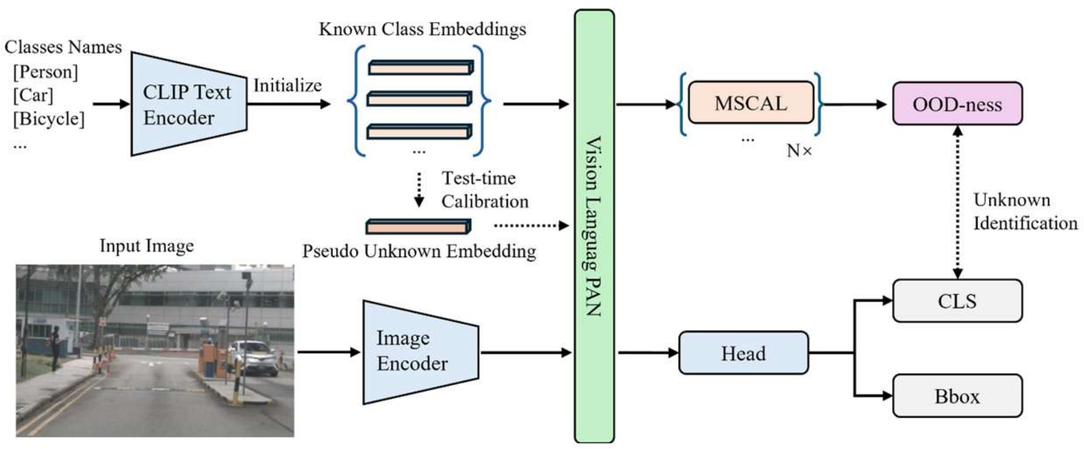
From Open Vocabulary to Open World: Teaching Vision Language Models to Detect Novel Objects

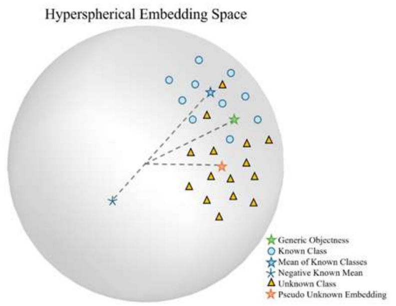
Zizhao Li, Zhengkang Xiang, Joseph West, Kourosh Khoshelham
Email: zizhao.li1@student.unimelb.edu.au

Introduction

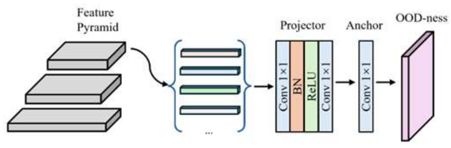
Object detection localizes and classifies objects but typically assumes all categories are known. In real-world settings like autonomous driving, this closed-set assumption is unsafe because unknown objects may be misclassified. Open world object detection addresses this by recognizing unknowns and learning new classes incrementally, yet existing replay-based methods are inefficient and have low unknown recall. Recent vision–language models enable open vocabulary detection via text prompts but struggle with semantically similar or unseen objects. We propose a unified framework with Open World Embedding Learning (OWEL) and Multi-Scale Contrastive Anchor Learning (MSCAL) to enhance unknown discovery and incremental learning.

Model Architecture





Pseudo Unknown Embedding



MSCAL Module

Overview of the proposed method. Known class embeddings are initialized from a pretrained CLIP text encoder. The image encoder extracts multi-scale features, which are fused with text features through RepVL-PAN to form feature pyramids. The detection head predicts classes via image–text similarity and regresses bounding boxes. MSCAL aligns anchors and features across scales to produce OOD score maps that reduce known–unknown confusion. Pseudo unknown embeddings are generated from optimized known embeddings and the generic "objectness" concept for unknown discovery.

Experimental Results

Experimental Results on OWOD benchmarks. Our method largely outperforms the SOTA methods in terms of both known mAP and unknown recall (U-Recall) on both benchmarks.

Task IDs (→)	Task 1		Task 2				Task 3				Task 4		
Method	U-Recall (↑)	mAP (↑) Current known	U-Recall (↑)	mAP (↑) Previously known	Current known	Both	U-Recall (↑)	mAP (↑) Previously known	Current known	Both	mAP (↑) Previously known	Current known	Both
ORE [18]	4.9	56.0	2.9	52.7	26.0	39.4	3.9	38.2	12.7	29.7	29.6	12.4	25.3
OST† [61]	-	56.2	-	53.4	26.5	39.9	-	38.0	12.8	29.6	30.1	13.3	25.9
OW-DETR [14]	7.5	59.2	6.2	53.6	33.5	42.9	5.7	38.3	15.8	30.8	31.4	17.1	27.8
UC-OWOD [60]	2.4	50.7	3.4	33.1	30.5	31.8	8.7	28.8	16.3	24.6	25.6	15.9	23.2
ALLOW [32]	13.6	59.3	10.0	53.2	34.0	45.6	14.3	42.6	26.7	38.0	33.5	21.8	30.6
PROB [67]	19.4	59.5	17.4	55.7	32.2	44.0	19.6	43.0	22.2	36.0	35.7	18.9	31.5
CAT [30]	23.7	60.0	19.1	55.5	32.7	44.1	24.4	42.8	18.7	34.8	34.4	16.6	29.9
RandBox [57]	10.6	61.8	6.3	-	-	45.3	7.8	-	-	39.4	-	-	35.4
EO-OWOD [51]	24.6	61.3	26.3	55.5	38.5	47.0	29.1	46.7	30.6	41.3	42.4	24.3	37.9
MAVL* [33]	50.1	64.0	49.5	61.6	30.8	46.2	50.9	43.8	22.7	36.8	36.2	20.6	32.3
SKDF* [31]	39.0	56.8	36.7	52.3	28.3	40.3	36.1	36.9	16.4	30.1	31.0	14.7	26.9
YOLO-World‡	16.6	71.9	16.1	71.8	48.1	60.0	13.0	60.0	40.7	53.6	53.7	33.9	48.7
Ours	**73.5**	**72.1**	**77.5**	**72.4**	**51.0**	**61.7**	**76.1**	**61.6**	**41.6**	**54.9**	**56.0**	**34.3**	**50.6**
ORE [18]	1.5	61.4	3.9	56.5	26.1	40.6	3.6	38.7	23.7	33.7	33.6	26.3	31.8
OW-DETR [14]	5.7	71.5	6.2	62.8	27.5	43.8	6.9	45.2	24.9	38.5	38.2	28.1	33.1
PROB [67]	17.6	73.4	22.3	66.3	36.0	50.4	24.8	47.8	30.4	42.0	42.6	31.7	39.9
CAT [30]	24.0	74.2	23.0	67.6	35.5	50.7	24.6	51.2	32.6	45.0	45.4	35.1	42.8
EO-OWOD [51]	24.6	71.6	27.9	64.0	39.9	51.3	31.9	52.1	42.2	48.8	48.7	38.8	46.2
SKDF* [31]	60.9	69.4	60.0	63.8	26.9	44.4	58.6	46.2	28.0	40.1	41.8	29.6	38.7
YOLO-World ‡	29.0	75.6	26.1	**75.7**	55.3	65.0	26.9	65.1	54.4	61.6	61.4	55.2	59.9
Ours	**71.3**	**76.4**	**74.4**	75.0	**59.8**	**67.0**	**74.6**	**67.0**	**53.8**	**62.6**	**65.5**	**56.9**	**63.4**

PROB EO-OWOD Ours

Qualitative results on COCO and nuScenes. Our method provides more meaningful bounding boxes for unknown classes, without significantly losing performance on known classes, which makes it easier to make trade-off between precision and recall in various applications.

Code available at: https://github.com/343gltysprk/ovow