Detecting Open World Objects via Partial Attribute Assignment

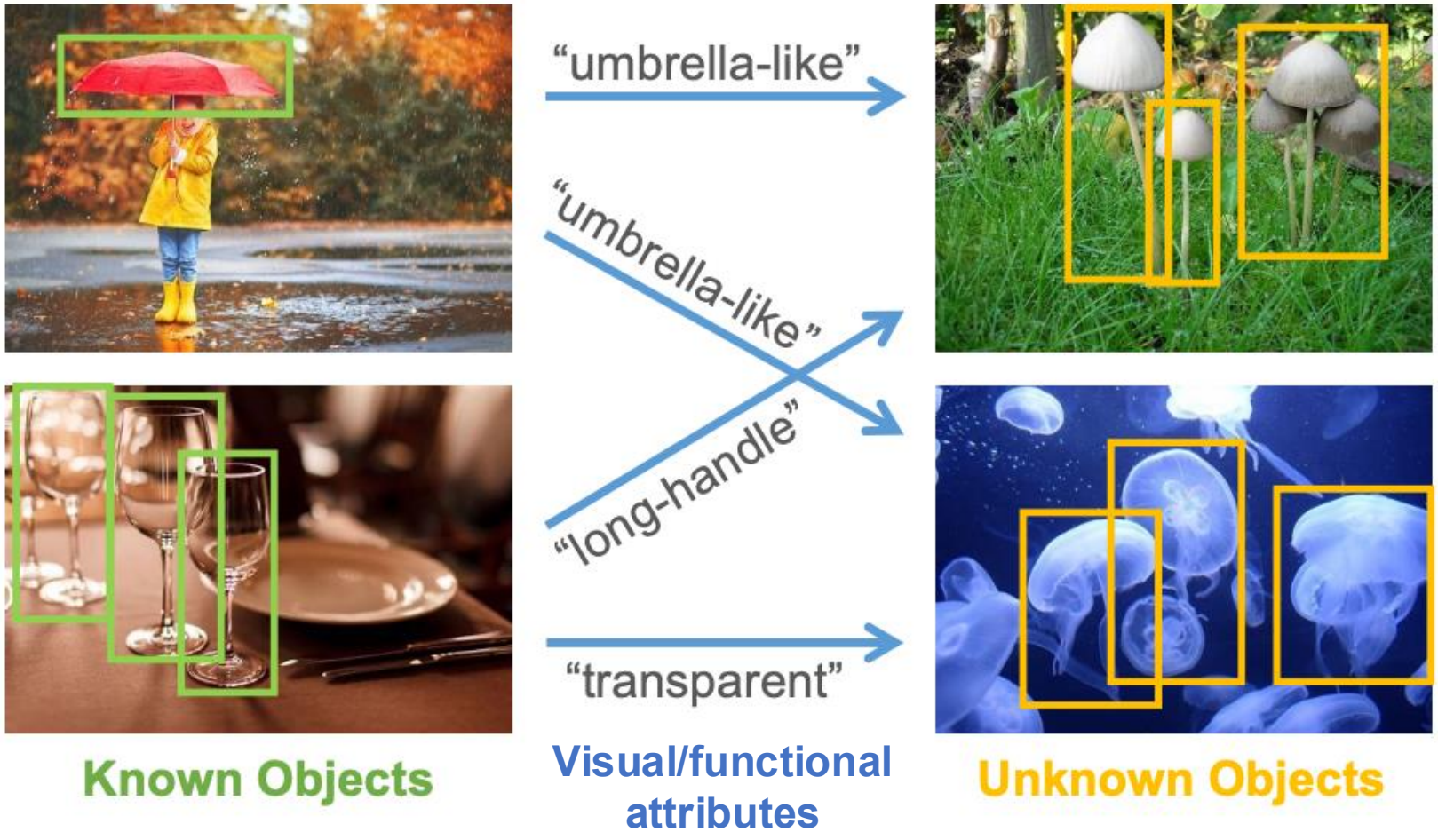
Muli Yang[1] Gabriel James Goenawan[1] Huaiyuan Qin[1]
Kai Han[2] Xi Peng[3] Yanhua Yang[4] Hongyuan Zhu[1]

*1: Institute for Infocomm Research (I^2R), A*STAR, Singapore*
2: The University of Hong Kong 3: Sichuan University 4: Xidian University

CVPR Nashville JUNE 11-15, 2025

Introduction

- *Open World Object Detection (OWOD)* aims to detect and continually learn previously unseen (unknown) objects.



Introduction

- Beginning with a large pool of potentially relevant attributes, the major challenge lies in effectively refining a subset of highly-relevant attributes, to enable reliable detection of both known and unknown objects.

?

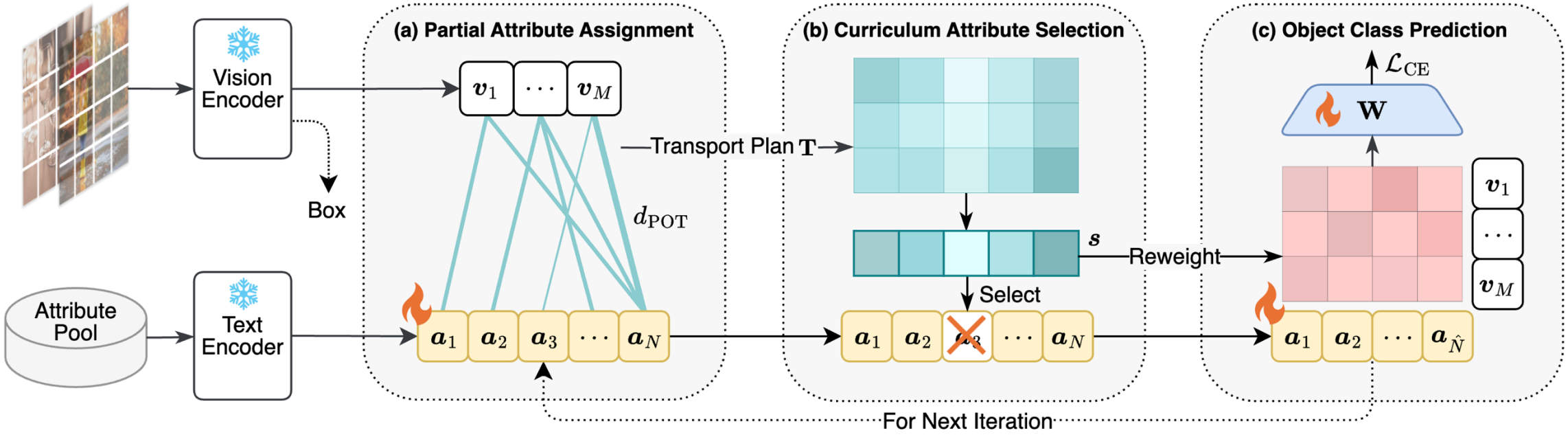
··· protective elegant yellow transparent elliptic sweet hexagonal ···

Introduction

- We model this problem as a *Partial Optimal Transport (POT)* problem, enabling end-to-end attribute selection and optimization for effectively detecting both known and unknown objects.

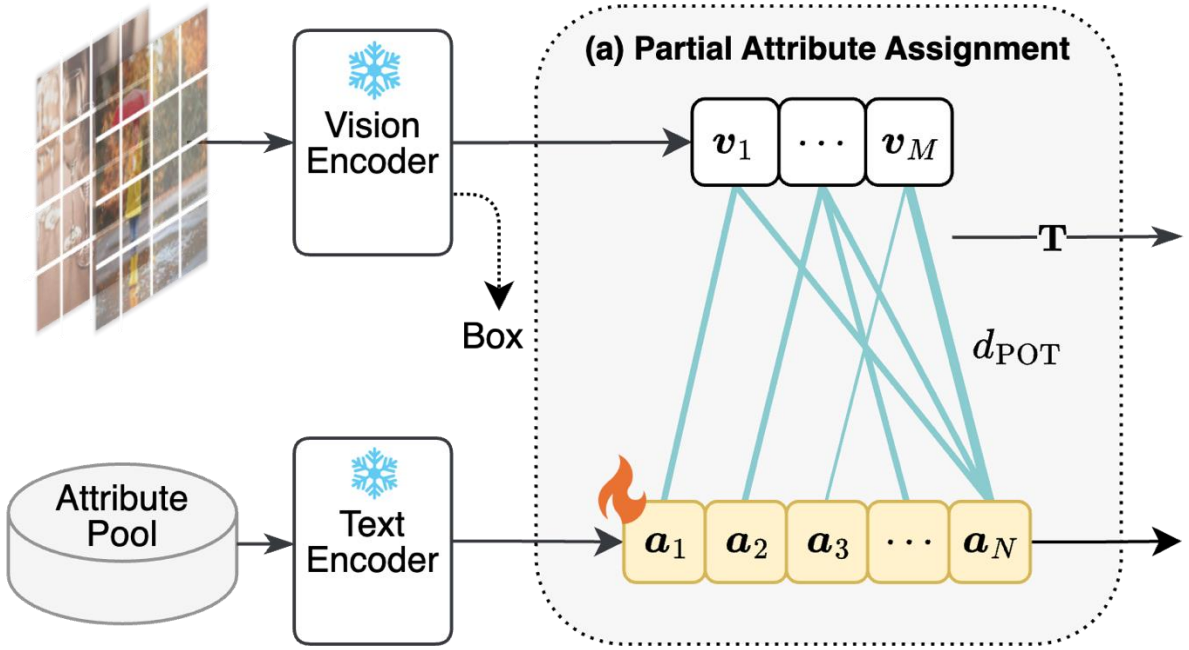
Method: overview

- We propose *Partial Attribute Assignment (PASS)*, a three-step algorithm that progressively selects and optimizes a targeted subset of relevant attributes throughout the training process in an end-to-end manner.



Step (a): Partial Attribute Assignment

- Initial optimization of attribute embeddings by solving the POT problem between visual and attribute embeddings



Optimize attribute embeddings by solving the following POT problem:

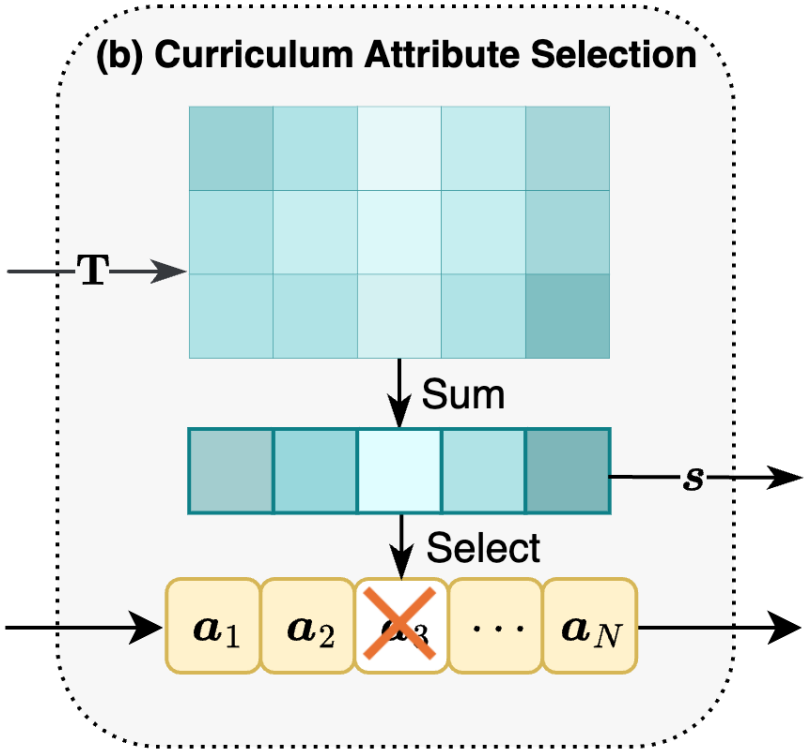
$$d_{\text{POT},\epsilon}(\mathcal{V},\mathcal{A};\mathbf{C}) \triangleq \min_{\mathbf{T}\in\Pi(\mathcal{V},\mathcal{A})} \langle \mathbf{T},\mathbf{C}\rangle_F - \epsilon h(\mathbf{T})$$

$$\mathcal{V} = \sum_{m=1}^{M} \frac{1}{M}\delta_{\boldsymbol{v}_m}, \qquad \mathcal{A} = \sum_{n=1}^{N} \frac{\alpha}{N}\delta_{\boldsymbol{a}_n}$$

$$\Pi(\mathcal{V},\mathcal{A}) \triangleq \{\mathbf{T}\in\mathbb{R}_+^{M\times N} | \mathbf{T}\mathbf{1}_N = \mathcal{V}, \mathbf{T}^\top \mathbf{1}_M \leqslant \mathcal{A}\}$$

Step (b): Curriculum Attribute Selection

- Gradually selecting a smaller subset of most relevant attributes based on relevance score originated from the POT plan



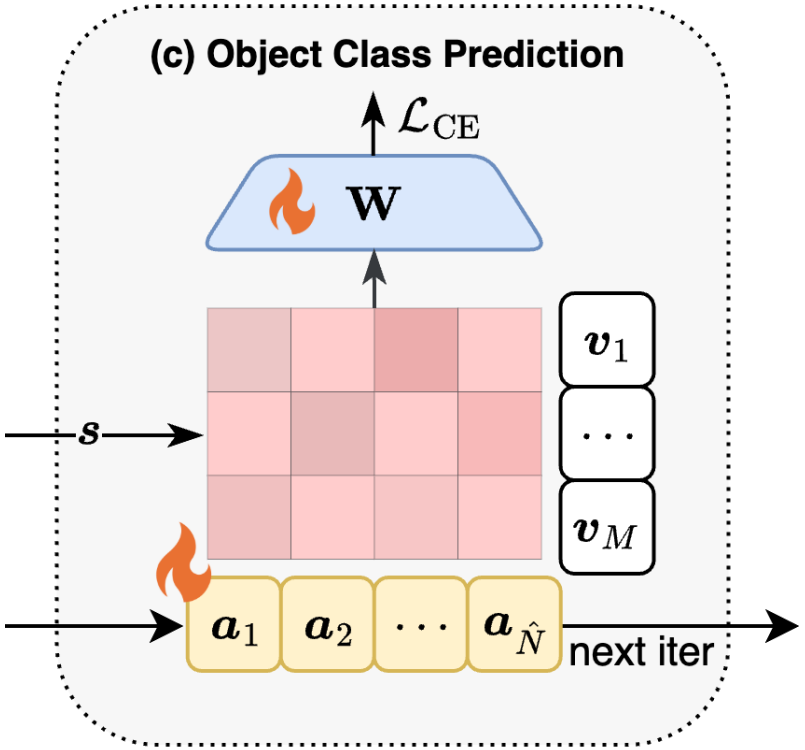
The POT plan $\mathbf{T}$ derived in step (a) manifests the correlation between individual visual and attribute embeddings. By summing along the column of $\mathbf{T}$, we define the relevance score of each attribute (*w.r.t.* visual objects) as

$$\boldsymbol{s} = [s_1, s_2, \ldots, s_N]^\top \qquad s_n = \frac{N}{\alpha} \sum_{m=1}^{M} T_{m,n}$$

Hence, we can gradually select a smaller subset of most relevant attributes based on $\boldsymbol{s}$ during training.

Step (c): Object Class Prediction

- Using the selected attributes to calculate the object class predictions of each visual input, and beginning the next three-step iteration



$\boldsymbol{s}$ can be further used to reweight the selected attributes when computing object class predictions, serving as a complementary "soft" filtering strategy:

$$p(O_k|\boldsymbol{v}) = \frac{\exp(\boldsymbol{w}_k^\top \mathbf{A}'^\top \boldsymbol{v})}{\sum_{k'=1}^{K} \exp(\boldsymbol{w}_{k'}^\top \mathbf{A}'^\top \boldsymbol{v})} \qquad \mathbf{A}' = [\boldsymbol{a}'_1, \boldsymbol{a}'_2, \ldots, \boldsymbol{a}'_{N'}]$$

$$\boldsymbol{a}'_n = s_n \hat{\boldsymbol{a}}_n$$

After optimized by the CE loss, attribute embeddings are fed back into step (a) for the next iteration.

Experiments: setup

- Classes in each dataset are halved into known and unknown classes

Dataset (↓)	Train Images	Test Images	Classes	Attributes
Aquatic	318	319	7 (4+3)	385
Aerial	5000	5000	20 (10+10)	1229
Game	788	787	59 (30+29)	390
Medical	93	89	12 (6+6)	390
Surgery	912	917	13 (6+7)	808

- Probability of unknown objects: *task relevance* (p_{ID}) × *unknownness* (p_{OOD})

$$p_{\text{ID}} = \max_{n \in \{1,\ldots,N'\}} \sigma(\boldsymbol{a}_n'^{\top} \boldsymbol{v}) \,,$$
$$p_{\text{OOD}} = 1 - \max_{k \in \{1,\ldots,K\}} p(O_k|\boldsymbol{v}) \,,$$

- Evaluation metric: mAP on known or unknown objects

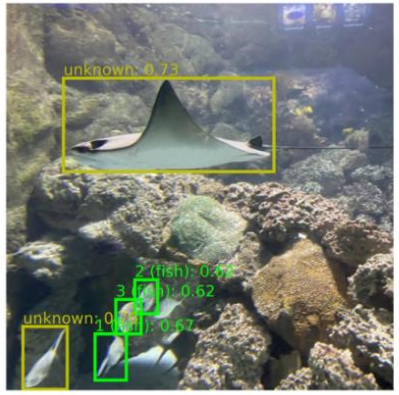
Experiments: comparison with SOTAs

Dataset (→)	Aquatic				Aerial				Game				Medical				Surgery				Overall			
Task ID (→)	Task 1		Task 2		Task 1		Task 2		Task 1		Task 2		Task 1		Task 2		Task 1		Task 2		Task 1		Task 2	
	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK
B/16 Backbone:																								
BASE-ZS+GT†	29.8	45.0	45.0	36.7	1.3	5.7	5.7	1.4	15.0	0.4	0.4	0.1	0.5	0.0	0.0	0.1	5.6	1.5	1.4	0.3	**10.4**	10.5	10.5	7.7
BASE-ZS	6.2	45.0	45.0	36.7	0.9	5.7	5.7	1.4	15.7	0.4	0.4	0.1	0.2	0.0	0.0	0.1	1.4	1.5	1.4	0.3	4.9	10.5	10.5	7.7
BASE-ZS+IN	26.5	45.1	45.1	36.7	1.9	5.7	5.7	1.4	2.4	0.3	0.3	0.0	0.6	0.0	0.0	0.1	1.7	1.4	1.0	0.3	6.6	10.5	10.4	7.7
BASE-ZS+LLM	24.7	45.1	45.1	36.5	1.4	5.7	5.7	1.4	15.1	0.4	0.4	0.1	0.6	0.0	0.0	0.1	8.9	1.5	1.3	0.3	10.2	10.5	10.5	7.7
BASE-FS	7.1	41.1	41.1	31.9	1.2	10.4	10.1	4.0	16.0	4.6	4.8	3.9	0.6	6.1	6.1	3.3	1.3	11.9	11.3	10.9	5.2	14.8	14.7	10.8
FOMO [78]	3.5	43.8	44.1	40.8	0.9	12.0	12.6	5.4	13.3	3.8	4.4	4.1	2.1	6.4	5.5	11.5	6.1	12.7	12.9	11.0	5.2	15.7	15.9	14.6
PASS (Ours)	5.2	43.4	43.2	46.6	1.9	14.0	16.0	7.0	21.5	10.0	7.7	9.0	4.9	8.4	6.8	12.1	14.3	15.6	13.1	14.7	9.6	**18.3**	**17.4**	**17.9**
Δ	+1.7	-0.4	-0.9	+5.8	+1.0	+2.0	+3.4	+1.6	+8.2	+6.2	+3.3	+4.9	+2.8	+2.0	+1.3	+0.6	+8.2	+2.9	+0.2	+3.7	+4.4	+2.5	+1.5	+2.3
L/14 Backbone:																								
BASE-ZS+GT†	34.8	36.0	36.0	42.3	1.0	7.9	7.2	0.8	12.4	0.9	0.8	0.3	2.4	0.2	0.2	0.3	2.4	0.2	2.6	1.3	10.6	9.0	9.4	9.0
BASE-ZS	0.7	35.9	36.0	42.3	9.1	8.2	7.2	0.8	6.8	0.9	0.8	0.3	0.0	0.2	0.2	0.3	3.6	2.9	2.6	1.3	4.1	9.6	9.4	9.0
BASE-ZS+IN	19.6	35.8	35.8	41.8	2.3	7.2	6.9	0.9	15.8	0.9	0.8	0.3	0.9	0.1	0.1	0.2	3.1	2.1	1.9	1.1	8.3	9.2	9.1	8.8
BASE-ZS+LLM	24.7	35.8	35.8	42.2	0.6	7.6	7.2	0.8	12.5	0.9	0.8	0.2	1.6	0.1	0.1	0.2	12.6	2.6	2.5	1.3	10.4	9.4	9.3	9.0
BASE-FS	2.4	43.6	42.9	42.8	9.7	23.7	21.9	13.0	8.2	10.4	10.2	13.4	1.1	23.2	21.7	24.2	3.6	26.0	25.0	7.4	5.0	25.4	24.3	20.2
FOMO [78]	18.2	50.1	48.1	47.1	6.0	25.3	23.7	16.0	30.4	10.7	9.9	11.2	9.4	21.8	19.9	34.6	12.0	29.0	28.9	8.5	15.2	27.4	26.1	23.5
PASS (Ours)	21.7	53.9	56.6	58.3	8.4	34.2	36.1	20.2	36.0	24.3	23.7	26.3	13.1	34.3	30.0	32.0	16.6	45.6	47.9	43.3	**19.1**	**38.5**	**38.9**	**36.0**
Δ	+3.5	+3.8	+8.5	+11.2	+2.4	+8.9	+12.4	+4.2	+5.6	+13.6	+13.8	+15.1	+3.7	+12.5	+10.1	-2.6	+4.6	+16.6	+19.0	+34.8	+3.9	+11.1	+12.8	+12.5

U*: Unknown mAP; **K: Known mAP; **PK**: Previously Known mAP; **CK**: Currently Known mAP*

Experiments: visualization

- Detection results and top attributes for detected objects

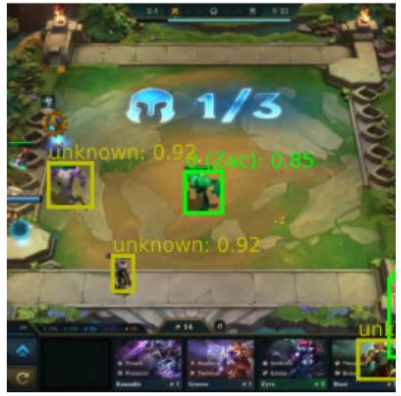


Top attributes for detected known classes:
- *Shape is truncate tail*
- *Shape is flexible*
- *Environment is tide pool*

Top attributes for detected unknown classes:
- *Shape is dorsal fin shape*
- *Texture is ridged*
- *Shape is adipose fin shape*

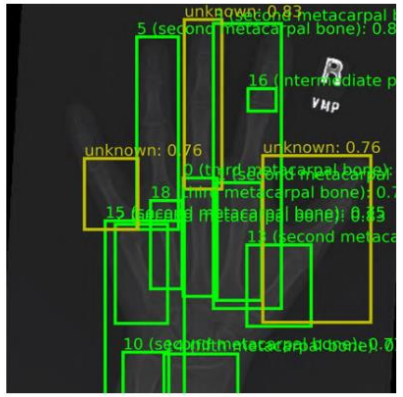
Top attributes for detected known classes:
- *Shape is stationary*
- *Shape is converging*
- *Behavior is tournament*

Top attributes for detected unknown classes:
- *Appearance is distinctive roof design*
- *Context is court lines*
- *Shape is stationary*



Top attributes for detected known classes:
- *Material is calcium*
- *Appearance is visible joints between bones*
- *Color is veins*

Top attributes for detected unknown classes:
- *Appearance is edges*
- *Shape is smooth*
- *Texture is striped*



Top attributes for detected known classes:
- *Shape is straight*
- *Shape is convex*
- *Shape is knobby*

Top attributes for detected unknown classes:
- *Shape is cracked*
- *Shape is straight*
- *Features is thumb metacarpal*

- Code: https://github.com/muliyangm/PASS
- Thank you for listening!