Zone-YOLO: Vision-Language Object Detection Using Zone Prompt

Jiaxiong Yang, Ning Jia, Xianhui Liu, Rui Fan, *Senior Member, IEEE*, Yougang Sun, *Senior Member, IEEE*, and Weidong Zhao

***Abstract*— Object detection in complex traffic scenarios is crucial for Intelligent Transportation Systems (ITS). At present, most real-time traffic object detection methods primarily rely on YOLO-style vision-only detectors, limiting their potential for further improvement. Vision-Language Object Detection (VLOD) has made promising progress currently, yet its adoption in the realm of ITS remains limited. Previous VLOD methods utilize text features in the classification task, without fully exploring their impact on the regression process for object localization. Besides, existing multi-modal fusion approaches fail to fuse text features with multi-scale image features at corresponding scales, which is detrimental to the representation capability of the model. In this work, we dive into the limitations above and introduce Zone-YOLO to improve the VLOD to a new level. Specifically, we propose Scale-Aware Modal Fusion (SAMF) to fully exploit the text and image features and learn to fuse the multi-modal representations seamlessly at different scales with channel- and modal-wise enhancement. Moreover, we present a novel Zone Prompt learning method to introduce text features into regression process and capture the zone-class-entity triple co-occurrence, which significantly improves the localization performance of the model. Extensive experiments show that Zone-YOLO outperforms the comparative methods by a considerable margin, achieving 55.1 AP, 72.1 AP_{50} and 71.2 AP_L on COCO. The competitive results on BDD100K and VisDrone2019 further demonstrate the superiority of Zone-YOLO on efficient traffic object detection.**

***Index Terms*— Traffic object detection, vision-language model, multi-modal feature fusion, prompt learning, YOLO.**

I. Introduction

OBJECT detection is capable of recognizing and locating the image's region of interest, such as cars and pedestrians, and is widely applied in traffic surveillance, Advanced Driving Assistance Systems (ADAS). Over the past decades, research on traffic object detection has achieved significant breakthroughs in terms of model structure [1], [2], [3], [4], data security [5], and interpretability [6], becoming the foundational building blocks for high-level decision-making and path planning capabilities in many ITS tasks. YOLO-style detectors [7], [8], [9], [10], [11] integrate the end-to-end architecture with a lightweight backbone, excelling in real-time traffic object detection. However, the majority of YOLO-style detectors to traffic object detection rely on vision-only features, which inherently struggle with a lack of semantic information, limiting their potential for further scalability and improvement. In traffic scenes, the varying sizes of objects and complicated background result in a large number of missed and false detections. Fig. 1 intuitively illustrates this phenomenon.

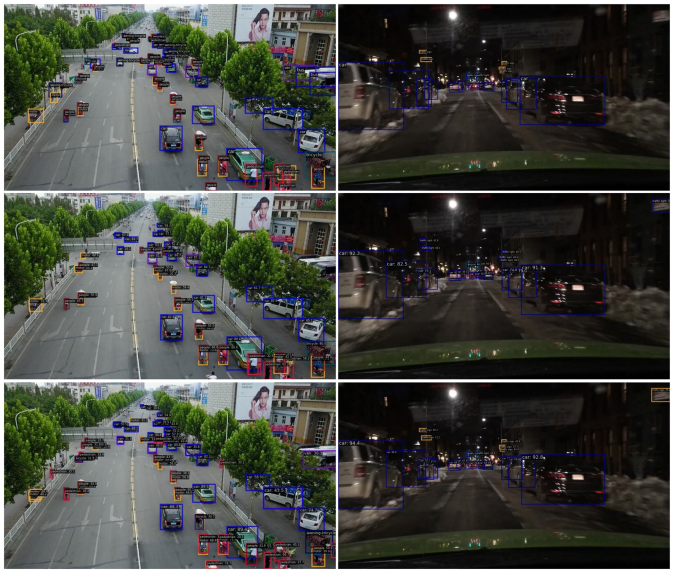


Fig. 1. Traffic object detection in complex scenes, from top to bottom are the results of Ground Truth, YOLOv8, and our Zone-YOLO. It can be observed that Zone-YOLO achieves higher recall and confidence scores, and exhibits better detection performance under challenges such as overlap, small size, and dim lighting condition.

Recently, Vision-Language Models (VLMs) [12], [13], [14], [15] have been extensively researched [16]. By fusing multi-modal information, VLMs are able to derive more general and robust feature representations. Inspired by this, several works [17], [18], [19], [20], [21], [22] integrated text encoders with object detectors, proposing VLM-based Object Detection (VLOD) methods. These methods leverage text encoders such as BERT [23] to extract more semantic information, significantly improving detection performance. There are two issues worth investigating in VLOD. First, the inconsistencies between different modalities inherent in VLMs. Therefore, the efficient fusion of image and text features is crucial to fully harness the feature information and subsequently improve the model's representational capabilities. Second, most VLOD methods primarily focus on utilizing text features

Received 21 August 2024; revised 9 November 2024; accepted 28 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62403361. The Associate Editor for this article was A. Al-Dulaimi. *(Corresponding author: Ning Jia.)*

Jiaxiong Yang, Ning Jia, Xianhui Liu, Rui Fan, and Weidong Zhao are with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: jiaxiongyang@tongji.edu.cn; jianing7072@tongji.edu.cn; xianhuiliu488@163.com; ranger_fan@outlook.com; wdzhaocad@163.com).

Yougang Sun is with the Institute of Rail Transit and the National Maglev Transportation Engineering Research and Development Center, Tongji University, Shanghai 201804, China (e-mail: 1989yoga@tongji.edu.cn).

Digital Object Identifier 10.1109/TITS.2024.3510117

for contrastive learning in the classification task without exploring their impact on the regression process.

Pertaining to the first issue, studies [14], [15] have utilized Multi-Head Self-Attention (MHSA) in Transformer [24] to capture semantic correlations between image and text features for multi-modal feature fusion. Some methods [15], [25] concatenate these two features and feed them into multi-layer Transformers to achieve modal fusion at an early stage, while Co-Attention [14] modifies the MHSA to compute mutual attention between two modalities. YOLO-World [22] distinctively uses Max-Sigmoid Attention to aggregate text features into image features. Although the aforementioned works have achieved significant results, they have failed to distinguish the macro and micro concepts within the text features. Integrating these concepts with different scales of image features indiscriminately may potentially disrupt the fusion features and exacerbate the modality gap [15]. For this reason, we propose the Scale-Aware Modal Fusion (SAMF) method that aligns image and text features at corresponding scales by updating a scale-aware query (SQ), thereby suppressing concept aliasing in feature fusion. The multi-modal attention mechanism utilized in SAMF can enhance the features of different modalities independently.

In response to the second issue, drawing on Prompt learning and the Adapter technique, this study innovatively proposes Zone Prompt, which introduces the regional information of text features into the regression task, aiming to improve the detection performance from another perspective. Some existing studies on regional prompts are available. PTP [26] numbers the region blocks of an image and then predicts the objects based on the given blocks. PEVL [27] reconstructs its objective with explicit object position modeling to generate the bounding box coordinates. However, these methods cannot be directly applied to object detection, and the class-specific prompts they use pose significant difficulties in category-to-object coreference. In this paper, we design class-agnostic zone prompts to avoid referential ambiguity and introduce an adapter to get class-specific zone embeddings that capture the co-occurrence information between categories and regions. Then, a brand-new Zone Head is built to fuse the image features with zone embeddings and achieve the interaction of zone-class-entity co-occurrence features, thus avoiding the direct matching of text and image features. An auxiliary branch is also included to resist potential information loss during prompt learning.

Incorporating the above two aspects, we propose Zone-YOLO, a VLM-based YOLO fashioned detector, and investigate its application in traffic object detection. Our main contributions are summarized into threefold:

1) We pioneered the scale-aware dual-stream multi-modal fusion method to fully exploit the text features and learn to fuse the multi-modal representations seamlessly at different scales with coarse-to-fine feature enhancement.
2) We presented a novel zone prompt learning approach to introduce text features into the regression head and capture the zone-class-entity triple co-occurrence for richer multi-modal information aggregation.
3) The proposed Zone-YOLO fine-tuned on two traffic benchmarks, BDD100K and VisDrone2019, has demonstrated its excellent detection capability in traffic scenarios. Experiments on two universal datasets, COCO and LVIS, showcase our superior performance among YOLO-style detectors. Compared with the baseline, Zone-YOLO achieves 55.1 AP on COCO by a large margin, and no catastrophic decline emerges on LVIS.

The rest of the paper is organized as follows. Section II introduces the works related to VLOD and the prompt learning methods, which offer us inspiration and serve as the theoretical foundation. After that, the proposed modules are presented and elaborated on in Section III. Experimental setup and results are exhibited in Section IV, highlighting the superior performance of Zone-YOLO on traffic object detection and verifying the effectiveness of proposed methods. Finally, the conclusions and prospects are summarized in Section V.

II. Related Works

A. Vision-Language Object Detection

VLOD is a novel trend in modern object detection that improves the models' performance and extends their generality by integrating multi-modal features. OVR-CNN [18] is the first to build VLOD using BERT [23] and Faster R-CNN [28]. It learns a visual-semantic feature space by pre-training on large scale image-caption pairs. However, since VLMs were trained to match whole images to text descriptions, a domain shift arises when attempting direct contrastive learning of region-text features. RegionCLIP [19] leverages the CLIP with template captions, and aligns image regions and textual concepts into the same feature space through knowledge distillation. Moreover, ViLD [20] feeds the category names into the text encoder to obtain the word embeddings, then distills the region embeddings to align the word-region features. One-stage detector DetCLIPv2 [21] employs a similarity optimal matching set between visual regions and word concepts to guide fine-grained contrastive learning. YOLO-World [22] further introduces an effective pre-training strategies to avoid the domain shift. The aforementioned methods address the multi-modal alignment problem by optimizing the training objective, but require considerable computational costs and labeled data.

Meanwhile, in recent years, multi-modal fusion methods have been extensively studied and recognized for their capacity to bridge the modality gap. These methods can be sorted into single-stream [15], [25], [26] and dual-stream [14], [17], [29], [30], [31] fusion according to their implement structure. Single-stream fusion directly concatenates the image and text features and sends them to the subsequent decoder. By contrast, the dual-stream architecture preserves the independence of each modality and interweaves them to achieve cross-modality interaction. Co-Attention [14] calculates the mutual attention of two modalities by exchanging key-value pairs in MHSA [24], while [29] multiplies the two outputs as modal mixed features, which may cause the loss of information. GLIP [17] and PVLR [30] design a deep fusion module that fuses visual and textual information in the last few encoding layers using Co-Attention. Unlike those that incorporate MHSA, YOLO-World [22] employs Max-Sigmoid Attention to aggregate text features into image features, and

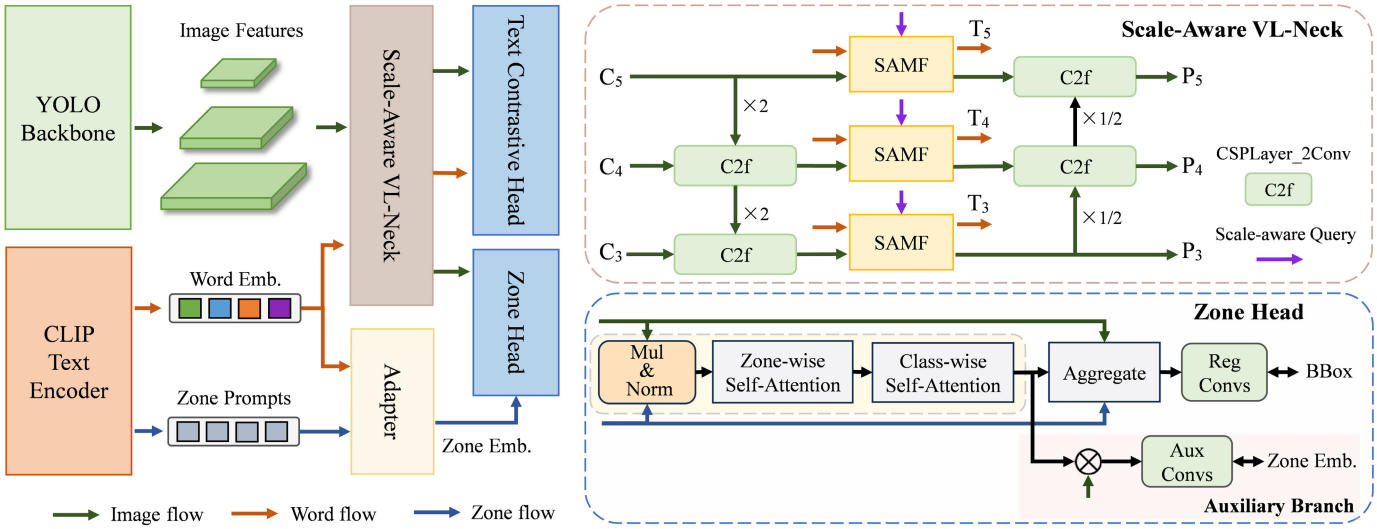


Fig. 2. Overall architecture of Zone-YOLO. Based on [22], Zone-YOLO is a Vision-Language detector. The Text Encoder encodes the class names and locality nouns into word embeddings and zone prompts, respectively. The Image Encoder encodes the input image into multi-scale image features. The proposed Scale-Aware VL-Neck exploits the cross-modality fusion at the corresponding scale. Innovatively, Zone Prompt learning is presented with an Adapter to weave in class-specific information and the Zone Head to capture zone-class-entity triple co-occurrence.

VTP-OVD [32] introduces an adapting stage and renders learnable prompts for fine-grained modal fusion.

However, previous works fused the text features with image features at different scales without discriminating between semantic differences, therefore, they potentially led to conceptual confusion [15]. To avoid this problem, the proposed SAMF constrains information interaction at the corresponding scale and optimizes the fused features from coarse to fine.

B. Prompt Learning

Prompt engineering can unlock additional representation capabilities of the model, facilitating easy transfer to downstream tasks and significantly boosting performance without laborious pre-training. CoOp [33] models the context of prompts as continuous representations and automates prompt engineering end-to-end for few-shot classification. DetPro [34] extends CoOp to VLOD by designing unique strategies to handle foreground and background proposals within images. PromptDet [35] demonstrates that the regional visual features are local and object-centric, proposing the regional prompt learning to steer the textual latent space for better alignment. MaPLe [36] designs prompts for both vision and language branches to model the stage-wise feature relationships between two modalities, while TaI [37] learns prompts with only text as images during training. DQ-DETR [38] and VTP-OVD [32] incorporate visual prompts with the existing text prompts to provide the prior task information for better downstream adaptation.

Most VLOD prompt learning is primarily tailored for contextual information, with little attention paid to positional information. PTP [26] numbers regions of an image and associates them with categories, not conforming to practical linguistic usage. PEVL [27] and BEV-InMLLM [39] reconstructs the training objective to regress the coordinates of bounding boxes. However, existing methods rely on class-specific prompts, which is suitable for visual grounding [17] task where text descriptions are available beforehand. On the other hand, one class often corresponds to many object entities, leading to referential ambiguity. Class-agnostic prompts could mitigate this issue, but they might sacrifice the co-occurrence information between categories and regions (e.g., boats tend to appear in the lower half of images). In this paper, we introduce Zone Prompt, which progressively integrates region and category information into image features, thereby providing the regression process with richer positional information.

III. Methodology

A. Model Architecture

The overall structure of Zone-YOLO is shown in Fig. 2, which consists of an image encoder from YOLOv8 [8] and a text encoder from CLIP [12], for feature extraction. Scale-Aware VL-Neck is designed for better feature fusion across different scales of multi-modal features. Moreover, an adapter is proposed to capture zone-class co-occurrence from word embeddings and zone prompts. Zone Head integrates zone embeddings into image features to capture the zone-class-entity co-occurrence, thereby guiding the bounding box regression process. Text Contrastive Head is consistent with YOLO-World [22] for the classification task.

Given the pre-defined class names, we adopt the pre-trained text encoder to extract the corresponding word embeddings $T \in \Re^{N \times D}$, where N is the number of classes, and D is the embedding dimension. Homologous, $I \in \Re^{C \times H \times W}$ represents the image features extracted by pre-trained image encoder, where C is the number of channels. In Scale-Aware VL-Neck, $\{C_3, C_4, C_5\}$ denotes multi-scale image features, while $\{P_3, P_4, P_5\}$ and $\{T_3, T_4, T_5\}$ represent image feature pyramids and word embeddings after modal fusion, respectively.

In essence, Zone-YOLO mitigates the modality gap by constructing co-occurrence feature spaces during fine-tuning to realize feature enhancement. In section III-B, we devise a scale-aware modal fusion approach to facilitate efficient multi-modal feature fusion. In section III-C, we design a zone

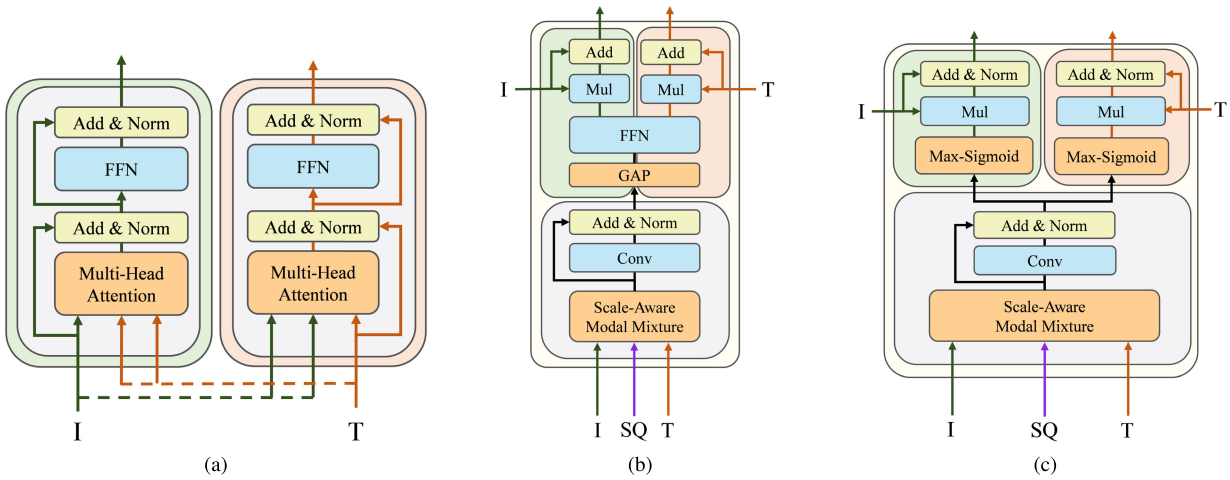


Fig. 3. Multi-modal fusion modules. (a) Co-Attention [14], (b) proposed SAMF with Channel Enhancement, and (c) proposed SAMF with Modal Enhancement, where I stands for image features and T for text features. In Co-Attention, multi-modal interaction is coupled with attention computation, while MI_{SAMF} in SAMF separates the process of information fusion and enhancement.

prompt method that constructs zone-class-entity co-occurrence to guide the learning of the regression process.

B. Scale-Aware Modal Fusion

Fusing multi-scale image features with text features is crucial for VLOD, as it can significantly improve the model's representation ability through multi-modal feature complementation and interaction. Based on this, we propose the Scale-Aware Modal Fusion (SAMF) method that aligns visual and textual semantic information at corresponding scales by updating a Scale-aware Query (SQ), suppressing the concept aliasing in feature fusion. Furthermore, we construct Scale-Aware VL-Neck by SAMF for early fusion, as shown in Fig. 2. In contrast to YOLO-World [22], SAMF is fully plug-and-play, where modal fusion occurs exclusively at lateral connections, without disturbing the image fusion process.

As shown in Fig. 3(b) and (c), SAMF mainly comprises two steps. First, we generate a scale-aware modal mixture matrix MI_{SAMF}. Afterwards, we utilize MI_{SAMF} to enhance the image and text features based on multi-modal attention.

1) Scale-Aware Modal Mixture: The scale-aware modal mixture operation is intuitive. First, the image and text features are projected into the same feature space with the channels aligned, creating a coarse-grained modal mixture matrix $MI_{SAMF} \in \Re^{C\times N\times HW}$. Here HW is the spatial dimension, N is the class dimension, and C is the number of channels. Next, we load the scale-aware query $SQ \in \Re^{1\times H\times W}$ onto MI_{SAMF} to obtain the scale-aware modal mixture matrix. The SQ serves as a mask, activating the information of the corresponding scale while suppressing that of irrelevant scales. The whole process can be described by:

$$MI_{SAMF} = \text{Conv}(\text{reshape}(SQ) \cdot (TW_v^\top \otimes IW_t^\top)) \quad (1)$$

where W_v and W_t are the projection matrices, $\otimes$ means matrix multiplication, and $\cdot$ means element-wise multiplication. Regularization and channel adjustment operations are omitted in the equation for brevity. SQ can be a learnable parameter tensor or the spatial attention [40] map of the image features. Subsequently, two multi-modal attention mechanisms are proposed, utilizing scale-aware MI_{SAMF} for feature enhancement in the channel and modal dimensions.

2) SAMF With Channel Enhancement: Inspired by Co-Attention [14] (Fig. 3(a)), we introduce a dual-stream channel-wise multi-modal attention mechanism that refines and fuses MI_{SAMF} into both text and image flows for adaptive feature enhancement, as shown in Fig. 3(b). Distinct from Cross-Attention, we utilize MI_{SAMF} to generate attention weights and separate the feature fusion and enhancement processes. SAMF initially performs scale-aware aggregation of multi-modal information, followed by refining the features of each modality individually.

Specifically, we first perform the channel attention on MI_{SAMF} similar to [41]. Global average pooling (GAP) compresses the spatial or class dimension. Then, a feed-forward network (FFN) with a sigmoid activation generates the attention vectors, which represents the global distribution of channel-wise responses. Finally, the attention vectors are weighted back to the image and text features by element-wise multiplication. SAMF with channel enhancement can be formulated as:

$$\begin{cases} I_l' = I_l \cdot \text{FFN}\left(\text{GAP}(MI_{SAMF})\right) + \theta_1 \cdot I_l \\ T_l' = T \cdot \text{FFN}\left(\text{GAP}(MI_{SAMF})\right) + \theta_2 \cdot T \end{cases} \quad (2)$$

where, FFN shares its parameters between two modalities, and l represents the level index in the Neck. We additionally introduce two learnable parameters $\theta \in [0, 1]$ to control the residual connections.

3) SAMF With Modal Enhancement: In an analogous manner to the SAMF with channel enhancement, SAMF with modal enhancement (Fig. 3(c)) incorporates the textual information from scale-aware MI_{SAMF} with image features, and integrates visual information identically with text features. In details, we select the maximum value of MI_{SAMF} in the class dimension and apply the sigmoid operation to generate attention maps to aggregate textual sematic information into image features. Symmetric operations are also performed for

text features, which can be expressed by:

$$\begin{cases} I_l' = I_l \cdot \delta\left(\underset{i\in\{1\ldots N\}}{\text{Max}}\left(MI_{SAMF}\right)\right) + \theta_1 \cdot I_l \\ T_l' = T \cdot \delta\left(\underset{j\in\{1\ldots HW\}}{\text{Max}}\left(MI_{SAMF}\right)\right) + \theta_2 \cdot T \end{cases} \tag{3}$$

where δ denotes sigmoid function, and two learnable parameter $\theta \in [0, 1]$ control the residual connections. The max-sigmoid attention is similar to the operations in [22], and the only difference is that we do not compress dimensions, thereby minimizing information loss. Zone-YOLO connects the SAMF modules in Fig. 3(b) and (c), while each module can be used independently in practice.

C. Zone Prompt

Conventional VLOD methods fail to consider the utilization of text features for bounding box regression. Few methods, such as [26] and [27], employ class-specific regional prompts to enhance the localization ability of visual grounding. However, they face two problems when migrating to object detection task: (i) Text descriptions are not provided beforehand in object detection. Therefore, we cannot know which objects to detect in advance. (ii) Random matching is suboptimal for resolving class-to-objects referential ambiguity, restricting the model to seeing partial positive samples at a time.

Herein, we propose Zone Prompt, which introduce the class-agnostic zone prompts to handle with referring difficulty, an Adapter to obtain class-specific zone embeddings to capture zone-class co-occurrence, the Zone Head to achieve the zone-class-entity triple co-occurrence, and finally, a self-supervised auxiliary branch to improve the stability of zone embeddings.

1) Class-Agnostic Zone Prompts: Like the word embeddings in [20], we adopt the fixed zone prompts to avoid the problem of referential ambiguity and meet the requirement of being class-agnostic for object detection. We partition the image into nine regions, describe them with fixed locality nouns, and feed them into the text encoder to generate class-agnostic zone prompts $P \in \Re^{9\times C}$, where C is the embedding dimension same as T. Considering that locality nouns contain more semantic information and are more suitable for language models, we did not assign region numbers to the zone prompts. The effects of various zone prompts and prompt tuning are assessed in the ablation experiment.

Although the aforementioned class-agnostic zone prompts enable concise application in VLOD, they apparently lack spatial information. Location-related low-context information is attached to the word embeddings and image features, so capturing both zone-class and zone-entity co-occurrence information is vital for object detection. We address this problem internally through the following components.

2) Class-Specific Zone Embedding: Exploring the contextual patterns of where categories appear in images facilitates object localization. We design a simple adapter, as shown in Fig 4, to leverage the strengths of the semantic representation ability of the text encoder and capture zone-class co-occurrence. The Adapter comprises a single-layer MHSA [24] and a Language Semantic Attention (LSA) module. Taking word embedding to the value and key matrix and zone prompts to the query matrix of MHAS, we first capture the intricate

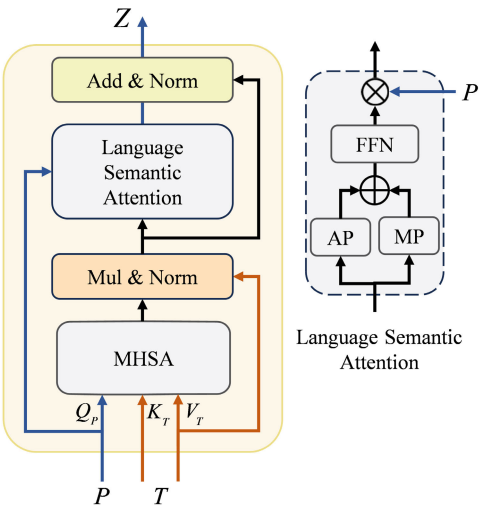


Fig. 4. Adapter for class-specific zone embeddings, where P stands for zone prompts and T for word embeddings. AP and MP in Language Semantic Attention (LSA) represent average pooling and mean pooling in the zone dimension. The adapter incorporates two forms of attentions: MHSA captures the inter-relationships between categories and regions, and LSA refines coarse MI_{Adp} by zone-wise weighting.

relationship between categories and regions in MI_{Adp}. After that, we aggregate zone prompts into MI_{Adp} via LSA to derive class-specific zone embeddings $Z \in \Re^{N\times 9\times C}$:

$$MI_{Adp} = T \otimes \text{MHSA}(P, T, T)^{\top} \tag{4}$$

$$Z = P \cdot \text{FFN}\left(\underset{k\in\{1\ldots 9\}}{\text{Max}}\left(MI_{Adp}\right) + \underset{k\in\{1\ldots 9\}}{\text{Mean}}\left(MI_{Adp}\right)\right) \tag{5}$$

where δ represents the sigmoid function. Element-wise multiplication requires tensor expansion and broadcasting, which inevitably leads to coarse-grained MI_{Adp}. Consequently, zone-wise attention LSA is proposed to refine these coarse-grained features by constraining the distribution of the zone dimension, thereby mitigating the feature aliasing problem.

3) Zone Head: In one-stage anchor-free detector [7], [8], [9], [10], [11], object detection is completely decoupled into classification and regression tasks. Each position of the feature maps represents an entity to be regressed, and the channel (typically, 4) represents the coordinates. We propose the novel Zone Head for the regression process, as illustrated in Fig. 2. By fusing entity information from image feature I with zone embedding Z, Zone Head captures zone-class-entity triple co-occurrence.

Referring to [12], [22], and [29], we eliminate the channel dimension as a trade-off in computational complexity when multiplying I and Z to obtain the triple co-occurrence matrix $MI_{Head} \in \Re^{N\times K\times HW}$:

$$MI_{Head} = ZW_z^{\top} \otimes IW_i^{\top} \tag{6}$$

where W_z and W_i are the projection matrices. Here, the zone dimension corresponds to K (9 in the paper), the class dimension corresponds to N, and the entity dimension corresponds to HW.

Eliminating the channels may aggravate the gap between multi-modal feature. Therefore, two self-attention operations, Zone-wise Self-Attention and Class-wise Self-Attention are proposed to realize the self-awareness informative interaction between entity and class, entity and region, guiding

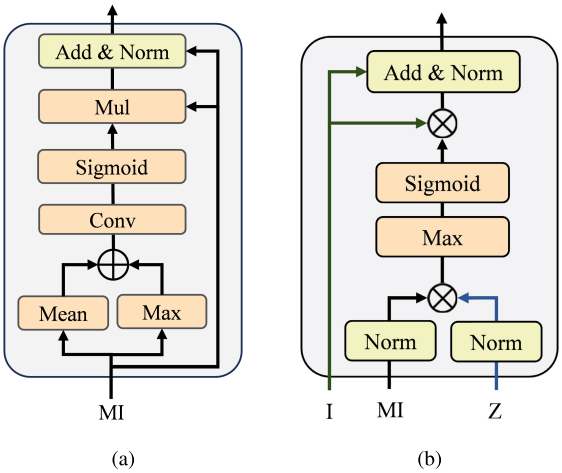


Fig. 5. (a) Self-Attention for MI_{Head}. Class-wise Self-Attention and Zone-wise Self-Attention facilitate informative interactions between entities and classes, entities and regions, thereby guiding the model to align the feature space of MI_{Head} in a self-aware manner. (b) Aggregate MI_{Head} into image features, where I stands for image features and Z for zone embeddings.

the model to align the feature space of MI_{Head}. Fig. 5(a) shows the details of these operations. Specifically, in Zone-wise Self-Attention, the attention weights pertinent to regions and entities are derived by compressing N. In Class-wise Self-Attention, the attention weights associated with categories and entities are attained by compressing K. The above self-attention mechanisms are formalized as follows:

$$MI'_{Head} = MI_{Head} \cdot \delta(\text{Conv}(\max_{i\in\{1\ldots N\}}(MI_{Head}) + \text{Mean}_{i\in\{1\ldots N\}}(MI_{Head}))) \quad (7)$$

$$MI'_{Head} = MI_{Head} \cdot \delta(\text{Conv}(\max_{k\in\{1\ldots K\}}(MI_{Head}) + \text{Mean}_{k\in\{1\ldots K\}}(MI_{Head}))) \quad (8)$$

where δ indicates the sigmoid function and *Conv* represents 1×1 convolution. It is worth noting that the regularization after matrix multiplication is crucial, and our self-attention operation employs layer normalization, which is omitted in the (8) and (7) for brevity. Unlike rendering MI_{SAMF} for modality enhancement in SAMF, here, the attention weights are loaded back to MI_{Head} to augment the co-occurrence information itself.

Ultimately, we aggregate the zone-class-entity triple co-occurrence into image features for final regression convolutions. As depicted in Fig. 5(b), we utilize zone embedding Z to eliminate the class dimension, and then adopt the max-sigmoid attention in (3) to incorporate the region information into the image features.

4) Auxiliary Branch: To encourage the network to learn more generalized co-occurrence information, and to be resilient to the loss of regional information caused by the Zone Head, an auxiliary branch is developed to provide explicit supervision for zone embedding refinement, as illustrated in Fig. 2. The output $Z' \in \Re^{C\times N\times 9}$ is acquired from MI_{Head} by eliminating Entity dimensions using image features. We resort to the mean squared error (MSE) loss to minimize the distance between Z' and original zone embedding Z:

$$MSE = \frac{1}{\prod}\sum_{i=1}^{\prod}(Z' - Z)^2 \quad (9)$$

By aligning the input and output zone embeddings, we can alleviate the sparsity issue in MI_{Head} during feature transformation, improve the stability of features after dimension reduction, and maximize the regularization effect within the Zone Head. We keep the contrastive loss and regression loss untouched, consistent with [22].

IV. Experiments

A. Dataset and Evaluation Metrics

MS-COCO: COCO [42] is a standard general dataset comprising 80 categories of common objects in natural context. It contains about 118k images for training and 5k images for validation, with bounding box and instance segmentation annotations.

BDD100K: BDD100K [43] is tailored for evaluating the robustness and performance of detectors in the context of autonomous driving. It boasts an impressive collection of over 100k diverse video sequences, offering a rich tapestry of real-world driving scenarios.

VisDrone2019: VisDrone2019 [44] is a renowned challenge in the realm of ITS. It encompasses 10 categories and comprises high-resolution images captured by drones, where the objects are relatively small. The training set contains 6471 images, and the validation set contains 548 images.

LVIS Dataset: LVIS [45] is a comprehensive dataset with a long-tail data distribution. It divides the 1203 categories into frequent, common, and rare according to their appearing frequency in the training set. The frequent and common classes compose LVIS-base and rare classes refer to LVIS-novel [20].

Evaluation Metrics: In line with [22], we use COCO average precision metrics [42] to evaluate the detection performance, which contains AP, AP_{50}, and AP_{75} for different IoU thresholds and AP_S, AP_M, and AP_L for different object size. We additionally reported AP_{85} and AP_{95} under larger IoU thresholds to better observe the location ability in ablation studies. For LVIS experiment, APr, APf, APc [45] and fixed AP [46] are reported to measure the model's generalization capability, and the maximum predictions is set to 1K.

B. Implementation Details

The Zone-YOLO is developed based on the MMYOLO toolbox [47]. We remove the VLPAN from YOLO-World [22] to build the baseline model, retaining the efficiency of large-scale pre-training on region-text pairs, and provide three variants of Zone-YOLO for fair comparison, e.g., small (S), medium (M), and large (L). We adopt the frozen CLIP [12] text encoder to encode the class names and locality nouns, and use the pre-trained weights from [22] to initialize Zone-YOLO. All models are fine-tuned for 80 epochs on 2 NVIDIA RTX4090 GPUs using AdamW [48] optimizer with a total batch size of 32. Following previous works [8], [22], the initial learning rate is set to 0.0002 and decays with the linear policy. Common data augmentations are used, and other

TABLE I

COMPARISON WITH DIFFERENT YOLO DETECTORS ON COCO. WE FINE-TUNE ZONE-YOLO ON COCO TRAIN2017 AND EVALUATE IT ON COCO VAL2017. ‡ DENOTES FINE-TUNING ON THE MODEL WEIGHTS THAT PRE-TRAINED WITH MORE DATA, ACCORDING TO [22]. THE BEST RESULTS ARE **BOLDED**. SINCE THE TEXT INPUT IS FIXED, #PARAM AND FLOPS ARE REPORTED AFTER REMOVING TEXT ENCODER

Method	Year	Arch.	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FLOPs	#Param
YOLOv6-S [7]	2022	YOLOv5 w/ RepBlock	43.7	60.9	46.8	23.7	48.9	59.9	44.2G	17.2M
YOLOv8-S [8]	2023	YOLOv5 w/ ELAN	45.1	61.7	48.9	26.0	50.2	61.4	28.6G	11.2M
YOLO-World-S [22]	2024	YOLOv8 w/ VLPAN	46.1	62.5	50.3	**27.3**	50.8	62.2	32.9G	12.8M
YOLOv9-S [9]	2024	YOLOv7 w/ GELAN	**46.8**	63.4	50.7	26.6	**56.0**	**64.5**	26.4G	7.1M
YOLOv10-S [10]	2024	YOLOv8 w/ mNMS	46.3	63.0	50.4	26.8	51.0	63.8	21.6G	7.2M
Zone-YOLO-S	2024	YOLOv8	**46.8**	**63.8**	**51.2**	26.9	52.0	63.2	36.2G	13.5M
YOLOv6-M [7]	2022	YOLOv5 w/ RepBlock	48.0	65.6	52.2	29.8	53.7	64.2	82.2G	34.3M
YOLOv8-M [8]	2023	YOLOv5 w/ ELAN	50.6	67.4	55.2	32.2	55.5	67.6	78.9G	25.9M
YOLO-World-M [22]	2024	YOLOv8 w/ VLPAN	51.0	67.7	55.7	32.9	55.8	66.9	96.2G	28.4M
YOLOv9-M [9]	2024	YOLOv7 w/ GELAN	51.4	68.1	56.1	33.6	**57.0**	**68.0**	76.3G	20.0M
YOLOv10-M [10]	2024	YOLOv8 w/ mNMS	51.1	68.1	55.8	33.8	56.5	67.0	59.1G	15.4M
Zone-YOLO-M	2024	YOLOv8	**51.5**	**68.4**	**56.7**	**34.5**	56.0	67.3	89.3G	29.4M
YOLOv6-L [7]	2022	YOLOv5 w/ RepBlock	50.8	68.3	54.9	32.3	56.1	67.0	144.0G	58.5M
YOLOv8-L [8]	2023	YOLOv5 w/ ELAN	52.9	69.8	57.8	35.3	58.1	69.6	165.2G	43.7M
YOLO-World-L [22]	2024	YOLOv8 w/ VLPAN	53.9	70.9	58.8	36.8	58.7	70.8	175.9G	47.6M
YOLO-World-L [22] (Baseline)	2024	YOLOv8 w/o VLPAN	53.4	70.2	58.3	36.0	58.5	69.4	168.0G	44.0M
YOLOv9-C [9]	2024	YOLOv7 w/ GELAN	53.0	70.2	57.8	36.2	58.5	69.3	102.1G	25.3M
YOLOv10-L [10]	2024	YOLOv8 w/ mNMS	52.5	69.6	57.2	35.1	57.8	68.5	120.3G	24.4M
Zone-YOLO-L	2024	YOLOv8	54.9	71.9	**60.6**	**37.0**	60.4	70.9	178.1G	47.9M
Zone-YOLO-L‡	2024	YOLOv8	**55.1**	**72.1**	60.5	36.9	**61.0**	**71.2**	178.1G	47.9M

configurations are unchanged. The nine locality nouns used in most experiments are: “top left”, “top center”, “top right”, “middle left”, “center”, “middle right”, “bottom left”, “bottom center”, and “bottom right”.

C. Main Result

1) Experiment on COCO Dataset: Table I shows the overall performance of Zone-YOLO and recent YOLO detectors [7], [8], [9], [10], [22] on the COCO benchmark. Zone-YOLO and YOLO-World are fine-tuned for 80 epochs based on pre-trained weights, while the others are trained from scratch for hundreds of epochs. For a fair comparison, all models were trained with mask-refine settings except YOLOv6 and results were obtained from MMYOLO [47] or their published code.

It is evident that Zone-YOLO-L attains a substantial improvement over the baseline, especially in AP_{75} and AP_M, which increased by 2.3 and 1.9, respectively. Compared with YOLO-World [22], which also employs Vision-Language Neck, Zone-YOLO surpasses it incontestably in all metrics. Benefit from proposed SAMF and Zone Prompt, Zone-YOLO-L exceeds YOLOv9 and YOLOv10 significantly in terms of AP by 1.9 and 2.4, AP_{75} by 2.8 and 3.4, and AP_M by 1.9 and 2.6, respectively. As expected, with more generalized pre-trained weights, Zone-YOLO-L‡ achieves the best results in terms of AP, AP_{50}, AP_M, and AP_L. Zone-YOLO in small and medium sizes also perform satisfactory results in AP, AP_{50}, and AP_{75}, although there remains a narrow gap in other metrics. Our Zone-YOLO exhibits slightly higher FLOPs and larger parameters, which is inevitable since we introduce additional components and operations. In summary, Zone-YOLO exhibits competitive performance compared with other YOLO detectors on general dataset and strikes a favorable trade-off between efficiency and performance.

2) Experiment on Traffic Dataset: Table II demonstrates the excellent detection performance of Zone-YOLO on two traffic datasets, BDD100K and VisDrone2019. We choose the latest five YOLO series detectors for comparison, and all of them use small size models to make the results more practical. FPS is obtained on a NVIDIA V100 GPU without employing any optimization strategies.

Zone-YOLO comprehensively outperforms these models and achieves significant improvements across multiple metrics. On the one hand, this is attributed to the broad applicability and vast knowledge reserve of the VLMs. On the other hand, the proposed Zone Prompt and SAMF can effectively assist Zone-YOLO in transferring to traffic scene contexts. Specifically, when facing real-world driving scenarios in BDD100K, Zone-YOLO ranks first in 4 out of 6 metrics, with a particular highlight on AP_L, surpassing the second-place YOLOv9 by 1.5. In VisDrone2019, which contains a large number of overlapping small objects, Zone-YOLO exceeds all counterparts, achieving improvements of 2.0, 3.1, and 2.7 over the second-best model in AP_{50}, AP_M, and AP_L respectively. Although Zone-YOLO falls slightly behind in terms of parameters and computational complexity, its inference speed is comparable to that of recent YOLO detectors. In general, the aforementioned experiments demonstrate that Zone-YOLO is

TABLE II

COMPARISON WITH YOLO DETECTORS ON BDD100K AND VISDRONE2019. WE FINE-TUNE ZONE-YOLO ON BOTH DATASETS FOR 80 EPOCHS IN LINE WITH [22]. THE BEST RESULTS ARE **BOLDED** AND THE SECOND ARE UNDERLINED. SINCE THE TEXT INPUT IS FIXED, #PARAM AND FLOPS ARE REPORTED AFTER REMOVING TEXT ENCODER. THE FPS IS MEASURED ON ONE NVIDIA V100

Method	BDD100K						VisDrone2019						FPS	FLOPs	#Param
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L			
PPYOLOE-S [11]	29.2	50.8	27.7	11.2	33.5	54.2	20.0	34.6	20.0	10.5	31.5	51.0	208.3	17.4G	7.9M
YOLOv6-S [7]	29.7	51.2	28.6	10.5	34.7	56.1	19.5	33.2	19.5	9.7	30.8	47.4	286.9	44.2G	17.2M
YOLOv8-S [8]	31.5	53.9	30.5	**12.5**	37.0	57.6	20.9	36.8	20.7	10.7	33.2	49.7	266.3	28.6G	11.2M
YOLOv9-S [9]	32.0	54.2	30.9	11.7	**37.7**	58.2	20.3	37.1	19.9	10.5	33.1	48.5	247.0	26.4G	7.1M
YOLOv10-S [10]	31.5	54.0	30.6	12.1	36.9	57.1	20.1	36.9	19.8	10.9	32.1	49.7	263.4	21.6G	7.2M
Zone-YOLO-S	**32.3**	**55.4**	**31.0**	12.0	37.2	**59.7**	**22.4**	**39.1**	**22.6**	**11.3**	**36.3**	**53.7**	262.8	36.2G	13.5M

Fig. 6. Visualization results of Zone-YOLO on traffic scenario under different angles, lighting conditions, and object densities. Zone-YOLO produces fitted bounding boxes with higher confidence scores, detects more small objects, and reduces the number of false detections.

well-suited for application in the ITS field, serving as a new foundational model for downstream tasks.

In order to vividly illustrate the superior performance of Zone-YOLO in traffic scenarios, we select some challenging images from BDD100K and VisDrone2019. Fig. 6 showcases the detection results of Zone-YOLO under different angles, lighting conditions, and object densities. Specifically, the first row reflects the ability for detecting small objects. In the left two columns, Zone-YOLO easily captures objects of different scales. Meanwhile, in the right two columns, which depict a dense prediction scenario, Zone-YOLO rarely has false detection. The second row is the results under dim lighting condition. Zone-YOLO consistently displays precise bounding boxes with higher confidence scores, though there is still a negligible number of wrong classifications for incomplete objects. The third row exhibits the results from different perspectives, revealing the problem in distinguishing overlapping small objects. Apart from the insufficient representation capabilities, the MI_{Head} in Zone Prompt assigns each entity to a single category and region, which may not be conducive to handling this situation. Overall, the qualitative analysis confirms the model's ability to accurately detect objects in complex traffic scenes, underscoring its robustness and effectiveness.

TABLE III

COMPARISON WITH DETECTORS ON LVIS. WE FINE-TUNE ZONE-YOLO ON THE LVIS-BASE SET AND EVALUATE IT ON LVIS MINIVAL. THE FINE-TUNING RESULTS OF YOLOV8 AND YOLO-WORLD ARE TAKEN FROM [22], AND OTHER OVOD RESULTS QUOTED FROM THEIR ORIGINAL PUBLICATIONS. MQDET WITH TEXTUAL PROMPTS IS REPORTED

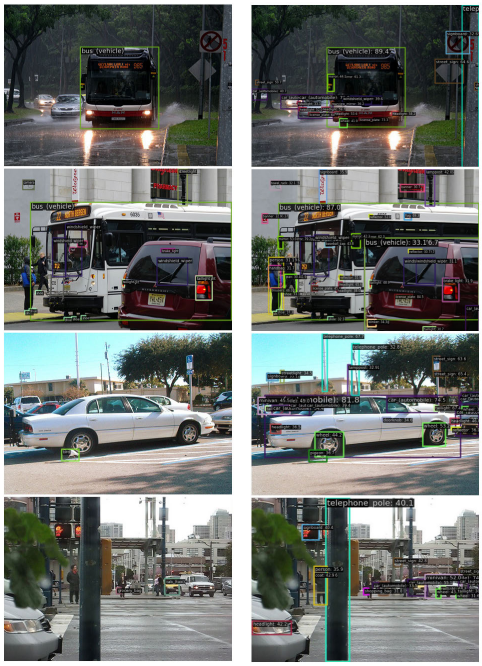
Method	APr	APc	APf	AP
ViLD [20]	16.7	26.5	34.2	27.8
RegionCLIP [19]	17.1	-	-	28.2
Detic [49]	17.8	-	-	26.8
MQdet [50]	20.8	21.4	31.0	26.0
DetPro [34]	20.8	27.8	32.4	28.4
YOLOv8-S [8]	7.4	17.4	27.0	19.4
YOLO-World-S [22]	12.8	20.4	32.7	23.9
Zone-YOLO-S	11.8	17.9	32.1	24.2
YOLOv8-M [8]	8.4	21.3	31.5	23.1
YOLO-World-M [22]	15.9	24.6	39.0	28.8
Zone-YOLO-M	16.6	28.4	41.8	33.9
YOLOv8-L [8]	10.2	25.4	35.8	26.9
YOLO-World-L [22]	20.4	31.1	43.5	34.1
Zone-YOLO-L	16.3	30.5	43.7	35.6

3) Experiment on LVIS Dataset: In Table III, we fine-tune our Zone-YOLO on LVIS-base and report the highest APr result. It was found that, despite the modifications made

TABLE IV

ABLATION FOR ALL PROPOSED COMPONENTS ON COCO VAL2017. EXPERIMENTS ARE CONDUCTED WITH ZONE-YOLO-L‡, WHERE **PROMPT** DENOTES EMPLOYING ZONE PROMPTS AND THE ZONE HEAD, **ADAPTER** REPRESENTS INTEGRATING ADAPTER, AND **AUX** SIGNIFIES UTILIZING THE AUXILIARY BRANCH. CORRESPONDING STRUCTURAL ADJUSTMENTS ARE TAKEN FOR DIFFERENT COMPONENTS

	SAMF	Prompt	Adapter	Aux	AP	AP_{50}	AP_{75}	AP_{85}	AP_{95}	AP_S	AP_M	AP_L
0					53.4	70.2	58.3	44.5	12.8	36.0	58.5	69.4
1	✓				54.3	71.3	59.5	45.2	12.9	36.2	59.4	71.4
2		✓			53.9	70.6	59.0	45.0	13.0	36.5	59.3	70.1
3		✓	✓		54.1	71.0	59.1	45.2	12.8	36.2	58.8	70.6
4		✓	✓	✓	54.3	71.0	59.4	45.5	13.0	36.4	59.8	71.0
5	✓	✓	✓		54.9	72.0	60.2	46.0	12.5	36.7	60.1	72.0
6	✓	✓	✓	✓	55.1	72.1	60.5	46.1	12.6	36.9	61.0	71.2



(a) GT (b) Ours

Fig. 7. Visualization results of Zone-YOLO-L‡ on LVIS.(a) Ground Truth. (b) results from Zone-YOLO-L‡. Zone-YOLO detects most of the novel objects, maintaining the generalization capability of the vision-language model.

to the baseline, fine-tuning Zone-YOLO can avoid catastrophic degradation in rare object detection. Compared with the two-stage OVOD works [19], [20], [34], [49], [50], our one-stage Zone-YOLO possesses a notable edge in APc, APf and AP. We must confess that, compared with YOLO-World [22], the Apr of Zone-YOLO-S and Zone-YOLO-L undergo a noticeable decline. The only difference is that we adjusted the model structures and added novel components, while [22] did not. It is believed that the structural inconsistencies during fine-tuning apparently lead to a deterioration in the generalization ability of the pre-trained model. Nonetheless, Zone-YOLO significantly outperforms YOLOv8, indicating that the impact of the pre-trained vision-language model still exists. Moreover, we select several images in traffic scenes for visualization to intuitively demonstrate the satisfactory performance of Zone-YOLO. In Fig. 7, it can be seen that Zone-YOLO correctly detects most of the novel class objects.

TABLE V

RESULTS OF DIFFERENT SQ PATTERNS ON COCO. EXPERIMENTS ARE CONDUCTED UNDER ZONE-YOLO-L‡. **SA** STANDS FOR SPATIAL ATTENTION IN [40], WHERE CROSS-MODAL SCALE-WISE CORRESPONDENCE OF INFORMATION IS IMPLICITLY ESTABLISHED

Scale-Aware Query	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	53.4	70.2	58.3	36.0	58.5	69.4
Learnable Param.	54.3	71.3	59.4	36.2	59.5	70.9
Image Feat. w/ SA	54.3	71.3	59.5	36.2	59.4	71.4
Image Feat. w/o SA	53.9	70.7	58.9	35.9	59.0	70.2

TABLE VI

ABLATION EXPERIMENTS ON SAMF ENHANCEMENT. THE NOTATIONS "CHANNEL" AND "MODAL" REPRESENT CHANNEL-WISE AND MODAL-WISE MULTI-MODAL ATTENTION RESPECTIVELY, WHILE "MODAL+CHANNEL" REFERS TO THE CONNECTION BETWEEN THE TWO MODULES

SAMF	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	53.4	70.2	58.3	36.0	58.5	69.4
w/ channel	53.7	70.6	58.8	35.9	58.9	70.3
w/ modal	54.1	71.0	59.3	35.7	59.2	71.0
w/ modal+channel	54.3	71.3	59.5	36.2	59.4	71.4

D. Ablation Studies

1) Ablations for Proposed Components: To verify the effectiveness of the proposed components, we report the ablation result in Table IV. It is important to noted that Zone Prompt comprises a series of components, with dependencies among them. The **Prompt** operation builds the basic form of the Zone Prompt, and the **Adapter** and **Aux** rely on it.

Comparing rows 0 and 1 in the table, SAMF comprehensively improves the baseline performance by 1.1, 1.2, and 2.0 on AP_{50}, AP_{75}, and AP_L, respectively. It reveals that SAMF indeed ameliorates the utilization of the multi-modal features through modal-mixture matrix MI_{SAMF} and coarse-to-fine enhancement. In addition, the model proves the feasibility of fine-tuning existing VLOD.

In rows 2, 3, and 4, we gradually apply zone prompts with zone head, adapter, and auxiliary learning branch. It can be seen that using solely zone prompts with zone head enhances the AP to 53.9 (+0.5), reaching the same level as YOLO-World. This underscores the efficacy of incorporating zone prompts into the regression process, while also reflecting the scalability of VLOD. In row 3, the Adapter further improves metrics such as AP, AP_{50}, and AP_L, but experiences a slight decrease in AP_{95}, AP_S, and AP_M. This result indicates that

TABLE VII

RESULTS OF DIFFERENT TYPES OF PROMPTING STRATEGIES ON COCO. EXPERIMENTS ARE CONDUCTED UNDER ZONE-YOLO-L‡. **ZP1** AND **ZP2** REPRESENT DIFFERENT ZONE WORDS ACCORDING TO DIFFERENT DISCOURSE CONVENTIONS, WHERE ZP1 USES TOP AND BOTTOM AND ZP2 USES THE COMPARATIVE FORM UPPER AND LOWER. **PROMPT TUNING** REPRESENTS LEARNABLE PROMPTS INITIALIZED BY ZP1

Prompting Strategy	AP	AP_{50}	AP_{75}	AP_{80}	AP_{95}	AP_S	AP_M	AP_L
Prompt tuning	55.1	72.0	60.3	46.7	12.9	36.6	60.7	72.0
ZP1(top & bottom)	55.1	72.1	60.5	46.1	12.6	36.9	61.0	71.2
ZP2(upper & lower)*	55.0	71.9	60.5	46.4	12.4	36.2	61.2	71.4

* replace the word "top" with "upper" and the word "bottom" with "lower" in ZP1.

additional information injection (from zone-entity to zone-class-entity) contributes to the overall improvement, but has a negative effect on the small object detection, as we assign each entity to a single category and region.

Auxiliary branch restored the AP_{95} and AP_S performance, as shown in row 5, demonstrating the efficacy of this concise self-supervision learning scheme. Rows 3, 4, and 6 repeat this phenomenon, as textual features could potentially hinder the AP_{95}, but auxiliary branch can alleviate this problem. As anticipated, SAMF and Zone Prompt can jointly improve the model performance in row 6, attaining optimal results on challenging metrics such as AP_{85} and AP_S.

2) Scale-Aware Query Pattern: Table V presents the effectiveness of different Scale-Aware Query (*SQ*) patterns, including learnable parameters, image features with and without spatial attention [40]. The native image features after channel averaging are used for "Image Feat. w/o SA" and for initializing the learnable parameters. The spatial attention maps derived from image features are designated as "Image Feat. w/ SA".

It can be observed that learnable parameters achieve the best result on AP_M, while "Image Feat. w/ SA" performs better on AP_{75} and AP_L, and both have similar results overall. Using image features alone also brings some improvement, but lower than other patterns. We suggest that *SQ* allows the model to adaptively align semantic information across modalities, whereas using only image features tends to obscure the distinction between conceptual and structural differences, thus weakening the alignment effect.

3) Ablations on SAMF Enhancement: Table VI shows the effect of different SAMF configurations on COCO. SAMF with Modal Enhancement gives the baseline a strong boost with 0.8 AP_{50}, 1.0 AP_{75} and 1.6 AP_L, while SAMF with Channel Enhancement provides a weak promotion with 0.4 AP_{50}, 0.5 AP_{75} and 0.9 AP_L. We believe that the information relate to modal in MI_{SAMF} dominates the multi-modal fusion, rather than the channel dimension. Besides, it is evident that using two modules together maximizes the improvements.

4) Prompting Strategy: We study the impact of different prompt strategies on Zone-YOLO in Table VII. Surprisingly, our findings show that prompt tuning with learnable parameters dose not yield superior results, with only moderately better than ZP1 on AP_{85}, AP_{95} and AP_L. We conjecture that zone prompts do not heavily rely on semantic information and will be optimized by the proposed Adapter and Zone Head. Despite utilizing distinct locality nouns, ZP1 and ZP2 yield comparable outcomes. ZP1 has slightly higher AP_{50}, AP_{95}, and AP_S compared to P2, while ZP2 exhibits a marginal advantage in terms of AP_{85}, AP_M, and AP_L, highlighting the robustness of the Zone Prompt.

V. CONCLUSION

In this work, we propose Zone-YOLO to improve the YOLO-style vision-language object detection fine-tuning to a new level, and demonstrate its application in traffic object detection. To address the defects of existing multi-modal fusion approaches, we propose SAMF to fully exploit the text features and learn to fuse the multi-modal representations seamlessly at different scales. We pioneer a novel Zone Prompt efficient fine-tuning method to introduce text features into regression process and capture the zone-class-entity triple co-occurrence, which significantly improves the localization performance of the model. Extensive experiments show that Zone-YOLO achieves competitive results in intricate traffic scenarios and demonstrate the superiority of fine-tuning on pre-trained VLOD. In the future, we will further investigate the effects of zone prompts, explore lightweight structures, and delve into the real-world ITS applications of Zone-YOLO as a foundational model.

REFERENCES

[1] Q. Ding et al., "CF-YOLO: Cross fusion YOLO for object detection in adverse weather with a high-quality real snow dataset," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 10, pp. 10749–10759, Oct. 2023.

[2] S. Liang et al., "Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 25345–25360, Dec. 2022.

[3] X. Zhou, G. Yang, Y. Chen, L. Li, and B. M. Chen, "VDTNet: A high-performance visual network for detecting and tracking of intruding drones," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 9828–9839, Aug. 2024.

[4] N. Jia, Y. Sun, and X. Liu, "TFGNet: Traffic salient object detection using a feature deep interaction and guidance fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 3020–3030, Mar. 2024.

[5] N. V. R. Guggilam, R. Chiramdasu, A. B. Nambur, N. Mikkineni, Y. Zhu, and T. R. Gadekallu, "An expert system for privacy-driven vessel detection harnessing YOLOv8 and strengthened by SHA-256," *Comput. Secur.*, vol. 143, Aug. 2024, Art. no. 103902.

[6] K. Fang, J. Chen, H. Zhu, T. R. Gadekallu, X. Wu, and W. Wang, "Explainable-AI-based two-stage solution for WSN object localization using zero-touch mobile transceivers," *Sci. China Inf. Sci.*, vol. 67, no. 7, Jul. 2024, Art. no. 170302.

[7] C. Li et al., "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[8] G. Jocher, A. Chaurasia, and J. Qiu. (Jan. 2023). *Ultralytics YOLO*. [Online]. Available: https://github.com/ultralytics/ultralytics

[9] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024, *arXiv:2402.13616*.

[10] A. Wang et al., "YOLOv10: Real-time end-to-end object detection," 2024, *arXiv:2405.14458*.

[11] S. Xu et al., "PP-YOLOE: An evolved version of YOLO," 2022, *arXiv:2203.16250*.

[12] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[13] M. Z. Hasan et al., "Vision-language models can identify distracted driver behavior from naturalistic videos," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 11602–11616, Sep. 2024.

[14] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2019, pp. 1–13.

[15] Q. Ye et al., "MPLUG-OwI2: Revolutionizing multi-modal large language model with modality collaboration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 13040–13051.

[16] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, Aug. 2024.

[17] L. H. Li et al., "Grounded language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2022, pp. 10965–10975.

[18] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14393–14402.

[19] Y. Zhong et al., "RegionCLIP: Region-based language-image pretraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16793–16803.

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

[21] L. Yao et al., "DetCLIPv2: Scalable open-vocabulary object detection pre-training via word-region alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23497–23506.

[22] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-world: Real-time open-vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16901–16911.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[24] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[25] C. Zhang et al., "All in one: Exploring unified vision-language tracking with multi-modal alignment," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5552–5561.

[26] J. Wang, P. Zhou, M. Z. Shou, and S. Yan, "Position-guided text prompt for vision-language pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23242–23251.

[27] Y. Yao et al., "PEVL: Position-enhanced pre-training and prompt tuning for vision-language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 11104–11117.

[28] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[29] C. Liu, H. Ding, Y. Zhang, and X. Jiang, "Multi-modal mutual attention and iterative interaction for referring image segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 3054–3065, 2023.

[30] H. Tan, Z. Tan, J. Li, J. Wan, and Z. Lei, "PVLR: Prompt-driven visual-linguistic representation learning for multi-label image recognition," 2024, *arXiv:2401.17881*.

[31] K. Dasgupta, A. Das, S. Das, U. Bhattacharya, and S. Yogamani, "Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15940–15950, Sep. 2022.

[32] Y. Long et al., "Fine-grained Visual–Text prompt-driven self-training for open-vocabulary object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 16277–16287, Nov. 2024.

[33] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.

[34] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14084–14093.

[35] C. Feng et al., "Promptdet: Towards open-vocabulary detection using uncurated images," in *Proc. 17th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 701–717.

[36] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MaPLe: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19113–19122.

[37] Z. Guo, B. Dong, Z. Ji, J. Bai, Y. Guo, and W. Zuo, "Texts as images in prompt tuning for multi-label image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2808–2817.

[38] S. Liu et al., "DQ-DETR: Dual query detection transformer for phrase extraction and grounding," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 1728–1736.

[39] X. Ding, J. Han, H. Xu, X. Liang, W. Zhang, and X. Li, "Holistic autonomous driving understanding by bird'view injected multi-modal large models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 13668–13677.

[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[42] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[43] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2636–2645.

[44] D. Du et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 213–226.

[45] A. Gupta, P. Dollár, and R. Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5356–5364.

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

[47] M. Contributors. (2022). *MMYOLO: OpenMMLab YOLO Series Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmyolo

[48] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–18.

[49] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Proc. Eur. Conf. Comput. Vis.*, vol. 13669. Cham, Switzerland: Springer, 2022, pp. 350–368.

[50] Y. Xu et al., "Multi-modal queried object detection in the wild," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 4452–4469.

Jiaxiong Yang received the B.S. degree from the College of Computer Science and Technology, Wuhan University of Technology, Wuhan, China, in 2022. He is currently pursuing the M.S. degree with the College of Electronic and Information Engineering, Department of Computer Science, Tongji University, Shanghai, China.

His research interests include pattern recognition, deep learning, and computer vision.

Ning Jia received the Ph.D. and Post-Doctoral degrees in computer science and technology from Tongji University, Shanghai, China, in 2019 and 2024, respectively.

He is currently an Associate Professor with the College of Electronic and Information Engineering, Tongji University. His research interests include computer vision, machine learning, and pattern recognition.

Xianhui Liu received the Ph.D. degree from Tongji University, Shanghai, China, in 2014.

He is currently an Associate Professor with the College of Electronic and Information Engineering, Tongji University. He is also the Deputy Director of the CAD Research Center, Tongji University. His research interests include machine learning, data mining, and networked manufacturing. He is a member of the Artificial Intelligence Committee of Shanghai Computer Association.

Rui Fan (Senior Member, IEEE) received the B.Eng. degree in automation from Harbin Institute of Technology, Harbin, China, in 2015, and the Ph.D. degree in electrical and electronic engineering from the University of Bristol, Bristol, U.K., in 2018.

He was a Research Associate with The Hong Kong University of Science and Technology, Hong Kong, from 2018 to 2020, and a Post-Doctoral Scholar- Employee at the University of California at San Diego, La Jolla, CA, USA, from 2020 to 2021. He is currently a Full Professor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, deep learning, and robotics. He served as an Associate Editor for ICRA'23 and IROS'23/24, the Area Chair for ICIP'24, and a Senior Program Committee Member for AAAI'23/24/25. He is the General Chair of the AVVision Community and organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21, and ECCV'22. He was honored by being included in Stanford University List of Top 2% Scientists Worldwide in both 2022 and 2023.

Yougang Sun (Senior Member, IEEE) received the Ph.D. degree in mechatronics engineering from Tongji University, Shanghai, China, in 2017.

From 2018 to 2021, he was a Post-Doctoral Fellow with the National Maglev Transportation Engineering Research and Development Center, Tongji University; and the CNERC Rail, The Hong Kong Polytechnic University, Hong Kong. He is currently an Associate Professor with the Institute of Rail Transit, Tongji University. His research interests include maglev trains, computer vision, and nonlinear control with applications to mechatronic systems. He served as an Associate Editor (an Editorial Board Member) for several journals, including *Journal of Magnetics*, *Traitement du Signal*, and *Advances in Mechanical Engineering*.

Weidong Zhao is currently a Professor with Tongji University. He is the Chief Expert of Information Technology at Sci-Tech Engineering of Manufacturing Industry, China, during the 11th five-year project of the Ministry of Science and Technology of China. His research interests include computer vision and machine learning.

Dr. Zhao is a member of China National Technical Committee for Industrial Automation Systems and Integration Standardization and the Team Leader of Information Technology Manufacturing Engineering of Shanghai Municipal Science and Technology Commission.