Textual Attention RPN for Open-Vocabulary Object Detection

Tae-Min Choi[1, 2]
tmchoi@kist.re.kr;tmchoi@rit.kaist.ac.kr
Inug Yoon[2]
iuyoon@rit.kaist.ac.kr
Jong-Hwan Kim[2]
johkim@rit.kaist.ac.kr
Juyoun Park[1]
juyounpark@kist.re.kr

[1] Korea Institute of Science and Technology (KIST)
Seoul, South Korea

[2] Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea

Abstract

Open-vocabulary object detection (OVD) is a computer vision task that detects and classifies objects from categories not seen during training. While recent OVD methods primarily focus on aligning region embeddings with visual-language pre-trained models like CLIP for classification, object detection requires effective localization as well. However, existing methods often use a proposal generator biased toward the training data, which creates a bottleneck in performance improvement. To address this challenge, we introduce the Textual Attention Region Proposal Network (TA-RPN). This network enhances proposal generation by integrating visual and textual features from the CLIP text encoder, utilizing pixel-wise attention for a comprehensive fusion across the image space. Our approach also incorporates prompt learning to optimize textual features for better localization. Evaluated on the COCO and LVIS benchmarks, TA-RPN outperforms existing state-of-the-art methods, demonstrating its effectiveness in detecting novel object categories.

1 Introduction

Object detection [8, 22] has been a crucial part of computer vision. However, traditional approaches depend on a fixed set of object categories, which makes them difficult to adapt to dynamic real-world scenarios. Real-world applications demand models capable of recognizing unseen objects beyond predefined categories. For these reasons, open-vocabulary object detection (OVD) [30] was recently introduced, which requires the capability to recognize and localize both base and novel classes objects during the training process.

Many OVD studies [9, 14, 28, 31] utilize the CLIP [20]. CLIP performs at zero-shot image classification because of its well-aligned image-text embedding space. To leverage CLIP's generalization capabilities for object detection, OVD methods have focused on distilling CLIP's vision-language knowledge to align cropped image features with the CLIP

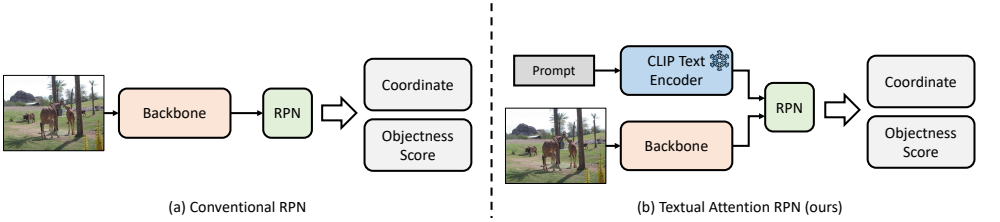

Figure 1: RPN predicts object's proposal coordinate and its objectness score. (a) Conventional RPN only leverages visual features from the backbone network. (b) Proposed Textual Attention RPN combines visual features to textual features from the CLIP text encoder. It improves generalizability and increases proposal generation ability for novel classes.

embedding space. Despite successes in alignment, a significant challenge arises because the proposal generator tends to be biased towards the base classes.

Existing OVD methods commonly use the Region Proposal Network (RPN) [22] to predict object proposals. During training, the RPN is exclusively trained with image features from the base classes, but in testing, it needs to generate proposals for both base and novel classes. Typically designed to predict object locations only based on visual features from the image, the RPN inevitably exhibits inherent bias since it does not encounter novel class samples during training. To address this, [6] proposed a Bias-Balanced RPN to debias the pretrained RPN. However, this solution is not ideal for OVD settings, where novel class samples are absent in training phases. Moreover, [28, 31] have identified that the RPN's performance often becomes a bottleneck for overall detection efficiency. A common solution in OVD is to use an external class-agnostic proposal generator [18] that creates pseudo-proposals for novel classes [7, 33], which are then incorporated into both original and pseudo-labeled datasets for training. While this method enhances detection of novel classes, it tends to result in noisy matching and depends on having known names for all categories during training. Additional strategies to reduce this bias have involved leveraging neighborhood sampling [28] and utilizing extra caption datasets [14, 31] to improve the RPN's generalization abilities. Despite these efforts, the RPN's dependency on visual features from the image backbone network remains a significant challenge.

We propose an enhancement to the traditional RPN, named the **T**extual **A**ttention **RPN** (**TA-RPN**), which integrates both visual and textual features. Unlike the conventional RPN that relies solely on visual features, as depicted in Figure 1, our model incorporates textual information from a CLIP text encoder, significantly enhancing its predictive capabilities for proposal generation. Specifically, TA-RPN aims to prevent bias that arises from using only the visual features of the base classes by leveraging the rich representation capabilities of the CLIP text encoder. This is facilitated by pixel-wise attention for the RPN, which bases its proposals on anchors defined at each pixel across the entire spatial domain, ensuring a comprehensive fusion of visual and textual data. Furthermore, inspired by CoOp [32], our model employs adaptive prompt learning instead of static prompts, such as "a photo of cls". This technique is crucial for effectively utilizing textual features for localization, offering a distinct advantage over traditional methods focused solely on classification.

To evaluate the effectiveness of the proposed method, we conduct extensive experiments using two benchmark datasets, COCO [15] and LVIS [10]. We combine our TA-RPN with BARON [28], based on the Faster R-CNN [22] with a full training backbone and RPN. Our

method consistently performs better than existing state-of-the-art methods in several settings. We achieve a 36.1 bbox average precision on COCO and a 21.5 bbox average precision on LVIS. It demonstrates that TA-RPN generates robust proposals using textual-visual feature fusion. To summarize, our contributions are three-fold: (1) We introduce the TA-RPN, a proposal generator that integrates visual and textual features within the open-vocabulary object detection framework. This approach allows the TA-RPN to robustly predict proposals by leveraging multi-modal features, enhancing detection capabilities beyond the traditional reliance on visual data alone. (2) We introduce prompt learning for localization, enabling the generation of textual features that are optimally suited for proposal prediction. (3) Our method demonstrates competitive performance compared to state-of-the-art benchmarks for open-vocabulary object detection on the COCO and LVIS datasets. These results show the practical effectiveness and advancement of our proposed TA-RPN in handling complex detection scenarios.

2 Related Works

Open-Vocabulary Object Detection. Traditional research in object detection has mostly relied on predefined categories, limiting flexibility in handling unseen objects. To address this limitation, open-vocabulary object detection (OVD) [30] was introduced, aiming to recognize and localize unseen categories. OVD seeks to identify and classify objects from an extensive range of categories without confinement to predefined classes.

OVR-CNN [30] initially demonstrated OVD's potential by pre-training on image-caption pair datasets to detect novel classes. Subsequently, extensive pre-trained Visual Language Models (VLMs) [20] have shown impressive zero-shot recognition capabilities. This advancement led to the introduction of several OVD methods based on VLMs. For instance, ViLD [9] leverages knowledge distillation from CLIP to enhance detection capabilities, utilizing CLIP's text encoder for classification. RegionCLIP [31] advances vision-language pretraining and region representation learning by generating region-text datasets with text descriptions using pre-trained VLMs. DetPro [5] introduces a learnable prompt in OVD, incorporating background interpretation and a context grading scheme. VLDet [14] designs an end-to-end framework that trains directly from image-text pairs, formulating the extraction of region-word pairs as a set matching problem. F-VLM [13] asserts the locality features of the CLIP image encoder are suitable for object detection, thus utilizing frozen CLIP to simplify the multi-stage training pipeline.

While most OVD methods have focused on aligning CLIP and RCNN features to enhance classification performance, they have encountered bottlenecks due to the RPN's bias toward base classes. Addressing this, BARON [28] introduced a neighborhood sampling strategy to generate a bag of regions, contributing to co-occurrence modeling. This strategy enables learning overall visual concepts as it uses expanded proposals for alignment. GOAT [24] proposed a generalized objectness assessment to enhance generalization for novel classes and employed open corpus concepts to improve generalization ability. Despite mitigating bias in base classes, these approaches required additional datasets, and the class-agnostic proposal generators still relied solely on visual features for proposal generation. In contrast, our proposed architecture fuses semantic information from the CLIP text encoder with visual features, enabling the proposal generator to leverage both sources of information to identify potential objects simultaneously. Although GOAT resolves the biased objectness score of RPN with an open corpus, our method with visual-textual fusion fundamentally addresses

the bias issue in base classes and enriches features to enhance the performance of novel classes.

Multi-Modal Feature Fusion. In OVD, most methods involve the CLIP text encoder to generate classifier weights. Unlike these methods, we aim to utilize the text encoder to predict both the coordinates and objectness scores of proposals by fusing visual and textual features. This approach of multi-modal feature fusion is also prevalent in tasks such as open-vocabulary segmentation and referring image segmentation, where pixel-wise classification is required. Unlike in object detection, feature fusion in these tasks typically occurs within both the encoder and decoder stages. For example, Hu et al. [11] introduced a concatenation-based fusion method tailored for segmentation. Similarly, KWAN [23] and ConvRNN [2] utilize text-image feature fusion via attention mechanisms to enhance the integration of modalities. More recently, LAVT [29] has integrated textual features into visual features using a pixel-word attention structure, aiming to densely align textual information with visual clues at each pixel. Additionally, DenseCLIP [21] addresses dense prediction tasks by transforming the image-text matching challenge into a pixel-text matching problem. Building on these developments, we adapt pixel-wise attention-based feature fusion methods to suit the structural needs of the RPN, allowing for handling of both visual and textual data.

Prompt Learning. Prompts are textual inputs used to guide models towards specific tasks or queries. They serve as cues, helping the model grasp the context or intent behind the image. Given the significant impact that prompt adjustments can have on downstream task performance, it is crucial to use prompts that are appropriate for the specific task at hand. Addressing this, CoOp [32] introduces a straightforward prompt learning strategy that optimizes prompts with common learnable vectors, trained through classification loss. This method has proven effective, particularly in enhancing performance in few-shot classification tasks, demonstrating that task-specific learnable prompts can yield superior results.

However, prompt learning strategies have focused on classification tasks. Our paper expands upon CoOp by adapting it to generate textual features for localization. While most prompt learning strategies have concentrated on optimizing context training specifically for classification, our approach introduces CoOp to improve localization capabilities for detecting novel classes.

3 Method

Our method, named Textual Attention RPN (TA-RPN), focuses on predicting an object's bounding box by integrating visual-textual features. We build on the foundational structure provided by BARON [28] with Faster R-CNN [22] as detailed in Section 3.1. The TA-RPN incorporates both pixel-wise and spatial attention modules to enhance feature integration, described in Section 3.2. Additionally, we utilize prompt learning to tailor textual features from the CLIP text encoder for improved localization accuracy, discussed in Section 3.3.

3.1 Preliminary

In OVD, the object detector is initially trained on a set of base classes, C_B, and is desgined to identify objects from an expanded set of classes, denoted as $C = C_B \cup C_N$, where C_N represents novel classes not seen during training. Additionally, we only utilize the names of base classes for training. We employ CLIP as our pre-trained vision-language model, where

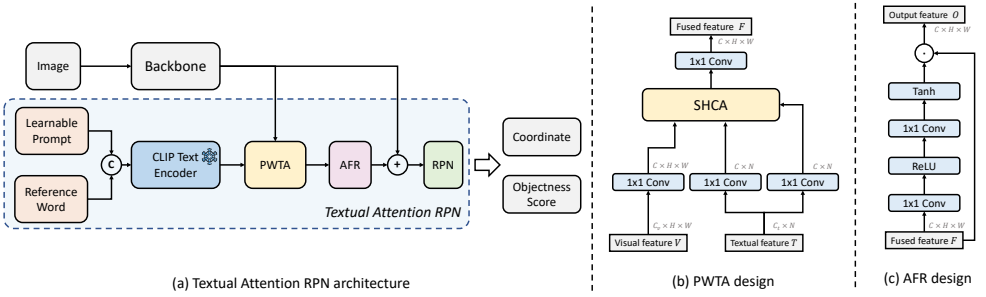

Figure 2: (a) The architecture of Text Attention RPN. It contains a CLIP text encoder, Pixel-Wise Textual Attention (PWTA) module, Adaptive Feature Refinement (AFR) module, and prompt learning structures. (b) PWTA module. It has a single-head cross attention module, which fuses visual and textual features for each pixel. (c) AFR module. It controls the effect of fused feature by making attention map with the same size as fused feature.

the text encoder, $\mathcal{T}(\cdot)$, generates text embeddings, and the image encoder, $\mathcal{I}(\cdot)$, produces image embeddings.

In this paper, our proposed architecture builds upon the RPN from Faster R-CNN. Traditionally, the RPN predicts an object's proposal coordinates and their objectness scores using only visual features derived from the backbone network. We extend this approach by incorporating textual features with visual features. In segmentation tasks, which generate pixel-wise masks, many methods integrate language features from large language models (e.g., BERT [3]) to predict pixel-wise category masks. Similarly, the RPN predicts the coordinates and objectness scores of proposals based on the features of each pixel. Inspired by these practices, we enhance our model by employing the CLIP text encoder to generate textual features. These are then combined with visual features to enrich the input to the RPN.

RPN is the most popular proposal generator used in two-stage object detectors. Thus, our approach, which leverages TA-RPN, could potentially extend to other RPN-based methods across various tasks. To evaluate TA-RPN's effectiveness on the OVD benchmarks, we choose not to rely on pre-trained RPN or additional datasets unlike several OVD methods [5, 7]. Instead, we base our OVD benchmarks on BARON [28], which employs a Faster R-CNN framework without relying on extra caption datasets and pre-trained RPN, allowing us to test TA-RPN under strict conditions.

BARON enhances the Faster R-CNN architecture by integrating the bag-of-regions concept. This approach replaces the traditional classifier with a linear layer that projects region features into pseudo words. These pseudo words are then input into the text encoder. For each category c, the text embedding f_c is derived by applying a fixed prompt template to the category names. The probability of region being classified into category c is calculated as $p_c = \frac{exp(\tau \cdot sim(\mathcal{T}(\omega), f_c))}{\Sigma_{i \in C} exp(\tau \cdot sim(\mathcal{T}(\omega), f_i))}$, where $sim(\cdot,\cdot)$, ω, and τ denote the cosine similarity, region's pseudo word, and temperature for scaling, respectively.

Our loss terms, borrowed from BARON, include the Faster R-CNN's regression and classification losses, as well as $\mathcal{L}_{individual}$ and $\mathcal{L}_{bag}$. $\mathcal{L}_{individual}$ is the InfoNCE loss [19] of current batch and $\mathcal{L}_{bag}$ is also InfoNCE loss between bag-of-regions embeddings. A more detailed description of our training procedure is available in the supplementary material.

3.2 Textual Attention RPN Architecture

In our model, given an input image and text, we generate proposal coordinates and their objectness scores. To extract visual features, we pass the image through a backbone network. The visual feature, denoted as $V \in \mathbb{R}^{C_v \times H \times W}$, where C_v, H, and W represent the channel size and spatial dimensions, respectively. For textual features, we use the CLIP text encoder to transform the input text into meaningful word vectors, denoted as $T \in \mathbb{R}^{C_t \times N}$, where C_t and N represent the channel size of textual features and number of reference words, respectively. After obtaining both visual and textual features, we proceed to feature fusion. As illustrated in Figure 2(a), we input both sets of features into the Pixel-Wise Textual Attention (PWTA) and Adaptive Feature Refinement (AFR) modules. The PWTA module injects textual attention into each pixel of the visual features, enhancing the local relevance of the text data. Simultaneously, the AFR module creates an attention map from the fused features, which is then used to multiply with the visual features for adaptive refinement.

Pixel-Wise Textual Attention. To accurately predict object coordinates and objectness scores, it is essential to integrate textual information with visual features. We utilize a pixel-wise attention structure, specifically designed to complement the functionality of the RPN. The RPN predicts object coordinates based on anchors at each pixel of the input feature and assesses the probability that an object exists within each proposal. By focusing on merging textual representation at the pixel level, our PWTA efficiently calculates attention weights between the two features, resulting in an effectively fused feature.

Figure 2(b) depicts PWTA design. Given an input vision feature $V \in \mathbb{R}^{C_v \times H \times W}$ and textual feature $T \in \mathbb{R}^{C_t \times N}$, we first transform both features to dimension C using three 1×1 convolution layers. Then, we pass them to Single-Head Cross Attention (SHCA) module, with the visual feature as the query and textual features as the key and the value. The output feature of SHCA is lastly passed through a 1×1 convolution, then we get the fused feature $F \in \mathbb{R}^{C \times H \times W}$. The 1×1 convolution belonging to PWTA is followed by instance normalization. Unlike other works that adopt multi-head attention module, we adopt the SHCA to balance the focus of attention. While multi-head attention can capture diverse relationships within complex datasets, single-head attention is particularly suitable for our structure. It offers a more focused and efficient approach, proving to be more effective in scenarios where the added complexity of multi-head attention does not yield significant performance gains. Notably, the simplicity of SHCA can be better generalization to novel classes, which is crucial for OVD. We show an ablation about attention head in the Section 4.3.

Adaptive Feature Refinement. Since predicting proposals predominantly relies on visual information, we introduced an additional refinement structure to ensure that fused features do not overwhelm the primary visual features. After processing through the PWTA, we combine the fused feature F with the visual feature V. Inspired by [27], we developed the Attention Feature Refinement (AFR) module to generate an attention map for refining these features. The AFR consists of a 1×1 convolution followed by ReLU and Tanh activations. Unlike [27], where max-pooling and average-pooling are used to generate a feature descriptor, we preserve the original size of the fused feature to maintain both channel-wise and spatial information. Furthermore, through adaptive refinement using the attention map, we highlight information for proposal generation. As illustrated in Figure 2(c), after the attention map is obtained, we apply element-wise multiplication to the fused features, enhancing the relevant visual data for improved proposal accuracy.

3.3 Prompt Learning

DetPro [5] employs learnable prompts based on CoOp [37], instead of hand-crafted fixed prompts. Learnable prompts have shown performance improvements classification performance of OVD task. Building on this success, we aim to adapt these prompts for predicting proposals in localization tasks. In CoOp, a prompt is generated by appending the class name at the end of the learnable context; this combination is then used as the classifier's weights. We refer to the word appended to the learnable context as the 'reference word'. To ensure the generation of textual features regardless of dataset changes or variations in the number of classes, we maintain a fixed number of reference words. Also, we conduct the ablation studies for this approach, which enables the model to dynamically adapt textual inputs to enhance spatial understanding, crucial for accurately localizing objects across diverse scenes. We demonstrate it by conducting an ablation study on the number of reference words and initialization method in Section 4.3.

4 Experiment Results

4.1 Experiment settings

Datasets. We evaluate our method using the COCO [15] and LVIS [10], following the open-vocabulary object detection setting from [1]. In this setting, the COCO 2017 dataset is restructured into 48 base classes and 17 novel classes. For LVIS, the frequent and common classes are designated as base classes (886 classes), and the rare classes are treated as novel classes (337 classes). We train our model using only the base classes in both datasets.

Metrics. We evaluate the detection performance using the average precision (AP) metric across both base and novel classes. For COCO, we specifically measure AP at the IoU threshold of 0.5, denoted as AP_{50}. Performance for novel and base classes is evaluated using AP_{50}^{N} and AP_{50}^{B}, respectively. Additionally, we assess the accuracy of proposal generation for novel classes with $AR^{N}@100$. For LVIS, our primary metric is AP for rare classes, denoted as AP_r, along with AP_c for common classes, AP_f for frequent classes, and overall AP.

Implementation Details. Initially, we apply the proposed TA-RPN to BARON [28] and carry out experimental evaluations. BARON is based on the Faster R-CNN architecture, which includes RPN. It is a suitable framework for our approach as it allows for concurrent learning of both the RPN and the detector. We develop our model using two different backbones depending on the dataset: ResNet50-C4 for COCO and ResNet50-FPN [16] for LVIS. Both backbones are initialized with pre-trained weights from SoCo [26]. For the pre-trained vision-language models (VLMs), we select the CLIP model based on ViT-B/32 [4]. By default, all experiments utilize hand-crafted prompts for category names from ViLD [9]. Our training process involves 90,000 iterations (1x schedule) for the COCO dataset and 180,000 iterations (2x schedule) for the LVIS dataset, both with a batch size of 16. In the pixel-wise text attention (PWTA) and the adaptive feature refinement (AFR) modules, we set the channel dimension, C, to 1024. The number of reference words is set at 48, initialized to the names of COCO base classes. Furthermore, we employ the same training strategies as BARON [28], which include their loss terms and the sampling settings for the bag-of-region embedding.

Method	Backbone	Extra Data	AP^N_{50}	AP^B_{50}	AP_{50}
OVR-CNN [30]	ResNet50-C4	✓	22.8	46.0	39.9
ViLD [9]	ResNet50-FPN	✗	27.6	59.5	51.3
RegionCLIP [51]	ResNet50-C4	✓	26.8	54.8	50.4
Detic [53]	ResNet50-C4	✓	27.8	47.1	45.0
VLDet [24]	ResNet50-C4	✓	32.0	50.6	45.8
F-VLM [13]	ResNet50-FPN	✗	28.0	-	39.6
OADP [25]	ResNet50-FPN	✗	30.0	53.3	47.2
CoDet [17]	ResNet50-C4	✗	30.6	52.3	46.6
ProxyDet [3]	ResNet50-C4	✗	30.4	52.6	46.8
BARON [28]	ResNet50-FPN	✗	34.0	60.4	53.5
Ours	ResNet50-C4	✗	**36.1**	53.1	43.3

Table 1: Comparison with existing OVD methods in terms of AP on the COCO dataset. We refer to the backbone structure and training source of each method.

Method	Backbone	Object Detection				Instance Segmentation			
		AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
ViLD [9]	ResNet50	16.1	20.0	28.3	22.5	16.3	21.2	31.6	24.4
RegionCLIP [51]	ResNet50	17.1	27.4	34.0	28.2	-	-	-	-
Detic [53]	ResNet50	-	-	-	-	17.8	26.3	31.6	26.8
DetPro [6]	ResNet50	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
F-VLM [13]	ResNet50	-	-	-	-	18.6	24.0	26.9	24.2
ProxyDet [3]	ResNet50	-	-	-	-	18.9	-	-	30.1
BARON* [28]	ResNet50	20.4	30.9	33.4	30.1	19.8	28.6	30.2	27.7
Ours	ResNet50	**21.5**	30.4	33.7	30.2	**20.7**	28.6	30.3	27.9

Table 2: Comparison with existing OVD methods in terms of bbox AP and mask AP on the LVIS dataset with the ResNet50 backbone. * indicates the re-implementation results.

4.2 Benchmark Results

Comparison on the COCO. In Table 1, we compare our method with state-of-the-art methods on the COCO dataset using a fair benchmark that includes several OVD methods based on the two-stage object detector, like Faster R-CNN. We also specify the structure of the backbone and training resources used. The results clearly show that our method significantly outperforms existing state-of-the-art methods. Notably, our method achieves better performance than others that utilize additional training resources, even though we rely solely on the COCO dataset, without the use of pseudo-labels, captions, or beyond the names of base categories. Specifically, our method exceeds BARON, our foundational method, by 2.1 in AP^N_{50}, highlighting the effectiveness of our TA-RPN which includes tailored modules for OVD.

Comparison on the LVIS. In Table 2, we compare our method against recent state-of-the-art OVD methods on the LVIS dataset, including those trained without extra resources. For fairness, we use FPN in the LVIS comparisons. Additionally, we incorporate an ensembling strategy and learned prompts as proposed by ViLD [9] and DetPro [6]. The results demonstrate that our method surpasses BARON by 1.1 in bbox AP_r and 0.9 in mask AP_r.

#	PWTA	AFR	Prompt learning	AP_{50}^{N}	AP_{50}^{B}	$AR^{N}@100$
1	✗	✗	✗	32.1	50.7	35.5
2	✓	✗	✗	34.3 (+2.2)	53.5	35.9 (+0.4)
3	✓	✓	✗	35.2 (+3.1)	53.1	36.1 (+0.6)
4	✓	✗	✓	34.7 (+2.6)	48.8	35.3 (-0.2)
5	✓	✓	✓	**36.1 (+4.0)**	53.1	**36.4 (+0.9)**

Table 3: Ablation study of PWTA, AFR, and prompt learning module of LADet

4.3 Ablation Study

Effect of Structures. In our ablation study, we empirically evaluate the effectiveness of the PWTA, AFR, and prompt learning in our model as shown in Table 3. We ablate our model by combining the ResNet50-C4 backbone with TA-RPN, matching the benchmarks set for the COCO dataset. The PWTA, which integrates visual and textual features, achieves a 34.3 AP_{50}^{N}, as indicated in Table 3 (#2). As further demonstrated in Table 3 (#3, #4), both the AFR and prompt learning significantly enhance performance when used PWTA. Finally, entry #5 in Table 3 shows that all components are complementary, collectively contributing to a total performance gain of 4.0 AP_{50}^{N}. Additionally, the results show the highest $AR^{N}@100$ when all modules are used. This indicates an enhanced detection capability for novel classes, demonstrating the effectiveness of TA-RPN.

Attention Structure. We demonstrate the effect of different types of attention heads in Table 4(a). We observe that the SHCA outperforms the Multi-Head Cross Attention (MHCA) modules. This results indicate that a simpler fusion structure is more suitable for the RPN, and it means that our structure sufficiently achieves visual-textual feature fusion in terms of localization. Additionally, to analyze the effectiveness of pixel-wise attention, we apply a 7×7 patch-wise attention to visual features. Experimental results indicate that pixel-wise attention is more suitable as it considers the structure of the RPN.

Reference Words. We demonstrate the effect of the number of reference words in Table 4(b), observing that AP_{50}^{N} increases as the number of reference words increases. This indicates that generating textual features with helpful knowledge for localization requires a sufficient number of words. Additionally, we explore the impact of different initialization settings for reference words in Table 4(c). Initially, we set the reference words to generate a random string of six characters. The second approach involves initializing them with the names of COCO base classes. The results show that initialization with COCO class names results in better than random initialization. This suggests that using reference words derived from COCO classes not only improves classification but also enhances localization.

Visualization results. As shown in Figure 3, we compare the qualitative results with the baseline, BARON. In the first two rows, it is clear that TA-RPN generates accurate proposals for novel classes (airplane, scissors, and skateboard). BARON generates multiple proposals for one object, but our results show the correct proposal. Furthermore, in the third and fourth rows, we can see that these accurate proposals are effectively used to classify novel classes even in complex scenes. BARON failed to detect the umbrella and airplane (which are novel classes) in the first and third scenes, respectively.

	AP_{50}^{N}	AP_{50}^{B}	AP_{50}
(a) type of attention structure study			
SHCA	**36.1**	53.1	43.3
MHCA (head=2)	35.0	53.5	43.5
MHCA (head=4)	34.7	53.4	43.1
MHCA (head=8)	34.7	53.2	43.1
Patch-wise	35.4	53.3	43.3
(b) number of reference word study			
12	34.3	53.2	42.9
24	34.5	53.1	42.8
36	35.3	53.2	43.3
48	**36.1**	53.1	43.3
(c) reference word initialization study			
Random	35.2	53.2	43.2
COCO	**36.1**	53.1	43.3

Table 4: Ablation studies on the COCO dataset.

Figure 3: Qualitative results

5 Conclusion

In this paper, we propose Textual Attention RPN (TA-RPN) by integrating textual information with visual features. TA-RPN employs Pixel-Wise Textual Attention (PWTA), Attention Feature Refinement (AFR), and the prompt learning. This integration improves proposal prediction, particularly for OVD. Our modifications allow for more precise proposal localization by maintaining the integrity of spatial and channel-wise information. The results from testing on the COCO and LVIS datasets demonstrate substantial improvements over existing methods, affirming the effectiveness of our multi-modal fusion approach.

6 Acknowledgment

This work was supported by Korea Institute of Science and Technology (KIST) Institutional Program [Project No.2E32992].

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 384–400, 2018.

[2] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7454–7463, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022.

[6] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4527–4536, 2021.

[7] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision*, pages 266–282. Springer, 2022.

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[9] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021.

[10] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[11] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016.

[12] Joonhyun Jeong, Geondo Park, Jayeon Yoo, Hyungsik Jung, and Heesu Kim. Proxydet: Synthesizing proxy novel classes via classwise mixup for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2462–2470, 2024.

[13] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[14] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *The Eleventh International Conference on Learning Representations*, 2022.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence 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*, pages 740–755. Springer, 2014.

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[17] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *European conference on computer vision*, pages 512–531. Springer, 2022.

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[23] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, 2018.

[24] Jiong Wang, Huiming Zhang, Haiwen Hong, Xuan Jin, Yuan He, Hui Xue, and Zhou Zhao. Open-vocabulary object detection with an open corpus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6759–6769, 2023.

[25] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2023.

[26] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021.

[27] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[28] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023.

[29] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[30] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[31] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[32] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[33] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.