UDLVM: Unknown Object Detection with Large Vision Models

Shulin Chen[1,2], Wangshu Yao[1,2,†], Shuxin Zhang[1,2], Xiaofeng Jiang[1,2]
[1]*School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu, China*
[2]*School of Software, Soochow University, Suzhou, Jiangsu, China*
20244227059@stu.suda.edu.cn, wshyao@suda.edu.cn, 20244227030@stu.suda.edu.cn, xfjiang@suda.edu.cn

***Abstract*—Open World Object Detection (OWOD) aims to detect objects of known and unknown categories in the real open scenarios, and can incrementally learn novel category knowledge under the guidance of humans simultaneously. However, due to the lack of unknown class labels, existing OWOD methods suffer from two main issues: semantic bias towards known objects and low recall of unknown objects. To address these issues, we propose a new and effective detection framework called "Unknown Object Detection with Large Vision Models (UDLVM)", which leverages two prevalent large vision models for the OWOD task, namely Visual State Space Model (VMamba) and Segment Anything Model (SAM). For one thing, a new VSSFPN module is designed, which incorporates the core of VMamba in the Feature Pyramid Network (FPN) for the better feature extraction to alleviate the semantic bias towards known objects. For another thing, a new DUSC method is developed, which utilizes the masks generated by SAM to assist in producing high-quality pseudo labels to improve the recall of unknown objects. Extensive experiments on public benchmark datasets show that our UDLVM framework can effectively alleviate the semantic bias issue and improve the unknown recall compared with existing methods.**

***Index Terms*—Open World Object Detection, Large Vision Model, Feature Extraction, Pseudo Label**

I. Introduction

Object detection is a classical task in the field of computer vision, which has been widely used in various domains, such as autonomous driving and remote sensing images. However, conventional object detection is restricted to a closed-set space, which can only detect categories learned in the training set and cannot identify novel categories appearing in the real open scenarios. Fortunately, Open World Object Detection (OWOD) [1] introduced by Joseph et al. extends this detection framework, which aims to not only detect both known and unknown objects in the real open scenarios, but also incrementally learn novel category knowledge under the guidance of humans. Recent years, OWOD has attracted extensive attention from researchers all over the world, and has made breakthrough progress in various aspects. Nevertheless, due to the lack of unknown class labels, existing OWOD methods suffer from two main issues: semantic bias towards known objects and low recall of unknown objects. For one thing, current methods tend to use a pseudo-label strategy to generate the supervision of unknown objects, but these pseudo labels are generated from the model that is trained only on known objects, thereby the model is inevitably biased towards known objects. For another thing, the recall of unknown objects is low because there are no ground-truth labels for unknown objects actually.

To address these issues, we propose a new and effective detection framework called "Unknown Object Detection with Large Vision Models (UDLVM)", which leverages two prevalent large vision models for the OWOD task, namely Visual State Space Model (VMamba) [2] and Segment Anything Model (SAM) [3]. Specifically, based on the core component "Visual State Space Block (VSSBlock)" in VMamba, which is particularly adept at capturing long-range dependencies in vision tasks, a "Visual State Space Feature Pyramid Network (VSSFPN)" module that incorporates VSSBlock in the Feature Pyramid Network (FPN) [4] is designed to alleviate the semantic bias issue. Then, leveraging SAM's ability to generate category-agnostic masks for all instances in an image, a new method named "Detecting Unknowns with SAM-assisted CROWD-D (DUSC)" is developed to assist CROWD-D (one module of our baseline) in producing high-quality pseudo-labels and improve the recall of unknown objects.

Fig. 1 illustrates the comparison between previous methods and our proposed method. Previous methods tend to use a common FPN for feature extraction and a simple strategy for pseudo-label generation, which selects top-k high-score prediction boxes according to "objectness" as pseudo labels where "objectness" indicates the confidence that a prediction box contains an unknown object. By contrast, our UDLVM framework incorporates VSSBlock in FPN for the better feature extraction and utilizes masks generated by SAM to assist CROWD-D in producing high-quality pseudo labels for unknown objects. The main contributions of this paper are summarized as follows:

- UDLVM incorporates VSSBlock in FPN for the better feature extraction (VSSFPN) so that it can learn more discriminative representations and alleviate the semantic bias issue. To the best of our knowledge, it is the first attempt to introduce visual SSM [5] into the OWOD task.
- UDLVM utilizes masks generated by SAM to assist CROWD-D in producing high-quality pseudo labels for unknown objects (DUSC), which effectively improves the recall of unknown objects.
- Extensive experiments on two public benchmark datasets, namely M-OWODB [1] and S-OWODB [6], show the effectiveness of our UDLVM framework.

† Corresponding author.

II. Related Work

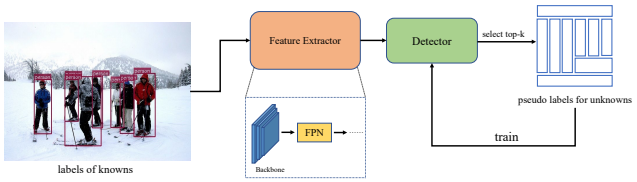
A. Open World Object Detection

Open World Object Detection [1] extends the conventional closed-set object detection task where it not only detects both known and unknown objects in the real open scenarios, but also incrementally learns novel category knowledge under the guidance of humans. Based on Faster R-CNN [7] framework, ORE [1] proposes an energy-based classifier and a contrastive clustering approach to identify unknown objects. Then OW-DETR [6] introduces Deformable DETR [8] framework into the OWOD task and designs an attention-based pseudo label scheme to distinguish unknown objects. Subsequently, CAT [9] decouples the localization and identification process via the cascade decoding manner and proposes a new self-adaptive pseudo-labelling mechanism to mine unknown objects. PROB [10] proposes a new probabilistic objectness-based approach via a Gaussian likelihood, which can better separate unknown objects from the background. ALLOW [11] designs a label-transfer learning strategy to accomplish the collaborative learning of known and unknown categories. Further, Hyp-OW [12] argues that there must exist a hierarchical structure between known and unknown objects, while MEPU [13] advocates that the foreground and background regions form two different distributions. Several notable works leveraging the large model for external supervision also have excellent performance. For example, SGROD [14] and KTCN [15] utilize SAM [3] to generate high-quality pseudo labels. Along with them, SKDF [16] distills and acquires knowledge from GLIP [17] to enhance the detection of unknown objects. Complementary to these, CROWD [18] designs a CROWD-D module to mine unknown objects and a CROWD-L module for representation learning with the help of submodular functions [19]. Based on CROWD, this paper leverages two large vision models, namely VMamba [2] and SAM for the further improvement.

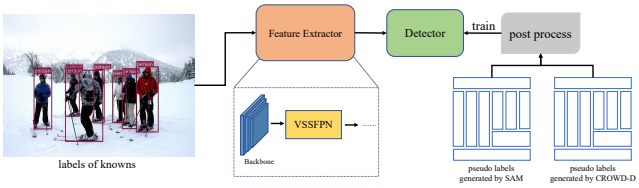
B. Large Vision Model

Large vision models with numerous parameters and pre-trained on huge amounts of data have attracted significant attention in the field of computer vision. CLIP [20] introduced by OpenAI is a vision-language model trained on large-scale image-text pairs via contrastive learning. GLIP [17] introduced by Microsoft unifies the phrase grounding and object detection tasks and SKDF [16] distills and acquires knowledge from it to enhance the detection of unknown objects. Moreover, SAM [3] can produce high-quality object masks from input prompts or generate masks for all instances in an image. Inspired by SGROD [14] and KTCN [15], this paper explores the application of SAM in the OWOD task.

Recent years, State Space Model (SSM) [5] has shown great potential in modeling long-range dependencies in the field of natural language processing. Vim [21] and VMamba [2] introduce SSM into the computer vision field to extract features for the first time. Mamba YOLO [22] designs a new backbone based on VMamba and YOLO [23], which achieves excellent performance in the conventional object detection task. Inspired by Mamba YOLO, we incorporate the core module of VMamba in FPN for the better feature extraction. And to the best of our knowledge, this is the first attempt to introduce SSM into the OWOD task.



(a) Previous Methods



(b) Our Method

Fig. 1: Comparison between previous methods and our proposed method. (a) Previous methods use a common FPN for feature extraction and a top-k selection strategy for pseudo label generation. (b) Our method incorporates VSSBlock in FPN for the better feature extraction and utilizes masks generated by SAM to assist CROWD-D in producing high-quality pseudo labels.

III. Method

A. Problem Definition

Given the task id t of the incremental learning, the set of known categories is defined as $\mathcal{K}^t = \{1, 2, ..., C\}$ and the set of unknown categories is defined as $\mathcal{U}^t = \{C+1, C+2, ...\}$ where C denotes the number of known categories that have been learned before. The known objects are labeled in the training set $\mathcal{D}^t = \{X^t, Y^t\}$ where $X^t = \{I_1, I_2, ..., I_M\}$ and $Y^t = \{L_1, L_2, ..., L_M\}$ represent the images and labels of M training samples respectively. If an image I_i contains N instances, then $L_i = \{l_1, l_2, ..., l_N\}$ includes the corresponding labels where $l_j = [c_j, x_j, y_j, w_j, h_j]$ denotes the known category id and the bounding box. It should be noted that there are no labels for unknown objects in $\mathcal{D}^t$ actually.

Our goal is to train a model $\mathcal{M}^t$ that can detect both known and unknown objects, alongside classify unknown objects as a novel category named "unknown". In the task t, humans will first annotate n novel categories from $\mathcal{U}^t$ for $\mathcal{M}^t$ to learn. In case of catastrophic forgetting [24], $\mathcal{M}^t$ must be fine-tuned on $\mathcal{K}^t$ and this newly annotated n categories subsequently. When the task t is finished and the task $t+1$ begins, $\mathcal{K}^t$ will be updated to $\mathcal{K}^{t+1} = \mathcal{K}^t \cup \{C+1, C+2, ..., C+n\}$ and $\mathcal{U}^t$ will be updated to $\mathcal{U}^{t+1} = \{C+n+1, C+n+2, ...\}$. At this point, our goal is to train a new model $\mathcal{M}^{t+1}$ through pretraining and fine-tuning without retraining from scratch on the whole dataset.

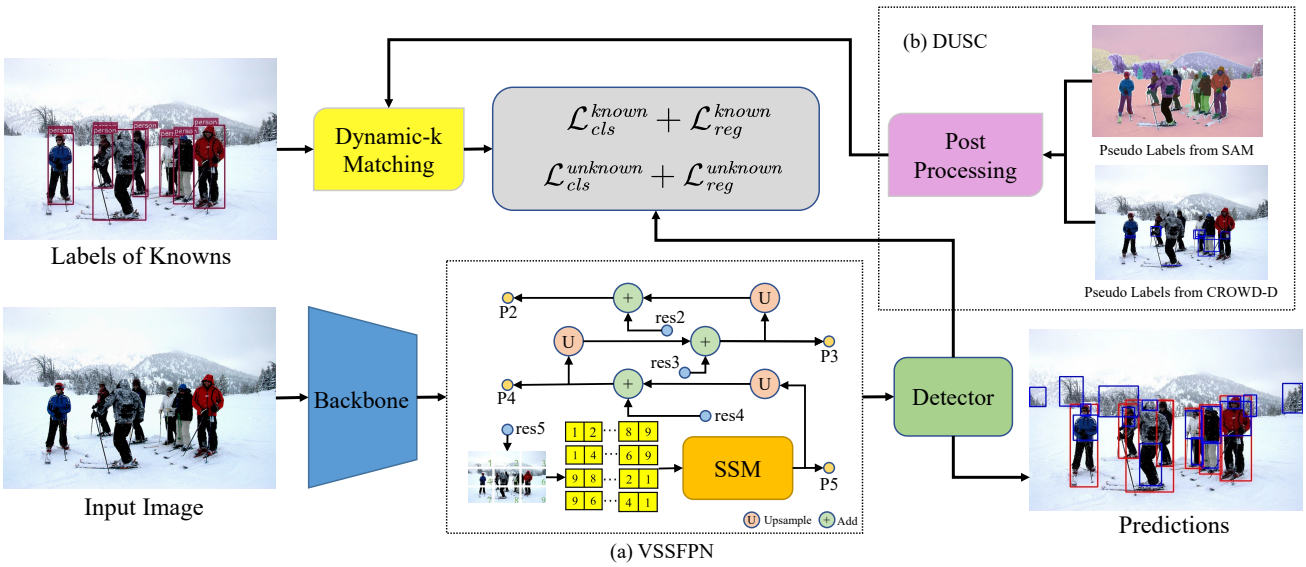


Fig. 2: Overall architecture of our proposed UDLVM. Firstly in (a), VSSFPN receives multi-scale feature maps from the backbone network and applies VSSBlock to the feature map of the lowest resolution for feature extraction. Then these feature maps are fused and fed into a detector to perform inference. Subsequently in (b), DUSC utilizes SAM to assist CROWD-D in generating high-quality pseudo labels for unknown objects. Finally, a dynamic-k matching algorithm is performed on all available labels with the initial predictions for sample selection and label assignment to calculate the total training loss.

B. Overall Architecture of UDLVM

Fig. 2 illustrates the overall architecture of our UDLVM, which provides a new solution to address semantic bias and low unknown recall issues in the OWOD task. Firstly, the backbone network extracts multi-scale feature maps from the input image. Assuming the input image $I \in \mathbb{R}^{H\times W\times 3}$, these feature maps have resolutions of $\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$, $\frac{H}{16}\times\frac{W}{16}$ and $\frac{H}{32}\times\frac{W}{32}$ respectively, along with channels of four different quantities:

$$[res_2, res_3, res_4, res_5] = Backbone\,(I) \tag{1}$$

where $res_i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C_i}$ denotes hierarchical representations from different backbone network stages. Subsequently in Fig. 2 (a), VSSFPN receives them and applies VSSBlock to the feature map of the lowest resolution for feature extraction (III-C). Then these feature maps are forwarded and fused to generate multi-scale outputs $[P_2, P_3, P_4, P_5]$. Next, a detector uses RoI Pooler [7] to process these multi-scale outputs:

$$f = RoI\ \ Pooler\,([P_2, P_3, P_4, P_5], randbox) \tag{2}$$

where f denotes the intermediate feature map and $randbox$ denotes random boxes generated by RandBox [25] replacing the conventional proposal boxes generated by RPN [7]. Then the detector exploits the detection head to perform inference, whose results includes three aspects, namely category logits, bounding boxe coordinates and objectness scores:

$$[cls\_logit, bbox, objectness] = head\,(f) \tag{3}$$

where the $objectness$ is defined as the sigmoid value of the cls_logit followed by an extra network branch. Afterwards in Fig. 2 (b), DUSC utilizes SAM to assist CROWD-D in generating high-quality pseudo labels for unknown objects, which need to be further refined through a post processing operation (III-D). And then, a dynamic-k matching algorithm [25] is performed on all available labels (pesudo labels and ground-truth labels) with the initial predictions to select samples and assign labels for the final loss calculation. Finally, the category loss $\mathcal{L}_{cls}$ and bounding box loss $\mathcal{L}_{reg}$ for known and unknown objects are calculated so that the total training loss function can be defined as:

$$\mathcal{L}^{total} = \mathcal{L}_{cls}^{known} + \mathcal{L}_{reg}^{known} + \mathcal{L}_{cls}^{unknown} + \mathcal{L}_{reg}^{unknown}. \tag{4}$$

When in the inference phase, we calculate the confidence score of each prediction box by:

$$score = softmax\,(cls\_logit)\cdot objectness, \tag{5}$$

and the prediction box whose score is lower than a certain threshold will be filtered out.

C. Visual State Space Feature Pyramid Network

The design of our VSSFPN module is disscussed in this section. Firstly, we unify the quantity of channels of $[res_2, res_3, res_4, res_5]$ to K through four different lateral convolutional layers:

$$C_i = lateral\_conv_i\,(res_i), i \in \{2,3,4,5\} \tag{6}$$

where $C_i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times K}$ denotes the feature map after the channel unification. Subsequently, considering the computational cost, we only apply VSSBlock to C_5 that possesses the lowest resolution for further feature extraction. VSSBlock

mainly consists of three parts: cross-scan, SSM and cross-merge. The cross-scan unfolds the 2D feature map into a 1D sequence through four different traversal paths. Then the SSM utilizes its strong long sequence modeling ability to process these 1D sequences, and the resultant sequences are converted back to 2D feature maps by the cross-merge. This process can be formulated as:

$$f_5 = cross\_merge(SSM(cross\_scan(C_5))) \tag{7}$$

where $f_5 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times K}$ denotes the intermediate feature map after VSSBlock. Finally, we use FPN to fuse these feature maps and generate multi-scale outputs:

$$P_i = output\_conv_i\left(f_i\right), i \in \{2,3,4,5\}, \tag{8}$$

$$f_i = C_i \oplus Upsample\left(f_{i+1}\right), i \in \{2,3,4\} \tag{9}$$

where $output\_conv_i$ denotes an output convolutional layer, the operator $\oplus$ denotes the addition operation of feature maps and $Upsample$ denotes a nearest neighbor upsampling operation with a scale factor of 2.

D. Detecting Unknowns with SAM-assisted CROWD-D

The details of our DUSC method is disscussed in this section. As shown in Algorithm 1, there are three phases in the process of our model training: pretraining, mining and fine-tuning. In the pretraining phase, only the ground-truth labels of known objects are used to train the model. Then following CROWD-D, the model parameters are freezed and unknown objects $bbox_{crowdd}$ are mined by maximizing Submodular Conditional Gain (SCG) functions [18] in the mining phase. Finally in the fine-tuning phase, the masks generated by SAM are utilized to assist CROWD-D in producing high-quality pseudo labels for unknown objects. SAM can generate category-agnostic masks for all instances in an image, then by using the masks of each instance to calculate the maximum bounding rectangle, the initial pseudo labels $bbox_1$ are obtained. Subsequently, $bbox_1$ are refined via Intersection over Union (IoU) to obtain $bbox_2$:

$$bbox_2 = \left\{ b \in bbox_1 \mid \max_{i \in [1,|gt|]} \{IoU\left(b, gt_i\right)\} \leq \lambda \right\} \tag{10}$$

where gt denotes the ground-truth boxes of known objects in the image. Next the boxes in $bbox_2$ that have an aspect ratio outside the range of $[a, b]$ are filtered out and random k boxes are sampled as the pseudo labels $bbox_3$. For $bbox_{crowdd}$ generated in the mining phase, the boxes that overlap too much amomg them are reduced via Non-Maximum Suppression (NMS). Afterwards, $bbox_{crowdd}$ is further refined following the operations that are used in processing $bbox_1$. Then the refined results are merged with $bbox_3$ to obtain the final high-quality pseudo labels $L^{'}$.

Algorithm 1: DUSC Algorithm.

Input: input image I, model $\mathcal{M}^t$, pretrained SAM model $\mathcal{M}_{\text{SAM}}$.
Output: high-quality pseudo labels $L^{'}$.
Step (1) in the pretraining phase:
  train $\mathcal{M}^t$ with ground-truth labels of knowns;
  obtain model $\mathcal{M}^t$ after pretraining.
Step (2) in the mining phase:
  freeze the parameters of $\mathcal{M}^t$;
  use $\mathcal{M}^t$ to mine unknown objects $bbox_{crowdd}$;
  obtain pseudo labels $bbox_{crowdd}$.
Step (3) in the fine-tuning phase:
  generate masks via SAM: $masks = \mathcal{M}_{SAM}\left(I\right)$;
  obtain initial pseudo labels $bbox_1$ via $masks$;
  obtain refined pseudo labels $bbox_2$ using (10);
  obtain final $bbox_3$ by filtering and sampling;
  reduce duplicate boxes in $bbox_{crowdd}$;
  refine $bbox_{crowdd}$ by repeating line 11 to 12;
  merge the refined results with $bbox_3$ to obtain $L^{'}$;
  return $L^{'}$.

IV. Experiment

A. Dataset

Our method is evaluated on two public OWOD benchmark datasets, namely M-OWODB [1] and S-OWODB [6]. M-OWODB divides images from MS-COCO and PASCAL-VOC datasets into four tasks, each of which contains 20 different categories. When the task t where $t \in [1, 4]$ arrives, the first $20 \cdot t$ categories are treated as known objects and the rest are treated as unknown objects. The model is only trained with labels of known objects whose category ids belong to $[20 \cdot (t-1), 20 \cdot t - 1]$, but identifies known objects whose ids belong to $[0, 20 \cdot t - 1]$ and classifies unknown objects as "unknown" denoted by id 80 during inference. S-OWODB is similar to this, but it only uses images from MS-COCO dataset and groups categories with the same semantic meaning into one task, which is more challenging than M-OWODB.

B. Metric

The mean Average Precision (mAP) is used to evaluate the detection performance of known objects. For unknown objects, the Unknown Recall (U-Recall) is used to measure the model's ability to discover unknowns. Besides, Wildness Impact (WI) and Absolute Open-Set Error (A-OSE) are employed to evaluate the open-world detection performance. WI measures the relative error in classifying unknown objects as known categories, while A-OSE measures the absolute error that the model misclassifies unknown objects as known categories.

C. Implementation Details

Following CROWD [18], our proposed framework is based on Faster R-CNN [7] with a pretrained ResNet-50 backbone. Our model is trained with batch size 12, optimizer AdamW, learning rate 2.5×10^{-5} and weight decay 1×10^{-4}. Besides, we set the channel quantity K of C_i to 256 in (6) and λ to 0.5 in (10). In III-D, the aspect ratio range $[a, b]$ is set to $[0.3, 4.0]$ and the sample quantity k is set to 10. All our experiments are conducted on single TeslaV100-SXM2-32GB GPU.

TABLE I: Experimental results on M-OWODB and S-OWODB benchmarks. (↑) denotes that the higher the better, while (↓) denotes that the lower the better. And * indicates that the results are reproduced by us. It is noted that U-Recall, WI and A-OSE are not computed in Task 4 because all categories are known in this task. The best results are highlighted in **bold**.

Task IDs(→)	Task 1				Task 2				Task 3				Task 4
Method	U-Recall (↑)	WI (↓)	A-OSE (↓)	mAP (↑)	U-Recall (↑)	WI (↓)	A-OSE (↓)	mAP (↑)	U-Recall (↑)	WI (↓)	A-OSE (↓)	mAP (↑)	mAP (↑)
M-OWODB Benchmark Results													
ORE [1]	4.9	0.0621	10459	56.0	2.9	0.0282	10448	39.4	3.9	0.0211	7990	29.7	25.3
OW-DETR [6]	7.5	0.0571	10240	59.2	6.2	0.0278	8441	42.9	5.7	0.0156	6803	30.8	27.8
ALLOW [11]	13.6	0.0564	46589	59.3	10.0	0.0274	24709	45.6	14.3	0.0194	14952	38.0	30.6
PROB [10]	19.4	0.0569	5195	59.5	17.4	0.0344	6452	44.0	19.6	0.0151	2641	36.0	31.5
CAT [9]	23.7	-	-	60.0	19.1	-	-	44.1	24.4	-	-	34.8	29.9
Hyp-OW [12]	23.5	-	11275	59.4	20.6	-	4805	44.4	26.3	-	3548	36.8	33.6
KTCN [15]	41.5	0.0809	13368	60.2	38.6	0.0439	11861	**46.0**	39.7	0.0250	6382	36.4	30.4
MEPU-FS [13]	31.6	0.0560	6050	60.2	30.9	0.0230	5925	44.8	30.1	0.0160	5159	35.4	30.9
SGROD [14]	34.3	0.0588	**4177**	59.8	32.6	0.0287	2415	44.9	32.7	0.0164	1526	36.0	31.2
SKDF [16]	39.0	-	-	56.8	36.7	-	-	40.3	36.1	-	-	30.1	26.9
CROWD* [18]	49.9	0.0347	4523	59.6	54.5	0.0133	2262	45.8	69.2	0.0099	1742	**39.9**	**38.1**
UDLVM (Ours)	**76.3**	**0.0323**	4448	**60.2**	**72.9**	**0.0115**	**2035**	44.9	**74.5**	**0.0092**	**1478**	39.0	37.8
S-OWODB Benchmark Results													
ORE [1]	1.5	-	-	61.4	3.9	-	-	40.6	3.6	-	-	33.7	31.8
OW-DETR [6]	5.7	-	-	71.5	6.2	-	-	43.8	6.9	-	-	38.5	33.1
CROWD* [18]	55.4	0.0219	9560	**71.7**	49.3	**0.0118**	**2170**	50.6	65.0	**0.0087**	**2000**	48.2	47.4
UDLVM (Ours)	**76.2**	**0.0210**	**8254**	70.6	**73.1**	0.0130	2477	**50.9**	**73.3**	0.0095	2283	**48.4**	**47.6**

D. Results on OWOD Benchmarks

The performance of our method is compared with several existing methods on M-OWODB and S-OWODB as shown in Table I. On the M-OWODB benchmark, except for the A-OSE metric of SGROD in Task 1, UDLVM surpasses all existing methods in terms of the U-Recall, WI and A-OSE metrics across all tasks. This can be attributed to contributions of VSSFPN that incorporates VSSBlock for the better feature extraction to alleviate semantic bias and DUSC that utilizes masks generated by SAM to assist in producing high-quality pseudo labels to improve unknown recall. But for the mAP metric, UDLVM overall has a slight drop compared with CROWD because DUSC introduces massive pseudo labels for unknown objects, which may interfere with the learning of known categories to some extent. On the S-OWODB benchmark, since most existing methods have not reported their results on the WI and A-OSE metrics, we mainly compare our method with CROWD. Benefiting from DUSC, UDLVM outperforms CROWD significantly in terms of the U-Recall metric across all tasks. Moreover, UDLVM achieves better performance on the WI and A-OSE metrics in Task 1, and improve the mAP metric from Task 2 to Task 4 slightly.

E. Ablation

Ablation experiments are conducted on the M-OWODB benchmark to analyze the contributions of VSSFPN and DUSC in our method as shown in Table II. When removing VSSFPN from UDLVM, the values of WI and A-OSE rise obviously across all tasks, which demonstrates that UDLVM suffers more from the semantic bias issue without VSSFPN. Similarly, when DUSC is removed from UDLVM, the value of U-Recall drops significantly across all tasks, which indicates that DUSC effectively improves the model's ability to discover unknown objects. Overall, both VSSFPN and DUSC contribute significantly to the performance of our method.

Moreover, further ablation experiments on the hyperparameters in DUSC are conducted as shown in Table III. When changing the sample quantity k in DUSC among $\{5, 10, 15\}$, U-Recall achieves the best performance when k is set to 15, while WI, A-OSE and mAP achieve the best performance when k is set to 5. Considering the overall performance, we select to set $k = 10$ as a tradeoff. When setting $k = 10$ and varying the IoU threshold λ in (10) among $\{0.5, 0.7\}$, the metrics except for A-OSE achieve the best performance when λ is set to 0.5. Overall, we choose to set $\lambda = 0.5$ and $k = 10$ in our final implementation.

V. Conclusion

In this paper, a new framework named UDLVM that consists of VSSFPN and DUSC is proposed for the OWOD task. VSSFPN incorporates VSSBlock, the core of VMamba into FPN for the better feature extraction to alleviate the semantic bias towards known objects. DUSC utilizes masks generated by SAM to assist in generating high-quality pseudo labels for unknown objects, which effectively improves the model's ability to discover unknown objects. Extensive experiments on two public OWOD benchmark datasets show the effectiveness of our proposed method. Despite the promising results of U-Recall, WI and A-OSE achieved by UDLVM, the mAP metric on known objects still has a slight drop compared with our baseline, indicating that the discovering of unknown objects may interfere with the learning of known categories to some extent, which needs to be further explored in the future work.

TABLE II: Ablation experiments on M-OWODB benchmark. We analyze the contributions of VSSFPN and DUSC in our method. UDLVM-VSSFPN and UDLVM-DUSC denote the variants without VSSFPN and DUSC respectively.

Task IDs(→)	Task 1				Task 2				Task 3				Task 4
Method	U-Recall (↑)	WI (↓)	A-OSE (↓)	mAP (↑)	U-Recall (↑)	WI (↓)	A-OSE (↓)	mAP (↑)	U-Recall (↑)	WI (↓)	A-OSE (↓)	mAP (↑)	mAP (↑)
UDLVM-VSSFPN	76.3	0.0331	4667	59.8	**74.2**	0.0136	2317	44.7	73.9	0.0097	1788	39.2	**38.4**
UDLVM-DUSC	51.3	**0.0321**	**4108**	60.0	52.2	0.0127	2042	**45.3**	68.6	**0.0090**	**1466**	**40.2**	37.9
UDLVM	**76.3**	0.0323	4448	**60.2**	72.9	**0.0115**	**2035**	44.9	**74.5**	0.0092	1478	39.0	37.8

TABLE III: Ablation experiments of hyperparameter in DUSC on the M-OWODB benchmark in Task 1.

Settings		U-Recall (↑)	WI (↓)	A-OSE (↓)	mAP (↑)
$\lambda = 0.5$	$k = 5$	73.0	**0.0310**	**4132**	**60.9**
	$k = 10$	76.3	0.0323	4448	60.2
	$k = 15$	**76.5**	0.0330	4367	58.7
$\lambda = 0.7$	$k = 5$	72.3	**0.0322**	**4265**	**60.9**
	$k = 10$	75.9	0.0326	4319	59.6
	$k = 15$	**76.8**	0.0353	4354	58.8

ACKNOWLEDGMENT

This work was partially supported by Suzhou science and technology plan project(SYG202349), Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] K. J. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5826–5836.

[2] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 103 031–103 063.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.

[4] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[5] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First Conference on Language Modeling*, 2024.

[6] A. Gupta, S. Narayan, K. J. Joseph, S. Khan, F. S. Khan, and M. Shah, "Ow-detr: Open-world detection transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9225–9234.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[8] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021.

[9] S. Ma, Y. Wang, Y. Wei, J. Fan, T. H. Li, H. Liu, and F. Lv, "Cat: Localization and identification cascade detection transformer for open-world object detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 681–19 690.

[10] O. Zohar, K.-C. Wang, and S. Yeung, "Prob: Probabilistic objectness for open world object detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 11 444–11 453.

[11] Y. Ma, H. Li, Z. Zhang, J. Guo, S. Zhang, R. Gong, and X. Liu, "Annealing-based label-transfer learning for open world object detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 11 454–11 463.

[12] T. Doan, X. Li, S. Behpour, W. He, L. Gou, and L. Ren, "Hypow: Exploiting hierarchical structure learning with hyperbolic distance enhances open world object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1555–1563.

[13] R. Fang, G. Pang, W. Miao, X. Bai, J. Zheng, and X. Ning, "Unsupervised recognition of unknown objects for open-world object detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 6, pp. 11 340–11 354, 2025.

[14] Y. He, W. Chen, S. Wang, T. Liu, and M. Wang, "Recalling unknowns without losing precision: An effective solution to large model-guided open world object detection," *IEEE Transactions on Image Processing*, vol. 34, pp. 729–742, 2025.

[15] X. Xi, Y. Huang, J. Lin, and R. Luo, "Ktcn: Enhancing open-world object detection with knowledge transfer and class-awareness neutralization," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 2024, pp. 1462–1470.

[16] S. Ma, Y. Wang, Y. Wei, E. Zhang, J. Fan, X. Sun, and P. Chen, "Skdf: A simple knowledge distillation framework for distilling open-vocabulary knowledge to open-world object detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 12, pp. 11 738–11 752, 2025.

[17] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 955–10 965.

[18] A. Majee, A. Gangrade, and R. K. Iyer, "Looking beyond the known: Towards a data discovery guided open-world object detection," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[19] S. Fujishige, *Submodular functions and optimization*. Elsevier, 2005, vol. 58.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 8748–8763.

[21] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," in *Forty-first International Conference on Machine Learning*, 2024.

[22] Z. Wang, C. Li, H. Xu, X. Zhu, and H. Li, "Mamba yolo: A simple baseline for object detection with state space model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8205–8213.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[24] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[25] Y. Wang, Z. Yue, X.-S. Hua, and H. Zhang, "Random boxes are open-world object detectors," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6210–6220.