Mamba-YOLO-World: Marrying YOLO-World with Mamba for Open-Vocabulary Detection

Haoxuan Wang[1], Qingdong He[2], Jinlong Peng[2], Hao Yang[3], Mingmin Chi[1,4*], Yabiao Wang[2]

[1]*School of computer science, Shanghai key laboratory of data science, Fudan University, China*
[2]*Tencent Youtu Lab, China* [3]*Shanghai Jiao Tong University, China*
[4]*Zhongshan PoolNet Technology Co., Ltd, Zhongshan Fudan Joint Innovation Center, China*

***Abstract*—Open-vocabulary detection (OVD) aims to detect objects beyond a predefined set of categories. As a pioneering model incorporating the YOLO series into OVD, YOLO-World is well-suited for scenarios prioritizing speed and efficiency. However, its performance is hindered by its neck feature fusion mechanism, which causes the quadratic complexity and the limited guided receptive fields. To address these limitations, we present Mamba-YOLO-World, a novel YOLO-based OVD model employing the proposed MambaFusion Path Aggregation Network (MambaFusion-PAN) as its neck architecture. Specifically, we introduce an innovative State Space Model-based feature fusion mechanism consisting of a Parallel-Guided Selective Scan algorithm and a Serial-Guided Selective Scan algorithm with linear complexity and globally guided receptive fields. It leverages multi-modal input sequences and mamba hidden states to guide the selective scanning process. Experiments demonstrate that our model outperforms the original YOLO-World on the COCO and LVIS benchmarks in both zero-shot and fine-tuning settings while maintaining comparable parameters and FLOPs. Additionally, it surpasses existing state-of-the-art OVD methods with fewer parameters and FLOPs.**

***Index Terms*—object detection, open-vocabulary, Mamba**

I. Introduction

Object detection, as a fundamental task in computer vision, plays a crucial role in various domains such as autonomous vehicles, personal electronic devices, healthcare, and security. The traditional methods [1]–[6] have made great progress in object detection. Nevertheless, these models are trained on closed-set datasets, limiting their capabilities to predefined categories (e.g., 80 categories in the COCO [7] dataset). To overcome such limitations, open-vocabulary detection (OVD) [8] has emerged as a new task that requires the model to detect objects beyond a predefined set of categories.

Some previous OVD works [10]–[14] attempt to leverage the inherent image-text alignment capabilities of pre-trained Vision-Language Models (VLMs). However, these VLMs are trained primarily at the image-text level, resulting in a lack of alignment capabilities at the region-text level. Recent works, such as MDETR [15], GLIP [16], DetClip [17], Grounding DINO [18], mm-Grounding-DINO [19] and YOLO-World [20] redefine OVD as a vision-language pre-training task, employing traditional object detectors to directly learn region-text level open-vocabulary alignment capability on large-scale datasets.

Fig. 1. Visualization Results of Zero-shot Inference on LVIS [9]. Our Mamba-YOLO-World significantly outperforms YOLO-World in terms of accuracy and generalization across small, medium, and large models.

According to the aforementioned related works, the key to converting a traditional object detector into an OVD model lies in implementing a visual-linguistic feature fusion mechanism that is adaptable to the existing neck structure of the model, such as the VL-PAN [20] in YOLO-World and the Feature-Enhancer [18] in Grounding-DINO. As a pioneering model incorporating the YOLO series into OVD, YOLO-World is well-suited for deployment in scenarios prioritizing speed and efficiency. Despite this, its performance is hindered by its VL-PAN feature fusion mechanism.

Specifically, the VL-PAN employs a max-sigmoid visual channel attention mechanism in text-to-image feature fusion flow and a multi-head cross-attention mechanism in image-to-text fusion flow, leading to several limitations. *Firstly*, the complexities of both fusion flows increase quadratically with

*Corresponding author (mmchi@fudan.edu.cn).
Code is available at: https://github.com/Xuan-World/Mamba-YOLO-World.

the product of image size and text length, due to the cross-modal attention mechanism. *Secondly*, the VL-PAN lacks globally guided receptive fields. On the one hand, the text-to-image fusion flow solely generates a visual channel weighting vector that lacks spatial guidance at the pixel level. On the other hand, the image-to-text fusion flow merely allows image information to guide each word individually, failing to leverage the contextual information within text descriptions.

To address the above limitations, we introduce Mamba-YOLO-World, a novel YOLO-based OVD model employing the proposed MambaFusion Path Aggregation Network (MambaFusion-PAN) as its neck architecture. Recently, Mamba [21], as an emerging State Space Model (SSM), has demonstrated its ability to avoid quadratic complexity and capture global receptive fields [22]–[26]. However, simply concatenating the multi-modal features in Mamba [27]–[29] results in a complexity of $O(N + M)$, which increases proportionally with the length of the concatenated sequence. This is particularly problematic for large vocabulary in OVD. Motivated by it, we propose a State Space Model-based feature fusion mechanism in MambaFusion-PAN. We use the mamba hidden state as an intermediary for feature fusion between different modalities, which incurs $O(N + 1)$ complexity and provides globally guided receptive fields. The visualization results shown in Fig. 1 demonstrate that our Mamba-YOLO-World significantly outperforms YOLO-World in terms of accuracy and generalization across all size variants.

Our contributions can be summarized as follows:

- We present Mamba-YOLO-World, a novel YOLO-based OVD model employing the proposed MambaFusion-PAN as its neck architecture.
- We introduce a State Space Model-based feature fusion mechanism consisting of a Parallel-Guided Selective Scan algorithm and a Serial-Guided Selective Scan algorithm, with $O(N + 1)$ complexity and globally guided receptive fields.
- Experiments demonstrate that our model outperforms the original YOLO-World while maintaining comparable parameters and FLOPs. Additionally, it surpasses existing state-of-the-art OVD methods with fewer parameters and FLOPs.

II. Method

Mamba-YOLO-World is mainly developed based on YOLOv8 [30], comprising a Darknet Backbone [3] and a CLIP [31] Text Encoder as model's backbone, our MambaFusion-PAN as model's neck, and a text contrastive classification head along with a bounding box regression head as model's heads, as depicted in Fig. 2.

A. Mamba Preliminaries

For a continuous input signal $u(t) \in \mathbb{R}$, SSM [32] maps it to a continuous output signal $y(t) \in \mathbb{R}$ through a hidden state $h(t) \in \mathbb{R}^E$.

$$h^{'}(t) = Ah(t) + Bu(t) \tag{1}$$

$$y(t) = Ch(t) \tag{2}$$

where E is the SSM state expansion factor, $A \in \mathbb{R}^{E\times E}$ is the state transition matrix, and $B \in \mathbb{R}^{E\times 1}$ and $C \in \mathbb{R}^{1\times E}$ are the input and output mapping matrices, respectively. Building on SSM, Mamba [21] introduces the Selective Scan algorithm, making A, B, and C functions of the input sequence.

B. MambaFusion-PAN

The MambaFusion-PAN is our proposed feature fusion network for replacing the Path Aggregation Feature Pyramid Network in YOLO. As shown in Fig. 2(a), the MambaFusion-PAN utilizes the proposed SSM-based *parallel* and *serial* feature fusion mechanism to aggregate multi-scale image features and enhance text features simultaneously through a three-stage feature fusion flow between visual and linguistic branch: Text-to-Image, Image-to-Text, and finally Text-to-Image. Specific components are detailed in the following parts of this section.

1) Mamba Hidden State: Currently, both Transformer-based and Mamba-based VLMs simply concatenate multi-modal features [18], [19], [27]–[29], [33], [34], leading to an inevitable increase in complexity as the text sequence length and image resolution grow. Although VL-PAN in YOLO-World employs unidirectional fusion without feature concatenation, it still results in $O(N^2)$ complexity. This is due to the visual channel attention mechanism in the text-to-image fusion flow and the multi-head cross-attention mechanism in the image-to-text fusion flow.

To address these issues, we propose extracting the compressed sequence information through the mamba hidden state $h(t) \in \mathbb{R}^{D\times E}$ to serve as an intermediary for feature fusion between different modalities, where D is the dimension of the input sequence and E is the SSM state expansion factor [21], [26]. Since both D and E are constants and not affected by the length of the sequences, our feature fusion mechanism has a complexity of $O(N + 1)$, where N comes from the input sequence of one modality and 1 comes from the mamba hidden state of another modality.

2) TextMambaBlock: The TextMambaBlock is composed of stacked Mamba layers. Given the text embeddings $w_0 \in \mathbb{R}^{L_t\times D_t}$ output from the CLIP Text Encoder, we adopt the TextMambaBlock depicted in Fig. 2(b) to extract not only the output text features $w_1 \in \mathbb{R}^{L_t\times D_t}$ but also the text hidden state $\mathit{THS} \in \mathbb{R}^{D_t\times E_t}$, which will be used for subsequent Text-to-Image feature fusion.

3) MF-CSPLayer: As shown in Fig. 2(c), we integrate *THS* with multi-scale image features through the MambaFusion CSPLayer (MF-CSPLayer). The MF-CSPLayer incorporates the proposed Parallel-Guided Selective Scan algorithm into a YOLO CSPLayer style network. After processing through MF-CSPLayer, we can obtain not only the output image features but also the image hidden state $\mathit{IHS} \in \mathbb{R}^{D_i\times E_i}$, which will be used for subsequent Image-to-Text feature fusion.

4) Parallel-Guided Selective Scan: The Mamba Selective Scan algorithm dynamically adjusts internal parameters based on the input sequence. Motivated by this, we innovatively propose the Parallel-Guided Selective Scan (PGSS) algorithm, which dynamically adjusts the values of Mamba internal

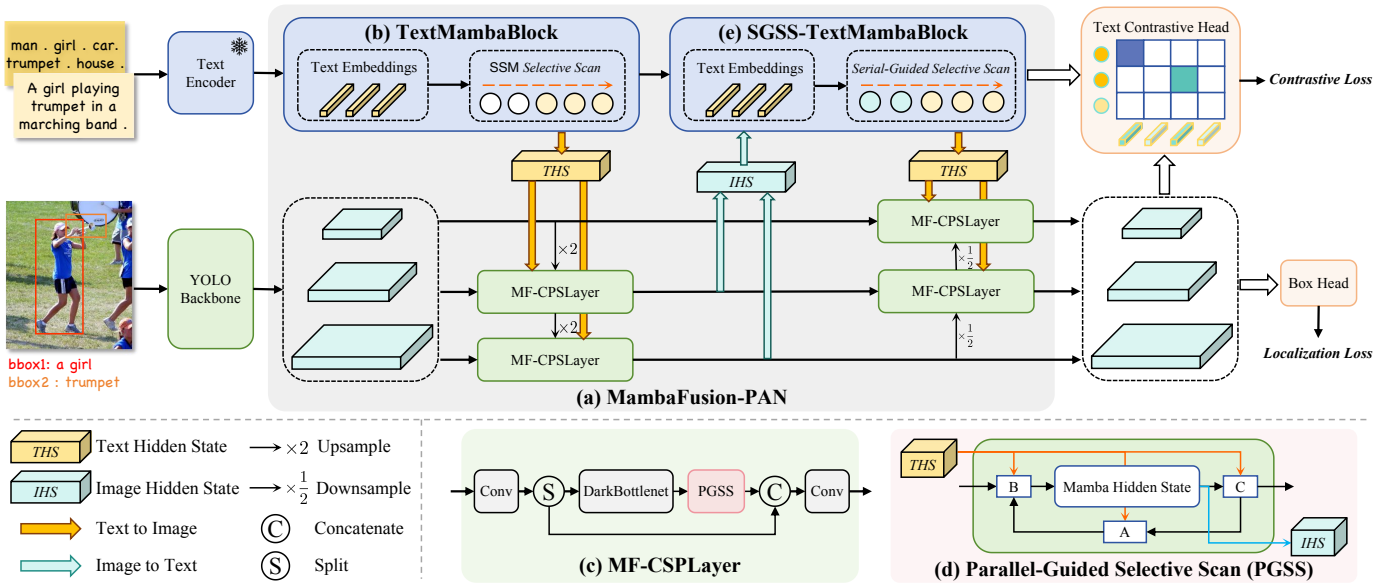

Fig. 2. Overall Architecture of Mamba-YOLO-World. It consists of five key components: **(a)** MambaFusion-PAN is our proposed feature fusion network for replacing the Path Aggregation Feature Pyramid Network in YOLO. **(b)** TextMambaBlock comprises stacked Mamba layers scanning the input text embeddings to extract the output text features and text hidden state (*THS*). **(c)** MF-CSPLayer incorporates the proposed PGSS algorithm into a YOLO CSPLayer style network. **(d)** In the Parallel-Guided Selective Scan (PGSS) algorithm, the compressed textual information *THS* is injected into Mamba parameters in *parallel* with the entire visual selective scanning process to extract the output image features and image hidden state (*IHS*). **(e)** SGSS-TextMambaBlock is a TextMambaBlock with a Serial-Guided Selective Scan algorithm. It adjusts Mamba parameters in *serial* by scanning the compressed visual information *IHS* before extracting the text features.

parameters (A, B, and C) based on both the input image sequence and *THS* during the scanning process, as illustrated in Fig. 2(d) and Algorithm 1. Therefore, the compressed textual information is injected into Mamba in *parallel* with the entire visual selective scanning process, enabling the multi-scale image features to be guided at the pixel level rather than the channel level. The outputs generated from it are passed to the subsequent layers of MF-CSPLayer. In the following, we refer to this part as the Text-to-Image feature fusion flow.

Algorithm 1: Parallel-Guided Selective Scan

Input: $\mathbf{X} \in \mathbb{R}^{L_i \times D_i}$, $\boldsymbol{THS} \in \mathbb{R}^{D_t \times E_t}$
Output: $\mathbf{Y} \in \mathbb{R}^{L_i \times D_i}$, $\boldsymbol{IHS} \in \mathbb{R}^{D_i \times E_i}$
$\mathbf{A} : size(D, E) \leftarrow \text{Parameter}$
$\mathbf{B} : size(L, E) \leftarrow \text{Linear}_{\text{B}}(\mathbf{X}, \boldsymbol{THS})$
$\mathbf{C} : size(L, E) \leftarrow \text{Linear}_{\text{C}}(\mathbf{X}, \boldsymbol{THS})$
$\boldsymbol{\Delta} : size(L, D) \leftarrow \text{softplus}(\text{Linear}_{\Delta}(\mathbf{X}, \boldsymbol{THS}) + \text{Param}_{\Delta})$
$\overline{\mathbf{A}}, \overline{\mathbf{B}} : size(L, D, E) \leftarrow \text{discretize}(\boldsymbol{\Delta}, \mathbf{A}, \mathbf{B})$
$\mathbf{Y}, \boldsymbol{IHS} \leftarrow \text{SSM}(\overline{\mathbf{A}}, \overline{\mathbf{B}}, \mathbf{C})(\mathbf{X})$
return $\mathbf{Y}, \boldsymbol{IHS}$

5) Serial-Guided Selective Scan: The Mamba Selective Scan algorithm continuously compresses information into $h(t)$ based on the input sequence. Motivated by this, we propose the Serial-Guided Selective Scan (SGSS) algorithm and combine it into the TextMambaBlock, as represented in Fig. 2(e). The SGSS aims to compress the prior knowledge from preceding sequences into $h(t)$ and use it as a guidance for the following sequences. Specifically, the SGSS-TextMambaBlock adjusts the values of Mamba internal parameters (A, B, and C) in *serial* by scanning the compressed visual information *IHS* before extracting the text features. In the following, we refer to this part as the Image-to-Text feature fusion flow.

III. Experiment

A. Implementation Details

Mamba-YOLO-World is developed based on the MMYOLO [35] toolbox and the MMDetection [36] toolbox. We provide three size variants, i.e., small (S), medium (M), and large (L). The experiments involve a pre-training stage followed by a fine-tuning stage. During the pre-training stage, we adopt the detection and grounding datasets including Objects365 (V1) [37], GQA [38], and Flickr30k [39]. In line with other OVD methods [15]–[20], the GQA and Flickr30k datasets are collectively designated as the GoldG [15] dataset after excluding images from COCO [7]. During the fine-tuning stage, we use the pre-trained Mamba-YOLO-World and fine-tune it on the downstream task datasets. Unless specified, we conduct the experiments following the settings of YOLO-World [20].

B. Zero-shot Results

After pre-training, we directly evaluate the proposed Mamba-YOLO-World on both LVIS [9] and COCO [7] benchmarks in a zero-shot manner and provide a comprehensive comparison with YOLO-World and other existing state-of-the-art methods.

1) Zero-shot Evaluation on LVIS: The LVIS dataset encompasses 1203 long-tail object categories. Following previous works [15]–[20], we use the *Fixed AP* [42] metric and report 1000 predictions per image on LVIS `minival` for a fair comparison. According to Table I, Mamba-YOLO-World achieves a $+1.5\%$ AP improvement for small variant and a $+1.8\%$ AP improvement for medium variant compared to YOLO-World

TABLE I
ZERO-SHOT EVALUATION ON LVIS MINIVAL (%)

Method	Backbone	Params	FLOPs	Pre-trained Data	AP	AP_r	AP_c	AP_f
MDETR [15]	R-101 [40]	169M	-	GoldG	16.7	11.2	14.6	19.5
GLIP-T [16]	Swin-T [41]	232M	-	O365,GoldG	24.9	17.7	19.5	31.0
Grounding-DINO-T [18]	Swin-T [41]	172M	-	O365,GoldG	25.6	14.4	19.6	32.2
DetCLIP-T [17]	Swin-T [41]	155M	-	O365,GoldG	34.4	26.9	33.9	36.3
mm-Grounding-DINO-T [19]	Swin-T [41]	173M	-	O365,GoldG	35.7	28.1	30.2	42.0
YOLO-World-S [20]	YOLOv8-S [30]	77M	297G	O365,GoldG	26.2	19.1	23.6	29.8
Mamba-YOLO-World-S (ours)	YOLOv8-S [30]	78M	297G	O365,GoldG	**27.7**	**19.5**	**27.0**	**29.9**
YOLO-World-M [20]	YOLOv8-M [30]	92M	324G	O365,GoldG	31.0	23.8	29.2	33.9
Mamba-YOLO-World-M (ours)	YOLOv8-M [30]	94M	324G	O365,GoldG	**32.8**	**27.0**	**31.9**	**34.8**
YOLO-World-L [20]	YOLOv8-L [30]	111M	370G	O365,GoldG	35.0	27.1	32.8	**38.3**
Mamba-YOLO-World-L (ours)	YOLOv8-L [30]	113M	369G	O365,GoldG	**35.0**	**29.3**	**34.2**	36.8

TABLE II
ZERO-SHOT EVALUATION ON COCO (%)

Method	Pre-train	AP	AP_{50}	AP_{75}
YOLO-World-S [20]	O365,GoldG	37.6	52.3	40.7
Mamba-YOLO-World-S (ours)	O365,GoldG	**38.0**	**52.9**	**41.0**
YOLO-World-M [20]	O365,GoldG	42.8	58.3	46.4
Mamba-YOLO-World-M (ours)	O365,GoldG	**43.2**	**58.8**	**46.6**
YOLO-World-L [20]	O365,GoldG	44.4	59.8	48.3
Mamba-YOLO-World-L (ours)	O365,GoldG	**45.4**	**61.3**	**49.4**

TABLE III
FINE-TUNING EVALUATION ON COCO (%)

Method	Pre-train	AP	AP_{50}	AP_{75}
YOLO-World-S [20]	O365,GoldG	45.9	62.3	50.1
Mamba-YOLO-World-S (ours)	O365,GoldG	**46.4**	**62.5**	**50.5**
YOLO-World-M [20]	O365,GoldG	51.2	68.1	55.9
Mamba-YOLO-World-M (ours)	O365,GoldG	**51.4**	**68.2**	**56.1**
YOLO-World-L [20]	O365,GoldG	53.3	70.1	58.2
Mamba-YOLO-World-L (ours)	O365,GoldG	**54.1**	**71.1**	**59.0**

while keeping comparable parameters and FLOPs. Moreover, it outperforms YOLO-World by $+0.4\% \sim +3.2\%$ AP_r and $+1.4\% \sim +3.4\%$ AP_c across all size variants. The Mamba-YOLO-World-L obtains superior results compared with previous state-of-the-art methods such as [15]–[18] with fewer parameters and FLOPs.

2) Zero-shot Evaluation on COCO: The COCO dataset contains 80 categories and is the most commonly used dataset for object detection. As illustrated in Table II, our Mamba-YOLO-World shows overall advantages, outperforming YOLO-World by $+0.4\% \sim +1\%$ AP across all size variants.

C. Fine-tuning Results

In Table III, we further evaluate the fine-tuning results on the COCO benchmark. After fine-tuning on COCO `train2017`, Mamba-YOLO-World achieves higher accuracy and consistently outperforms the fine-tuned YOLO-World by $+0.2\% \sim +0.8\%$ AP across all size variants.

D. Ablation Studies

In Table IV, we conduct ablation experiments to analyze the impact of the MambaFusion-PAN Text-to-Image and Image-to-Text feature fusion flow based on Mamba-YOLO-World-S. The zero-shot evaluation results on the COCO benchmark indicate that both our parallel (Text→Image) and serial (Image→Text) feature fusion methods effectively boost the performance without increasing the parameters or FLOPs.

TABLE IV
ABLATION ON MAMBAFUSION-PAN

Text→Image	Image→Text	Params	FLOPs	AP	AP_{50}	AP_{75}
✗	✗	78M	297G	37.1	51.7	39.9
✗	✓	78M	297G	37.3	52.0	40.2
✓	✗	78M	297G	37.8	52.5	40.7
✓	✓	78M	297G	**38.0**	**52.9**	**41.0**

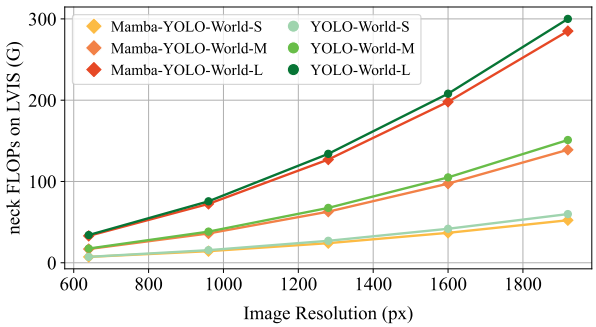

Fig. 3. Comparison of Neck FLOPs Across Different Image Resolutions

Additionally, we analyze the changes in computational cost as the input image resolution increases. As illustrated in Fig. 3, the MambaFusion-PAN (neck of Mamba-YOLO-World) consumes up to 15% fewer FLOPs than the VL-PAN (neck of YOLO-World) across all size variants, indicating a lower model complexity of our MambaFusion-PAN.

IV. CONCLUSION

In this paper, we present Mamba-YOLO-World for open-vocabulary object detection. We introduce an innovative State Space Model-based feature fusion mechanism and integrate it into MambaFusion-PAN. Experimental results demonstrate that Mamba-YOLO-World outperforms the original YOLO-World with comparable parameters and FLOPs. We hope this work will bring new insights into the multi-modal Mamba architecture and encourage further exploration for open-vocabulary vision tasks.

REFERENCES

[1] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[5] 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.

[6] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *The Eleventh International Conference on Learning Representations*, 2023.

[7] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[8] 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.

[9] 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.

[10] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *International Conference on Learning Representations*, 2022.

[11] 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.

[12] W. Kuo, Y. Cui, X. Gu, A. Piergiovanni, and A. Angelova, "Open-vocabulary object detection upon frozen vision and language models," in *The Eleventh International Conference on Learning Representations*, 2023.

[13] S. Xu, X. Li, S. Wu, W. Zhang, Y. Li, G. Cheng, Y. Tong, K. Chen, and C. C. Loy, "Dst-det: Simple dynamic self-training for open-vocabulary object detection," *arXiv preprint arXiv:2310.01393*, 2023.

[14] X. Wu, F. Zhu, R. Zhao, and H. Li, "Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7031–7040.

[15] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1780–1790.

[16] 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.

[17] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. Xu, and H. Xu, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9125–9138, 2022.

[18] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[19] X. Zhao, Y. Chen, S. Xu, X. Li, X. Wang, Y. Li, and H. Huang, "An open and comprehensive pipeline for unified object grounding and detection," *arXiv preprint arXiv:2401.02361*, 2024.

[20] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 901–16 911.

[21] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[22] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[23] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[24] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," *arXiv preprint arXiv:2407.08083*, 2024.

[25] L. Ren, Y. Liu, Y. Lu, Y. Shen, C. Liang, and W. Chen, "Samba: Simple hybrid state space models for efficient unlimited context language modeling," *arXiv preprint arXiv:2406.07522*, 2024.

[26] T. Dao and A. Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," *arXiv preprint arXiv:2405.21060*, 2024.

[27] Y. Qiao, Z. Yu, L. Guo, S. Chen, Z. Zhao, M. Sun, Q. Wu, and J. Liu, "Vl-mamba: Exploring state space models for multimodal learning," *arXiv preprint arXiv:2403.13600*, 2024.

[28] B.-K. Lee, C. W. Kim, B. Park, and Y. M. Ro, "Meteor: Mamba-based traversal of rationale for large language and vision models," *arXiv preprint arXiv:2405.15574*, 2024.

[29] H. Zhao, M. Zhang, W. Zhao, P. Ding, S. Huang, and D. Wang, "Cobra: Extending mamba to multi-modal large language model for efficient inference," *arXiv preprint arXiv:2403.14520*, 2024.

[30] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[32] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *The International Conference on Learning Representations (ICLR)*, 2022.

[33] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[34] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[35] M. Contributors, "MMYOLO: OpenMMLab YOLO series toolbox and benchmark," https://github.com/open-mmlab/mmyolo, 2022.

[36] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[37] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8430–8439.

[38] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[39] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[42] A. Dave, P. Dollár, D. Ramanan, A. Kirillov, and R. Girshick, "Evaluating large-vocabulary object detectors: The devil is in the details," *arXiv preprint arXiv:2102.01066*, 2021.