Semantic-Guided Weakly Supervised Open-Vocabulary Object Detection

1st Chao Ma
School of Computer Engineering and Science
Shanghai University
Shanghai, China
machao23@shu.edu.cn

2nd Liyan Ma
School of Computer Engineering and Science
Shanghai University
Shanghai, China
liyanma@shu.edu.cn

3rd Xiangfeng Luo
School of Computer Engineering and Science
Shanghai University
Shanghai, China
luoxf@shu.edu.cn

4th Shaorong Xie*
School of Computer Engineering and Science
Shanghai University
Shanghai, China
srxie@shu.edu.cn

5th Yantao Shang
School of Computer Engineering and Science
Shanghai University
Shanghai, China
2118537431@shu.edu.cn

6th Wendi Rao
School of Computer Engineering and Science
Shanghai University
Shanghai, China
raowendi@shu.edu.cn

***Abstract*—Weakly Supervised Open-Vocabulary Object Detection (WOVOD) focuses on detecting novel classes with only image-level labels of base classes. However, existing methods suffer from three main limitations: insufficient semantic guidance for localization tasks, underutilized features, and noise in pseudo-labels. To address these issues, we propose a Semantic-Guided Weakly Supervised Open-Vocabulary Object Detection (SGWOVD) framework. Our method includes three key innovations: (1) a Semantic-Guided SAM Proposal Generator uses CLIP's semantic understanding to guide SAM in generating semantic proposals; (2) a Semantic Gating Proposal Feature Enhancement module combines image, text, proposal and dataset-aware features to enhance the semantic representation of targets; and (3) a Proposal Noise-Aware Multi-Instance Learning module refines pseudo-label quality by assessing noise across multiple detection heads, alongside spatial and semantic dimensions. Extensive evaluations on OV-COCO, OV-LVIS, and Pascal VOC 2007 demonstrate that SGWOVD consistently outperforms existing state-of-the-art methods. These results validate SGWOVD's strong semantic understanding capability and robustness to noise.**

***Index Terms*—weakly supervised open-vocabulary object detection, semantic guidance, CLIP, SAM**

I. Introduction

Object detection, which has been widely applied in autonomous driving and robotics in recent years, often relies on object-level annotations, which limits its deployment in real-world scenarios. To address this, Weakly Supervised Object Detection (WSOD) [1]–[6] and Open-Vocabulary Object Detection (OVOD) [7]–[14] emerged. The former achieves detection with image-level supervision, but lacks the expansion for open-set objects. The latter has the ability of open-vocabulary detection, but still relies on base object-level supervision.

To combine the advantages of the above two methods, Weakly Supervised Open-Vocabulary Object Detection (WOVOD) has emerged. While WSOVOD [15] has made some progress in this area, there are still three major issues that need to be addressed: 1) insufficient semantic guidance for novel class localization; 2) inappropriate feature utilization in classification; and 3) high background noise in pseudo-labels.

To address these challenges, we propose the Semantic-Guided Weakly Supervised Open-Vocabulary Object Detection (SGWOVD) framework (Fig. 1). SGWOVD is mainly composed of three parts: (1) a Semantic-Guided SAM Proposal Generator uses the CLIP [16] semantic prior as prompts to guide SAM [17] in generating high-value proposals; (2) a Semantic Gating Proposal Feature Enhancement module uses semantic gating to transfer multimodal features to enhance the quality of proposal features for classification; and (3) a Proposal Noise-Aware Multi-Instance Learning module applies MIL [1] and SAM-based pseudo-labels to iteratively clean up noise, improving weakly-supervised detection accuracy.

This study makes several important contributions, including:

- We uniquely introduce visual-text alignment prior into a category-agnostic segmentation model to generate semantic-guided proposals, reducing the semantic gap between base classes and novel classes in the OVOD task.
- We strategically incorporate multimodal and dataset-level features into the proposal features through semantic gating, thereby providing reliable feature support for the open vocabulary classification head.
- We iteratively refine the SAM-based pseudo-labels through a noise-aware instance refinement, effectively filtering out the noise of the pseudo-labels and improving the detection efficiency.

II. Related Work

A. Weakly Supervised Object Detection

WSOD aims to train detectors using only image-level labels, typically formulated as a multiple instance learning (MIL) problem [1]–[6]. WSDDN [1] established an alignment relationship between labels and regional features based on MIL, while D2F2WOD [2] enhanced the localization ability through progressive domain adaptation. Even with these improvements, WSOD still lags behind fully supervised methods due to the sparsity guided by space. In contrast, our method employs category-agnostic segmentation techniques and visual-text large models to enhance the quality of proposals and uses open-vocabulary detection heads to achieve open-set detection.

B. Open-Vocabulary Object Detection

OVOD can detect novel classes by training with labeled data of base classes. Fully supervised OVOD (FSOVOD) primarily focuses on region-text alignment, large-scale pre-training, and knowledge distillation [7]–[14], [18]. ViLD [9] aligns the region-text embeddings, while OADP [13] utilizes a distillation pyramid to transfer knowledge from CLIP. Recently, Visual-RFT [14] introduced reinforcement fine-tuning to reduce data scarcity. However, FSOVOD still requires expensive base classes annotations and struggles with semantically distinct novel classes.

WOVOD combines weak supervision with open-vocabulary learning. WSOVOD [15] generates pseudo ground truth (PGT) using feature adaptation with SAM; however, it only utilizes the spatial layout of SAM and lacks semantic understanding of novel classes. Our method improves WOVOD by leveraging the semantic capabilities of CLIP to SAM, combining the proposed score-weighting mechanism, and filtering out the false label noise in the image-level annotations. This framework narrows the gap between base classes and novel classes, thereby generating more accurate detection results.

III. Method

As shown in Fig. 1, given an input image I, we extract global visual features X_{img} through the backbone network. Based on X_{img}, the Date-aware Feature Extractor (DAFE) [15] produces X_{daf} by integrating dataset attribute prototypes. Meanwhile, the Semantic-Guided SAM Proposal Generator injects the frozen CLIP semantic features into the SAM to generate proposals with semantic guidance. These proposals are then fused with the local proposals. Based on these proposals, the Semantic Gating Proposal Feature Enhancement adopts a Gated Feature Alignment, using X_{img} and E_t to enhance the proposal feature X_{prop} and generate a fused representation X_{fuse}, which is further augmented with X_{daf} to yield X_{full}. Finally, through the Proposal Noise-Aware Multi-Instance Learning module, X_{full} is used in the multi-branch open vocabulary detection head, and iteratively filters out the noise of PGT, achieving the final detection result. The overall training loss $\mathcal{L}_{SGWOVD}$ is:

$$\mathcal{L}_{SGWOVD} = \mathcal{L}_{PG} + \mathcal{L}_{OM} + \mathcal{L}_{NAIR}, \tag{1}$$

where $\mathcal{L}_{PG}$ is the proposal generation loss, $\mathcal{L}_{OM}$ and $\mathcal{L}_{NAIR}$ are the object mining and noise-aware instance refinement losses in Proposal Noise-Aware Multi-Instance Learning.

A. Semantic-Guided SAM Proposal Generator (SSPG)

Traditional Region Proposal Networks (RPN) [19] heavily rely on object-level annotations, which severely limits their generalization to novel classes. In contrast, while SAM [17] can generate class-agnostic proposals, it is highly sensitive to prompt accuracy and lacks category-level semantic guidance for localization, making it unsuitable for open-vocabulary detection. To address these limitations, we propose the SSPG method, as shown in Fig. 2, which combines these two methods and fills the semantic gaps present in SAM, generating semantic-guided proposals.

Specifically, we employ Weakly Supervised Region Proposal Network (WSRPN), which is based on the Faster R-CNN [19] architecture, to generate local proposals $Prop_{loc}$. Meanwhile, the image I is cropped into K overlapping patches and processed by a frozen CLIP visual encoder to generate visual embedding vectors $E_v \in \mathbb{R}^{B\times K\times D}$, where D is the embedding dimension. For the text part, each target class is transformed into a predefined prompt template and encoded by the frozen CLIP text encoder to generate text embeddings $E_t \in \mathbb{R}^{B\times C\times D}$, where C represents the number of categories.

To refine embeddings for point sampling, we first apply self-attention residual blocks to E_v and E_t to obtain E'_v and E'_t. Subsequently, a cross-attention mechanism recalibrates the modalities by treating E'_v as the query and E'_t as the key and value, followed by a feed-forward layer to generate the vision-aware representation E_{t2v}. The base map $M_{base} = E_{t2v} \cdot (E'_t)^\top$ is derived through matrix multiplication. Next, we employ a self-attention mechanism followed by a softmax on M_{base} to obtain the scores for each patch, and after a dimension reshape, we use bilinear interpolation to expand the sparse distribution to the original image size, obtaining a dense and spatially continuous saliency heatmap $M_s \in \mathbb{R}^{B\times H\times W}$.

In Grid-based Spatial Voting (GSV), M_s is divided into $n \times n$ grid cells $\mathcal{G} = \{g_{1,1}, g_{1,2}, \ldots, g_{n,n}\}$. For each cell $g_{i,j}$, if the regional average confidence $S_{i,j}$ exceeds the predefined threshold τ_v, we select the pixel with the highest semantic response as the candidate anchor point $P_{i,j}$:

$$P_{i,j} = \arg \max_{(x,y)\in g_{i,j}} (M_s(x, y)), \tag{2}$$

these points serve as point prompts P, enabling SAM to generate semantic-guided proposals $Prop_{sam}$. The hyperparameters n and τ_v will be verified in the ablation experiments.

To fix the RPN underfitting in early stages and SAM's semantic drift in later stages, we add a Fusion-Weight to mix $Prop_{loc}$ and $Prop_{sam}$. Specifically, we use a progress-aware weight $W_{prog} = W_t - (W_t - W_b) \cdot t/T$ to indicate the training progress. where t is the current iteration and $T = 40$k is the total iteration. Furthermore, we also use a quality-aware weight $W_{qa} = \mathbb{E}[Prop_{sam}]/(\mathbb{E}[Prop_{loc}] + \mathbb{E}[Prop_{sam}])$ to measure how reliable each proposal source is, where $\mathbb{E}(\cdot)$

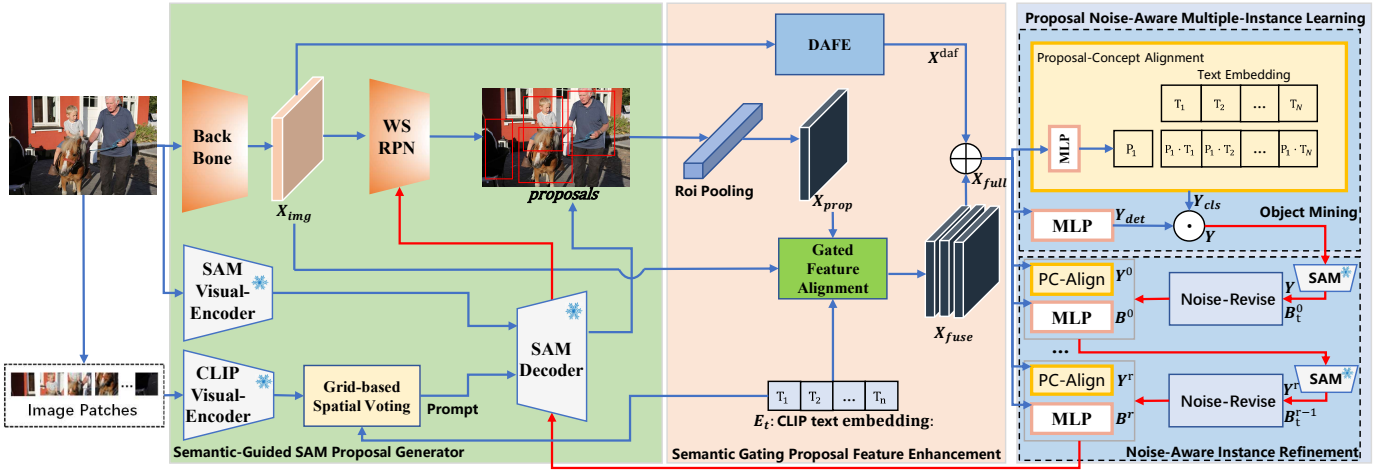

Fig. 1. Framework of the proposed SGWOVD method. The Semantic-Guided SAM Proposal Generator combines WSRPN and semantic-guided SAM-generated candidate regions through proposal score-weighted fusion, which includes subsequent objects for object mining. The Semantic Gating Proposal Feature Enhancement module integrates X_{img}, X_{prop}, E_t, and X_{daf} to obtain full-scale target features. The Proposal Noise-Aware Multi-Instance Learning module filters the noise of PGT, and discovers image-level label vocabulary that matches the underlying target.

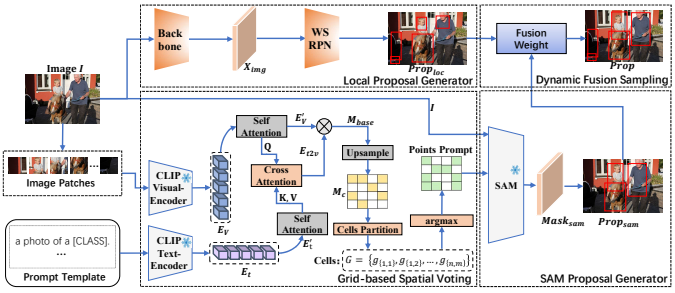

Fig. 2. SSPG dynamically fuses local proposals and CLIP-guided SAM proposals via grid-based spatial voting to generate semantically guided proposals.

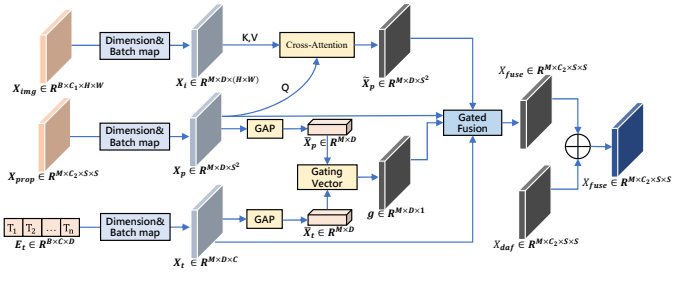

Fig. 3. SPFE utilizes the gating mechanism to integrate multimodal features for enhancing proposal features.

gives the average proposal score. W_{qa} adaptively weights the more accurate source to ensure robust supervision. The final dynamic fusion weight $W_d = W_{qa} \cdot (1 - W_{prog})/(1 + W_{prog})$ controls how we sample from both sources to build the final set $Prop$.

Following WSOVOD, we adopt an anchor-free paradigm: each location predicts a center c, a quadruple distance vector $d(l, r, t, b)$, and a foreground probability p, without using anchor offsets. $D_v(d, c)$ converts distance vectors to bounding boxes. Using the framework's predictions as PGT (p^*, c^*, d^*, D_v^*), the module is optimized via:

$$\mathcal{L}_{PG} = L_{cls}(p, p^*) + L_{reg}(d, d^*) + (1 - L_{iou}(D_v, D_v^*)), \quad (3)$$

where L_{cls}, L_{reg}, and L_{iou} represent binary cross-entropy, distance regression, and IoU loss, respectively.

B. Semantic Gating Proposal Feature Enhancement (SPFE)

While SSPG initiates semantic-guided proposals, serialization processing often limits the detection of novel classes. To address potential semantic biases in proposals and to enhance feature guidance, we propose the SPFE module (Fig. 3). By combining the Gated Feature Alignment with DAFE, SPFE constructs a semantic rectification field that enriches target-specific information and refines proposal quality.

In the Gated Feature Alignment, the three primary features (X_{prop}, X_{img}, E_t) are projected into a consistent dimension D and mapped from B batches to M proposals, generating: $X_p \in \mathbb{R}^{M \times D \times S^2}$, $X_i \in \mathbb{R}^{M \times D \times (H \cdot W)}$, and $X_t \in \mathbb{R}^{M \times D \times C}$ respectively, where S is the spatial scale after RoIPooling. To evaluate the correlation between X_p and X_t, global average pooling (GAP) is applied to X_p and X_t respectively to obtain $\bar{X}_p \in \mathbb{R}^{M \times D}$ and $\bar{X}_t \in \mathbb{R}^{M \times D}$; Subsequently, the adaptive gating vector g is introduced to modulate the information flow:

$$g = \sigma(W_g \cdot \text{Concat}(\bar{X}_p, \bar{X}_t) + b_g), \quad (4)$$

where $g \in \mathbb{R}^{M \times D \times 1}$ is the gating weight calculated through the Sigmoid function σ, W_g is a learnable parameter and b_g is the bias. This gating mechanism serves as a filter.

Meanwhile, we employ an attention mechanism where X_p serves as the query to retrieve contextual information from the global image features X_i (key, value), thereby obtaining $\tilde{X}_p \in \mathbb{R}^{M \times D \times S^2}$. Subsequently, these features are integrated through the Gated Fusion module to generate the final fused

representation $X_{fuse} \in \mathbb{R}^{M\times C_2\times S\times S}$:

$$X_{fuse} = Exp((1-g)\odot X_p + g\odot\left(\tilde{X}_p + \text{FFN}(X_t)\right)), \quad (5)$$

where $(\odot)$ is element-wise multiplication, $\text{Exp}(\cdot)$ is dimensional expansion operation. Finally, X_{img} is fed into DAFE to generate the data-aware feature X_{daf}. X_{full} is formed by the aggregation: $X_{full} = X_{fuse} + X_{daf}$, which contains both textual semantic and dataset prototypes.

C. Proposal Noise-Aware Multi-Instance Learning (PNML)

The PNML module aligns visual features with textual semantics using only image-level labels, addressing the core challenge of WOVOD, through a dual-stage approach: Object Mining (OM) and Noise-Aware Instance Refinement (NAIR).

OM. Following WSOVOD, we treat images as bags of candidate regions. X_{full} is mapped into category-agnostic detection scores Y_{det} and CLIP-based classification scores Y_{cls} (both $\in \mathbb{R}^{M\times C}$). We obtain the final score Y for assigning category c to region m through an element-wise product:

$$Y = \text{Softmax}(Y_{cls}) \odot \text{Softmax}(Y_{det}), \quad (6)$$

the class probability $\psi_c = \sum_{m=1}^{M} Y_c^m$ is supervised by the binary cross-entropy loss $\mathcal{L}_{OM}$:

$$\mathcal{L}_{OM} = \sum_{c=1}^{C} t_c \ln(\psi_c) + (1-t_c)\ln(1-\psi_c), \quad (7)$$

where $t \in \{0,1\}^C$ represents the one-hot category label that marks whether category c exists at the image level.

NAIR. To refine localization, following WSOVOD, we utilize multi-branch classification heads where detection boxes from preceding branches serve as PGT and are fed into SAM as box prompts to supervise the subsequent one. To prevent noise amplification, we employ a Momentum Update mechanism to maintain a historical state memory: $B_t = \lambda B_{t-1} + (1-\lambda)b_t$, where B_t and b_t denote memorized and current boxes, with B_1 initialized as b_1. Following common practice in momentum-based updates, λ is empirically set to 0.1 to ensure a stable evolution of the proposal memory.

We align historical and current boxes via truncation-padding and compute their IoU: $IoU_I = \max IoU(b_t, B_{t-1})$, assessing spatial consistency. Combined with X_{full} features, an MLP generates noise-aware weights W_n^t for each branch t. The refinement objective $\mathcal{L}_{NAIR}$ is the weighted sum of classification and regression losses:

$$\mathcal{L}_{NAIR} = \sum_{t=1}^{T} W_n^t \left(L_{cls}^t + L_{reg}^t\right), \quad (8)$$

where L_{cls}^t and L_{reg}^t denote cross-entropy and smooth $L1$ losses, respectively.

IV. Experiments

A. Datasets

We evaluate the performance of our proposed SGWOVD framework on MS-COCO [24] and LVIS [25], which are widely used in open-vocabulary object detection tasks. The MS-COCO dataset is split into 48 base classes and 17 novel classes, denoted as OV-COCO. For LVIS, 337 rare classes are treated as novel classes and the other 866 classes as base classes, denoted as OV-LVIS. Additionally, we validate the effectiveness of our method in the WSOD task on the Pascal VOC 2007 dataset [26].

B. Evaluation Metric

In the OVOD task, for the OV-COCO dataset, we employ mAP_{50}^N, mAP_{50}^B, and mAP_{50} as metrics for detection accuracy. These metrics respectively calculate the mean average precision (mAP) for novel classes, base classes, and all classes when the intersection over union (IoU) threshold is 0.5. For OV-LVIS, the primary evaluation metrics include AP_r, AP_c, AP_f, and AP, which represent the average precision of rare classes, common classes, frequent classes, and all classes, respectively. For the WSOD task on VOC2007, we evaluate our method's experimental performance using two metrics: CorLoc, which quantifies correct localization accuracy as the proportion of images with IoU at least 0.5 and mAP_{50}.

C. Implementation Details

We adopt RN50-WS-MRRP [27] as the visual backbone, which is initialized with weights pre-trained on the ILSVRC dataset. For the multimodal components, we utilize the pre-trained CLIP with a ViT-L/14 backbone [16] and SAM with a ViT-H backbone [17]. The parameters of both CLIP and SAM are **kept frozen** throughout the training process to preserve their extensive zero-shot knowledge.

Model training is performed using Synchronized Stochastic Gradient Descent (Synchronized SGD) on 2 Nvidia 4090 GPUs, with a batch size of 2, 1 image per GPU. During training, we set the learning rate to 5×10^{-3} and adopt unified hyperparameters for all network components, including a momentum of 0.9 to control the direction of gradient updates, a 50% Dropout rate to prevent overfitting, and a learning rate decay with a coefficient of 0.1. Task-specific strategies are as follows: for WSOD, the parameters of the backbone network are kept fixed; for WOVOD, the backbone network is allowed to be fine-tuned with an extremely low learning rate 1×10^{-5}. For the proposal generation and filtering stage, we maintain a consistent set of 2,000 proposals per image across all datasets. To reduce redundancy, a class-agnostic NMS with a threshold of 0.9 is applied to SAM masks before merging with local proposals. During the final inference phase, NMS is applied with a threshold of 0.7 to produce the final detection boxes.

D. OVOD Task

To verify the effectiveness of SGWOVD, we conduct comprehensive evaluations on the OV-COCO and OV-LVIS.

OV-COCO: We compare SGWOVD with state-of-the-art methods in Tab. I. The first part lists FSOVOD methods utilizing both image-level and object-level supervision. While SGDN [22] achieves 37.5% mAP_{50}^N, it relies heavily on costly bounding-box annotations. In the second part, under the strictly weakly supervised setting, SGWOVD achieves 36.1%

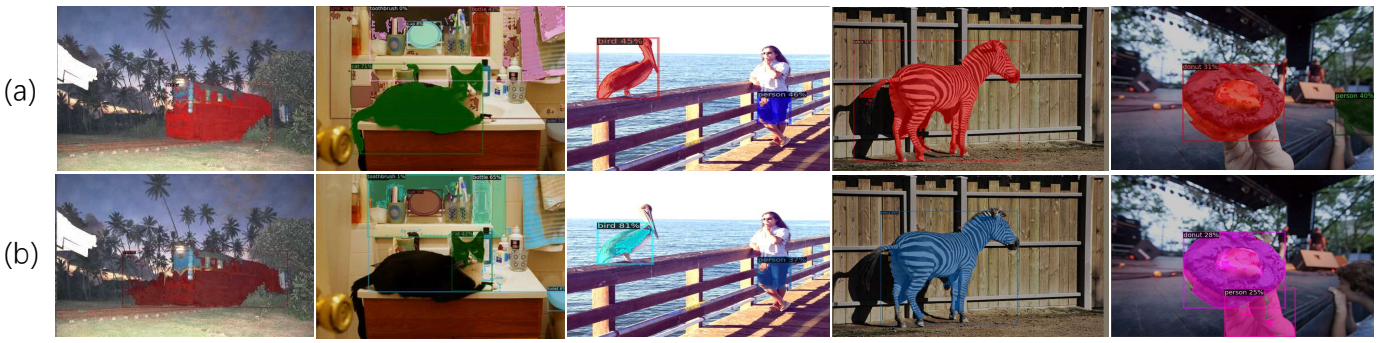

Fig. 4. Visualization of detection results of SGWOVD and WSOVOD methods. (a) Object detection and segmentation results of SGWOVD. (b) Object detection and segmentation results of WSOVOD.

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART OVOD METHODS ON OV-COCO

Method	Backbone	Train Supervision		OV-COCO		
		Image-level	Object-level	mAP_{50}^{N}	mAP_{50}^{B}	mAP_{50}
OVR-CNN [20]	RN50-C4	COCOcaption	COCO48cls.	22.8	46.0	39.9
VILD [9]	RN50-FPN	CLIP400M	COCO48cls.	27.6	59.5	51.3
RKDWTF [21]	RN50-C4	COCOcaption	COCO48cls.	36.6	54.0	49.4
SGDN [22]	RN50	Flickr30K,VG	COCO48cls.	37.5	**61.0**	**54.9**
OADP [13]	RN50	COCOcaption	COCO48cls.	32.7	53.3	47.2
Baron-KD [12]	RN50	COCOcaption	COCO48cls.	34.0	55.9	50.2
LLMDet-KD [23]	Swin-T	COCOcaption	COCO48cls.	35.6	57.3	53.3
CCKT-DET [11]	RN50	CLIP400M	COCO48cls.	36.4	58.4	54.5
WSOVOD [15]	RN50-WS-MRRP	COCO	–	35.0	27.9	29.8
WSOVOD [15]	RN50-WS-MRRP	COCO,LVIS	–	36.7	28.4	30.3
Ours	RN50-WS-MRRP	COCO	–	36.1	30.2	32.3
Ours	RN50-WS-MRRP	COCO,LVIS	–	**37.7**	31.5	33.9

mAP_{50}^{N} using only image-level labels, rivaling several fully supervised counterparts. When trained jointly on OV-COCO and OV-LVIS, SGWOVD reaches 37.7% mAP_{50}^{N}, 1.6% higher than the single-dataset baseline. Our method maintains a steady lead over WSOVOD in both single and cross-dataset evaluations, demonstrating its stronger ability to exploit semantic details and its inherent resilience to noise.

OV-LVIS: We further extended SGWOVD on OV-LVIS, which presents a more challenging long-tailed distribution (Tab. II). Even with only base classes image-level supervision, SGWOVD proves particularly effective on rare categories, reaching 23.8% AP_r. This result represents a 1.6% improvement over WSOVOD and is comparable to the fully-supervised (FSOVOD) method SGDN (23.6%). Similar gains are evident in the common and overall categories, where SGWOVD reaches 28.1% and 30.2% respectively, consistently outpacing WSOVOD. These findings suggest that our approach is both robust and adaptable to complex, large-vocabulary scales.

TABLE II
COMPARISON RESULTS OF SGWOVD AND OTHER OVOD METHODS ON THE OV-LVIS DATASET

Method	AP_r	AP_c	AP_f	AP
VILD [9]	16.7	26.5	32.3	27.8
OADP [13]	21.9	28.4	32.0	28.7
Baron-KD [12]	22.8	27.8	33.4	29.4
SGDN [22]	23.6	**29.0**	**34.3**	**31.1**
WSOVOD [15]	22.2	26.6	31.9	28.5
Ours	**23.8**	28.1	33.1	30.2

TABLE III
COMPARISON RESULTS OF SGWOVD AND OTHER WSOD METHODS ON THE VOC 2007 DATASET

Method	Backbone	VOC 2007	
		mAP_{50}	CorLoc
OICR [4]	VGG16	41.2	60.6
UWSOD [3]	RN18-WS-MRRP	45.0	63.8
SPE [5]	VGG16	51.0	70.4
Seo et al. [6]	VGG16	58.7	69.8
WSOVOD [15]	RN50-WS-MRRP	63.4	80.1
Ours	RN50-WS-MRRP	**65.2**	**83.6**

E. WSOD Task

To evaluate the effectiveness of SGWOVD in WSOD, we conducted experiments on the Pascal VOC 2007 dataset. As described in Table III, our method achieved competitive results, with a mAP_{50} of 65.2% and a CorLoc of 83.6% on the test set. Although both SGWOVD and WSOVOD used the RN50-WS-MRRP backbone network, SGWOVD performed better. Moreover, SGWOVD also had significant performance advantages compared to frameworks based on the classic VGG16 [4]–[6]. These results collectively demonstrate the advantages of our proposed method in WSOD tasks.

TABLE IV
ABLATION STUDY OF EACH COMPONENT

SSPG	SPFE	PNML	mAP^N_{50}	mAP^B_{50}
			32.4	26.9
✓			34.1 (↑ 1.7)	28.7 (↑ 1.8)
	✓		33.5 (↑ 1.1)	27.9 (↑ 1.0)
		✓	33.2 (↑ 0.8)	27.5 (↑ 0.6)
✓		✓	34.9 (↑ 2.5)	29.3 (↑ 2.4)
✓	✓	✓	**36.1** (↑ 3.7)	**30.2** (↑ 3.3)

F. Qualitative Results

As shown in Fig. 4, we compared the visual results of SGWOVD and WSOVOD on the OV-COCO dataset. From the visual perspective, SGWOVD outperforms WSOVOD in both localization and category recognition, as its results are more reliable. This advantage is particularly evident in the bounding box prediction, where the bounding boxes of SGWOVD closely fit the objects, avoiding redundant or missed areas, while the bounding boxes of WSOVOD often have obvious offsets or inaccurate boundaries. This performance gap is mainly attributed by the semantic-guided features in SGWOVD. The obtained segmentation masks provide pixel-level spatial cues, aligning the bounding boxes with the target contours, and enabling more accurate classification.

G. Ablation Study

To evaluate the effectiveness of each module in SGWOVD, we conducted a series of ablation studies. Firstly, to evaluate the individual contribution of each component, we conducted progressive ablation experiments on OV-COCO (Tab. IV). The baseline model consists of a basic OVOD framework, a simple RPN proposal generator, and a MIL classification head based on CLIP text embeddings. We successively integrated PNML, SSPG, and SPFE. The experimental results show that each module achieved performance improvements. Among them, SSPG resulted in significant improvements, increasing mAP^N_{50} by 1.7% and mAP^B_{50} by 1.8%. This is attributed to its ability to adjust the SAM guidance through the CLIP prior, thereby enhancing the semantic space localization capability.

Secondly, to determine the optimal configuration of GSV, we conducted a grid search on the OV-COCO test set to quantitatively analyze the coupling effect between the grid size and the confidence threshold (Fig. 5). The experiment shows that when the grid size is set to 32×32 and the threshold $\tau_v = 0.3$, the model achieves the best localization performance, with CorLoc reaching a peak of 55.3%. Specifically, an excessively low threshold will cause SAM to focus on redundant background noise, while an overly high threshold will significantly reduce the localization accuracy due to excessive filtering. The 32×32 grid setting effectively avoids noise caused by high-density sampling while maintaining localization accuracy.

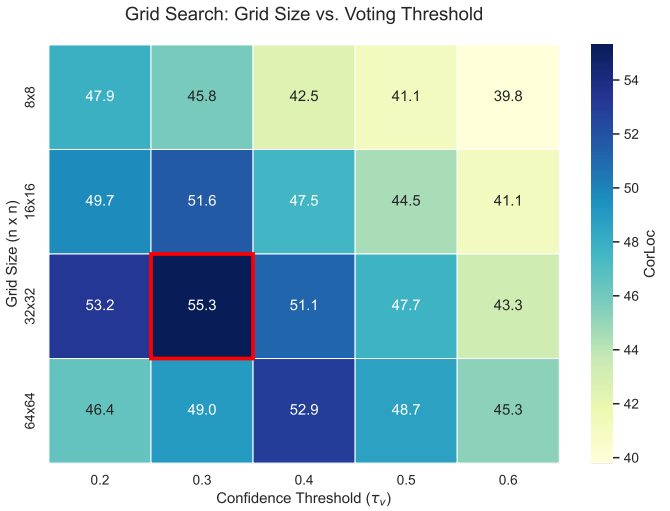

Fig. 5. Hyperparameter Grid Search Ablation for Grid-based Spatial Voting.

V. CONCLUSION

We present a new WOVOD framework named SGWOVD. It aims to leverage the semantic understanding and category-agnostic capabilities of large models to boost the perception of novel classes, and focuses on mitigating noise in pseudo-supervision, effectively addressing the semantic generalization and noise interference issues in the WOVOD task. Experimental results show that SGWOVD performs remarkably well on mainstream datasets: even with only image-level annotations, it rivals certain object-level supervised methods, and even performs better in fine-grained and long-tail category detection. Moreover, the effectiveness of each module has also been verified, providing a new solution for reducing annotation costs and breaking category limitations.

REFERENCES

[1] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2846–2854.

[2] Y. Wang, R. Guerrero, and V. Pavlovic, "D2f2wod: Learning object proposals for weakly-supervised object detection via progressive domain adaptation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 22–31.

[3] Y. Shen, R. Ji, Z. Chen, Y. Wu, and F. Huang, "Uwsod: Toward fully-supervised-level capacity weakly supervised object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7005–7019, 2020.

[4] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2843–2851.

[5] M. Liao, F. Wan, Y. Yao, Z. Han, J. Zou, Y. Wang, B. Feng, P. Yuan, and Q. Ye, "End-to-end weakly supervised object detection with sparse proposal evolution," in *European Conference on Computer Vision*. Springer, 2022, pp. 210–226.

[6] J. Seo, W. Bae, D. J. Sutherland, J. Noh, and D. Kim, "Object discovery via contrastive learning for weakly supervised object detection," in *European conference on computer vision*. Springer, 2022, pp. 312–329.

[7] D. Kim, A. Angelova, and W. Kuo, "Region-centric image-language pretraining for open-vocabulary detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 162–179.

[8] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 965–10 975.

[9] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[10] H. Wang, Q. He, J. Peng, H. Yang, M. Chi, and Y. Wang, "Mamba-yolo-world: marrying yolo-world with mamba for open-vocabulary detection," in *ICASSP 2025-2025 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[11] C. Zhang, C. Zhu, P. Dong, L. Chen, and D. Zhang, "Cyclic contrastive knowledge transfer for open-vocabulary object detection," *arXiv preprint arXiv:2503.11005*, 2025.

[12] S. Wu, W. Zhang, S. Jin, W. Liu, and C. C. Loy, "Aligning bag of regions for open-vocabulary object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 254–15 264.

[13] L. Wang, Y. Liu, P. Du, Z. Ding, Y. Liao, Q. Qi, B. Chen, and S. Liu, "Object-aware distillation pyramid for open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 186–11 196.

[14] W. Zhang, M. Li, H. Wang, J. Liu, and X. Chen, "Visual-rft: Reinforcement fine-tuning for visual-language models in open-vocabulary detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 12 345–12 354.

[15] J. Lin, Y. Shen, B. Wang, S. Lin, K. Li, and L. Cao, "Weakly supervised open-vocabulary object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3404–3412.

[16] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[18] L. Liu, J. Feng, H. Chen, A. Wang, L. Song, J. Han, and G. Ding, "Yolo-uniow: Efficient universal open-world object detection," *arXiv preprint arXiv:2412.20645*, 2024.

[19] 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*, 2015.

[20] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 393–14 402.

[21] H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan, and F. Shahbaz Khan, "Bridging the gap between object and image-level representations for open-vocabulary detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 781–33 794, 2022.

[22] H. Shi, M. Hayat, and J. Cai, "Open-vocabulary object detection via scene graph discovery," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4012–4021.

[23] S. Fu, Q. Yang, Q. Mo, J. Yan, X. Wei, J. Meng, X. Xie, and W.-S. Zheng, "Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 987–14 997.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[25] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5356–5364.

[26] M. Everingham, L. Van Gool, C. K. Williams *et al.*, "The pascal visual object classes challenge 2007 (voc2007) results," in *PASCAL VOC Workshop*, 2007.

[27] Y. Shen, R. Ji, Y. Wang *et al.*, "Enabling deep residual networks for weakly supervised object detection," in *European Conference on Computer Vision*, 2020.