ScaleDet: A Scalable Multi-Dataset Object Detector

Yanbei Chen, Manchen Wang, Abhay Mittal, Zhenlin Xu,

Paolo Favaro, Joseph Tighe, Davide Modolo

AWS AI Labs, Amazon

Poster Session: TUE-PM-302

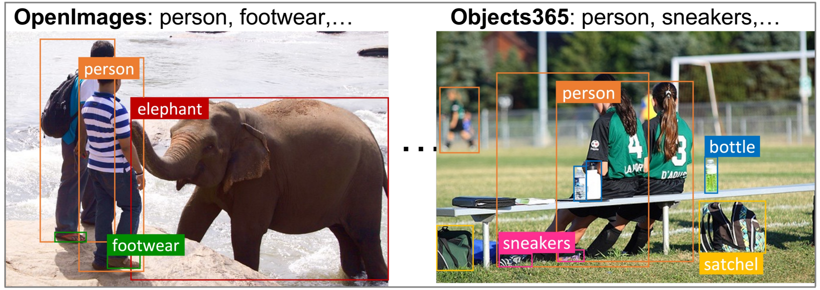
JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

Overview

train across multiple upstream datasets



During training:

Label space of OpenImages = {person, footwear,…}

Label space of Objects365 = {person, sneakers,…}

❑ Problem

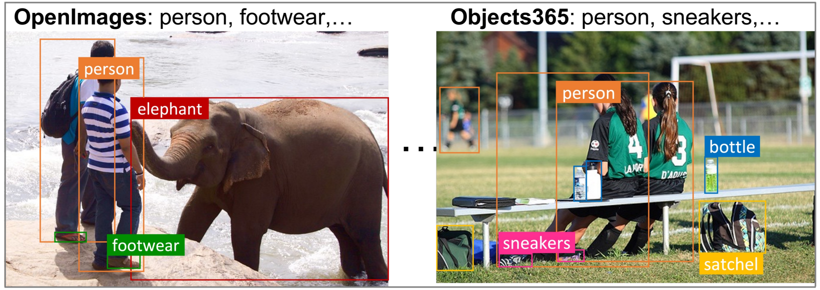
- **Multi-dataset object detection**

❑ Challenges

- train across multiple upstream datasets with heterogenous label spaces

Overview

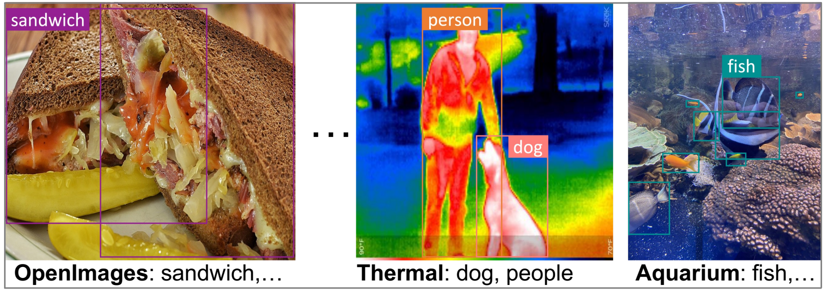
train across multiple upstream datasets



At test time:

Dataset Thermal (from an unseen domain)

Dataset Aquarium (with unseen classes)



test on any upstream or downstream dataset

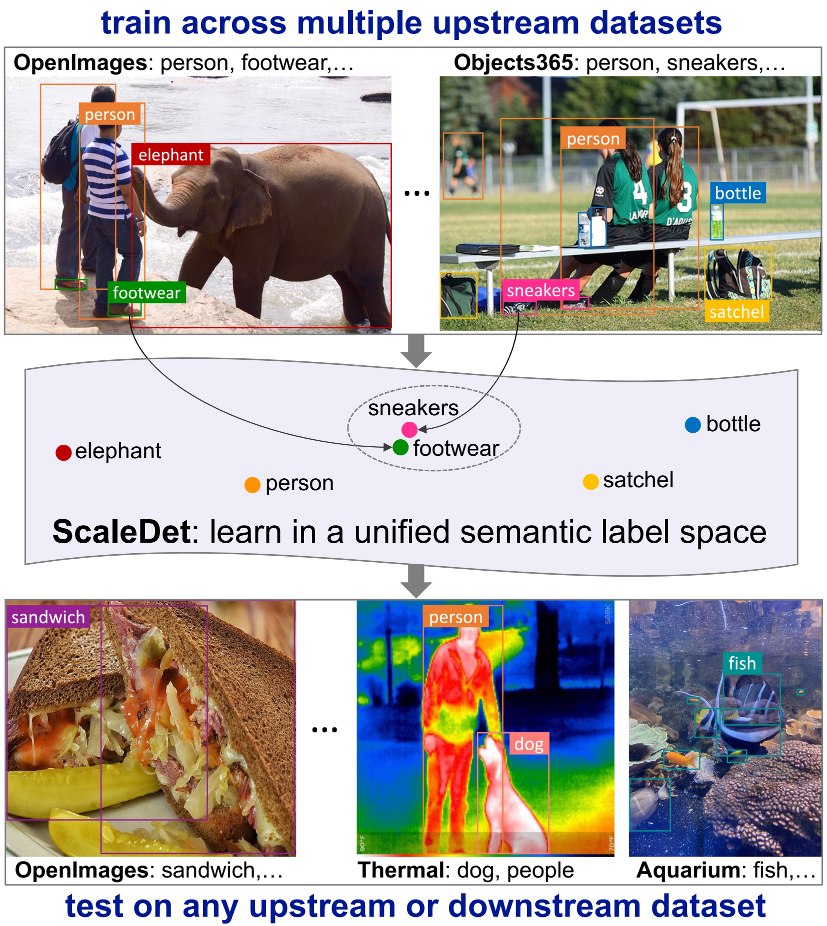
❑ Problem

- **Multi-dataset object detection**

❑ Challenges

- train across multiple upstream datasets with heterogenous label spaces
- generalize well to any given upstream and downstream datasets (which contain both seen and unseen classes/domains)

Overview



❑ Problem

- **Multi-dataset object detection**

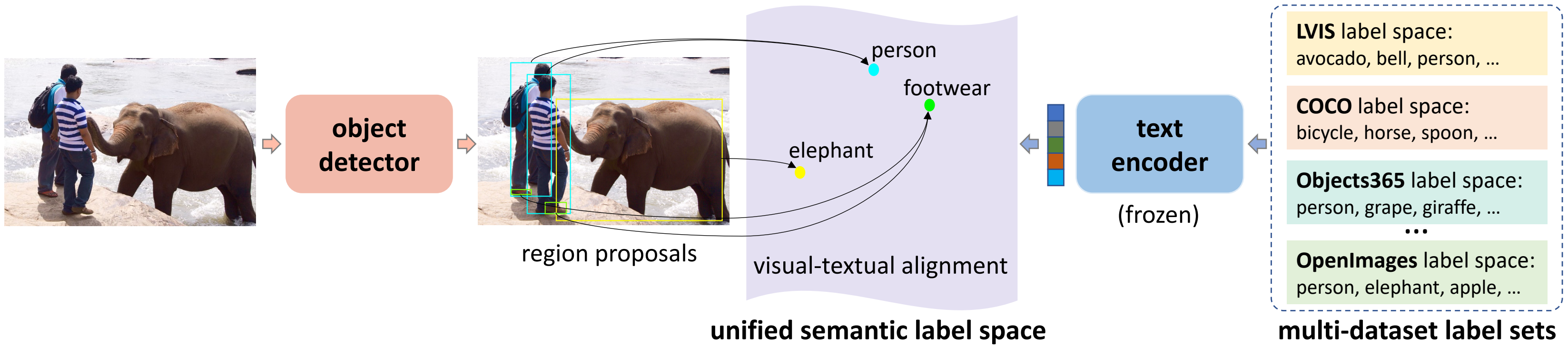
❑ Challenges

- train across multiple upstream datasets with heterogenous label spaces
- generalize well to any given upstream and downstream datasets (which contain both seen and unseen classes/domains)

❑ Proposed approach – ScaleDet

- **A scalable multi-dataset object detector**

Proposed approach – ScaleDet

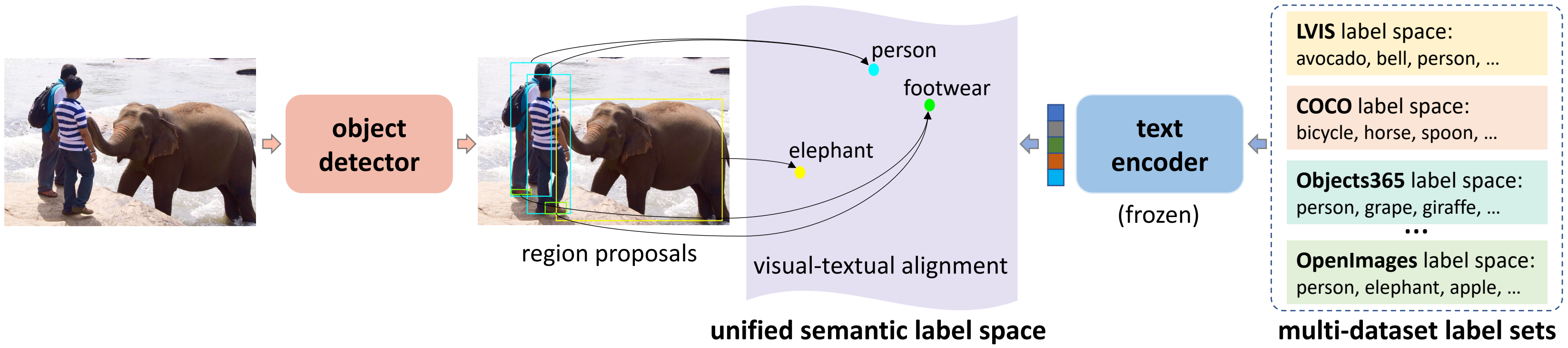


❑ Scalable unification of multi-dataset label space

- *encode class labels as text embeddings*
- *unify label spaces by taking their disjoint union*

$$L = L_1 \coprod \ldots \coprod L_K = \{l_{1,1}, l_{1,2}, \ldots, l_{K,1}, l_{K,2}, \ldots\}$$

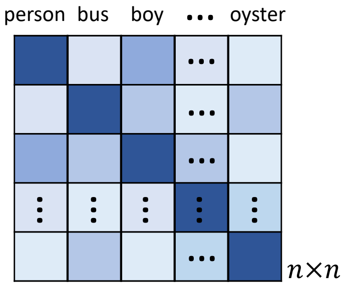
Proposed approach – ScaleDet



❑ Scalable unification of multi-dataset label space

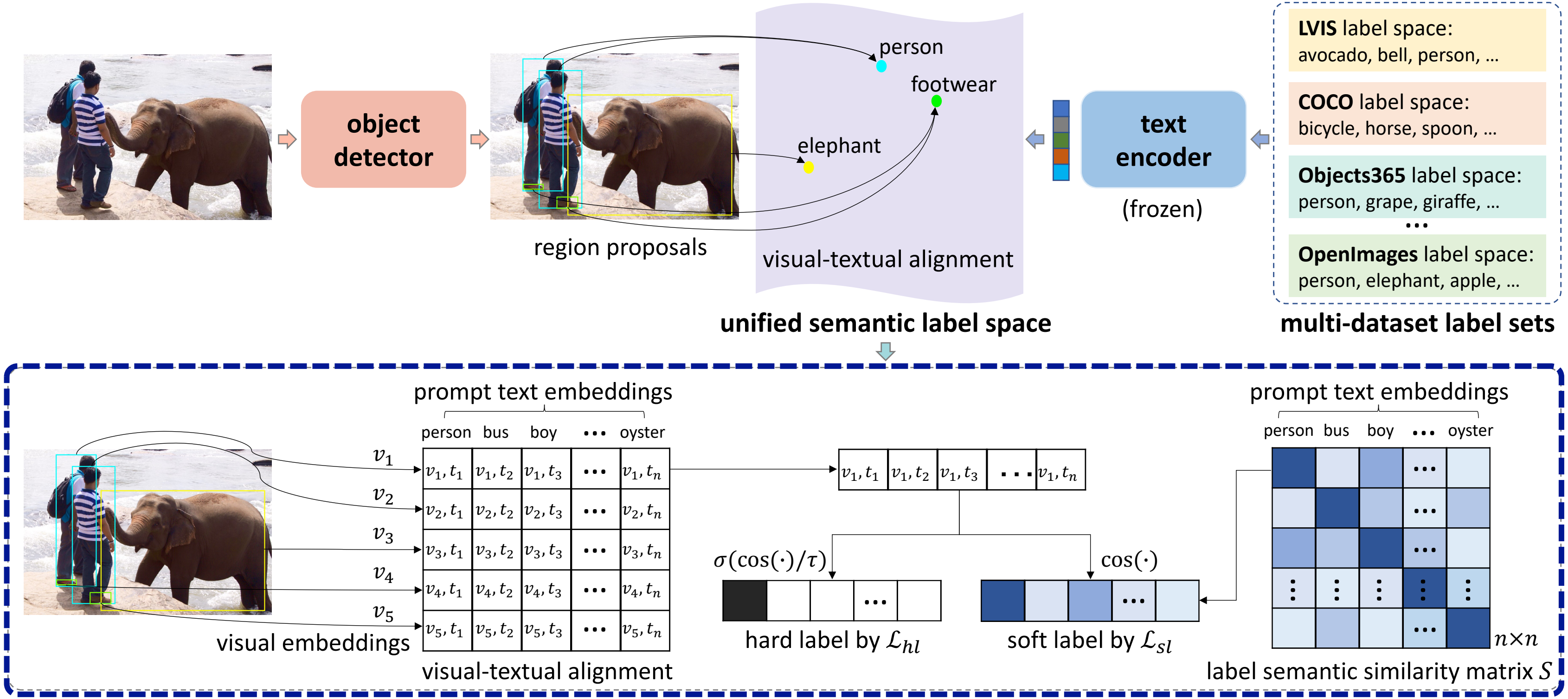
- *encode class labels as text embeddings*
- *unify label spaces by taking their disjoint union*

$$L = L_1 \coprod \dots \coprod L_K = \{l_{1,1}, l_{1,2}, \dots, l_{K,1}, l_{K,2}, \dots\}$$



label semantic similarity matrix S

Proposed approach – ScaleDet



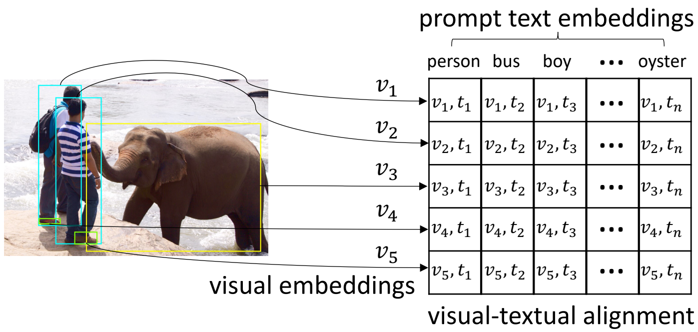
- ❑ **Training by aligning visual and textual embeddings**

Proposed approach – ScaleDet

❑ Training by aligning visual and textual embeddings

- *compute the visual-language similarities*

$$\mathbf{c}_i = [\cos(v_i, t_1), \cos(v_i, t_2), \ldots, \cos(v_i, t_n)].$$

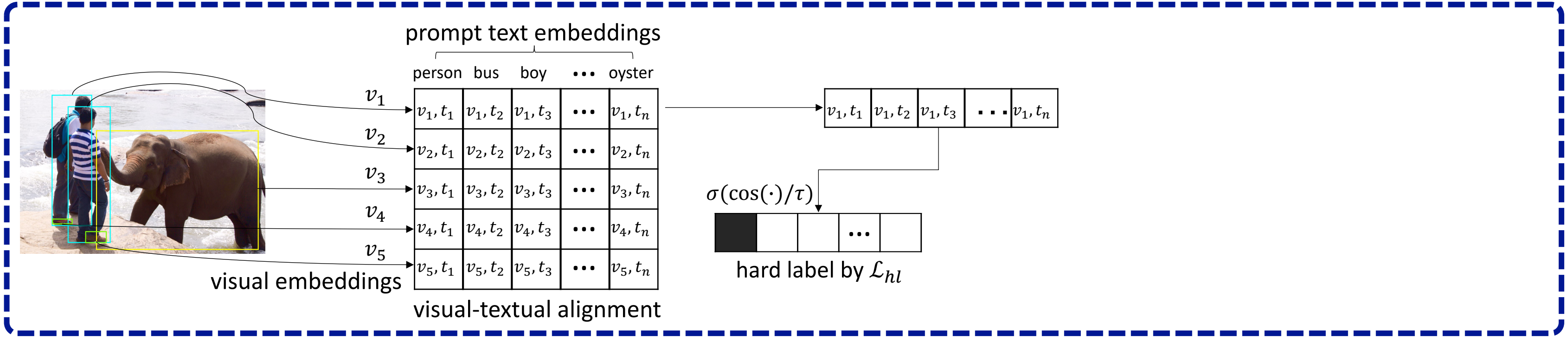


Proposed approach – ScaleDet

❑ Training by aligning visual and textual embeddings

- *compute the visual-language similarities*
- *compute the hard label assignment loss*

$$\mathcal{L}_{hl} = \mathrm{BCE}(\sigma_{sg}(\mathbf{c}_i/\tau), l_i)$$

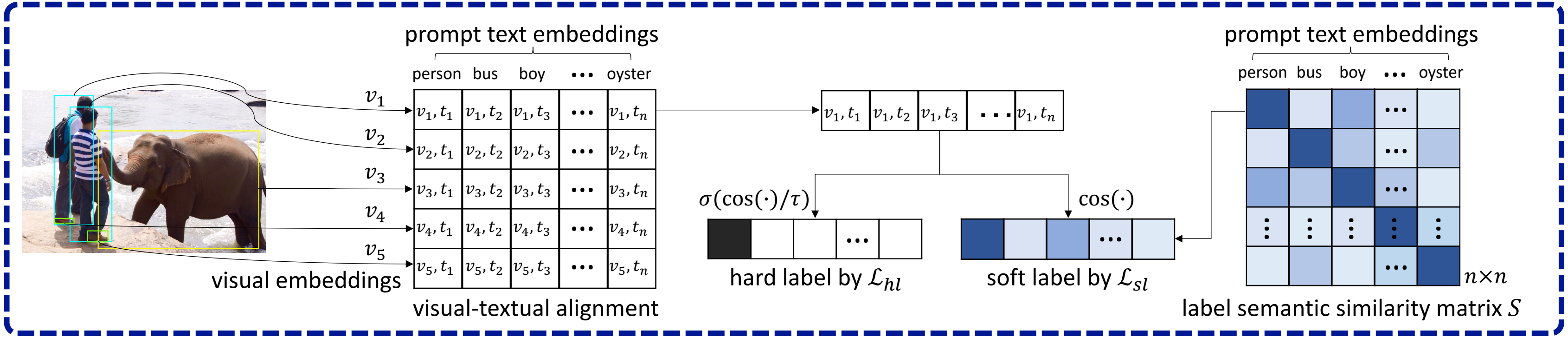


Proposed approach – ScaleDet

❑ Training by aligning visual and textual embeddings

- *compute the visual-language similarities*
- *compute the hard label assignment loss*
- *compute the soft label assignment loss*

$$\mathcal{L}_{sl} = \mathrm{MSE}(\mathbf{c}_i, \mathbf{s}_i)$$



Experiments

❑ Evaluation 1. Training with a growing number of datasets

- **Upstream datasets** (for training and testing) – 4 datasets
  - *LVIS (L), COCO (C), Objects365 (O365), OpenImages (OID)*
- **Downstream datasets** (for testing) – 13 datasets
  - *Object Detection in the Wild (ODinW)*

Experiments

❑ Evaluation 1. Training with a growing number of datasets

- **Upstream datasets** (for training and testing) – 4 datasets
  - *LVIS (L), COCO (C), Objects365 (O365), OpenImages (OID)*

Model	Dataset(s)	L	C	O365	OID	mAP
baseline	L	33.1	37.0	15.2	41.5	31.7
	C	11.0	46.8	7.9	33.1	24.7
	O365	19.2	39.8	28.8	47.6	33.9
	OID	15.7	31.3	14.1	69.3	32.6
ScaleDet	L,C	33.3	44.9	15.9	43.7	34.5
	L,C,O365	36.5	47.0	**31.2**	44.9	39.9
	L,C,O365,OID	**36.8**	**47.1**	30.6	**69.4**	**46.0**

Table. Evaluation on upstream datasets.

L: LVIS. C: COCO. O365: Objects365. OID: OpenImages.

Experiments

❑ Evaluation 1. Training with a growing number of datasets

- **Downstream datasets** (for testing) – 13 datasets
  - *Object Detection in the Wild (ODinW)*

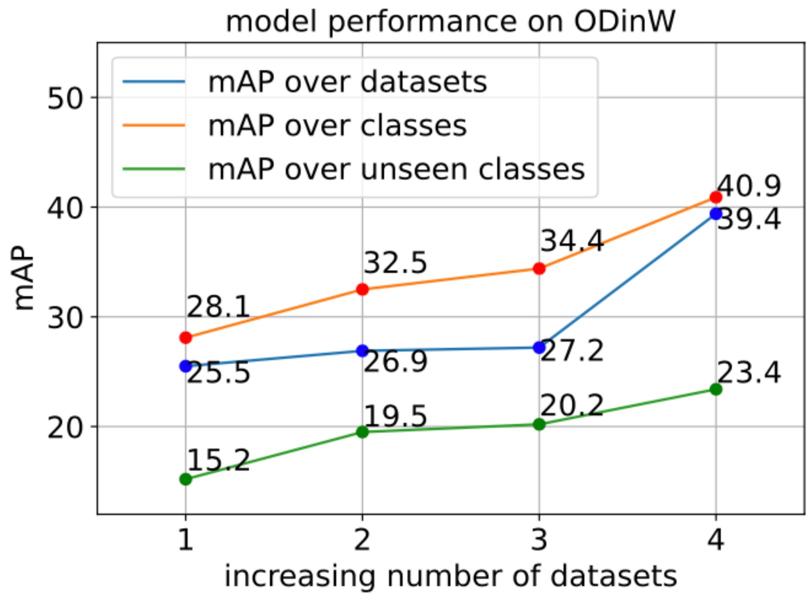


Table. Evaluation on downstream datasets: ODinW.

Experiments

❑ Evaluation 2. Comparison to SOTA on upstream datasets

- **State-of-the-art methods:** UniDet, Detic

	Model	Dataset(s)	COCO	O365	OID	mAP
1	UniDet	single	42.5	24.9	65.7	44.3
2		multiple	**45.5**	24.6	66.0	45.4
3	**ScaleDet**	single	42.1	26.5	66.6	45.1
4		multiple	**45.5**	**27.9**	**69.6**	**47.7**

Table. Comparison to UniDet on multi-dataset training on COCO, O365, OID.

	Model	Datasets	LVIS	COCO	mAP
1	Detic [49]	L,C	33.0	43.9	38.4
2	**ScaleDet**	L,C	**33.3**	**44.9**	**39.1**
3	Detic [49]	L,C,IN21k	35.4	42.4	38.9
4	**ScaleDet**	L,C,O365	36.5	47.0	41.7
5	**ScaleDet**	L,C,O365,OID	**36.8**	**47.1**	**41.9**

Table. Comparison to Detic on multi-dataset training on LVIS, COCO.

Experiments

❑ Evaluation 2. Comparison to SOTA on upstream datasets

- **State-of-the-art methods:** UniDet, Detic, and others

	Model	Model Type	mAP
1	Faster RCNN [32]	single-dataset detection	37.9
2	Mask RCNN [15]		39.8
3	CenterNet [52]		40.2
4	CascadeRCNN [4]		41.6
5	DETR [5]		42.0
6	CenterNet2 [50]		42.9
7	UniT [17]	detection + understanding	42.3
8	RegionCLIP [48]		42.7
9	Detic [49]	detection + classification	42.4
10	UniDet [51]	multi-dataset detection	45.5
11	**ScaleDet**		**47.1**

Table. Comparison on COCO with ResNet50 backbone.

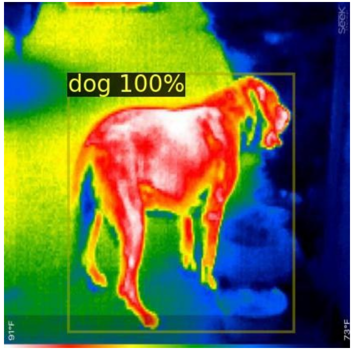
	Model	Model Type	mAP
1	Faster RCNN-T [32]	single-dataset detection	46.0
2	DyHead-T [8]		49.7
3	CascadeRCNN-T [4]		50.4
4	GLIP-T [26]	detection + understanding	55.2
5	GLIPv2-T [46]		55.5
6	GLIPv2-B [46]		**58.8**
7	Detic-B [49]	detection + classification	54.9
8	**ScaleDet-B**	multi-dataset detection	**58.8**

Table. Comparison on COCO with Swin Transformer backbone.

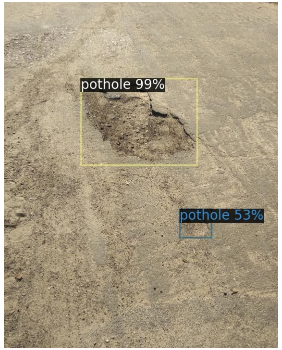
Experiments

❑ Evaluation 3. Comparison to SOTA on downstream datasets

- **State-of-the-art methods:** GLIP, GLIPv2, Detic
- Datasets: 13 downstream datasets on Object detection in the Wild (ODinW)



(a) A rare domain: thermal

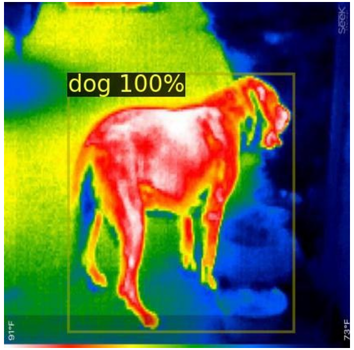


(b) A rare class label: pothole

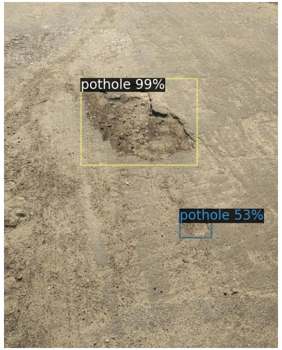
Experiments

❑ Evaluation 3. Comparison to SOTA on downstream datasets

- **State-of-the-art methods:** GLIP, GLIPv2, Detic



(a) A rare domain: thermal



(b) A rare class label: pothole

Model	Model Type	#Data	ODinW direct	ODinW fine-tune
GLIP-T [26]	detection + understanding	5.5M	46.5	64.9
GLIPv2-T [46]		5.5M	48.5	66.5
GLIPv2-B [46]		20.5M	**54.2**	69.4
Detic-R [49]	detection + classification	12.6M	29.4	64.4
Detic-B [49]		12.6M	38.7	70.1
ScaleDet-R	detection	3.6M	39.4	68.5
ScaleDet-T		3.6M	44.3	70.4
ScaleDet-B		3.6M	47.3	**71.8**

Table. Comparison to GLIP, GLIPv2, Detic on downstream datasets ODinW.

Summary of contribution

- ❑ We propose ScaleDet - A scalable multi-dataset detector to train across different datasets, and test on any given upstream and downstream datasets.

- ❑ We propose to train the multi-dataset detector by aligning the visual and text embeddings using hard label and soft label assignment losses.

- ❑ We demonstrate the state-of-the-art performance in multi-dataset training, and show the state-of-the-art generalization on Object Detection in the Wild.

Thank you for your attention!