Tensor Decomposition and Language Description for Open-Vocabulary Object Detection

Qiuyu Liang, Yongqiang Zhang*

College of Computer Science, Inner Mongolia University, Hohhot, China
liangqiuyu2024@163.com, zhangyongqiang@imu.edu.cn

Abstract

Open-vocabulary object detection (OVOD) aims at detecting and recognizing objects beyond a fixed set of classes. Although region-word alignment and knowledge distillation have been explored for training a strong open-vocabulary detector, our analysis reveals three main issues (inaccurate alignment, redundant distillation, and low-quality class embedding) that limit OVOD's performance. In this paper, we explore the well-designed **T**ensor decomposition and **L**anguage descriptions for open-vocabulary object **Det**ection (called TLDet). Proposals with the highest similarity score often correspond to discriminative but incomplete regions (*e.g.*, object heads), resulting in inaccurate region-word alignment. To mitigate this issue, we propose a low-rank proposal filtering module that quantitatively assesses the completeness of each proposal by performing singular value decomposition and computing the sum of its singular values. This allows the model to reduce discriminative proposals and enhance the precision of alignment between visual regions and textual concepts. Furthermore, to mitigate redundant knowledge transfer, we introduce a core tensor distillation approach that decomposes teacher and student features into core tensors via Tucker decomposition and performs distillation through optimized tensor alignment. This ensures that the student acquires the most essential knowledge from the teacher. Finally, to improve the quality of class embedding, a language description enhancement method is proposed by exploring the knowledge of LLM to enrich the representations of categories during inference. Extensive experiments on popular datasets demonstrate the superior performance of our TLDet, achieving 36.1% mAP on COCO and 30.1% mask mAP on LVIS, and outperforming existing methods on novel categories.

Introduction

Traditional object detectors (Ren et al. 2016; Zhang et al. 2024; Zhu et al. 2022; Zhang et al. 2020), which rely on a fixed set of predefined categories, have achieved remarkable performance. However, these detectors perform poorly in real-world scenarios, where object categories exhibit long-tailed and open-ended distributions. To address this limitation, researchers have shifted from closed-set detection to a more general paradigm, *i.e.*, open-vocabulary object detection (OVOD) (Zareian et al. 2021). This task aims to empower object detectors with the ability to recognize and localize arbitrary objects, including unseen categories during training. Benefiting from advances in large-scale vision language pretraining models, recent research (Zhong et al. 2022) leverages CLIP (Radford et al. 2021) to construct a shared semantic space that bridges visual region and category embeddings in the OVOD task.

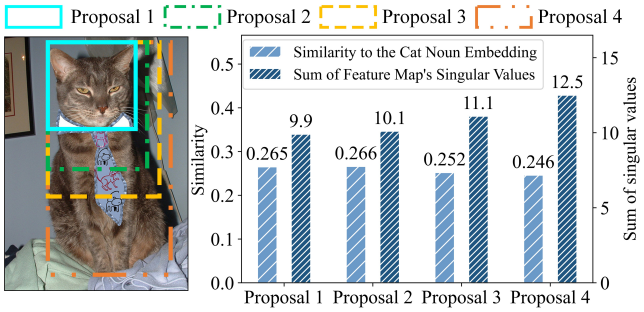

Figure 1: The left vertical axis indicates the similarity between the proposal features and the 'Cat' noun embedding, and the right vertical axis shows the sum of the singular values of the proposal features. Proposal 1–4 indicate the boxes arranged from the top-left to the bottom-right of the image.

Achieving precise visual regions and word embeddings alignment remains a critical challenge in training robust object detectors (Ma et al. 2023), especially when dealing with unseen categories. To address this issue, VLDet (Lin et al. 2023) formulates alignment as a bipartite matching problem (Kuhn 1955) between regions and word embeddings, assigning a proposal to each word. The latest approach (Jin et al. 2024) uses large language models (Brown et al. 2020) to generate fine-grained descriptions of categories, which are used to enhance region-word alignment. The fundamental mechanism of alignment centers on measuring the similarity between image regions and text embeddings.

However, we find that the alignment process preferentially assigns high similarity scores to the most discriminative but incomplete region of an object, as shown in Figure 1. Aligning the discriminative and partial regions (*i.e.*, the cat's head) with the word embedding of 'Cat' obtains the highest similarity score. Intuitively, when bounding boxes for novel classes are unavailable, aligning with complete ob-

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ject regions can be beneficial for training a robust detector. Unfortunately, the aforementioned matched partial regions (*i.e.*, inaccurate alignment) may not meet the requirement of IoU greater than 0.5 in the object detection evaluation criteria, which hinders the performance of OVOD. To alleviate this issue, we propose a Low-Rank Proposal Filtering (LRPF) method from a tensor decomposition perspective. Our key motivation stems from recent studies (Chen et al. 2019; Lin et al. 2020; Chen et al. 2022), which demonstrate that larger singular values in feature representations capture more informative content, while smaller ones correspond to low-contribution information. As illustrated in Figure 1, the region with the highest similarity (the most discriminative region) does not necessarily exhibit the largest sum of singular values, and the entire region obtains the largest sums of singular values. Therefore, before aligning proposal features with category embeddings, we first perform singular value decomposition (SVD) on proposal feature maps and compute the sum of singular values as a completeness metric. Based on the principle that larger sums of singular values indicate more complete regions, we filter out discriminative proposals and retain the proposal features of complete regions for alignment with category embeddings.

Recent OVOD methods (Ma et al. 2022; Bangalath et al. 2022; Wu et al. 2023a) use knowledge distillation to transfer knowledge from a powerful teacher model (Jia et al. 2021; Radford et al. 2021) to a target model (student model). Specifically, the teacher model processes cropped image regions to generate high-quality semantic features, whereas the student model extracts region-specific features via RoIAlign from the feature map. However, architectural discrepancies between two models often result in suboptimal knowledge distillation, where redundant distillation causes the student model to receive irrelevant features (Mirzadeh et al. 2020). Consequently, the detection performance of the student model is compromised due to inefficient feature transfer. To overcome this issue, we propose a Core Tensor Knowledge Distillation (CTKD) method that aligns essential feature components through tensor decomposition between teacher and student networks and matches their core tensors. CTKD ensures that the student model acquires optimally distilled knowledge from the teacher.

Although Region Proposal Networks (RPNs) have proven effective in covering potential novel categories (Gu et al. 2022; Zhou et al. 2022; Wu et al. 2023a), existing models still exhibit poor inference performance on unseen classes. We attribute this limitation to low-quality class embeddings, resulting in weak activation (*i.e.*, low similarity) between novel regions and their corresponding class embeddings. To alleviate this issue, we propose a Language Description Enhancement (LDE) module that enriches the representation of category by generating language descriptions using LLMs, which are subsequently integrated with the original embeddings to improve inference performance.

Finally, extensive experiments on two popular datasets, COCO and LVIS, demonstrate that our proposed method outperforms existing OVOD methods by a large margin on novel categories. The key contributions of this paper can be summarized as follows:

- We reveal three main issues faced by existing methods in open-vocabulary object detection, a novel **TLDet** is proposed to alleviate these issues from the perspectives of **T**ensor decomposition and **L**anguage description.
- A low-rank proposal filtering module is proposed to alleviate the effect of redundant proposals in region-word alignment.
- A core tensor knowledge distillation method is proposed to perform efficient knowledge transfer.
- A language description enhancement method is proposed to improve the quality of category embedding during inference.
- Extensive experiments conducted on two widely used benchmarks and transfer datasets demonstrate that TLDet outperforms existing open-vocabulary object detection approaches, particularly on novel categories.

Related Works

Open-Vocabulary Object Detection

Open-vocabulary object detection aims to detect objects beyond a closed set of categories. OVR-CNN (Zareian et al. 2021) is one of the earliest works to introduce the open-vocabulary detection task, achieving notable performance by combining bounding-box annotations and image-text pairs. This paradigm has catalyzed significant advances in OVOD methods that utilize large-scale image-text pairs such as CC3M (Sharma et al. 2018) to expand detection lexicons. For example, VLDet (Lin et al. 2023) aligns regions with textual descriptions by finding an optimal bipartite matching (Kuhn 1955). RALF (Kim et al. 2024) retrieves related negative classes and augments loss functions from the real-world to optimize the open-vocabulary object detector. In addition, knowledge distillation has proven to be an effective technique for improving model performance by transferring knowledge from teacher networks to students. For example, HierKD (Ma et al. 2022) focuses on hierarchical global local distillation, and RKD (Bangalath et al. 2022) explores region-based knowledge distillation to improve alignment between region-level and image-level embeddings. Similarly, BARON (Wu et al. 2023a) leverages bag-of-regions embedding for knowledge extraction from pre-trained vision-language image encoders (Jia et al. 2021; Radford et al. 2021).

Although recent OVOD methods have achieved promising results, three key limitations remain unaddressed: (1) inaccurate alignment between visual regions and category embeddings, (2) inefficient knowledge transfer caused by redundant distillation, and (3) low-quality category embeddings during inference. In this paper, we propose a novel TLDet to mitigate the above issues.

Tensor Decomposition

Tensor decomposition is a powerful tool for analyzing and compressing high-dimensional visual representations in deep networks. For example, Denton et al. (2014) uses low-rank approximations to compress CNN weight matrices without losing accuracy. Lee, Kim, and Song (2018) applies

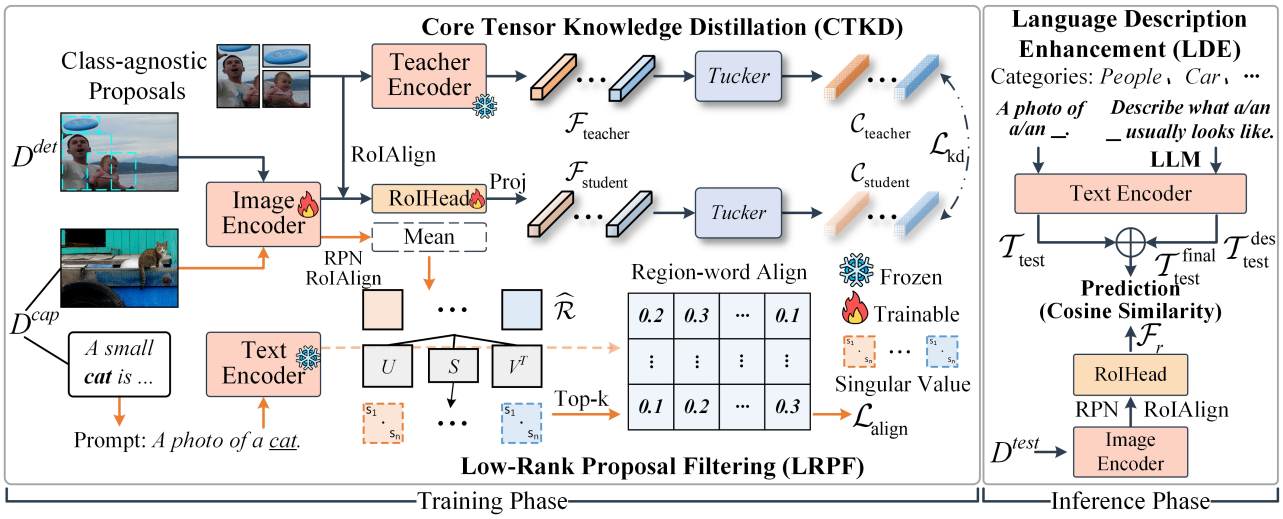

Figure 2: An overall architecture of our proposed method. In the LRPF module, the feature maps of proposals are decomposed using SVD, and the top k proposals with the largest sums of singular values are selected for alignment with text embeddings. In the CTKD module, class-agnostic proposals are passed through the teacher encoder to obtain teacher features, which are then decomposed using $Tucker$ decomposition. The student features undergo identical decomposition, enabling core tensor-to-tensor distillation. Finally, LDE module processes LLM-generated descriptions via text encoder, then fuses them with original category embeddings through weighted summation to produce enhanced class embeddings for prediction.

SVD to reduce spatial redundancy while retaining important semantics. Wang et al. (2024a) and Wang et al. (2025) leverage SVD-based reconstruction to effectively capture global feature representations. Prior pruning methods such as HRank (Lin et al. 2020) and FPC (Chen et al. 2022) prove that low-rank feature maps hold less information and removing them minimally impacts the final performance. In few shot object detection (Wu et al. 2021), SVD is used to enhance the generalizability of the detector on new objects, and EigenGAN (Kas et al. 2024) uses tensor decomposition to generate robust images by capturing the structure of the feature space.

Building on these advances, we explore tensor decomposition as an effective tool to filter proposals and improve region-word alignment in OVOD.

Large Language Models

Large Language Models encapsulate a vast amount of knowledge and can generate detailed descriptions of detected objects, making them valuable for downstream visual tasks. For example, Menon and Vondrick (2023) leverages the linguistic knowledge embedded in pre-trained LLMs to generate descriptors for visual categories, enriching vision-language models without the need for additional training or labeling. A recent OVOD approach (Jin et al. 2024) uses GPT-3 (Brown et al. 2020) to generate fine-grained descriptions during training, improving region word alignment.

However, frequent interactions with LLMs during training add computational overhead. In contrast, in this paper, our language description enhancement method can be applied offline to significantly improve category representations and boost the performance of OVOD.

Methods

Problem Setup

Open-vocabulary object detection (OVOD) aims to extend the detection capability beyond a fixed set of categories by leveraging image-text pairs. Following the classic OVOD setting (Zareian et al. 2021), only base categories C^{base} are annotated with bounding boxes during training. At test time, the detector is expected to generalize beyond C^{base} and localize and classify objects from novel categories C^{novel}, where $C^{\text{base}} \cap C^{\text{novel}} = \varnothing$ and $C^{\text{test}} = C^{\text{base}} \cup C^{\text{novel}}$. In this setting, novel categories C^{novel} are not known in advance to simulate open-vocabulary scenarios. Meanwhile, image-text pairs are used to extend the detector's lexicon with an open-ended vocabulary C^{open} that includes both seen and unseen categories.

Overview

Our TLDet is built on the two-stage detection framework of Faster R-CNN (Ren et al. 2016), and is designed to address the three challenges (inaccurate alignment, redundant distillation, and low-quality class embedding) of open-vocabulary object detection through three new designed modules: **Low-Rank Proposal Filtering (LRPF)**, **Core Tensor Knowledge Distillation (CTKD)**, and **Language Description Enhancement (LDE)**. We illustrate the overall architecture of our method in Figure 2. Consistent with standard OVOD practices, our framework takes two types of data as inputs: (1) **Detection data** ($\mathcal{D}^{\text{det}}$), which includes images and annotated base categories for supervised learning, and (2) **Caption data** ($\mathcal{D}^{\text{cap}}$), which consists of image-text pairs for region-word alignment.

Low-Rank Proposal Filtering

Accurate alignment between proposal features and category embeddings is crucial to building robust open-vocabulary object detection. However, our analysis reveals that proposals with the highest similarity scores to text embeddings often fail to encapsulate complete object regions (as discussed in the Introduction and Figure 1), resulting in inaccurate alignment during the region-word alignment phase. In this paper, we address this issue by retaining the proposals with the highest completeness among the candidate proposals from a tensor decomposition perspective.

Specifically, given an input image $I \in \mathcal{D}^{\text{cap}}$, visual features are first extracted using an image encoder. These features are passed through RPN and RoIAlign to generate proposal feature maps $\mathcal{R} = \{R_1, R_2, \ldots, R_n\}$, where $R_i \in \mathbb{R}^{1\times c\times h\times w}$ denotes the feature vector of the i-th proposal, and c, h and w represent the output channel, height, and width of the image encoder. Then, we perform channel-wise averaging on these feature maps to obtain the aggregated feature map $\hat{\mathcal{R}} \in \mathbb{R}^{n\times h\times w}$. Subsequently, we apply low-rank singular value decomposition (SVD) on $\hat{\mathcal{R}}$ to obtain a compressed approximation, which is formulated as:

$$\hat{\mathcal{R}} \approx \mathbf{U}\mathbf{S}\mathbf{V}^{\top}, \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{n\times h\times r}$, $\mathbf{S} \in \mathbb{R}^{n\times r\times r}$, and $\mathbf{V} \in \mathbb{R}^{n\times w\times r}$ correspond to the left singular vectors, the diagonal matrix of singular values, and the right singular vectors, respectively. Here, r represents the number of singular values.

Recent research (Chen et al. 2019) demonstrated that feature representations can be decomposed into a set of eigenvectors, and their relative importance can be quantified by the corresponding singular values. Therefore, to effectively evaluate the contribution of the proposal feature map $R_i \in \mathcal{R}$, we utilize the singular values in the diagonal matrix $\mathbf{S}_i \in \mathbb{R}^{n\times r\times r}$ and compute their sum as a completeness metric of the proposals, *i.e.*, $G_i = \sum_j^r s_j$, where larger G_i indicates that the corresponding feature map represents a more complete region. Feature maps with smaller G_i will be filtered out. Finally, we retain the top-k proposal features for subsequent region-word alignment (*i.e.*, $\mathcal{L}_{\text{align}}$ in Figure 2), and k is a hyperparameter.

Core Tensor Knowledge Distillation

Some OVOD frameworks leverage knowledge distillation to transfer visual knowledge from pre-trained foundation models (Radford et al. 2021). However, recent research indicates that architectural discrepancies between the teacher and student models lead to redundant knowledge distillation, hindering the student's ability to effectively learn relevant features. To alleviate this issue, we propose a core tensor knowledge distillation method from a decomposition perspective.

Following previous methods (Bangalath et al. 2022), we offline obtain category-agnostic proposals for the input image $I \in D^{det}$ by prompting MViT (Maaz et al. 2022) with the query 'all objects'. Then, the corresponding regions are cropped from the original image and passed through a frozen image encoder (*i.e.*, the teacher model) to generate teacher features $\mathcal{F}_{\text{teacher}}$. In parallel, these proposals undergo RoIAlign on the feature maps produced by our image encoder (*i.e.*, the student model) to extract student features $\mathcal{F}_{\text{student}}$. To obtain core knowledge, we perform the Tucker decomposition (Tucker 1966) on teacher and student proposal features to obtain their core tensor representations:

$$\begin{aligned} \mathcal{C}_{\text{teacher}} &= \text{Tucker}(\mathcal{F}_{\text{teacher}}), \\ \mathcal{C}_{\text{student}} &= \text{Tucker}(\mathcal{F}_{\text{student}}). \end{aligned} \tag{2}$$

The core tensors $\mathcal{C}_{\text{teacher}}$ and $\mathcal{C}_{\text{student}}$ capture the essential information components in a compressed subspace and remove redundant features (Zeng et al. 2020). Then a ℓ_2 normalization is applied to the tensors along the feature dimension, and the L_1 loss is used as the distillation objective:

$$\mathcal{L}_{\text{kd}} = \left\| \ell_2(\mathcal{C}_{\text{teacher}}) - \ell_2(\mathcal{C}_{\text{student}}) \right\|_1, \tag{3}$$

which guides the student to mimic the semantic structure encoded by the teacher in a low-rank space, thereby enhancing the model's ability to learn principal knowledge while mitigating redundancy from the teacher model.

Language Description Enhancement

In Faster-RCNN based OVOD models, RPN has been demonstrated to generate region proposals that cover potential objects for novel categories. However, during the inference phase, the performance in recognizing novel categories remains unsatisfactory. This discrepancy primarily stems from low-quality class embeddings, which yield weaker similarity scores when matched against novel region proposal features. To alleviate this problem, we enhance the responsiveness of novel categories by using large language models (LLMs) during the inference phase.

Previous approaches generate the textual description for each category using the "*A photo of a/an* {*category name*}." prompt template, which is subsequently fed into the CLIP text encoder to obtain the corresponding embeddings $\mathcal{T}_{\text{test}}$. In this paper, inspired by the ability of LLMs to generate fine-grained appearance descriptions (Xu et al. 2023), we introduce LLMs into the inference stage of open-vocabulary object detection. Specifically, we use the prompt *"Describe what a/an* {*category name*} *usually looks like."* to query the LLMs, which returns a detailed textual description for each category. These descriptions are then encoded by the CLIP text encoder to obtain a new set of enhanced textual embeddings, denoted as $\mathcal{T}_{\text{test}}^{\text{des}}$. Subsequently, the original category embeddings $\mathcal{T}_{\text{test}}$ and the enhanced description-based embeddings $\mathcal{T}_{\text{test}}^{\text{des}}$ are combined via a weighted summation to produce the final class embeddings $\mathcal{T}_{\text{test}}^{\text{final}}$:

$$\mathcal{T}_{\text{test}}^{\text{final}} = \text{mean}(\mathcal{T}_{\text{test}}, \alpha \cdot \mathcal{T}_{\text{test}}^{\text{des}}), \tag{4}$$

where α is a hyper-parameter. Note that the embeddings $\mathcal{T}_{\text{test}}$ and $\mathcal{T}_{\text{test}}^{\text{des}}$ are precomputed offline, so our LDE module brings no extra computation during inference. Finally, the inference process is conducted as follows:

$$p(r, c) = \frac{\exp(\cos(\mathcal{F}_r, t_c)/\tau)}{\sum_{c' \in C^{\text{test}}} \exp(\cos(\mathcal{F}_r, t_{c'})/\tau)}, \tag{5}$$

where $t \in \mathcal{T}_{\text{test}}^{\text{final}}$, and $\mathcal{F}_r$ denotes visual features of the image encoder after the RoI Head, as shown in Figure 2.

Method	Backbone	Detector	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
OVR-CNN (Zareian et al. 2021)	RN50	Faster-RCNN	22.8	46.0	39.9
Detic (Zhou et al. 2022)	RN50	Faster-RCNN	27.8	47.1	45.0
RegionCLIP (Zhong et al. 2022)	RN50	Faster-RCNN	26.8	54.8	47.7
ViLD (Gu et al. 2022)	RN50	Faster-RCNN	27.6	59.5	51.3
GOAT (Wang et al. 2023a)	RN50	Faster-RCNN	31.7	51.3	46.1
CoDet (Ma et al. 2023)	RN50	Faster-RCNN	30.6	52.3	46.6
OADP (Wang et al. 2023b)	RN50	Faster-RCNN	30.0	53.3	47.2
BARON (Wu et al. 2023a)	RN50	Faster-RCNN	33.1	54.8	49.1
UniDetector (Wang et al. 2023c)	RN50	Faster-RCNN	35.2	56.8	51.2
CondHead (Wang 2023)	RN50x4	Faster-RCNN	33.7	58.0	51.7
VLDet (Lin et al. 2023)	RN50	Faster-RCNN	32.0	50.6	45.8
MMC-Det (Xu et al. 2024)	RN50	Faster-RCNN	33.7	-	47.7
ProxyDet (Jeong et al. 2024)	RN50	Faster-RCNN	30.4	52.6	46.8
RALF (Kim et al. 2024)	RN50	Faster-RCNN	33.4	54.5	49.0
DVDet (Jin et al. 2024)	RN50	Faster-RCNN	34.6	52.8	48.0
OV-DETR (Zang et al. 2022)	ViT-B/32	Deform-DETR	29.4	61.0	52.7
Ro-ViT (Kim, Angelova, and Kuo 2023)	ViT-B/16	OLN-RPN	30.2	-	41.5
CORA (Wu et al. 2023b)	RN50	DAB-DETR	35.1	35.5	35.4
Prompt-OVD (Song and Bang 2024)	ViT-B/16	Deform-DETR	30.6	63.5	54.9
SIA (Wang et al. 2024b)	RN50	DAB-DETR	35.5	40.3	39.3
Baseline (Lin et al. 2023)	RN50	Faster-RCNN	32.0	50.6	45.8
TLDet (Ours)	RN50	Faster-RCNN	**36.1**	54.5	49.7

Table 1: Comparison with other state-of-the-art methods on OV-COCO dataset. The best results are shown in bold numbers.

Training Process

We train our method within a classical two-stage detection framework, following previous work (Lin et al. 2023; Wang et al. 2023a; Liang and Zhang 2025), we treat each image and its corresponding caption as a positive pair, while all other captions within the same batch are considered as negative samples. A binary cross-entropy loss $\mathcal{L}_{\text{image-word}}$ is applied to supervise image-word alignment. The overall training objective is defined as:

$$\mathcal{L}(I) = \begin{cases} \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{image-word}}, & \text{if } I \in D^{\text{cap}}, \\ \mathcal{L}_{\text{kd}} + \mathcal{L}_{\text{rpn}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cls}}, & \text{if } I \in D^{\text{det}}, \end{cases} \quad (6)$$

where $\mathcal{L}_{\text{align}}$ denotes the region-word alignment loss computed after applying our low-rank proposal filtering strategy, and $\mathcal{L}_{\text{kd}}$ represents the core tensor-based knowledge distillation loss. $\mathcal{L}_{\text{rpn}}$, $\mathcal{L}_{\text{reg}}$, and $\mathcal{L}_{\text{cls}}$ correspond to the region proposal network loss, bounding box regression loss, and classification loss, respectively, as defined in Faster-RCNN.

Experiments

Experimental Settings

Datasets. We conduct experiments on two widely used open-vocabulary object detection benchmarks (*i.e.*, COCO (Lin et al. 2014) and LVIS (Gupta, Dollar, and Girshick 2019)). For COCO, we follow the category split defined by OVR-CNN (Zareian et al. 2021), which partitions the object categories into 48 base and 17 novel classes. To provide supervision from image-text pairs, we adopt the COCO Captions (Chen et al. 2015) dataset, which offers five human-written descriptions per image. For LVIS, which contains a large and diverse vocabulary of 1,203 object categories, we follow the protocol in (Gu et al. 2022) by designating the 337 rare categories as novel classes, while the remaining common and frequent categories are treated as base classes. For image-text pairs, we use the CC3M dataset (Sharma et al. 2018), which consists of 2.8 million freely collected image-text pairs from the Web.

Evaluation Metrics. Following the standard OVOD evaluation protocol, we primarily report the box AP at an IoU threshold of 0.5 for novel categories on COCO dataset, denoted as Novel AP_{50}. For LVIS dataset, we focus on the mask AP for rare categories, denoted as $\text{mAP}_r^{\text{Mask}}$. In addition, we also provide Base AP_{50} and Overall AP_{50} for COCO, as well as $\text{mAP}_c^{\text{Mask}}$, $\text{mAP}_f^{\text{Mask}}$, and $\text{mAP}_{\text{all}}^{\text{Mask}}$ for LVIS, where the subscripts c and f represent common and frequent categories, respectively.

Implementation Details. We adopt a frozen CLIP (Radford et al. 2021) text encoder to embed captions and object labels, and use the CLIP ViT-B/32 image encoder as the teacher model for core tensor knowledge distillation. DeepSeek (DeepSeek-AI 2025) is used in our language description enhancement module to generate descriptions for each category. To accelerate convergence, we initialize the model using weights from a detector pre-trained on fully supervised base-category annotations, following the procedure in (Zhou et al. 2022). The top k values for low-rank proposal filtering on COCO and LVIS are set to 15 and 3, and α for COCO and LVIS in the language description enhancement module (*i.e.*, Eq. 4) are set to 0.25 and 0.35, respectively. All experiments are conducted with 8 NVIDIA A6000 GPUs.

On COCO dataset, we follow the training protocol without applying any data augmentation. Our model is based on Faster R-CNN (Ren et al. 2016) with a ResNet50-C4 (He

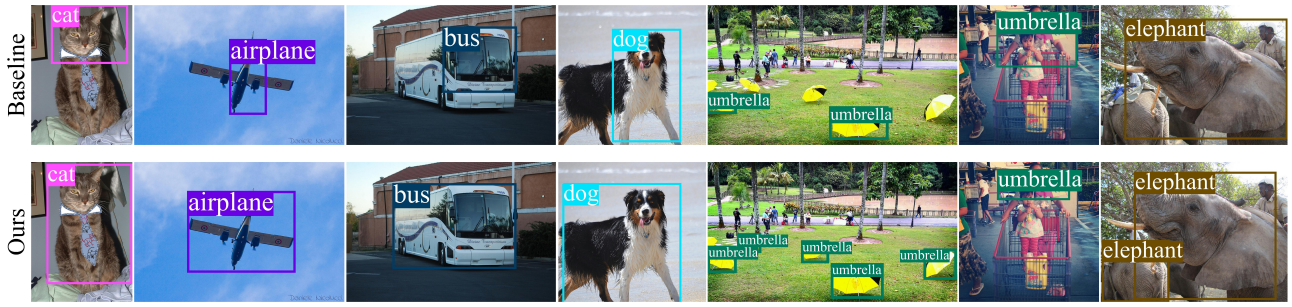

Figure 3: Qualitative results on COCO dataset. The first line is the detection results of baseline (Lin et al. 2023), and the second line shows the detection results of our TLDet. Detection confidence threshold is set to 0.5.

Method	Backbone	mAP_r^{mask}	mAP_{all}^{mask}
RegionCLIP	RN50	17.1	28.2
ViLD	RN50	16.6	25.5
Detic	RN50	19.5	30.9
CoDet	RN50	23.4	30.7
OADP	RN50	21.7	26.6
BARON	RN50	19.2	26.5
ProxyDet	RN50	18.9	30.1
MMC-Det	RN50	21.1	31.0
RALF	RN50	21.9	26.6
CORA	RN50x4	22.2	-
CondHead	RN50x4	24.4	32.0
Detic	Swin-B	23.9	38.4
UniDetector	Swin-B	26.5	32.5
GOAT	Swin-B	27.4	38.5
DVDet	Swin-B	27.5	40.2
Prompt-OVD	ViT-B/16	23.1	24.2
Baseline	Swin-B	26.3	38.1
TLDet (Ours)	Swin-B	**30.1**	39.1

Table 2: Comparison (rare and all metrics) with SOTA open-vocabulary object detection methods on LVIS benchmark.

et al. 2016) backbone. During the warm-up phase, the learning rate is linearly increased from 0 to 0.002 over the first 1,000 iterations. Training is conducted for 90,000 iterations using the SGD optimizer with a batch size of 8, and the learning rate is decayed by a factor of 10 at the 60,000 and 80,000 iteration marks. On LVIS, we adopt the training strategy of Detic (Zhou et al. 2022), using CenterNet2 (Zhou, Koltun, and Krähenbühl 2021) with Swin-B (Liu et al. 2021) as the backbone. The learning rate increases linearly from 0 to 2e-4 during the first 1,000 warm-up iterations. We train the model for 90,000 iterations using the Adam optimizer with a batch size of 16.

Main Results

COCO Result. As shown in Table 1, we compare our proposed TLDet with existing OVOD methods, categorizing the models into two main groups: Faster/Mask R-CNN-like detectors and DETR-like detectors. The results in Table 1 highlight that TLDet outperforms the others in terms of novel AP. Specifically, it improves the performance of novel categories by +1.5% (from 34.6% to 36.1%) compared to the top-performing Faster RCNN-like method DVDet. TLDet and GOAT are both based on the same VLDet architecture, TLDet demonstrates superior performance in detecting novel objects, boosting the novel AP from 31.7% to 36.1%. Compared to DETR-like detectors, TLDet also achieves the best performance. Notably, it surpasses the latest SIA model by +0.6% (from 35.5% to 36.1%) in novel AP. Furthermore, compared to the baseline (32.0%), TLDet obtains a large gain of +4.1% (from 32.0% to 36.1%).

LRPF	CTKD	LDE	Novel AP	Base AP	Overall AP
			32.0	50.6	45.8
✓			34.2	51.3	46.9
	✓		33.8	54.4	49.0
		✓	33.1	50.4	45.9
✓	✓		35.2	54.2	49.2
✓		✓	35.0	51.6	47.2
	✓	✓	34.4	54.4	49.1
✓	✓	✓	**36.1**	54.5	49.7

Table 3: Ablation studies of each component in TLDet on COCO. We report mAP at an IoU of 0.5.

LVIS Result. Table 2 shows a comparison of our TLDet with other methods on LVIS benchmark. Our method achieves 30.1% mAP on novel categories, surpassing all previous approaches across RN50, ViT, and Swin-B backbones. Compared to GOAT (Wang et al. 2023a), TLDet achieves a +2.2% (from 27.4% to 29.6%) improvement on novel categories, further demonstrating its scalability for large-vocabulary detection. Furthermore, TLDet outperforms baseline by +3.8% (from 26.3% to 30.1%) on novel categories, highlighting the effectiveness of our approach.

Ablation Study

Ablation Study for Each Component. Table 3 presents a detailed ablation analysis of three key components in our proposed method, *i.e.*, low-rank proposal filtering (LRPF), core tensor knowledge distillation (CTKD), and language description enhancement (LDE), on COCO dataset. We observe a consistent improvement in novel AP after integrating LRPF, which we attribute to its ability to filter out proposals with minimal contribution. Adding CTKD to the framework

Method	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
TKD	33.2	53.3	48.0
CTKD	**33.8**	54.4	49.0

Table 4: Comparison of CTKD and TKD on COCO.

further boosts both novel AP and base AP, enhancing overall base class performance. Meanwhile, LDE leverages external knowledge from LLMs to construct rich class embeddings, strengthening category representations and improving the ability of model to recognize novel categories. The combination of LRPF, CTKD and LDE leads to the highest novel detection performance (*i.e.*, 36. 1%), highlighting the effectiveness of each component.

Core Tensor Knowledge Distillation *vs.* Traditional Knowledge Distillation. Table 4 presents a comparison between Traditional Knowledge Distillation (TKD) and our proposed CTKD on COCO dataset. The results demonstrate that CTKD consistently outperforms traditional distillation in all metrics. This improvement stems from the ability of core tensor decomposition to suppress redundant feature responses and enhance the compactness of representations.

Efficiency Analysis on LRPF. To investigate the computational overhead introduced by SVD on regional features, we conduct a detailed efficiency analysis of the proposed LRPF module. Specifically, on an NVIDIA A6000 GPU, a training epoch takes 0.618179 seconds, whereas the SVD operation within LRPF requires only 0.000238 seconds, which is less than 0.04% of the total training time. The GPU memory usage remains nearly constant (5267.30 MB $\rightarrow$ 5267.32 MB), indicating an insignificant additional memory footprint.

Fusion Strategy Analysis on LDE. To examine semantic drift arising from the fusion of original category embeddings and LLM-enhanced description-based embeddings in the LDE module, we perform experiments to evaluate two alternative parameter-free variants: (i) *cosine-similarity adaptive fusion*, which dynamically adjusts the fusion weight according to the cosine similarity between embeddings; and (ii) *KL-consistency fusion*, which constrains the fused embedding distribution to remain consistent with the original embedding via Kullback–Leibler divergence regularization. All parameter-free variants achieve comparable performance, *i.e.*, Cosine-adaptive (35.6% mAP) , KL-consistency fusion (35.9% mAP) and our Static mean (36.1% mAP), confirming that our fusion strategy is semantically stable.

Visualization and Analysis

Novel Categories Visualization. We illustrate the detection results of the novel categories of baseline (Lin et al. 2023) and our TLDet on COCO dataset in Figure 3. Compared with baseline, our model effectively detects the entire object rather than just its discriminative regions. The main reason is that LRPF filters out low-contribution proposals, thereby enhancing the precise alignment of regions and words.

Visualization for LDE. To further validate the effectiveness of our LDE during inference, we visualize the detection results before and after its application, as shown in Figure 4.

Figure 4: Qualitative results of novel categories for Language Description Enhancement (LDE) in our method.

Method	PASCAL VOC	LVIS
RegionCLIP (Zhong et al. 2022)	46.9	6.1
OVR-CNN (Zareian et al. 2021)	52.9	5.2
PB-OVD (Gao et al. 2022)	59.2	8.0
Baseline (Lin et al. 2023)	61.7	10.0
TLDet (Ours)	**63.7**	**12.3**

Table 5: Transfer to other datasets. We evaluated COCO-trained model on PASCAL VOC test set and LVIS validation set without re-training. We report mAP at an IoU of 0.5.

The results clearly demonstrate that LDE enhances the prediction scores. We attribute this to the fact that although RPN can generalize to novel classes, current class embedding fails to produce strong activation with novel regions. In particular, LDE improves novel objects with scores below 0.5, which are highlighted with the red box in Figure 4.

Transfer To Other Datasets

To validate the generalization capability of TLDet, we conduct experiments by transferring a COCO-trained model to the PASCAL VOC2007 test set (Everingham et al. 2010) and the LVIS validation set, without any additional re-training. As shown in Table 5, compared to baseline, our model achieves improvements of 2.0% and 2.3% on VOC test and LVIS validation sets, respectively. These improvements demonstrate the effectiveness of TLDet across varied image domains and language vocabularies.

Conclusion

In this paper, we analyze three challenges in current region-word alignment and distillation based OVOD methods: inaccurate alignment, redundant distillation, and low-quality class embeddings. To address these problems, we propose to utilize the tensor decompositions and language descriptions from LLMs for the task of open-vocabulary object detection. Extensive experiments on multiple benchmark datasets show that our approach obviously outperforms current state-of-the-art methods. Meanwhile, we provide in-depth ablation studies and qualitative analyzes to demonstrate the effectiveness of each module in our proposed TLDet.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62206077), the Inner Mongolia Natural Science Foundation for Distinguished Young Scholars (No. 2025JQ009), the Inner Mongolia Talent Development Project for Outstanding Young Talents, and the Inner Mongolia University Graduate Research Innovation Foundation (No. 11200-5253731).

References

Bangalath, H.; Maaz, M.; Khattak, M. U.; Khan, S. H.; and Shahbaz Khan, F. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *NeurIPS*, 35: 33781–33794.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:1504.00325.

Chen, X.; Wang, S.; Long, M.; and Wang, J. 2019. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, 1081–1090. PMLR.

Chen, Y.; Wen, X.; Zhang, Y.; and He, Q. 2022. FPC: Filter pruning via the contribution of output feature map for deep convolutional neural networks acceleration. *Knowledge-Based Systems*, 238: 107876.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Denton, E. L.; Zaremba, W.; Bruna, J.; LeCun, Y.; and Fergus, R. 2014. Exploiting linear structure within convolutional networks for efficient evaluation. *NeurIPS*, 27.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*, 88: 303–338.

Gao, M.; Xing, C.; Niebles, J. C.; Li, J.; Xu, R.; Liu, W.; and Xiong, C. 2022. Open vocabulary object detection with pseudo bounding-box labels. In *ECCV*, 266–282. Springer.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2022. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 5356–5364.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Jeong, J.; Park, G.; Yoo, J.; Jung, H.; and Kim, H. 2024. ProxyDet: Synthesizing Proxy Novel Classes via Classwise Mixup for Open-Vocabulary Object Detection. In *AAAI*, volume 38, 2462–2470.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 4904–4916. PMLR.

Jin, S.; Jiang, X.; Huang, J.; Lu, L.; and Lu, S. 2024. LLMs Meet VLMs: Boost Open Vocabulary Object Detection with Fine-grained Descriptors. In *ICLR*.

Kas, M.; Chahi, A.; Kajo, I.; and Ruichek, Y. 2024. Eigen-GAN: An SVD subspace-based learning for image generation using Conditional GAN. *Knowledge-Based Systems*, 293: 111691.

Kim, D.; Angelova, A.; and Kuo, W. 2023. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *CVPR*, 11144–11154.

Kim, J.; Cho, E.; Kim, S.; and Kim, H. J. 2024. Retrieval-Augmented Open-Vocabulary Object Detection. In *CVPR*, 17427–17436.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Lee, S. H.; Kim, D. H.; and Song, B. C. 2018. Self-supervised Knowledge Distillation Using Singular Value Decomposition. In *ECCV*.

Liang, Q.; and Zhang, Y. 2025. SAM based Region-Word Clustering and Inference Score Adjusting for Open-Vocabulary Object Detection. In *ACM Multimedia 2025*, 2596–2605.

Lin, C.; Sun, P.; Jiang, Y.; Luo, P.; Qu, L.; Haffari, G.; Yuan, Z.; and Cai, J. 2023. Learning Object-Language Alignments for Open-Vocabulary Object Detection. In *ICLR*.

Lin, M.; Ji, R.; Wang, Y.; Zhang, Y.; Zhang, B.; Tian, Y.; and Shao, L. 2020. Hrank: Filter pruning using high-rank feature map. In *CVPR*, 1529–1538.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

Ma, C.; Jiang, Y.; Wen, X.; Yuan, Z.; and Qi, X. 2023. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *NeurIPS*, 36: 71078–71094.

Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; and Hu, W. 2022. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *CVPR*, 14074–14083.

Maaz, M.; Rasheed, H.; Khan, S.; Khan, F. S.; Anwer, R. M.; and Yang, M.-H. 2022. Class-agnostic object detection with multi-modal transformer. In *ECCV*, 512–531. Springer.

Menon, S.; and Vondrick, C. 2023. Visual Classification via Description from Large Language Models. In *ICLR*.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *AAAI*, volume 34, 5191–5198.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2556–2565.

Song, H.; and Bang, J. 2024. Prompt-Guided DETR with RoI-pruned masked attention for open-vocabulary object detection. *Pattern Recognition*, 155: 110648.

Tucker, L. R. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311.

Wang, C.; Wang, W.; Liang, Q.; and Gao, G. 2025. Local and Global Structure-Aware Contrastive Framework for Entity alignment. *Neurocomputing*, 129445.

Wang, C.; Wang, W.; Liang, Q.; Yu, J.; and Gao, G. 2024a. Gsea: Global structure-aware graph neural networks for entity alignment. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 187–199. Springer.

Wang, J.; Zhang, H.; Hong, H.; Jin, X.; He, Y.; Xue, H.; and Zhao, Z. 2023a. Open-vocabulary object detection with an open corpus. In *ICCV*, 6759–6769.

Wang, L.; Liu, Y.; Du, P.; Ding, Z.; Liao, Y.; Qi, Q.; Chen, B.; and Liu, S. 2023b. Object-aware distillation pyramid for open-vocabulary object detection. In *CVPR*, 11186–11196.

Wang, T. 2023. Learning to detect and segment for open vocabulary object detection. In *CVPR*, 7051–7060.

Wang, Z.; Li, Y.; Chen, X.; Lim, S.-N.; Torralba, A.; Zhao, H.; and Wang, S. 2023c. Detecting everything in the open world: Towards universal object detection. In *CVPR*, 11433–11443.

Wang, Z.; Zhou, W.; Xu, J.; and Peng, Y. 2024b. SIA-OVD: Shape-Invariant Adapter for Bridging the Image-Region Gap in Open-Vocabulary Detection. In *ACM Multimedia 2024*.

Wu, A.; Zhao, S.; Deng, C.; and Liu, W. 2021. Generalized and discriminative few-shot object detection via svd-dictionary enhancement. *NeurIPS*, 34: 6353–6364.

Wu, S.; Zhang, W.; Jin, S.; Liu, W.; and Loy, C. C. 2023a. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, 15254–15264.

Wu, X.; Zhu, F.; Zhao, R.; and Li, H. 2023b. CORA: Adapting CLIP for Open-Vocabulary Detection with Region Prompting and Anchor Pre-Matching. In *CVPR*, 7031–7040.

Xu, Y.; Zhang, M.; Fu, C.; Chen, P.; Yang, X.; Li, K.; and Xu, C. 2023. Multi-modal queried object detection in the wild. *NeurIPS*, 36: 4452–4469.

Xu, Y.; Zhang, M.; Yang, X.; and Xu, C. 2024. Exploring multi-modal contextual knowledge for open-vocabulary object detection. *IEEE Transactions on Image Processing*.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Open-vocabulary detr with conditional matching. In *ECCV*, 106–122. Springer.

Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. In *CVPR*, 14393–14402.

Zeng, H.; Xie, X.; Cui, H.; Zhao, Y.; and Ning, J. 2020. Hyperspectral image restoration via CNN denoiser prior regularized low-rank tensor recovery. *Computer Vision and Image Understanding*, 197: 103004.

Zhang, Y.; Bai, Y.; Ding, M.; and Ghanem, B. 2020. Multi-task generative adversarial network for detecting small objects in the wild. *International Journal of Computer Vision*, 128(6): 1810–1828.

Zhang, Y.; Tian, R.; Zhang, Y.; Zhang, Z.; Bai, Y.; Ding, M.; and Zuo, W. 2024. R-CCF: region-aware continual contrastive fusion for weakly supervised object detection. *Applied Intelligence*, 54(6): 4689–4712.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *CVPR*, 16793–16803.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 350–368. Springer.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2021. Probabilistic two-stage detection. arXiv:2103.07461.

Zhu, Y.; Zhang, Y.; Ding, M.; and Zuo, W. 2022. Uncertainty-aware graph-guided weakly supervised object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7): 3257–3269.