62.7	57.2	60.9
VehiclesOpenImages	52.2	53.6	51.2	54.3	40.5	55.5	57.1	57.7	50.9	50.8	57.9	59.2	54.4	51.8	56.9	58.2	59.3	59.4	59.1	57.5
MountainDewCommercial	21.8	18.5	20.9	20.0	25.9	18.8	19.8	18.3	23.1	24.4	23.3	24.5	21.3	25.1	20.3	28.2	22.0	34.0	27.9	27.9
PASCAL VOC	43.0	45.2	44.8	48.2	47.0	49.7	48.2	50.3	49.0	49.1	50.3	49.4	50.2	51.9	49.1	52.2	52.1	52.2	53.0	52.9
OxfordPets(species)	37.3	33.8	38.5	22.5	36.9	45.8	54.7	46.6	49.9	46.3	57.5	59.0	59.9	60.0	59.1	65.9	61.5	63.8	63.3	63.7
WildfireSmoke	1.9	0.6	4.3	17.9	9.1	6.4	28.8	16.2	23.7	29.4	16.9	31.1	24.7	27.7	27.5	27.4	28.2	36.3	35.5	36.3
Aquarium	27.5	24.2	23.2	24.2	27.0	36.4	29.2	29.6	28.3	34.2	41.5	34.9	36.6	37.2	39.8	45.7	42.3	43.4	44.1	42.8
ShellfishOpenImages	23.5	15.7	12.9	17.4	21.8	22.2	16.4	20.1	23.0	22.6	30.3	22.5	27.7	21.9	26.5	26.5	26.3	26.2	31.3	30.8
AerialMaritimeDrone(large)	15.3	22.1	15.4	16.1	17.7	25.4	22.4	22.0	21.9	20.7	25.0	24.2	25.9	24.2	23.8	27.5	22.4	31.5	38.2	26.2
AerialMaritimeDrone(tiled)	18.8	20.2	18.1	23.3	18.8	24.5	30.1	31.6	30.7	25.6	27.0	34.3	29.5	31.9	27.2	33.5	35.3	37.9	32.8	33.4
SelfDrivingCar	11.5	12.3	13.5	11.5	11.4	12.9	15.8	15.2	15.8	15.6	15.7	17.2	17.9	19.2	18.2	20.2	23.6	21.9	20.0	22.2
Plantdoc	8.7	4.0	9.9	7.6	9.4	15.9	12.4	19.8	15.4	17.1	19.8	18.7	20.8	19.5	16.9	28.1	29.6	29.0	30.9	31.8
BrackishUnderwater	10.4	7.2	12.4	10.1	15.7	21.4	23.4	18.2	21.7	22.2	25.6	26.2	23.3	31.4	32.3	34.1	31.2	33.0	33.5	36.4
Pistols	43.9	36.2	45.1	34.2	30.9	47.6	44.4	50.1	47.5	49.3	52.9	37.1	47.3	52.7	46.3	54.1	49.7	48.5	51.0	46.6
MaskWearing	26.2	22.9	33.4	33.1	31.2	45.8	32.3	49.5	37.8	42.7	55.8	47.6	49.7	46.4	54.6	56.3	44.3	55.5	50.1	55.6
EgoHands(generic)	56.5	56.9	39.8	53.9	57.0	59.8	57.9	58.5	59.8	62.2	60.9	63.2	60.2	59.9	62.2	64.6	63.3	61.5	64.8	66.3
ChessPiecesPieces	70.4	76.0	65.0	67.4	62.2	75.8	75.8	74.1	75.8	77.9	74.2	75.2	75.1	76.4	77.3	77.2	77.2	75.8	76.6	77.3
ThermalCheetah	46.4	45.3	45.0	39.1	29.3	59.3	62.8	53.5	58.6	50.5	52.7	60.5	66.6	61.5	54.0	59.8	59.1	63.6	66.5	63.3
DroneControl	26.6	27.8	23.1	26.3	26.4	35.1	36.7	40.2	42.9	38.8	44.1	41.3	45.4	45.0	42.9	51.9	50.3	52.4	48.8	54.6
HardHatWorkers	37.4	35.3	31.5	22.2	19.4	38.3	39.6	38.9	37.5	38.2	38.1	40.4	40.2	38.3	38.4	40.0	41.6	42.0	40.7	39.7
AmericanSignLanguageLetters	25.2	18.9	20.1	23.8	20.8	55.5	43.0	53.6	45.5	50.9	56.0	59.7	56.9	60.9	57.5	69.5	66.2	64.8	71.7	62.2
UnoCards	40.2	30.8	39.0	31.9	28.6	63.0	57.1	63.2	61.0	55.5	72.4	72.0	74.0	73.0	73.9	81.5	81.2	80.0	78.6	80.2
BoggleBoards	42.9	48.3	30.9	55.6	42.7	63.1	67.0	65.7	66.6	63.1	71.3	71.5	69.2	70.3	65.0	70.9	71.8	72.1	72.8	73.2
WebsiteScreenshots	6.0	9.9	8.2	5.5	5.1	10.8	9.9	10.8	8.1	11.2	11.1	10.8	11.9	9.5	11.8	13.8	13.1	13.0	11.4	14.0
ThermalDogsAndPeople	56.0	53.9	44.6	63.9	55.8	60.5	68.3	67.2	67.4	60.6	65.1	59.7	70.9	64.4	61.6	70.6	66.2	75.1	73.6	77.2
BCCD	51.2	42.7	38.8	51.3	39.4	57.2	51.8	54.6	54.0	55.9	58.7	57.3	55.2	57.3	52.4	59.5	58.5	55.5	56.8	58.1
Dice	3.4	8.3	6.5	13.1	4.3	7.3	13.7	15.6	18.8	12.8	22.9	21.1	13.2	21.6	16.9	30.1	22.4	21.0	30.1	23.2
OpenPoetryVision	1.8	2.8	3.4	3.1	2.2	5.8	8.1	5.1	7.2	5.6	9.8	11.4	11.5	11.8	13.0	20.7	24.5	23.6	24.3	22.9
PKLot640	11.9	27.2	26.0	11.3	26.8	41.3	44.6	39.5	41.2	39.9	43.9	49.5	38.3	48.5	40.3	60.3	54.5	59.5	59.9	55.5
EgoHands(specific)	11.8	19.7	17.4	12.1	18.6	26.3	34.4	22.8	24.0	28.5	32.2	35.6	30.4	30.2	40.1	36.7	40.3	33.4	40.7	42.3

Table 9: Detection performance for GLIP. All methods are finetuned with Full-FT approach

	K=1					K=3					K=5					K=10				
Seed	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
CottontailRabbits	69.0	69.0	70.1	69.0	70.6	66.7	63.0	69.0	71.0	70.0	69.0	71.4	67.3	71.9	70.8	72.2	73.7	69.0	71.5	69.0
Packages	64.0	60.9	60.5	63.9	75.2	67.7	65.4	69.6	67.5	76.2	74.2	64.9	67.6	65.9	72.3	74.2	64.0	74.7	72.7	69.3
Raccoon	59.2	53.6	56.5	59.5	54.0	56.9	60.0	56.8	62.0	52.6	60.1	57.1	64.5	55.0	59.3	58.1	63.2	62.0	56.6	67.6
NorthAmericaMushrooms	76.3	60.4	38.4	65.6	59.9	84.5	73.0	87.7	64.1	75.4	74.6	78.1	88.4	85.7	68.2	88.1	72.8	88.1	83.6	83.7
Pothole	25.9	31.9	23.3	29.5	25.3	34.5	34.5	32.5	32.3	28.9	40.1	28.4	34.3	33.3	33.5	42.7	36.8	41.7	34.2	42.6
OxfordPets(breed)	11.1	13.7	7.9	13.4	11.8	26.4	29.6	21.4	28.6	23.9	38.3	38.6	39.6	40.5	37.8	56.8	60.9	58.3	60.1	56.0
VehiclesOpenImages	61.8	60.4	60.3	60.0	57.1	60.5	61.5	58.1	62.7	61.0	58.8	64.8	63.5	59.7	66.4	63.5	66.0	64.9	64.7	66.9
MountainDewCommercial	31.5	25.9	26.3	29.1	27.6	21.5	20.9	26.6	20.1	28.7	25.2	29.5	31.3	26.2	26.7	26.1	28.1	30.8	28.6	25.7
PASCAL VOC	54.0	55.3	57.5	56.5	58.0	58.0	55.6	59.3	57.4	56.5	57.8	56.6	59.8	57.0	57.4	59.4	59.3	58.4	59.8	59.8
OxfordPets(species)	35.2	41.3	63.0	28.5	57.6	64.3	69.7	64.9	66.0	66.5	72.3	69.4	64.7	68.2	70.6	71.8	70.5	72.1	69.0	73.7
WildfireSmoke	19.9	18.7	27.1	16.6	15.4	25.9	32.7	31.6	28.8	34.5	31.6	37.7	33.0	34.9	39.6	34.6	37.2	42.6	39.7	40.5
Aquarium	32.7	31.3	26.5	28.2	33.1	38.5	36.4	31.1	37.3	38.6	46.6	37.3	39.9	41.5	41.7	47.1	45.0	44.3	46.7	42.8
ShellfishOpenImages	27.2	19.4	25.0	23.5	22.4	28.2	25.4	22.2	41.7	21.3	33.7	26.4	33.2	31.7	28.4	35.3	37.4	30.3	36.9	34.2
AerialMaritimeDrone(large)	18.3	21.4	14.3	18.8	19.7	24.8	23.2	25.6	19.5	22.4	27.3	23.7	23.8	25.5	24.7	28.8	25.3	27.4	28.0	27.6
AerialMaritimeDrone(tiled)	19.8	20.9	18.6	23.5	23.3	22.7	29.7	27.0	28.5	31.4	27.7	32.8	32.9	33.5	26.4	30.2	36.8	32.3	33.3	28.1
SelfDrivingCar	11.7	13.5	13.5	12.3	12.9	14.4	15.8	16.5	17.4	16.7	16.1	16.7	19.3	19.1	20.6	20.8	20.5	22.0	20.9	22.3
Plantdoc	7.5	5.4	8.6	7.9	11.1	18.0	15.3	16.5	16.8	17.7	18.9	19.4	23.4	19.5	20.6	24.2	27.7	28.0	27.4	23.6
BrackishUnderwater	7.2	5.4	15.8	12.8	17.3	24.2	25.6	18.3	26.1	26.7	25.0	27.7	28.2	32.0	31.1	34.3	33.5	39.1	35.1	39.2
Pistols	56.7	49.9	59.8	51.3	53.4	55.7	49.9	58.0	56.7	57.0	62.2	47.5	59.3	60.8	54.4	61.5	59.5	61.3	59.8	59.7
MaskWearing	31.8	30.6	39.0	30.5	37.3	54.0	39.7	47.7	48.8	45.9	46.3	50.1	40.4	47.3	45.4	49.4	49.0	56.7	53.7	53.9
EgoHands(generic)	66.3	67.3	63.9	62.7	61.5	66.9	65.7	67.2	67.0	65.9	66.7	68.4	68.3	64.8	67.3	67.2	69.2	67.3	67.1	69.4
ChessPiecesPieces	65.4	76.3	60.0	62.6	60.0	76.0	73.1	75.0	77.1	76.3	77.5	74.5	77.8	78.7	78.7	78.6	80.2	79.8	80.3	81.2
ThermalCheetah	52.4	62.0	49.9	54.3	50.2	57.7	55.8	63.0	59.4	50.0	60.1	57.7	54.8	66.4	59.7	63.3	58.2	56.3	67.0	55.1
DroneControl	21.6	28.3	22.2	22.6	24.4	35.6	33.7	36.0	40.0	39.6	39.8	40.4	45.5	46.3	40.7	51.1	50.1	54.1	49.5	54.7
HardHatWorkers	37.0	33.9	34.8	32.3	34.9	39.7	40.4	40.2	36.8	39.4	39.1	41.3	41.6	41.0	40.2	40.6	41.3	42.7	41.5	41.0
AmericanSignLanguageLetters	13.2	13.7	9.2	11.8	12.7	34.7	22.9	24.0	35.0	30.6	46.0	32.3	37.4	45.8	38.7	63.6	46.8	59.2	67.1	48.5
UnoCards	17.0	21.5	25.1	17.3	20.4	42.7	42.4	41.2	39.9	32.3	54.6	49.1	52.4	57.1	53.5	76.3	72.7	73.4	75.8	73.0
BoggleBoards	40.7	35.9	32.6	45.3	41.3	53.1	54.9	59.5	59.6	55.9	62.2	65.7	65.9	65.0	59.3	71.3	69.6	72.5	66.8	70.6
WebsiteScreenshots	5.2	10.4	5.5	5.9	4.4	9.0	10.9	9.8	7.9	10.7	8.4	10.1	10.7	7.9	10.8	13.1	11.5	12.4	10.8	12.7
ThermalDogsAndPeople	58.7	47.8	54.8	57.4	67.6	58.9	71.1	67.2	65.0	65.1	70.1	72.8	64.2	69.5	74.6	77.8	79.3	74.6	74.4	77.7
BCCD	49.2	30.8	48.4	50.7	44.8	54.9	51.1	53.1	53.2	54.7	54.7	55.6	54.7	55.1	52.3	59.0	57.0	57.0	56.6	57.2
Dice	5.1	8.3	6.5	14.2	5.1	7.7	13.7	13.3	18.4	14.7	23.1	24.7	11.1	22.5	23.9	26.9	28.0	17.1	29.9	24.9
OpenPoetryVision	2.3	2.0	2.8	1.8	2.6	3.3	3.5	3.8	3.8	3.4	5.1	5.5	4.7	4.3	4.2	8.7	7.9	7.6	8.8	9.1
PKLot640	12.4	20.3	25.0	8.9	23.4	41.6	46.6	38.6	39.1	42.1	43.6	48.1	44.6	49.8	41.3	57.6	53.5	58.4	56.6	55.6
EgoHands(specific)	22.7	26.7	25.8	27.6	23.8	37.3	31.6	23.8	33.4	30.6	32.1	29.7	34.1	39.7	41.7	46.5	45.7	40.5	39.7	45.7

Table 10: Detection performance for Faster RCNN. All methods are finetuned with Full-FT approach.

	K=1					K=3					K=5					K=10				
Seed	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
CottontailRabbits	44.0	16.5	41.9	36.4	48.4	56.9	41.9	57.0	45.1	49.3	54.7	59.3	43.3	55.2	54.8	62.7	38.5	47.0	39.1	53.2
Packages	11.1	59.9	40.4	44.0	64.3	46.6	60.6	66.3	55.7	65.3	36.0	67.7	45.6	49.5	55.2	62.1	70.3	48.1	63.1	67.6
Raccoon	30.2	20.5	17.5	51.3	26.2	31.8	38.6	7.0	45.2	30.6	47.4	38.5	13.9	40.6	30.7	33.3	47.6	35.5	44.6	35.4
NorthAmericaMushrooms	56.6	35.5	28.6	41.5	69.3	70.3	61.5	86.3	83.7	70.5	76.8	62.8	81.1	84.5	66.5	90.5	66.1	81.5	80.1	74.4
Pothole	3.8	15.0	3.5	6.3	10.5	19.7	18.1	12.7	27.3	8.3	22.3	21.7	20.1	27.3	13.8	26.9	21.7	26.7	33.0	26.0
OxfordPets(breed)	10.3	13.5	14.2	13.4	11.3	35.6	37.6	35.7	33.2	36.3	47.5	50.1	47.4	45.7	48.4	60.3	62.0	60.2	60.0	60.8
VehiclesOpenImages	22.0	31.1	23.8	27.8	19.0	42.2	43.5	28.7	37.1	36.7	46.2	44.5	40.7	51.3	49.4	50.3	50.9	50.4	51.4	47.8
MountainDewCommercial	7.3	7.9	5.2	8.4	20.5	16.9	21.4	13.4	26.6	17.2	21.4	21.8	17.9	24.9	21.1	24.6	21.8	28.6	22.0	32.5
PASCAL VOC	11.3	19.0	19.0	13.5	15.8	30.7	31.7	33.6	32.2	33.4	35.1	36.4	38.0	37.3	35.5	40.2	42.1	40.8	42.7	40.0
OxfordPets(species)	16.7	9.2	21.1	6.0	19.4	30.5	28.7	33.9	38.9	42.5	41.5	48.1	41.9	51.0	49.2	53.9	42.5	53.9	52.6	53.0
WildfireSmoke	0.6	0.2	3.2	15.9	2.6	5.3	18.0	23.4	19.9	26.2	11.3	15.7	24.0	20.0	17.9	9.8	23.6	32.4	33.5	27.3
Aquarium	20.5	16.4	13.9	10.1	11.5	28.8	20.4	22.4	21.7	28.6	35.0	24.7	29.4	30.2	35.8	40.6	38.5	41.2	36.6	35.4
ShellfishOpenImages	4.5	4.8	2.8	3.2	3.8	6.3	5.7	9.4	14.5	8.9	12.0	11.1	13.1	18.2	13.1	20.4	21.7	17.5	21.8	16.9
AerialMaritimeDrone(large)	10.0	12.9	12.5	18.8	12.9	19.5	23.5	21.6	21.7	24.0	24.9	20.7	29.2	27.1	27.9	27.4	25.6	31.5	28.8	27.2
AerialMaritimeDrone(tiled)	7.7	17.1	7.9	9.8	17.5	20.3	29.0	31.3	16.0	26.5	26.5	29.0	31.6	29.3	29.8	31.2	35.0	35.4	28.4	36.7
SelfDrivingCar	8.3	8.9	10.5	9.1	10.3	13.3	13.4	14.3	13.9	15.9	17.2	15.8	18.4	19.6	21.0	17.0	23.7	21.0	22.1	24.6
Plantdoc	3.7	4.4	10.0	9.0	7.3	16.9	11.9	17.5	15.0	13.6	21.4	17.0	21.2	18.5	19.4	28.7	22.6	29.5	30.7	25.5
BrackishUnderwater	5.8	4.5	7.1	4.0	12.1	16.8	20.0	11.2	18.9	14.7	21.8	22.1	16.5	24.9	26.8	28.7	28.7	32.0	29.9	29.8
Pistols	18.2	18.9	20.8	4.8	0.6	26.0	21.8	28.8	27.2	33.6	41.2	27.6	38.3	32.4	32.3	39.6	26.4	40.2	35.5	44.7
MaskWearing	34.4	11.6	39.3	22.4	18.1	37.2	23.2	41.7	39.1	44.4	43.6	50.5	41.3	38.5	40.8	43.9	38.4	45.9	46.7	41.8
EgoHands(generic)	29.7	40.5	33.4	37.4	46.9	50.3	49.7	48.7	49.1	53.2	48.3	52.3	54.0	51.4	56.0	57.3	60.5	54.9	54.8	59.4
ChessPiecesPieces	57.6	69.9	51.3	50.7	47.0	72.2	71.7	71.8	72.9	75.3	75.5	74.5	74.5	75.7	76.9	76.9	75.6	78.1	73.9	78.3
ThermalCheetah	34.9	42.7	19.0	30.2	26.7	55.0	52.3	54.9	57.1	44.5	61.5	58.0	50.5	55.3	58.0	58.4	50.2	50.2	58.1	51.5
DroneControl	26.2	22.9	26.3	24.0	22.4	29.9	33.5	38.6	41.1	36.0	39.4	40.2	44.4	43.6	38.5	52.7	51.0	51.4	47.7	50.8
HardHatWorkers	30.3	27.7	19.2	15.1	17.1	36.3	35.6	32.2	31.5	35.3	35.9	37.5	37.6	34.9	36.8	38.2	39.2	40.1	38.2	38.2
AmericanSignLanguageLetters	31.2	22.1	32.3	29.0	20.4	57.7	44.9	55.9	55.2	58.9	61.4	59.1	61.9	58.4	60.2	69.0	70.4	69.1	71.5	62.6
UnoCards	36.1	28.3	36.1	28.7	28.0	69.7	65.7	66.3	62.8	60.7	75.8	74.7	75.1	75.8	74.1	80.9	79.5	80.3	78.9	80.6
BoggleBoards	44.5	52.8	37.3	57.2	47.8	63.2	67.2	67.6	65.4	64.1	70.0	67.9	68.7	69.7	67.3	71.8	71.1	69.5	72.4	72.4
WebsiteScreenshots	6.2	9.5	6.5	6.2	4.3	10.1	9.7	11.1	7.6	10.8	13.2	11.1	12.8	9.8	13.8	14.9	13.1	14.2	12.5	15.8
ThermalDogsAndPeople	25.0	35.9	12.7	23.8	26.2	57.0	54.8	61.5	48.1	49.4	47.2	55.5	68.8	58.0	51.0	72.4	57.9	64.8	70.7	68.9
BCCD	52.8	28.4	45.1	49.6	38.4	54.2	49.0	47.8	51.5	53.3	56.1	55.0	48.8	54.9	51.6	58.9	55.2	54.6	57.2	55.8
Dice	4.0	4.7	7.3	14.0	3.7	12.8	12.0	13.1	20.4	21.9	26.4	22.3	15.4	30.0	29.5	39.1	36.1	26.2	47.4	30.8
OpenPoetryVision	2.8	2.2	3.7	3.8	3.0	8.3	7.3	8.2	10.0	9.3	15.5	12.1	15.6	15.1	17.1	27.3	27.6	30.2	30.2	24.7
PKLot640	11.1	23.9	18.9	7.3	21.0	40.3	41.1	38.6	35.1	36.0	46.2	43.1	40.2	47.5	33.9	54.8	35.9	44.3	51.7	53.1
EgoHands(specific)	7.9	15.1	14.2	12.0	11.5	21.3	23.8	18.9	18.2	25.3	27.2	27.8	22.6	24.8	30.4	31.8	28.9	29.4	30.2	29.2

Table 11: Detection performance for F-ViT. All methods are finetuned with Full-FT approach.

	K=1					K=3					K=5					K=10				
Seed	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
CottontailRabbits	67.2	68.4	66.3	68.8	64.7	65.8	67.0	64.6	64.7	70.5	67.7	67.9	72.8	70.1	68.8	66.4	74.6	64.7	69.4	64.3
Packages	67.0	60.6	58.3	48.0	67.2	59.7	57.0	67.6	53.4	68.2	65.4	63.1	59.9	63.9	62.0	68.6	70.2	70.9	63.4	68.6
Raccoon	54.9	61.5	62.5	57.8	57.6	49.0	55.8	57.5	46.8	59.4	49.8	56.3	58.0	49.6	56.9	40.1	53.7	53.6	54.7	58.9
NorthAmericaMushrooms	67.5	47.6	64.1	65.4	64.4	81.9	74.7	81.4	67.9	61.4	77.4	67.4	85.2	83.6	66.6	84.8	69.3	76.9	85.1	81.8
Pothole	10.6	8.4	12.8	13.3	20.0	19.6	16.3	19.1	23.6	19.3	26.7	15.7	27.7	26.8	26.8	25.9	28.7	31.1	30.5	27.7
OxfordPets(breed)	49.4	51.2	53.5	51.8	47.8	66.9	68.1	67.8	66.3	64.7	71.0	69.3	70.5	70.1	70.5	70.3	73.2	74.7	74.8	73.0
VehiclesOpenImages	43.8	47.2	43.4	49.1	42.7	47.4	47.6	51.0	45.9	39.2	48.5	44.3	49.1	53.3	44.1	53.8	47.9	48.5	50.2	44.9
MountainDewCommercial	11.5	2.6	6.9	7.9	13.9	11.1	16.0	4.8	16.2	15.2	13.9	18.6	4.7	14.4	15.3	16.2	14.9	13.7	15.5	18.5
PASCAL VOC	47.4	47.1	46.9	46.8	48.1	48.3	47.6	48.4	48.7	48.6	47.4	47.8	47.1	48.0	48.5	47.9	46.7	48.6	48.1	49.2
OxfordPets(species)	45.7	50.9	56.8	61.3	57.3	65.0	68.5	65.5	69.5	69.3	69.7	69.5	69.2	68.4	72.2	74.0	72.1	71.8	71.9	73.4
WildfireSmoke	4.4	10.3	3.7	19.9	8.2	6.2	24.2	22.5	22.8	32.9	26.6	25.4	28.4	23.2	28.8	25.5	31.3	31.3	31.5	30.9
Aquarium	32.3	27.1	26.7	24.9	27.4	36.1	29.2	31.7	34.2	32.8	37.2	34.0	34.4	35.7	37.5	42.3	30.0	36.1	38.9	37.5
ShellfishOpenImages	21.5	26.9	22.3	16.9	33.4	33.2	28.7	25.1	27.2	35.2	29.2	33.3	30.1	34.0	32.7	32.8	39.1	34.5	35.3	38.5
AerialMaritimeDrone(large)	14.9	12.2	12.7	13.9	15.8	17.7	15.9	19.8	18.3	21.3	18.5	16.7	16.4	21.3	22.3	18.5	20.2	20.8	20.3	23.8
AerialMaritimeDrone(tiled)	19.7	21.5	21.3	22.0	21.5	25.4	28.7	28.8	30.1	23.7	27.6	31.2	26.9	32.5	26.4	29.6	32.1	29.6	30.0	27.2
SelfDrivingCar	12.0	9.8	10.9	9.7	9.9	12.4	13.0	11.6	11.9	11.2	13.9	14.1	11.1	12.1	11.8	12.4	14.2	13.3	13.1	12.6
Plantdoc	14.1	13.6	16.5	14.0	16.1	21.4	26.7	21.7	24.5	20.9	23.0	23.7	25.3	23.8	24.3	28.8	28.0	28.9	24.3	28.2
BrackishUnderwater	8.1	8.6	10.4	8.8	12.0	17.5	20.2	12.9	20.3	20.1	17.1	19.5	19.2	23.4	23.8	27.2	25.0	22.5	25.7	26.0
Pistols	29.4	27.9	29.4	29.3	26.9	34.5	34.1	35.5	29.1	29.2	40.7	37.8	31.0	39.8	37.9	42.4	40.4	39.0	34.9	40.5
MaskWearing	27.7	21.1	31.9	29.4	25.9	28.8	44.9	34.6	41.8	41.8	34.9	43.5	45.3	47.9	38.5	39.6	38.1	39.7	48.7	35.1
EgoHands(generic)	53.3	53.8	47.0	53.1	49.1	57.3	58.1	53.8	57.9	54.8	58.9	58.9	59.0	58.6	56.1	62.0	60.9	61.7	61.6	61.2
ChessPiecesPieces	59.2	70.4	53.2	50.7	50.8	71.2	70.5	69.6	72.0	69.6	72.5	70.4	73.8	72.4	71.8	72.4	72.2	74.7	71.3	75.4
ThermalCheetah	30.1	29.4	35.1	45.1	37.0	51.8	46.2	41.9	52.4	44.0	46.9	45.7	39.1	47.5	44.8	47.0	45.5	34.8	46.8	42.1
DroneControl	27.6	30.2	28.0	27.6	29.9	31.8	34.0	36.0	37.5	38.1	42.9	41.8	44.2	41.9	44.2	49.0	43.6	50.2	41.4	51.2
HardHatWorkers	31.1	28.6	35.0	35.9	36.4	35.6	35.9	36.3	36.9	35.1	37.3	37.6	36.9	36.2	37.8	38.1	37.7	37.7	38.4	37.8
AmericanSignLanguageLetters	37.8	27.6	29.2	32.7	30.6	52.5	40.2	49.2	55.1	54.5	60.1	53.3	55.9	59.0	59.2	64.0	49.3	64.2	69.7	63.8
UnoCards	36.3	30.8	42.0	35.5	31.5	55.4	46.9	54.6	55.6	51.7	59.3	60.5	66.6	64.1	67.1	65.3	68.7	67.7	72.7	74.2
BoggleBoards	47.3	46.5	54.2	48.8	49.2	58.7	59.6	60.6	57.9	58.2	53.2	57.0	61.0	61.1	56.2	58.6	59.5	62.5	60.7	60.0
WebsiteScreenshots	3.3	7.1	3.6	3.3	3.0	6.0	7.3	6.5	4.8	8.5	6.5	7.8	8.9	6.0	10.1	8.8	8.4	9.2	7.7	9.5
ThermalDogsAndPeople	52.7	54.9	54.0	49.5	58.8	63.2	63.0	56.5	60.3	61.8	58.2	64.6	61.2	58.4	64.0	67.0	62.1	60.4	60.4	63.1
BCCD	42.8	35.4	40.5	41.8	39.8	52.1	44.9	50.0	47.4	51.5	50.2	49.4	49.7	47.6	51.4	53.6	51.4	54.7	51.7	50.6
Dice	5.6	9.2	10.3	11.0	6.3	12.3	12.0	11.6	16.2	15.4	13.7	21.8	23.4	21.5	16.8	17.6	17.3	29.8	16.2	24.1
OpenPoetryVision	2.3	2.1	0.6	2.1	0.0	6.0	2.9	3.4	1.7	2.7	8.3	6.1	5.3	6.5	4.6	12.2	5.5	7.3	15.6	12.2
PKLot640	9.4	19.4	17.0	5.8	13.6	29.1	33.6	33.1	32.5	30.5	36.4	39.5	36.7	37.3	32.3	47.9	37.6	38.3	45.2	41.6
EgoHands(specific)	17.3	23.7	16.0	23.1	24.5	27.4	25.5	24.5	23.8	31.2	31.5	24.3	25.8	28.0	30.8	30.7	29.5	29.5	29.3	30.8

Table 12: Detection performance for DyHead with TFA.

Seed	0	1	2	3	4
CottontailRabbits	60.6	61.8	57.8	52.3	62.4
Packages	51.9	65.4	70.2	61.8	75.6
Raccoon	58.6	57.5	50.8	56.2	57.6
NorthAmericaMushrooms	70.2	72.5	72.8	76.1	70.7
Pothole	1.1	5.3	5.2	9.8	3.4
OxfordPets(breed)	6.2	6.2	6.9	5.5	6.5
VehiclesOpenImages	54.8	57.7	59.7	55.8	52.7
MountainDewCommercial	15.2	11.7	15.3	15.0	22.5
PASCAL VOC	55.6	55.7	56.5	57.1	56.4
OxfordPets(species)	18.4	18.7	15.4	16.1	14.5
WildfireSmoke	1.7	0.0	0.7	0.0	0.5
Aquarium	30.7	24.2	23.0	25.3	27.5
ShellfishOpenImages	21.1	11.1	12.1	15.0	12.2
AerialMaritimeDrone(large)	12.9	15.1	13.4	8.9	16.2
AerialMaritimeDrone(tiled)	24.5	24.1	25.7	21.6	19.4
SelfDrivingCar	12.0	12.6	12.8	12.5	12.7
Plantdoc	4.8	5.3	7.6	5.4	7.4
BrackishUnderwater	6.2	11.8	7.1	9.3	11.6
Pistols	35.3	23.6	40.0	33.7	40.5
MaskWearing	20.4	18.8	18.6	22.3	24.9
EgoHands(generic)	27.3	22.7	26.0	27.5	26.2
ChessPiecesPieces	61.1	60.2	61.9	61.2	62.7
ThermalCheetah	31.8	25.5	29.2	31.1	26.5
DroneControl	25.6	21.6	28.7	28.5	27.3
HardHatWorkers	25.6	24.4	26.2	21.0	16.4
AmericanSignLanguageLetters	23.1	15.5	14.0	17.3	14.5
UnoCards	12.7	12.1	10.5	13.6	12.9
BoggleBoards	1.5	1.2	1.7	1.3	1.3
WebsiteScreenshots	1.4	2.1	1.7	1.7	2.0
ThermalDogsAndPeople	60.4	63.9	57.7	58.7	53.4
BCCD	36.1	33.8	29.0	31.0	30.5
Dice	4.3	4.4	5.9	6.4	5.0
OpenPoetryVision	0.0	0.0	0.0	0.0	0.0
PKLot640	8.6	9.2	11.8	10.5	14.2
EgoHands(specific)	13.0	13.1	14.2	14.2	14.9

Table 13: Detection performance for GLIP(A) with TFA.

Seed	0	1	2	3	4
CottontailRabbits	61.1	60.1	60.2	59.9	60.1
Packages	54.0	62.3	64.3	60.3	66.6
Raccoon	58.6	59.6	49.9	56.5	59.1
NorthAmericaMushrooms	58.2	59.6	64.7	59.1	60.3
Pothole	6.3	2.8	6.7	11.4	7.1
OxfordPets(breed)	3.9	4.1	3.9	3.0	3.8
VehiclesOpenImages	55.3	53.4	55.4	54.2	54.6
MountainDewCommercial	16.4	16.5	17.7	18.3	18.0
PASCAL VOC	54.1	52.5	52.9	54.8	53.7
OxfordPets(species)	17.4	19.8	12.5	16.6	14.8
WildfireSmoke	3.1	0.1	0.1	0.1	0.0
Aquarium	27.9	21.9	16.4	20.1	24.7
ShellfishOpenImages	16.3	16.1	16.8	15.8	16.4
AerialMaritimeDrone(large)	13.8	14.7	14.0	10.7	14.6
AerialMaritimeDrone(tiled)	18.6	21.7	18.0	19.9	18.8
SelfDrivingCar	11.5	11.7	12.1	11.8	11.7
Plantdoc	3.7	4.0	3.6	3.6	4.1
BrackishUnderwater	7.1	7.4	5.6	7.8	8.3
Pistols	33.6	34.5	35.8	34.5	34.0
MaskWearing	9.6	9.6	7.1	7.3	9.5
EgoHands(generic)	26.3	22.9	26.0	24.4	27.5
ChessPiecesPieces	31.9	43.8	42.0	42.0	44.7
ThermalCheetah	31.0	33.9	30.1	28.0	28.7
DroneControl	16.9	20.0	19.6	19.2	18.8
HardHatWorkers	19.0	16.6	16.4	15.6	13.2
AmericanSignLanguageLetters	13.7	11.4	12.9	13.0	9.0
UnoCards	7.7	5.3	4.1	7.9	6.3
BoggleBoards	1.0	1.0	1.0	0.9	0.9
WebsiteScreenshots	1.5	2.1	1.8	2.0	1.8
ThermalDogsAndPeople	47.5	51.8	45.9	45.4	47.9
BCCD	9.8	12.9	15.7	12.4	9.1
Dice	3.2	4.9	4.8	3.3	4.3
OpenPoetryVision	0.0	0.0	0.0	0.0	0.0
PKLot640	0.6	0.7	12.5	0.7	12.6
EgoHands(specific)	10.1	10.4	12.1	12.1	9.4

Table 14: Detection performance for Faster RCNN with TFA.

Seed	0	1	2	3	4
CottontailRabbits	55.8	38.9	37.2	36.5	46.2
Packages	44.0	59.5	63.8	58.0	67.2
Raccoon	50.4	54.6	54.4	46.6	54.1
NorthAmericaMushrooms	58.0	56.1	54.4	50.5	60.8
Pothole	1.7	4.9	1.4	6.0	3.6
OxfordPets(breed)	7.9	7.7	7.6	6.7	8.6
VehiclesOpenImages	47.8	51.3	49.0	51.7	49.8
MountainDewCommercial	9.6	11.7	7.7	12.8	17.0
PASCAL VOC	46.6	48.2	47.1	48.2	47.5
OxfordPets(species)	18.6	0.0	0.0	0.0	0.0
WildfireSmoke	0.0	0.0	0.0	0.0	0.0
Aquarium	28.0	23.2	21.6	23.9	27.3
ShellfishOpenImages	14.8	6.1	8.6	13.5	12.2
AerialMaritimeDrone(large)	12.8	15.0	8.8	7.1	8.8
AerialMaritimeDrone(tiled)	16.8	20.6	16.5	22.9	19.2
SelfDrivingCar	11.4	11.5	12.9	11.7	12.1
Plantdoc	9.4	5.6	9.0	6.7	5.8
BrackishUnderwater	11.1	11.9	10.1	9.6	10.4
Pistols	21.6	19.6	26.3	21.2	25.9
MaskWearing	23.4	21.7	22.3	24.3	24.7
EgoHands(generic)	28.1	23.7	11.5	23.4	24.2
ChessPiecesPieces	63.2	61.6	62.5	63.8	64.1
ThermalCheetah	36.9	38.4	31.2	39.2	35.0
DroneControl	30.2	23.4	22.3	27.7	23.5
HardHatWorkers	24.4	22.7	23.6	22.2	22.1
AmericanSignLanguageLetters	20.9	15.4	16.6	22.2	16.6
UnoCards	16.7	15.2	15.5	13.6	15.2
BoggleBoards	9.1	5.0	5.9	6.1	6.9
WebsiteScreenshots	1.9	1.5	2.1	2.2	2.3
ThermalDogsAndPeople	49.4	55.4	50.4	54.3	52.2
BCCD	33.8	8.9	25.4	19.2	29.0
Dice	4.1	5.7	5.7	4.6	7.9
OpenPoetryVision	0.0	0.0	0.0	0.0	0.0
PKLot640	10.2	10.8	11.3	8.1	12.1
EgoHands(specific)	11.2	13.1	11.2	13.7	13.9

Table 15: Detection performance for F-ViT with TFA.

Seed	0	1	2	3	4
CottontailRabbits	42.9	41.6	51.2	49.5	50.8
Packages	48.7	51.2	53.8	48.7	51.2
Raccoon	56.1	54.3	58.3	55.5	57.3
NorthAmericaMushrooms	21.7	15.6	15.0	24.2	19.2
Pothole	0.2	0.0	0.0	0.0	0.1
OxfordPets(breed)	2.7	2.5	2.6	2.6	2.6
VehiclesOpenImages	31.1	35.6	33.5	34.6	32.0
MountainDewCommercial	1.7	3.9	2.0	4.2	1.5
PASCAL VOC	48.2	48.2	48.8	48.7	48.5
OxfordPets(species)	2.0	2.0	1.9	1.7	2.7
WildfireSmoke	0.8	0.7	0.5	0.8	0.7
Aquarium	17.9	21.2	19.7	21.4	20.7
ShellfishOpenImages	23.4	23.3	24.3	23.2	22.8
AerialMaritimeDrone(large)	11.4	10.1	11.5	11.9	12.0
AerialMaritimeDrone(tiled)	16.7	14.9	15.2	15.7	16.2
SelfDrivingCar	8.1	7.9	8.0	8.1	8.1
Plantdoc	1.4	1.3	1.4	1.6	1.4
BrackishUnderwater	3.6	3.7	3.8	3.8	3.8
Pistols	17.2	16.7	19.1	19.3	20.4
MaskWearing	0.2	0.2	0.1	0.2	0.2
EgoHands(generic)	4.0	4.7	4.5	5.0	4.7
ChessPiecesPieces	2.1	2.3	2.2	1.7	2.3
ThermalCheetah	10.6	9.6	9.5	7.5	8.2
DroneControl	7.3	7.3	7.3	7.3	7.2
HardHatWorkers	2.1	2.1	2.0	2.0	2.0
AmericanSignLanguageLetters	1.3	2.0	1.4	2.0	2.2
UnoCards	0.0	0.0	0.0	0.0	0.0
BoggleBoards	0.0	0.0	0.0	0.0	0.0
WebsiteScreenshots	0.1	0.1	0.1	0.1	0.1
ThermalDogsAndPeople	40.5	41.6	42.6	41.1	41.5
BCCD	4.9	5.5	3.5	3.7	5.1
Dice	0.1	0.1	0.2	0.2	0.1
OpenPoetryVision	0.0	0.0	0.0	0.0	0.0
PKLot640	2.4	2.4	2.5	2.5	2.6
EgoHands(specific)	2.2	2.4	2.5	2.1	2.4

Table 16: Detection performance for Faster RCNN with FSCE.

Seed	0	1	2	3	4
CottontailRabbits	45.2	61.5	62.8	56.1	58.9
Packages	56.7	61.0	66.3	54.6	62.3
Raccoon	53.5	50.8	43.5	48.1	56.0
NorthAmericaMushrooms	73.5	58.6	69.2	84.2	59.2
Pothole	2.7	10.1	3.6	13.3	4.1
OxfordPets(breed)	28.8	33.6	26.8	24.6	30.6
VehiclesOpenImages	46.4	49.0	44.3	46.5	40.6
MountainDewCommercial	17.9	20.5	15.7	15.4	26.9
PASCAL VOC	41.1	43.8	42.9	43.6	38.5
OxfordPets(species)	31.1	30.9	32.9	36.6	36.7
WildfireSmoke	3.8	1.8	3.9	3.6	0.9
Aquarium	30.4	23.4	19.8	22.9	29.9
ShellfishOpenImages	14.8	4.4	6.5	12.1	9.2
AerialMaritimeDrone(large)	17.1	21.0	18.7	17.6	21.0
AerialMaritimeDrone(tiled)	20.4	21.6	23.5	26.0	20.6
SelfDrivingCar	13.8	14.2	15.8	4.8	17.2
Plantdoc	14.7	7.1	16.1	10.0	9.6
BrackishUnderwater	15.3	21.2	12.0	19.2	17.8
Pistols	20.0	16.1	29.4	27.5	33.4
MaskWearing	35.7	29.7	30.4	34.1	37.8
EgoHands(generic)	41.3	43.2	41.8	45.0	42.7
ChessPiecesPieces	71.3	71.5	5.5	71.1	74.4
ThermalCheetah	51.2	56.1	47.7	52.9	46.0
DroneControl	33.8	29.3	34.3	37.6	31.2
HardHatWorkers	37.6	36.9	37.6	36.6	36.4
AmericanSignLanguageLetters	37.4	31.7	36.0	41.2	39.0
UnoCards	52.2	51.1	45.9	51.1	44.2
BoggleBoards	51.0	1.3	35.5	54.5	35.2
WebsiteScreenshots	6.4	5.8	7.8	4.8	7.5
ThermalDogsAndPeople	50.1	59.3	59.9	59.4	52.1
BCCD	43.7	45.6	43.8	46.9	47.4
Dice	10.3	13.8	13.7	17.7	16.8
OpenPoetryVision	2.2	1.3	1.6	2.2	2.6
PKLot640	32.9	35.4	29.2	33.5	29.7
EgoHands(specific)	19.8	21.1	20.5	21.2	21.2

Table 17: Detection performance for F-ViT with FSCE.

Seed	0	1	2	3	4
CottontailRabbits	55.8	59.2	56.6	48.5	58.3
Packages	60.0	55.8	64.3	48.8	58.4
Raccoon	49.9	58.1	56.7	54.1	59.9
NorthAmericaMushrooms	76.5	60.2	70.7	54.0	67.7
Pothole	11.1	22.7	23.6	19.6	20.8
OxfordPets(breed)	65.4	68.1	62.3	67.3	62.8
VehiclesOpenImages	46.3	41.6	46.8	46.0	45.2
MountainDewCommercial	10.5	16.1	11.0	14.6	8.7
PASCAL VOC	45.2	44.7	44.8	45.2	44.1
OxfordPets(species)	61.0	66.0	63.6	66.8	69.6
WildfireSmoke	9.1	15.5	21.7	19.3	16.8
Aquarium	35.2	29.8	30.7	31.3	35.7
ShellfishOpenImages	31.2	26.7	24.0	24.7	31.1
AerialMaritimeDrone(large)	20.5	15.3	16.3	19.1	17.5
AerialMaritimeDrone(tiled)	22.1	27.7	27.9	29.5	27.0
SelfDrivingCar	11.7	12.9	12.0	11.4	11.3
Plantdoc	20.7	22.6	17.4	23.4	18.4
BrackishUnderwater	16.8	17.4	12.4	18.4	20.2
Pistols	29.8	26.6	25.4	30.6	30.8
MaskWearing	35.2	39.1	36.5	43.7	36.1
EgoHands(generic)	54.8	54.9	53.9	57.9	55.1
ChessPiecesPieces	68.6	68.6	67.4	70.6	69.2
ThermalCheetah	45.2	44.5	35.1	42.2	41.0
DroneControl	32.6	33.2	34.3	37.9	34.3
HardHatWorkers	36.3	35.1	36.8	36.7	35.5
AmericanSignLanguageLetters	53.8	43.7	49.3	54.5	49.2
UnoCards	57.7	56.5	49.9	55.2	54.8
BoggleBoards	54.0	62.6	59.9	58.5	57.6
WebsiteScreenshots	5.9	6.6	6.3	4.6	7.0
ThermalDogsAndPeople	56.0	65.1	59.4	54.8	49.6
BCCD	49.8	45.8	45.9	45.9	47.2
Dice	12.3	8.9	9.8	14.0	10.3
OpenPoetryVision	3.9	3.2	2.1	3.2	4.7
PKLot640	26.9	27.7	31.4	25.3	27.4
EgoHands(specific)	29.2	22.7	23.0	23.6	28.4

Table 18: All results for combined datasets with COCO and O365, used in Sec. 4.3.3 in the main paper. All methods are trained with $K = 3$ setting.

#Images	2K (1%)					20K (10%)					0.10M (50%)					0.20M (100%)				
Seed	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
CottontailRabbits	51.8	37.3	54.9	53.7	46.4	58.5	38.3	52.9	52.5	48.7	65.4	44.6	55.0	63.1	60.5	64.9	57.0	60.7	61.4	67.9
Packages	38.9	31.3	60.7	33.1	67.0	42.5	53.1	47.7	57.4	69.2	47.5	60.0	74.3	43.8	69.7	41.5	63.7	50.9	45.7	56.3
Raccoon	27.5	35.5	29.6	37.5	29.2	35.1	43.8	43.4	32.6	36.4	42.3	40.4	49.9	48.8	36.8	48.6	49.2	51.8	50.6	34.5
NorthAmericaMushrooms	82.6	61.7	82.8	80.7	66.5	83.1	67.4	81.5	82.1	68.9	78.6	63.2	85.0	85.9	70.5	81.4	62.4	85.7	84.5	64.0
Pothole	11.3	9.3	12.5	13.1	8.8	13.6	11.0	14.3	10.8	8.7	23.1	13.9	20.9	20.2	18.0	22.4	13.2	17.3	16.9	17.4
OxfordPets(breed)	24.9	27.6	26.9	28.3	27.6	29.6	31.5	28.2	28.1	29.7	29.4	32.2	35.2	36.7	34.3	35.5	35.5	33.1	34.2	32.7
VehiclesOpenImages	27.1	29.2	25.6	28.8	25.9	40.7	43.6	37.2	38.1	40.0	44.0	49.4	46.4	47.6	43.4	45.2	53.2	48.0	48.8	46.1
MountainDewCommercial	9.3	17.3	10.7	17.4	14.7	11.6	22.7	6.7	18.3	19.3	12.5	20.9	3.9	19.2	17.0	6.5	11.6	15.7	20.3	18.9
PASCAL VOC	12.8	15.0	14.5	15.2	14.0	25.8	27.6	28.1	27.5	23.6	36.3	40.0	36.7	38.7	37.3	38.3	40.8	42.5	39.5	41.1
OxfordPets(species)	31.0	36.6	35.8	35.7	45.5	46.4	44.4	45.7	46.7	55.3	51.4	53.3	51.6	50.5	61.4	51.7	59.6	50.0	57.1	62.6
WildfireSmoke	4.4	5.9	5.0	9.3	8.0	4.6	5.0	10.5	19.0	14.0	6.9	17.8	13.0	16.1	17.9	7.4	11.3	23.8	21.1	19.6
Aquarium	20.0	13.7	12.5	14.5	22.3	25.6	17.5	17.3	18.1	25.9	29.3	20.3	21.1	24.3	27.7	30.2	21.6	23.5	23.9	26.9
ShellfishOpenImages	4.4	5.9	6.6	7.0	7.0	12.2	7.0	14.2	12.9	11.6	19.9	14.0	13.4	16.3	17.8	22.9	14.7	18.3	19.7	16.0
AerialMaritimeDrone(large)	15.6	12.0	18.2	13.3	16.2	20.3	17.6	26.2	13.3	19.8	20.1	19.1	25.5	14.7	22.3	16.7	19.2	27.0	17.0	22.8
AerialMaritimeDrone(tiled)	18.7	17.9	20.2	13.3	16.6	16.6	24.4	24.8	15.9	22.0	22.6	20.7	29.5	22.7	23.6	22.3	23.8	25.3	20.1	23.6
SelfDrivingCar	8.6	10.4	11.5	12.7	12.2	10.6	11.9	14.1	13.7	14.5	12.4	13.3	14.8	15.6	15.9	12.5	14.3	14.9	16.3	17.0
Plantdoc	14.0	9.5	15.5	13.7	14.9	15.2	13.4	18.2	11.9	15.7	18.0	16.9	19.9	16.3	21.6	20.8	15.8	20.9	13.6	20.8
BrackishUnderwater	15.8	18.8	12.1	16.8	17.1	12.4	17.1	12.9	19.9	17.6	17.4	19.1	16.1	22.2	18.4	17.2	19.7	15.9	23.7	21.5
Pistols	28.9	21.1	30.2	25.2	22.1	26.7	25.4	34.4	18.9	25.3	38.1	29.0	34.5	34.2	39.8	36.0	27.9	32.5	35.0	40.0
MaskWearing	30.3	28.3	34.2	23.7	28.7	25.3	38.6	29.0	30.3	33.1	42.2	42.7	40.7	33.3	31.2	40.5	38.5	50.8	37.9	43.6
EgoHands(generic)	39.8	41.2	40.7	37.2	48.2	44.1	46.3	45.3	46.5	50.4	56.7	47.5	52.4	53.7	56.9	56.2	47.7	53.3	50.0	57.1
ChessPiecesPieces	71.5	71.7	70.7	73.1	75.1	73.2	73.9	73.1	74.0	75.3	74.3	72.2	73.3	73.7	73.7	74.2	71.8	73.1	73.3	74.2
ThermalCheetah	38.1	53.2	41.8	53.6	43.5	46.9	56.5	51.7	57.6	38.1	54.9	56.1	53.2	59.6	48.5	53.7	55.4	51.9	58.7	45.3
DroneControl	28.9	25.0	28.8	38.7	29.9	28.9	33.8	34.0	37.4	34.5	30.9	33.7	41.1	41.5	36.4	32.1	34.1	40.3	40.5	37.3
HardHatWorkers	31.9	32.3	30.4	27.5	25.5	34.6	36.4	30.3	34.4	32.7	37.0	37.1	34.6	34.3	36.1	37.9	38.4	36.4	36.2	37.7
AmericanSignLanguageLetters	52.8	40.6	45.3	49.2	50.1	55.6	45.2	47.1	51.4	47.4	57.6	46.1	52.1	53.0	49.4	58.6	48.5	56.5	57.2	56.8
UnoCards	37.8	35.8	32.4	40.7	31.3	51.7	42.6	41.5	46.9	39.0	57.0	53.7	50.1	56.4	51.4	59.5	55.2	51.6	57.9	54.0
BoggleBoards	50.9	55.0	53.5	53.3	52.8	55.6	60.3	58.1	60.3	58.1	57.7	66.8	67.2	67.2	64.4	62.6	65.6	62.8	66.7	65.0
WebsiteScreenshots	6.6	7.3	6.6	5.1	9.3	8.4	9.7	6.8	5.3	11.4	8.3	10.4	8.9	7.6	12.0	9.1	11.4	10.1	7.7	13.1
ThermalDogsAndPeople	47.2	44.7	67.8	42.9	46.1	55.3	49.5	59.5	36.9	48.7	57.0	53.3	58.5	48.1	43.3	60.7	55.7	64.2	48.0	46.0
BCCD	53.3	50.4	49.8	52.2	50.8	50.4	52.3	51.2	52.1	50.1	56.1	51.7	51.5	53.0	52.0	56.2	54.1	51.7	54.2	55.2
Dice	7.3	11.8	6.3	13.0	9.9	6.5	12.0	10.2	13.5	12.2	9.2	13.0	9.6	16.6	14.2	9.4	12.2	12.8	20.6	15.6
OpenPoetryVision	2.4	2.3	3.1	2.5	1.8	3.5	3.6	3.5	3.8	3.2	4.3	6.5	4.5	7.3	6.2	5.6	6.5	5.8	7.3	7.2
PKLot640	34.5	38.7	36.1	35.4	33.6	33.6	40.6	37.3	32.0	37.1	35.4	40.9	37.3	35.2	37.9	35.0	40.0	35.8	37.0	39.5
EgoHands(specific)	17.2	19.2	21.4	20.3	24.0	25.2	28.4	19.5	21.3	27.7	24.6	28.2	24.3	28.5	28.9	25.4	29.2	24.4	26.3	28.1

References

[1] C. Li*, H. Liu*, L. H. Li, P. Zhang, J. Aneja, J. Yang, P. Jin, Y. J. Lee, H. Hu, Z. Liu, et al. ELEVATER: A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models. In *Proc. NeurIPS*, 2022.

[2] L. H. Li*, P. Zhang*, H. Zhang*, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao. Grounded Language-Image Pre-training. In *Proc. CVPR*, 2022.

[3] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *Proc. ICLR*, 2019.

[4] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang. FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding. In *Proc. CVPR*, 2021.

[5] X. Wang, T. Huang, J. Gonzalez, T. Darrell, and F. Yu. Frustratingly Simple Few-Shot Object Detection. In *Proc. ICML*, 2020.

[6] S. Wu, W. Zhang, L. Xu, S. Jin, X. Li, W. Liu, and C. C. Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *Proc. ICLR*, 2024.