Multi-peak Graph-based Multi-instance Learning for Weakly Supervised Object Detection

RUYI JI, Institute of Software, Chinese Academy of Sciences & University of Chinese Academy of Sciences, China
ZEYU LIU, Department of Automation, China University of Petroleum, Beijing, China
LIBO ZHANG, Institute of Software, Chinese Academy of Sciences, China
JIANWEI LIU and XIN ZUO, Department of Automation, China University of Petroleum, Beijing, China
YANJUN WU and CHEN ZHAO, Institute of Software, Chinese Academy of Sciences, China
HAOFENG WANG and LIN YANG, Beijing Institute of Computer Technology and Applications, China

Weakly supervised object detection (WSOD), aiming to detect objects with only image-level annotations, has become one of the research hotspots over the past few years. Recently, much effort has been devoted to WSOD for the simple yet effective architecture and remarkable improvements have been achieved. Existing approaches using multiple-instance learning usually pay more attention to the proposals individually, ignoring relation information between proposals. Besides, to obtain pseudo-ground-truth boxes for WSOD, MIL-based methods tend to select the region with the highest confidence score and regard those with small overlap as background category, which leads to mislabeled instances. As a result, these methods suffer from mislabeling instances and lacking relations between proposals, degrading the performance of WSOD. To tackle these issues, this article introduces a multi-peak graph-based model for WSOD. Specifically, we use the instance graph to model the relations between proposals, which reinforces multiple-instance learning process. In addition, a multi-peak discovery strategy is designed to avert mislabeling instances. The proposed model is trained by stochastic gradients decent optimizer using back-propagation in an end-to-end manner. Extensive quantitative and qualitative evaluations on two publicly challenging benchmarks, PASCAL VOC 2007 and PASCAL VOC 2012, demonstrate the superiority and effectiveness of the proposed approach.

CCS Concepts: • **Computing methodologies** → **Object detection**; **Instance-based learning**;

Additional Key Words and Phrases: Weakly supervised object detection, multi-instance learning, context information, graph neural network

Ruyi Ji and Zeyu Liu are contributed equally to this research.

This work was supported by the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBS-LY-JSC038, the National Natural Science Foundation of China, Grant No. 61807033, National Key Research and Development Program of China No. 2017YFB0801900, and Tencent Youtu Lab. Libo Zhang was supported by Youth Innovation Promotion Association, CAS (2020111), and Outstanding Youth Scientist Project of ISCAS.

Authors' addresses: R. Ji, L. Zhang (corresponding author), Y. Wu, and C. Zhao, Institute of Software, Chinese Academy of Sciences, South Fourth Street, Zhong Guan Cun, Haidian, Beijing, China, 100190; emails: ruyi2017@iscas.ac.cn, libo@iscas.ac.cn, yanjun@iscas.ac.cn, zhaochen@iscas.ac.cn; Z. Liu, J. Liu, and X. Zuo, Department of Automation, China University of Petroleum, Beijing, Changping, Beijing, China, 102249; emails: logonod@gmail.com, liujw@cup.edu.cn, zuox@cup.edu.cn; H. Wang and L. Yang, Beijing Institute of Computer Technology and Applications, Yongding Road, Beijing, China, 100854; emails: wanghaofeng@sina.com, hsjyl@126.com.

1551-6857/2021/06-ART70 $15.00
https://doi.org/10.1145/3432861

ACM Reference format:
Ruyi Ji, Zeyu Liu, Libo Zhang, Jianwei Liu, Xin Zuo, Yanjun Wu, Chen Zhao, Haofeng Wang, and Lin Yang. 2021. Multi-peak Graph-based Multi-instance Learning for Weakly Supervised Object Detection. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 2s, Article 70 (June 2021), 21 pages.
https://doi.org/10.1145/3432861

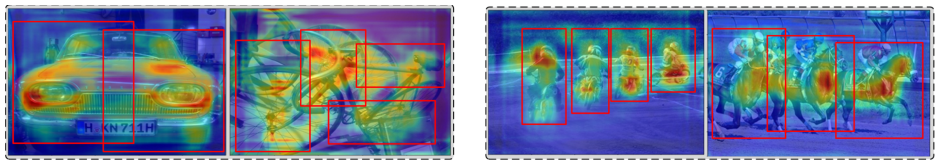
1 INTRODUCTION

Object detection is an important and fundamental task in the fields of computer vision. With the development of **convolutional neural networks (CNNs)**, the performance of object detection has been forwarded rapidly. Compared to the era of conventional hand-crafted methods, deep learning techniques, especially those based on CNNs, promote cutting-edge methods significantly for object detection [25, 48, 58, 59]. However, these methods usually work under the supervision of bounding boxes, which require large-scale manually annotated datasets [61]. The high cost of labor-intensive and time-consuming accurate bounding box annotations has hindered the wide applications of object detection technologies in real scenarios.

To circumvent the forementioned limitation, **weakly supervised object detection (WSOD)** technology, which only needs image-level annotations to train a model, has been proposed and investigated in works [2, 13, 27, 38, 46, 51, 64, 72, 90, 91]. Although many object detection methods achieve promising results, there is still a huge performance gap between WSOD and fully supervised methods due to lacking bounding box annotations.

It has been demonstrated that contextual information, or relations between object instances, could be beneficial to object recognition [8, 15, 21, 22, 36, 37, 59, 67, 75, 76]. Much effort has been devoted to the research area before the prevalence of deep learning. With the predominance of deep learning technologies, WSOD methods enter the rapid development period, but there is no significant process regarding incorporating object relations in WSOD task. Previous approaches consider object instance individually, ignoring relations between each other. One possible reason could be that relations between object instances are hard to learn. Moreover, the location, scale, and number of object instances are various across different scenarios. It is common to deploy a single neural network with a regular architecture [28, 30] for WSOD. Incorporating relations between proposals is still underexplored. Thus, we argue that incorporating the distributed discriminative information in correlated regions could be beneficial for object instance detection.

Notably, the paradigm of combining **multiple-instance learning (MIL)** with CNNs is the current mainstream. This paradigm usually tends to select the proposals with the highest confidence score from bags for object estimation, ignoring the distribution of discriminative features, regardless of the number of object instances appearing in an image. From the Figure 1, the images with the visualized activation map indicate that the complementary discriminative features scatter in multiple regions both in the cases of single instance and multiple instances. Here are two typical cases that describe this problem very well. In the first case, for the images without multiple object instances from the same bag, the target object's discriminative regions usually involve multiple proposals. The methods, which only use the highest confidence score proposal and surrounding candidate proposals, tend to mislabel discriminative regions with small overlaps as the background category. In the second case, for multiple object instances with the same class label, the previous methods tend to take proposals with lower-class scores as background, which leads to limited performance. The images containing more than one object instances for the same class are ubiquitous and natural in the challenging PASCAL VOC datasets. For instance, VOC2007 *trainval* set is formed by 7,913 image-level object labels and 15,662 annotated object instances, indicating at least 7,749 instances not selected during training. In this case, the selected



(a) Single instance. (b) Multiple instances.

Fig. 1. Visualization of the distributed response in different samples in PASCAL VOC 2007.

object instances with relatively limited scales and appearance variations, may be insufficient for training a discriminative classifier. Moreover, the missing instances may be mislabeled as background category during training, degrading the CNN-based classifier's discriminative capability.

On the whole, at least two key problems are underexplored for this line of research:

Question 1: how to model the relations between instances to enhance the performance of multiple-instance learning process;

Question 2: how to discover the multiple valid response peaks for front categories in an image and label object instances correctly.

To address the two questions mentioned above, an end-to-end multi-peak graph-based instance discovery framework is proposed to reinforce the performance for WSOD. Our approach is inspired by the success of graph neural networks [24, 62, 84]. Graph neural networks can affect every single element (e.g., an atom in the target chemistry molecular in predicting the properties of molecules and materials) by aggregating information (or features) from correlated elements (e.g., neighboring atoms in the molecular structure). Moreover, the aggregation strategy is automatically learned, driven by the predefined learning target. Dependencies between correlated elements can be mined by graph neural networks, without making additional assumptions on data distributions. Thus, we adopt the graph neural network to model the relations between object instances. WOSD's rationality is supported by two essential assumptions: (1) the top-scoring proposal along with neighbouring highly overlapped proposals are likely to have the same class label; (2) the instance with discriminative part(s) should have response peak(s) on the feature map. A multi-peak-based discovery strategy is proposed to satisfy the above two assumptions. All possible object instances present in an image can be automatically discovered effectively. Specifically, the multi-peak-based discovery strategy mines the response peaks based on multiple-instance learning and employs the *IoU* metric to model the spatial relationships between response peaks and surrounding proposals. By integrating multi-peak-based instance discovery strategy into the iterative training process, the proposed model can gradually locate object instance with image-level label. Local optima situation can be prevented by the proposed method, because more object instances from the same bag are used for training.

In principle, our approach is significantly different from the previous method [46]. Lin et al. propose **object instance mining (OIM)** using spatial similarity and appearance similarity. Though their method focuses on the relations between top-scoring object instance and surrounding candidates, there is still only one top-scoring object instance from each bag. In contrast, our proposed method can discover multiple discriminative response peaks within the same bag of instances.

To summarize, the key contributions can be highlighted into threefold:

- An instance graph based on overlap and center relations is designed to reinforce the performance of multiple-instance learning.

- A novel **multi-peak-based instance discovery (MPID)** strategy is proposed. It can effectively reduce the number of mislabeled instances.
- Extensive quantitative and qualitative evaluations demonstrate that our proposed method performs favorably against state-of-the-art methods.

The rest of this article is organized as follows. Section 2 describes the related works. The details of our proposed multi-peak object instance discovery strategy and object instance graph are elaborated in Section 3. Experimental results and discussions are given in Section 4. Section 5 gives the conclusion and future work.

2 RELATED WORK

This article is related to several research strands, including object proposals, deep networks for object detection, weakly supervised object detection, multiple-instance learning and graph neural networks.

2.1 Multiple-instance Learning

Unlike common recognition task, MIL considers training instances as bags and assigns a bag of instances to a single label. Some studies [14, 50] follow supervised learning pipeline. Others incorporate weakly supervised learning and provide a promising way to explore abundant image-level annotated data. Recently, many approaches have been proposed and remarkable improvement has been achieved. These methods can be split into three categories, including MIL with instance space methods, MIL with neural network methods, and MIL with attention mechanism.

For MIL with instance space methods, to incorporate feature extracted from bags of instances, these methods use aggregation functions to assign labels on bag level. Zhou et al. [92] explicitly construct a graph with bag labels and design graph kernel to perform binary classification. Kotzias et al. [43] propose an objective function to infer bag labels and instance-level labels based on similarity. MIRank proposed in Reference [3] reinforces nonlinear capability of loss function by novel kernel functions and shows efficacy in ranking to drug's bio-availability task. Peng et al. [56] propose a novel algorithm, aiming to predict instance-level label in MIL by designing a loss function defined at the instance level.

For MIL with neural network methods, many of them solve weakly supervised problem with MIL and define the problem learning form bags of instances, where the probability of the label is predicted by neural networks. Zhou et al. [93] present the BP-MIP algorithm, introducing neural network into multiple-instance learning, which uses the back propagation algorithm to optimize the loss function. Bilen et al. [4] deploy CNNs-based model pre-trained on large-scale datasets to facilitate regional selection and classification. Wu et al. [86] consider a dual multi-instance set, including proposals and text annotations involved in each image, and develop a deep neural model with the weakly supervised setting for MIL task. Different from Wu et al., Pinheiro et al. [57] introduce a weakly supervised framework for semantic segmentation, and design a MIL-based pooling function to classify image from pixel-level scores to avoid the necessity of using the high-cost segmentation datasets.

For MIL with attention mechanism, the insight is to apply attention mechanism to consider the weights between different instances. Ilse et al. [34] deploy a deep neural network to learn Bernoulli distribution for MIL problem. The model first encodes multiple instances in low-dimensional embedding, and explores gated attended weight embedding to predict instance-level labels. Pappas et al. [54] develop a model with MIL to predict aspect rating and deploy importance weight mechanism to balance between different sentences, which leads to different contributions

to the final aspect ratings. Kong et al. [42] integrate attention mechanism to MIL and design various attention neural networks to explore audio tagging task. Angelidis et al. [1] introduce a new dataset for evaluation and propose a method utilizing MIL network based on attention mechanism for segment classification.

To tackle the lack of localization data problem in object detection task, we follow MIL with weakly supervised pipeline to explore abundant image label data.

2.2 Object Proposals

Object proposal methods generate candidate bounding boxes from images depending on detailed information. There are plenty of methods for object proposal task. References [7, 31, 32] make detailed comparisons and discussions about object proposal methods. Current related methods can be split into super-pixel grouping method, sliding-window-based method and neural-network-based method, respectively. Super-pixel grouping methods, e.g., selective search [77] uses hierarchical grouping algorithm to provide class-agnostic and high-quality candidate boxes. However, sliding-window-based methods, e.g., Edgebox [94] utilizes a sliding window approach with coarse to fine search based on edge group method. Object proposal methods are a fundamental part of object detection as it provides a coarse level selection of bounding boxes. Neural-network-based methods, e.g., **Region Proposal Network (RPN)** [59] takes an image with arbitrary size as input and outputs object proposals of different scales, and assigns an objective score to each proposal. The RPN network and detection network share embedding features, leading to an efficient end-to-end paradigm. Wang et al. [83] design an alternative strategy, namely, Guided Anchoring, to optimize the anchoring procedure by semantic features that can jointly predict locations and centers of regions of interest.

For effectiveness and implementation convenience, we follow Reference [77] to generate regional object proposals.

2.3 Deep Networks for Object Detection

Recently, proposed methods [25, 26, 59] have shown promising results with accurate results and real-time speed. The mainstream methods vary from multiple stages to single stage pipelines. In single stage pipeline, e.g., YOLO [58], deploys a one-stage neural network and defines object detection as a regression problem to predict bounding boxes along with class probabilities, leading to super fast and real-time performance. However, in multiple stage pipelines, e.g., R-CNN [26] formulates training as a multi-stage task. R-CNN first trains a convolutional layer on object proposals. Then SVM is deployed to make predictions. Finally, bounding-box regressors are trained to predict the locations and the classes of proposal. Fast-RCNN [25] uses spatial pyramid pooling networks to get fixed-scale features from **regions of interest (RoI)** and trains the neural network using a multi-task loss in an end-to-end fashion. Faster-RCNN [59] further merges RPN and Fast R-CNN into a single network by sharing computation, which leads to near real-time frame rate object detection. He et al. [29] present a flexible framework named Mask R-CNN, which introduces an additional stream to predict object masks simultaneously. The framework is able to be trained for both object detection and instance segmentation task. Liu et al. [47] propose a method named SSD for single-shot object detection task. The model aggregates multi-scale feature maps and predicts box adjustments for bounding boxes, which brings improvements in detection speed. Deconvolutional single-shot detector [20] integrates additional deconvolutional layer to further improve SSD method with more context information. The model replaces VGGNet with ResNet-101 and adds an additional deconvolutional layer to help integrate information from earlier feature maps. Zhang et al. [89] propose a novel single-shot-based detector, called RefineDet. The model uses anchor

refinement module to select positive anchor features and uses object detection module to make predictions. He et al. [33] propose an object relation module to model relations between different objects, different from He et al., we take graph convolutional neural network to consider relations between proposals and use proposal-proposal level features to further boost weakly supervised object detection performance.

2.4 Graph Neural Network

Graphs naturally appear to model data relations in a wide range of domains, such as social media, bio-informatics, and computer vision. Graph networks can be split into five categories, including graph convolution neural network, graph attention network, graph auto-encoder network, graph generative network and graph spatial-temporal network.

Graph convolutional neural networks extend the operation of convolution from regular data to graph data. A number of works explore graph neural network to model the arbitrary relations between structured data. Some of them achieve promising results. Kipf et al. [41] present a semi-supervised learning to set with a variant of convolutional neural network. Bresson et al. [11] design graph convolutional neural networks for high-dimensional data. They explore the generalization of convolutional filters devoted to graph networks and investigate graph coarsening to model relations between data.

Graph Attention Network (GAT) [79] assumes that contributions from neighboring nodes are not equal and adopts attention mechanism to explicitly align between different nodes. Beyond single head attention, GAT also employs multiple head attention mechanism to improve model performance. Different from GAT, Gated Attention Network [88] takes different accounts for multiple heads attention.

Graph Auto-Encoders are usually unsupervised learning methods that reconstruct data relations from inner representation. The auto-encoders try to encode data in low dimensional space and disentangle representation through the data transformation process. Kipf et al. [40] use a graph auto-encoder based on variational auto-encoder to learn in embedding space. The model consists of a graph convolutional encoder and a simple inner product decoder. Berg et al. [78] consider movie ratings of recommender systems as nodes and adjacency matrix on graphs. The model shows competitive performance on collaborative filtering benchmarks. Pan et al. [52] propose a novel adversarial graph embedding framework that consists of an encoder to capture topological structure and a decoder to reconstruct graph structure.

Unlike common graph network, graph generative networks draw attention to generative tasks such as discovering structures and constructing knowledge graphs. The first generative graph model [5] considers the graph generation problem as learning distribution of random walks over the input graph. This method can be applied to a wide range of domains. GraphRNN [87] proposes a deep auto-regressive model to address graph distribution modeling problem. GraphRNN considers the problem as sequence modeling task and deploys recurrent neural network to generate adjacency vector step by step. Instead of step by step generation, MolGAN [6] predicts graph structure in one step and combines reinforcement learning objective to further improve generation performance.

Graph spatial-temporal networks build dynamic graph and capture spatial and temporal information to model the inner data pattern. DCRNN [45] uses encoder-decoder framework with multiple layers of diffusion convolutional recurrent layer, each layer first explores graph convolution network to collect spatial information and then uses recurrent network to further capture temporal information. Structural-RNN [35] combines high-level spatio-temporal graphs with sequence-to-sequence model. Structural-RNN uses graph network to encode, and feeds latent variables to both edge RNN and node RNN networks to generate features.

A lot of graph neural-network-based models utilize a supervised setting for graph relations modeling. To this end, we explore graph neural network methods for WSOD task.

2.5 Weakly Supervised Object Detection

The development of weakly supervised method alleviates the need for large scale human annotated data. For object detection, image-level data is much easier to obtain than bounding box level data. WSOD provides a promising way to explore large volumes of image-level annotations. There are bunches of studies on WSOD problem, many of them define the problem as a weakly supervised MIL task.

Many of early researches try to solve WSOD problem with discriminative models, such as SVM, CRF or Bayesian methods. Chum et al. [9] introduce an exemplar model that learns and generates regions of interest for each class instance with only image-level annotations. The model first samples regional object proposals and then uses SVM-based regional classifier to predict classes. Deselaers et al. [12] present a method to learn generic priors from meta-training data, and utilize conditional random field for WSOD task. The proposed method localizes object instances and predicts class labels at the same time. Pandey et al. [53] address WSOD problem by applying **deformable part-based models (DPM)** with latent SVM training to weakly supervised task. The model first trains DPM detectors and uses SVM to search the latent space for potential object locations. Shi et al. [66] design a novel method derived from Bayesian model and train an object classifier jointly with weakly-supervised object localization model. Song et al. [70] first find a discriminative set of regional boxes that co-occur in the labeled image dataset, and then mine positive boxes with SVM-based detector on all selective proposals. Wang et al. [82] propose the latent category learning method, which is based on **probabilistic latent semantic analysis (pLSA)** and category's discriminative model. The model first extracts candidate regions and then deploys pLSA model to do category learning. Huang et al. [60] use neural network pre-trained on large-scale data as general prior knowledge to extract high-level regional proposal features and train multiple-instance SVM model to classify on the regional proposals.

With the revolution of deep neural network, many works combine WSOD task with deep network, which leads to an end-to-end training. Cinbis et al. [10] propose a multifold MIL approach to train the detector and infer the object locations. The proposed method explores high dimensional CNN-based representations and window refinement strategy to train the model. Shi et al. [65] propose a method that uses appearance and semantic similarity to transfer source knowledge to target domain. Tang et al. [73] design a framework for the weakly supervised multiple-instance learning, regarding images as bags and patches as instances. The model first uses CNN backbone to generate convolutional features and deploys spatial pyramid pooling layer to produce fixed-size proposal features. Patches are fed to classification network and then the discovery network generates classification scores and bounding boxes. Kantorov et al. [39] aim to localize objects using context-aware-guided neural network, which follows weakly supervised setting. The model highlights the predicted object instance, distinguished from its surrounding regions. Tang et al. [71] design an iterative method to cluster object proposals by pseudo-instances and learn to detect target instances gradually with online setting. Lin et al. [46] introduce an end-to-end object instance mining method with spatial and appearance graphs to discover potential instances. The presented method is based on an assumption that the instances of similar appearance should belong to the same class and build graph to model relations between the same cluster's instances. Wan et al. [80] introduce a continuation multiple-instance learning (C-MIL) method to choose pseudo-object instances from subsets for guiding object detector training.

3 THE PROPOSED METHOD

3.1 Feature Extraction

Given a dataset $\mathcal{I}$ with C classes, including M images, we denote the dataset as $\mathcal{I} = \{(I^1, y^1), \ldots, (I^M, y^M)\}$ where I^m represents the images and $y^m = [y_1, \ldots, y_C] \in \{0, 1\}^C (m = 1, \ldots, M)$ are multi-hot annotations, indicating the presence or absence of each class in a given image. The proposal generation methods predict objective scores of candidate proposals and select out the proposals that correspond to salient regions. All candidate proposals are extracted using the selective search [77] method in our experimental setting. Formally, $P = \{p_i\}_{i=1}^N$ of an image I is a selection of candidate proposals by sliding window boxes with different sizes and aspect ratios. To enhance the robustness of the proposed model, we follow the paradigm of transfer learning. The image is fed into the backbone network (a truncated VGG16 network) to extract features. RoI pooling operation takes extracted features and selected proposals as input, and outputs fixed-size feature for each proposal. Notably, many CNN-based pre-trained networks could be chosen as alternates, such as ResNet-101 and Inception.

3.2 Object Instance Graph

In this section, we aim to deal with ***Question 1***. Specifically, we employ graph neural network to model the relations between proposals of the same bag and facilitate multiple-instance learning process.

It is well acknowledged that shallow layers of CNN-based networks focus on the detailed information, such as edges or small parts while deep layers pay more attention to generating semantic information. Thus, we perform the **global average pooling (GAP)** on features extracted from the backbone to obtain the global context information and fuse the global context information into features of each proposal by the concatenation operation in a skip-connection fashion. The fusion operation provides additional clues for object detection. Thus, we use $V = \{v_i \in \mathbb{R}^d\}_{i=1}^N$ to denote the set of features from proposals and $|V| = N$.

In this article, we design an object instance graph method to model interrelationships between different object instances and apply graph convolutional network to learn the data representation for object instances. Formally, let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a graph built on object instances, where $\mathcal{V}$ denotes the set of nodes $v_i \in \mathcal{V}$ and $|\mathcal{V}| = N$. $\mathcal{E}$ denotes the set of edges $e_{ij} = (v_i, v_j) \in \mathcal{E}$. $A \in \mathbb{R}^{N \times N}$ denotes the adjacency matrix of $\mathcal{G}$. The idea behind the graph built on regional proposals is to model interactions between correlated object instances with high overlaps or near distances. In our case, each proposal presents a node and the edges in $\mathcal{E}$ present relations between correlated proposals. The reason for applying graph neural network is that nodes within the graph can perform the message passing with neighboring nodes in the local regions. In this way, each proposal aggregates context information from correlated proposals of the same bag, boosting the multiple-instance learning and detection performance eventually.

For building the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, a straightforward and intuitive way to construct edges is to connect all nodes with each other. The disadvantages are obvious, as too much message passing will lead to unbearable computational costs. Moreover, unrelated proposals result in redundant or noisy information in message aggregation. For our work, a novel approach is used to model relations between proposals by exploring spatial and appearance relevance. Specifically, we introduce two types of edges to mine the correlated proposals, i.e., the overlap edges and the center edges, respectively. The appearance relevance is used to make quantitative analysis in similarity between correlated proposals.

Overlap Edges. To establish overlap edges, we first sort regional proposals based on Equation (1) for each proposal p_j. And then we select n_o regional proposals as nearest neighbors for

each node. Here, $IoU(p_i, p_j)$ denotes the overlap relevance between proposals p_i and p_j, which is formulated as follows:

$$IoU(p_i, p_j) = \frac{I(p_i, p_j)}{U(p_i, p_j)}, \tag{1}$$

where $I(p_i, p_j)$ and $U(p_i, p_j)$ represent the intersection and union of the two proposals, respectively. In terms of the proposal p_i, the top k overlap proposals will be selected as its neighborhoods, as they are likely to add complementary features. Exploring the additional contextual information will help network to refine the detection process and enhance performance [15]. Hence, we exploit graph convolutional network to mine all complementary information.

Center Edges. The distinct but center nearby proposals could also be complementary to each other, and the message passing between correlated proposals will facilitate the multiple-instance learning. This kind of neighborhoods is also beneficial to collecting contextual information from nearby object instances. To handle such kinds of relations, we introduce center edges, which could be computed with the following distance formula:

$$d(p_i, p_j) = \frac{|c_i - c_j|}{U(p_i, p_j)}. \tag{2}$$

Similar to overlap edge, we first sort regional proposals based on Equation (2) and then select n_d regional proposals as neighbors for each node. In Equation (2), c_i (or c_j) represents the center coordinate of p_i (or p_j). As a complement of overlap edges, the center edges enable the feature aggregation from distinct but related instances.

Appearance Relevance. After determining relationships between proposals, instead of treating every node equally for feature aggregation, we introduce appearance relevance to compute values of relations between object instances and build an asymmetric adjacency matrix. Specifically, the cosine function is leveraged to estimate $S_{ij} = cosine(x_i, x_j)$, where S_{ij} denotes the similarity between node v_i and node v_j, x_i and x_j denote the feature vectors of proposals. To emphasize the central role of node v_i, we assign different weights to edges connected to v_i. For edge $e_{ij} = (v_i, v_i) \in \mathcal{E}$, we apply the weight w_{ij} of 1 while for other edge $e_{ij} = (v_i, v_j) \in \mathcal{E}$, where $i \neq j$ the weight w_{ij} is $\frac{1}{N}$, where N is the degree of node v_i. Finally, the adjacency matrix $A_{i,j}$ is computed by the element-wise multiplication of S_{ij} and w_{ij}.

After constructing the adjacency matrix, we build a K-layer graph convolution network. The proposed graph-based model is formulated as follows:

$$V^k = AV^{(k-1)}W^k, \tag{3}$$

where A represents the adjacency matrix, and $W^{(k)} \in \mathbb{R}^{d \times d}$ is the learnable parameters in kth layer. $V^{(k)} \in \mathbb{R}^{n \times d}$ is the output of graph convolution operation in the kth layer; the graph convolution layer consists of one linear layer followed by one ReLU layer.

3.3 Multiple-instance Learning

The purpose of the multiple-instances learning stream is to generate proposal confidence scores for each class $\{c_i\}_{i=1}^{C}$ from the outputs V^k of object instance graph network. Following Reference [4], the graph-based proposal features are split into two branches to output two matrices $\mathbf{x}^c, \mathbf{x}^d \in \mathbb{R}^{C \times N}$ by two fully connected layers, where C denotes the number of classes. Then, we apply softmax operations along two different dimensions to the matrices

$$\{\sigma(\mathbf{x}^c)\}_{ij} = \frac{e^{x_{ij}^c}}{\sum_{k=1}^{C} e^{x_{kj}^c}}, \{\sigma(\mathbf{x}^d)\}_{ij} = \frac{e^{x_{ij}^d}}{\sum_{k=1}^{N} e^{x_{ik}^d}}. \tag{4}$$

The first stream contributes to the classification score of each proposal, while the second stream computes contribution of proposals to multiple-instance learning. The element-wise Hadamard product is used to compute the final score of each proposal:

$$\mathbf{x}^r = \sigma(\mathbf{x}^c)\sigma(\mathbf{x}^d) \in \mathbb{R}^{C\times N}. \tag{5}$$

Finally, the classification scores $\{\phi_c\}_{c=1}^{C}$ for each category are obtained by summing $\mathbf{x}^r$ along proposal dimension. We employ multi-class cross entropy loss to train the instance classifier, defined as

$$\mathcal{L}_{MIL} = -\sum_{c=1}^{C} y_c \log\phi_c + (1-y_c)\log(1-\phi_c), \tag{6}$$

where $y_c \in [0,1]$ indicates if the image contains any instance of class c in the image.

3.4 Multi-Peak-based Instance Discovery

To deal with ***Question 2***, we propose a MPID strategy. Here, we give the definition of peak as the most discriminative region of feature map, which is important to characterize the target. Specifically, we employ the following the two criteria to mine the response peaks for front categories:
Criteria 1: The peak in the image should have a high response for the specific front category.
Criteria 2: There should be small overlaps between different peaks.
Different from the previous methods, top k scoring proposals are selected for each front category. We apply the threshold λ_s to the selected top k proposals to satisfy the criteria 1,

$$topk\_score > top0\_score - \lambda_s, \tag{7}$$

ALGORITHM 1: Multi-peak-based Object Instance Discovery

Input: Image I, regions proposals $P = \{p_1, \ldots, p_N\}$, image label $Y = \{y_1, y_2, \ldots, y_c\}$
Output:

```
Feed image I along with its proposals into the network and output feature vectors F = {f_1, ..., f_N}
Compute topk_score, the top K scoring proposals for each class
for c in C, C denotes the number of training data category do
    for i = 1 to K do
        if topk_score[i, c] > topk_score[0, c] − λ_s then
            Select the box p_{c_0} for top 0 proposal
            Select the box p_{c_i} for top i proposal
            Compute IoU(p_{c_0}, p_{c_i})
            if overlap < λ_o then
                Set p_{c_i} to be a valid peak for c class
            end if
        end if
    end for
end for
for j = 1 to N do
    Compute the overlaps between p_j and p_{peaks}
    Sort (descent) the overlaps
    Set the label of p_j based on the max overlap
end for
```

and the proposals with overlap larger than the threshold λ_o should be filtered out, because a proposal that has a large overlap with the chosen peaks is not likely to be a new peak. The criteria 2 can be formulated as follows:

$$overlaps(peak_i, peak_j) < \lambda_o. \tag{8}$$

We present the statement about the loss function for MPID refinement streams here. Suppose an image with label Y and predicted label $Y_j = [y_{0,j}, y_{1,j}, \ldots, y_{C,j}]^T \in \mathbb{R}^{(C+1)\times 1}$ for the jth proposal, where $y_{c,j} = 1$ or 0 denotes the proposal belonging to class c or not, and background class is indicated by the index of $c = 0$. $x_{c,j}$ with class label c, are the proposals used for training. For each MPID refinement stream, we use confidence score ω_j as loss weight for jth proposal, which calculated from max correlated class scores. We apply negative log-likelihood loss to optimize the instance classifier, defined as

$$\mathcal{L}^k_{MPID} = -\frac{1}{|P|}\sum_{j=1}^{|P|}\sum_{c=1}^{C+1}\omega_j y_{c,j}\log x_{c,j}. \tag{9}$$

3.5 End-to-End Training

The loss for proposed method consists of two parts, i.e., losses for MIL and MPID. The overall loss can be formulated as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_{MIL} + \sum_{k=1}^{K}\mathcal{L}^k_{MPID}, \tag{10}$$

where $\mathcal{L}_{MIL}$ denotes the multi-class cross entropy loss for the MIL stream, and $\mathcal{L}^k_{MPID}$ denotes the negative log-likelihood loss for the kth MPID refinement stream.

4 EXPERIMENTS

4.1 Datasets

Following References [2, 13, 27, 38, 46, 51, 64, 72, 90, 91], we evaluate the proposed method on two public challenging datasets: the PASCAL VOC 2007 [17] and PASCAL VOC 2012 [18] datasets. The PASCAL VOC 2007 dataset defines 20 categories, including 9,963 images and 24,640 annotated objects. The PASCAL VOC 2012 dataset is formed by 22,531 images with publicly available annotations, divided into 20 classes. Two datasets share the same annotation categories, e.g., aeroplane, bicycle. For fair comparison, we follow the configuration used in Reference [71] to preprocess data. Specifically, datasets are divided into *train*, *val*, and *test* parts. The *trainval* set consists of 5,011 images for VOC 2007 dataset and 11,540 images for VOC 2012 dataset. As we explore weakly supervised methods, only image-level annotations are involved in our experimental setting.

4.2 Evaluation Metrics

We follow common evaluation criteria [16] to evaluate detection performance. For *trainval* set, we use **correct localization (CorLoc)** metric to evaluate detection performance. And for *test* set, we utilize **average precision (AP)** [16] and **mean average precision (mAP)** to measure localization accuracy. The former metric provides a measure of how well the detector adapts to all instances, while the latter metric indicates whether the detection is a good match.

4.3 Implementation Details

The proposed method is implemented in PyTorch [55] deep learning library. All experiments are carried out on a server with a 3.26 GHz Intel processor, 32 GB memory and one Nvidia Tesla V100 GPU. Following the previous works, the VGG16 [68] model pre-trained on ImageNet classification

Table 1. Comparison with the State-of-the-art Methods in Terms of Detection Performance AP(%) on the PASCAL VOC 2007 Test Set

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	soft	train	tv	mAP
WSDDN [4]	46.4	58.3	35.5	25.9	14.0	66.7	53.0	39.2	8.9	41.8	26.6	38.6	44.7	59.0	10.8	17.3	40.7	49.6	56.9	50.8	39.2
OICR [72]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
WCCN [13]	49.5	60.6	38.6	29.2	16.2	70.8	56.9	42.5	10.9	44.1	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	60.8	54.5	42.8
PCL [71]	54.4	69.0	39.3	19.2	15.7	62.9	64.4	30.0	25.1	52.5	44.4	19.6	39.3	67.7	17.8	22.9	46.6	57.5	58.6	63.0	43.5
TS^2C [85]	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
WeakRPN [74]	57.9	70.5	37.8	5.7	21.0	66.1	69.2	59.4	3.4	57.1	**57.3**	35.2	**64.2**	68.6	**32.8**	**28.6**	50.8	49.5	41.1	30.0	45.3
C-WSL* [23]	62.9	64.8	39.8	28.1	16.4	69.5	68.2	47.0	**27.9**	55.8	43.7	31.2	43.8	65.0	10.9	26.1	52.7	55.3	60.2	66.6	46.8
MELM [81]	55.6	66.9	34.2	29.1	16.4	68.8	68.1	43.0	25.0	65.6	45.3	53.2	49.6	68.6	2.0	25.4	52.5	56.8	62.1	57.1	47.3
OICR+W-RPN [69]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.9
SDCN [44]	59.8	67.1	32.0	**34.7**	22.8	67.1	63.8	67.9	22.5	48.9	47.8	**60.5**	51.7	65.2	11.8	20.6	42.1	54.7	60.8	64.3	48.3
WS-JDS [63]	52.0	64.5	45.5	26.7	**27.9**	60.5	47.8	59.7	13.0	50.4	46.4	56.3	49.6	60.7	25.4	28.2	50.0	51.4	66.5	29.7	45.6
Boosted-OICR [19]	**68.6**	62.4	**55.5**	27.2	21.4	**71.1**	**71.6**	56.7	24.7	60.3	47.4	56.1	46.4	69.2	2.7	22.9	41.5	47.7	**71.1**	69.8	49.7
OIM [46]	62.2	67.2	48.0	29.6	23.5	68.7	69.3	64.3	22.8	59.6	39.6	30.7	42.7	69.8	3.1	23.3	**57.9**	55.4	63.4	63.5	48.2
OIM+IR [46]	55.6	67.0	45.8	27.9	21.1	69.0	68.3	**70.5**	21.3	60.2	40.3	54.5	56.5	70.1	12.5	25.0	52.9	55.2	65.0	63.7	50.1
Ours	66.4	**71.1**	48.4	29.6	20.5	70.6	67.6	67.1	23.6	**66.1**	45.1	56.0	47.6	**70.2**	1.3	23.5	48.4	51.6	64.1	**70.0**	**50.4**
C-WSL* +FRCNN [23]	62.9	68.3	52.9	25.8	16.5	71.1	69.5	48.2	26.0	58.6	44.5	28.2	49.6	66.4	10.2	26.4	55.3	59.9	61.6	62.2	48.2
SDCN+FRCNN [44]	61.1	70.6	40.2	**32.8**	23.9	63.4	68.9	68.2	18.3	60.2	**53.5**	63.6	53.6	66.1	14.6	21.8	50.5	56.7	62.4	67.9	51.0
WS-JDS+FRCNN [63]	64.8	70.7	51.5	25.1	**29.0**	74.1	69.7	69.6	12.7	69.5	43.9	54.9	39.3	**71.3**	**32.6**	**29.8**	57.0	**61.0**	66.6	57.4	52.5
Pred Net(FRCNN) [2]	66.7	69.5	52.8	31.4	24.7	**74.5**	**74.1**	67.3	14.6	53.0	46.1	52.9	**69.9**	70.8	18.5	28.4	54.6	60.7	67.1	60.4	52.9
Boosted-OICR+FRCNN [19]	65.8	58.6	**55.0**	32.4	19.5	74.2	71.4	70.9	19.2	54.8	46.2	**67.5**	57.0	65.6	1.4	16.7	40.4	53.0	69.5	61.1	50.0
OIM+FRCNN [46]	53.4	**72.0**	51.4	26.0	27.7	69.8	69.7	**74.8**	21.4	67.1	45.7	63.7	63.7	67.4	10.9	25.3	53.5	60.4	**70.8**	58.1	52.6
Ours+FRCNN	**68.2**	68.3	49.4	32.1	28.3	67.7	69.1	71.3	**27.3**	**70.0**	46.0	65.7	55.0	68.1	2.1	28.7	**57.7**	55.2	65.0	**64.5**	**53.0**

dataset [61] is truncated from *conv1* to *conv5* as backbone. We initialize model parameters from norm distribution $\mathcal{N}(\mu, \sigma^2)$. We set $\mu = 0, \sigma = 0.01$ experimentally. The training data is augmented by strategy of normalization, random flipping and re-sizing. During training, stochastic gradient descent optimizer is used to train proposed model with momentum of 0.9, weight decay of 0.0005. The learning rates are initialized to 0.0005 and 0.0001 for main iterations and the warm-up iterations, respectively. We train each network with 30K and 90K iterations for the PASCAL VOC 2007, 2012 datasets, respectively, where warm-up period is included to bootstrap learning process. To reduce redundancy, we apply non-maximum suppression to filter out duplicated boxes. For comparison, a fully supervised Fast-RCNN [25] detection network is trained, where weakly supervised bounding boxes are regarded as pseudo-ground truths. We adopt the same data augmentation strategy in works [23, 46, 71, 72], and set threshold to be 0.3 for confidence score and 0.3 in terms of *IoU* to choose the regional boxes.

4.4 Results and Evaluations

In this section, we compare our method with state-of-the-art methods for WSOD and present the visualized results for qualitative analysis. The comparisons between the proposed method and state-of-the-art methods on the PASCAL VOC 2007 and 2012 datasets are shown in Table 1, Table 2, Table 3, and Table 4.

As presented in Table 1, we can conclude that the proposed method achieves the best method improves the original OICR [72] by 9.2% and outperforms other approaches such as WSDDN [4] (by 11.2%), PCL [71] (by 6.9%), WeakRPN [74] (by 5.1%), SDCN [44] (by 2.1%), Boosted OICR [19] (by 0.7%), and WS-JDS [63] (by 4.8%). Besides, our method presents the best performance in some categories, which shows the effectiveness of our graph-based network with multi-peak strategy. In addition to weakly supervised training, we also re-trained a Fast-RCNN detector using the learned pseudo-objects as ground-truth. It is clear that the results are boosted further, which achieves 53.0% mAP, as shown in Table 1.

Table 2. Comparison with the State-of-the-art Methods in Terms of Localization Performance (%) on the PASCAL VOC 2007 Trainval Set

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	soft	train	tv	mAP
WSDDN [4]	65.1	58.8	58.5	33.1	39.8	68.3	**60.2**	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
TST [65]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	59.5
OICR [72]	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	81.4	60.6
WCCN [13]	83.9	72.8	64.5	44.1	40.1	65.7	82.5	58.9	33.7	72.5	25.6	53.7	67.4	77.4	26.8	49.1	68.1	27.9	64.5	55.7	56.7
TS^2C [85]	84.2	74.1	61.3	52.1	32.1	76.7	82.9	66.6	42.3	70.6	39.5	57.0	61.2	88.4	9.3	54.6	72.2	60.0	65.0	70.3	61.0
C-WSL [23]	85.8	81.2	64.9	50.5	32.1	**84.3**	85.9	54.7	43.4	80.1	42.2	42.6	60.5	90.4	13.7	57.5	**82.5**	61.8	74.1	82.4	63.5
WeakRPN [74]	77.5	81.2	55.3	19.7	44.3	80.2	86.6	69.5	10.1	**87.7**	**68.4**	52.1	**84.4**	**91.6**	**57.4**	**63.4**	77.3	58.1	57.0	53.8	63.8
WS-JDS [63]	82.9	74.0	**73.4**	47.1	**60.9**	80.4	77.5	**78.8**	18.6	70.0	56.7	67.0	64.5	84.0	47.0	50.1	71.9	57.6	**83.3**	43.5	64.5
PCL [71]	79.6	**85.5**	62.2	47.9	37.0	83.8	83.4	43.0	38.3	80.1	50.6	30.9	57.8	90.8	27.0	58.2	75.3	**68.5**	75.7	78.9	62.7
MELM [81]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.4
Boosted-OICR [19]	86.7	73.3	72.4	55.3	46.9	83.2	**87.5**	64.5	44.6	76.7	46.4	**70.9**	67.0	88.0	9.6	56.4	69.1	52.4	79.8	82.8	65.7
C-MIL [80]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.0
OIM+IR [46]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**67.2**
Ours	**86.7**	80.4	66.1	52.7	47.3	81.7	83.4	74.4	**45.6**	82.2	49.4	68.6	72.8	88.7	73.0	57.9	76.3	60.5	75.3	**84.2**	67.1

Table 3. Comparison with the State-of-the-art Methods in Terms of Detection Performance (%) on the PASCAL VOC 2012 Test Set

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	soft	train	tv	mAP
WSDDN [4]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.9
WSDDN+context [4]	64.0	54.9	36.4	8.1	12.6	53.1	40.5	28.4	6.6	35.3	**34.4**	49.1	42.6	62.4	**19.8**	15.2	27.0	33.1	33.0	50.0	35.3
OICR [72]	67.7	61.2	41.5	25.6	22.2	54.6	49.7	25.4	19.9	47.0	18.1	26.0	38.9	67.7	2.0	22.6	41.1	34.3	37.9	55.3	37.9
WCCN [13]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.9
PCL [71]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.6
TS^2C [85]	67.4	57.0	37.7	23.7	15.2	56.9	49.1	**64.8**	15.1	39.4	19.3	48.4	44.5	67.2	2.1	23.3	35.1	40.2	46.6	45.8	40.0
WeakRPN [74]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.8
C-WSL* [23]	**74.0**	67.3	45.6	29.2	26.8	**62.5**	54.8	21.5	**22.6**	50.6	24.7	25.6	57.4	71.0	2.4	22.8	44.5	**44.2**	45.2	**66.9**	43.0
MELM [81]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	42.4
OICR+W-RPN [69]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	43.2
SDCN [44]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	43.5
WS-JDS [63]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.1
C-MIL [80]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.7
OIM+IR [46]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	44.4
Ours	73.1	**70.0**	**53.4**	**33.2**	**29.3**	59.6	**59.4**	41.7	19.8	**60.0**	23.4	**51.2**	**63.4**	**72.2**	4.3	**26.2**	**53.4**	25.6	**55.9**	62.8	**46.9**
C-WSL*+FRCNN [23]	**75.3**	**71.6**	**52.6**	32.5	29.9	**62.9**	**56.9**	16.9	24.5	**59.0**	**28.9**	27.6	**65.4**	**72.6**	1.4	23.0	49.4	**52.3**	42.4	**62.2**	45.4
SDCN+FRCNN [44]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.7
WS-JDS+FRCNN [63]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.1
Pred Net(FRCNN) [2]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	48.4
C-MIL+FRCNN [80]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.7
OIM+FRCNN [46]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.4
Ours+FRCNN	62.7	58.5	50.8	**35.1**	**32.7**	59.0	55.1	**67.4**	**27.2**	58.3	27.9	**65.4**	62.9	71.3	**5.1**	**26.0**	**57.6**	37.2	**60.4**	49.3	**48.5**

The proposed method achieves the comparable performance with state-of-the-art performance on the Pascal VOC 2007 *train* − *val* set, as shown in Table 2. Specifically, the proposed method outperforms OICR [72] (by 6.5%), WSDDN [4] (by 13.6%), TST [65] (by 7.6%), WeakRPN [74] (by 3.3%), PCL [71] (by 4.4%), MELM [81] (by 5.7%), Boosted-OICR [19] (by 1.4%), and C-MIL [80] 449 (by 2.1%). It is well verified that multi-peak strategy is a feasible way to deal with the issue of mislabeling on multiple instances in the same bag, which enhances the final detection performance. Our method achieves a comparable performance with OIM [46], i.e., 67.1 vs. 67.2. As we use graph neural network to aggregate complementary information from neighboring proposals, our method presents the best CorLoc results in some classes, which indicates graph-based network with multi-peak strategy can facilitate accurate result.

Furthermore, the detection and localization performances on PASCAL VOC 2012 dataset are reported in Tables 3 and 4. Our method presents a competitive Corloc performance in PASCAL

Table 4. Comparison with the State-of-the-art Methods in Terms of Localization Performance (%) on the PASCAL VOC 2012 Trainval Set

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	soft	train	tv	CorLoc
WSDDN+context [4]	78.3	70.8	52.5	34.7	36.6	80.0	58.7	38.6	27.7	71.2	32.3	48.7	76.2	77.4	16.0	48.4	69.9	47.5	66.9	62.9	54.8
OICR [72]	86.2	**84.2**	68.7	55.4	46.5	82.8	74.9	32.2	46.7	82.8	42.9	41.0	68.1	89.6	9.2	53.9	**81.0**	52.9	59.5	83.2	62.1
PCL [71]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.2
TS^{2}C [85]	79.1	83.9	64.6	50.6	37.8	87.4	74.0	**74.1**	40.4	80.6	42.6	53.6	66.5	88.8	**18.8**	54.9	80.4	60.4	70.7	79.3	64.4
WeakRPN [74]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.9
C-WSL* [23]	86.6	80.8	**73.9**	43.2	44.4	87.7	76.2	32.2	34.0	**87.1**	**49.1**	46.2	**88.2**	**91.2**	12.1	57.1	78.4	**65.5**	65.1	**85.3**	64.2
MELM [81]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.4
OICR+W-RPN [69]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**67.5**
SDCN [44]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**67.9**
WS-JDS [63]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.5
Boosted-OICR [19]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	66.3
C-MIL [80]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**67.4**
OIM+IR [46]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.1
Ours	**87.6**	83.7	72.7	**59.7**	**53.0**	**88.6**	**77.1**	49.1	**48.7**	84.5	48.8	**63.8**	85.3	90.7	13.8	**59.9**	80.7	41.6	**73.6**	84.9	**67.4**

VOC 2012 trainval set, outperforming OICR [72] (by 5.3%), WSDDN [4] (by 12.6%), TS^{2}C [85] (by 3.0%), WeakRPN [74] (by 2.5%), PCL [71] (by 4.2%), MELM [81] (by 6.0%), Boosted-OICR [19] (by 1.1%), and OIM [46] (by 0.3%). Regarding the mAP metric on Pascal VOC 2012 test set, the proposed method achieves the best mAP result,[1] which outperforms OICR [72] (by 9.0%), WSDDN [4] (by 9.0%), TS^{2}C [85] (by 6.9%), WeakRPN [74] (by 6.1%), PCL [71] (by 6.3%), MELM [81] (by 4.5%), C-MIL [80] (by 0.2%), and OIM [46] (by 2.5%). The proposed method achieves the best AP results in a lot of categories.

Figure 3 depicts some qualitative visualized results generated by the proposed method. It is well observed that our method is robust to variations of the size and scale, especially for rigid objects. Besides, our model is capable to deal with localization in different scenes, for example, multiple instances from the same bag co-existing in an image or various objects from different classes in relatively complicated scenarios. We also analyze some failure cases (Figure 3, last column), which can be roughly divided into two groups: (1) the most common failure for our model is that partial instance (e.g., person face, horse body, chair top) is easily localized, compared with the entire object. Reasons behind this can be explained that partial regions with less variable appearance such as "face" are more distinguishable than the other parts of the object instance. This is because we define the object regions as the most discriminative parts of the instance instead of the whole object; (2) instance overlaps confuse the detector when it predicts the multiple potted plants on the grass (Figure 3 third row rightmost subfigure). The model locates overlapped boxes that not only encircle one object but also include multiple same class instances as a big bounding box. Thus, our detector tends to regard them as a whole object instead of individual instances.

4.5 Ablation Study

We conduct ablation studies on the PASCAL VOC 2007 dataset to analyze the influence of important parameters and different modules in the proposed methods. Specifically, the effectiveness of global context information, object instance graph, multi-peak-based instance discovery strategy and the influence of super-parameters will be discussed in detail.

Global context information. It is well recognized that modeling the global contextual representations can obtain richer local and non-local information of target objects [49]. As can be seen in Figure 1 a proposal with road background is helpful to predict vehicle classes. To verify

[1]We submmit our results for VOC 2012 to the evaluation server, the anonymous result link is http://host.robots.ox.ac.uk:8080/anonymous/LYGIGK.html.

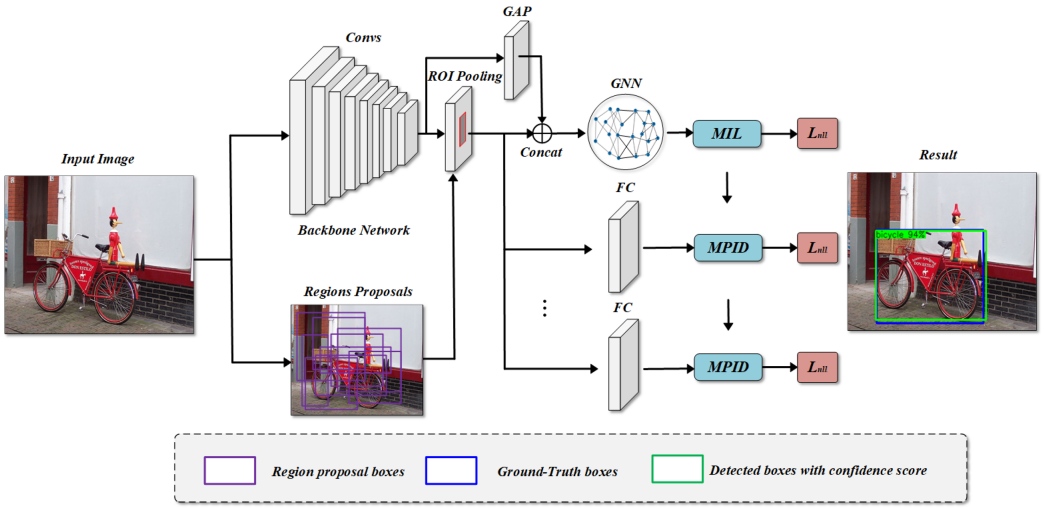


Fig. 2. The overview of our proposed model architecture, formed by (a) backbone network for feature extraction, (b) multiple-instance detection stream with graph neural network, (c) multi-peak-based instance discovery refinement streams. Best visualization in color.

Table 5. Ablation Experiments of Detection Performance (mAP%) and Localization Performance (CorLoc%) on the VOC2007 for Using Different Combinations of the Proposed Method

Method	CorLoc (%)	mAP (%)
Ours w/o GAP	62.4	45.6
Ours w/o graph	64.8	48.3
Ours w/o MPID	66.2	49.8
Ours	67.1	50.4

the effectiveness of global context information in the proposed method, we construct a variant of the proposed method, i.e., "ours w/o GAP." As shown in Figure 2, "ours w/o GAP" indicates that we construct instance graph based on proposals directly, without combination with the global context information. As shown in Table 5, there is a sharp decrease of mAP score, i.e., 50.4% vs. 45.6%. We can conclude that it is beneficial to achieve the accurate result based on global context information.

Pooling strategy. Moreover, we also analyze the influence of the pooling strategy to get global context information in Table 6. We can figure out that incorporating the global max-pooling information instead of the GAP information, resulting in 3.4% performance drop in terms of mAP score. We argue that the GAP operation facilitates the convolutional filter focusing on high average response patches instead of only the outstanding features, which is able to integrate more global context information for better results. It is essential to locate object instance and predict the class of instance based on global context information.

Object instance graph. To verify the effect of object instance graph in our method, we remove the object instance graph from our proposed method and term it as "ours w/o graph." As

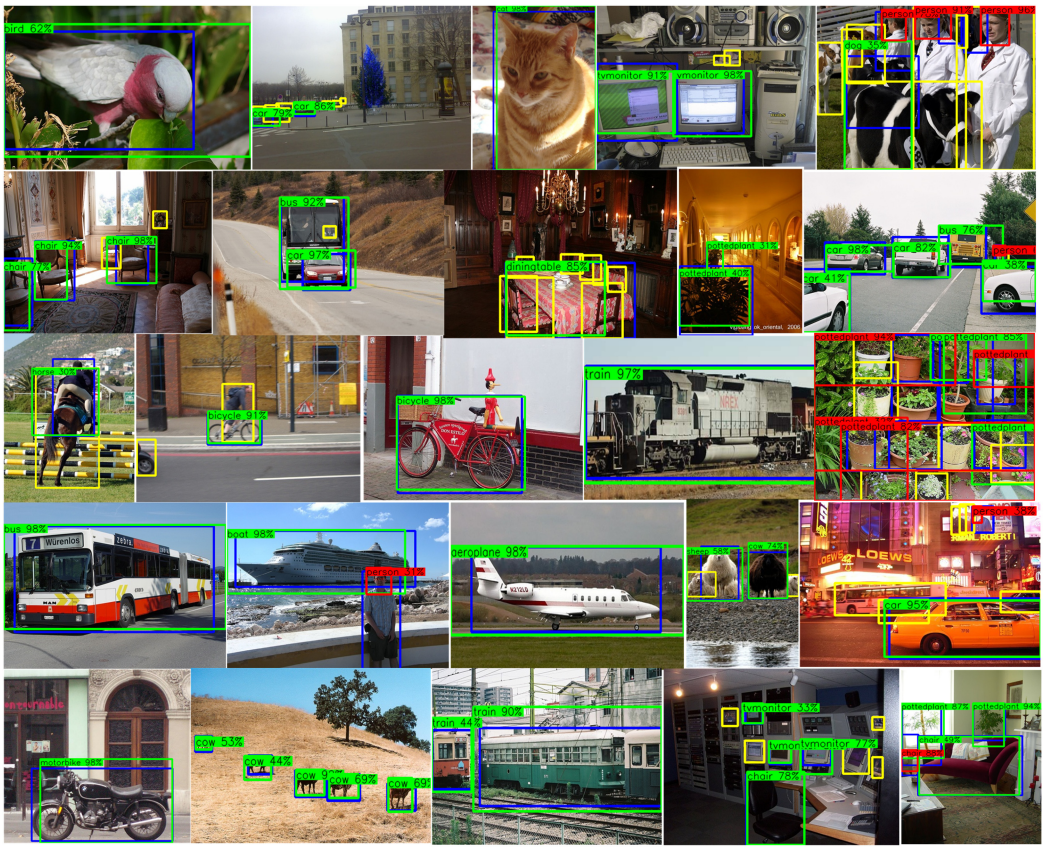


Fig. 3. Detection examples for Pascal VOC 2007 dataset. Blue rectangles are ground-truth boxes that have at least one detection with $IoU > 0$, and yellow ones are ground-truth with no detection intersection. Green boxes are correctly detected ($IoU > 0.5$ with ground truth), and red boxes are wrongly detected. The label in each detection box is the class label and confidence score of the detection.

Table 6. Ablation Experiments of Detection Performance (mAP%) and Localization Performance (CorLoc%) on the VOC2007 for Using Different Pooling Strategy

Pooling	CorLoc (%)	mAP (%)
GAP	67.1	50.4
GMP	65.4	47.0

presented in Table 5, we can observe that object instance graph can bring 2.1% mAP improvement. We attribute this improvement to the capability of modeling relations between instances, as instance graph can mine the relations between correlated object instances. The aggregation of correlated feature can better characterize the target and provide a more discriminative feature for multiple-instance learning. Our graph-based model confirms the benefit of graph-based MIL for WSOD task.

Table 7. The Influence of Related Super-parameters

Parameter	2	4	8
n_o	49.2	50.4	50.1

(a) The influence of n_o.

Parameter	0	2	8
n_d	49.0	50.4	46.6

(b) The influence of n_d.

Parameter	0.1	0.15	0.2
λ_o	50.4	48.9	48.3

(c) The influence of λ_o.

Parameter	0.1	0.15	0.2
λ_s	50.4	47.3	46.8

(d) The effectiveness of λ_s.

Multi-peak strategy. As illustrated in Section 3.4, we design a multi-peak-based instance discovery strategy to avoid mislabeling object instance into background category. To demonstrate the influence of multi-peak based instance discovery strategy, we construct a variant of the proposed method by abandoning the multi-peak based instance discovery strategy, i.e., "ours w/o MPID." From the Table 5, it can be seen that the mAP score of "ours w/o MPID" on PASCAL VOC 2007 test set drops 0.6% compared to the proposed method. We think degradation is caused by the mislabeled instances in the training process, which demonstrates the effectiveness of multi-peak-based instance discovery strategy.

The influence of super-parameters. We choose empirical values of n_o and n_d for each node to control numbers of overlap edges and center edges, as illustrated in Section 3.2. Increasing the number of center edges and overlap edges leads to a large volume of correlated neighborhoods, introducing the noisy information and redundant computation while decreasing the values of n_o and n_d makes a sharp decrease of number of edges within graph, which simplifies the relationships between object instances and leads to limited performance. In a summary, an appropriate number of edges improves the detection performance. Numerous neighborhoods will lead to overwhelming computation and do harm to model performance while too small values of n_o and n_d lead to too simple graph structure, which degrades the performance.

Furthermore, we also study the influence of λ_o and λ_s for multi-peak-based instance discovery strategy. As demonstrated in Section 3.4, we use the threshold λ_s to filter out instances with a low confidence score and employ the threshold λ_o to select proposals with a small overlap. The influence of different λ values is presented in Table 7. Higher values of λ lead to loosed criteria for selecting multiple peaks, too much noisy information involved, resulting in limited performance, while smaller values of λ lead to valid peaks filtered out, decreasing the volumes of object instance for training, which impacts the final performance.

5 CONCLUSION

In this article, we propose two improvements to boost the weakly supervised object detection. First, we propose a multi-peak-based instance discovery methodology that discovers multiple peaks from the same bag. Second, we propose a graph-based MIL method that aggregates information from correlated neighborhoods within the same bag of instances. Extensive quantitative and qualitative evaluations on two public challenging datasets, PASCAL VOC 2007 and 2012, demonstrate the superiority and effectiveness of the proposed method. In the future, we plan to optimize the multi-peak graph-based model for better performance and explore RPN-like networks as an alternative way to the selective search method.

REFERENCES

[1] Stefanos Angelidis and Mirella Lapata. 2018. Multiple-instance learning networks for fine-grained sentiment analysis. *Trans. Assoc. Comput. Linguist.* 6 (2018), 17–31. DOI: https://doi.org/10.1162/tacl_a_00002

[2] Aditya Arun, C. V. Jawahar, and M. Pawan Kumar. 2018. Dissimilarity coefficient-based weakly supervised object detection. Retrieved from http://arxiv.org/abs/1811.10016.

[3] C. Bergeron, G. Moore, J. Zaretzki, C. M. Breneman, and K. P. Bennett. 2012. Fast bundle algorithm for multiple-instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 6 (2012), 1068–1079.

[4] Hakan Bilen and Andrea Vedaldi. 2015. Weakly supervised deep detection networks. Retrieved from http://arxiv.org/abs/1511.02853.

[5] Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. 2018. NetGAN: Generating graphs via random walks. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 610–619. Retrieved from http://proceedings.mlr.press/v80/bojchevski18a.html.

[6] Nicola De Cao and Thomas Kipf. 2018. MolGAN: An implicit generative model for small molecular graphs. Retrieved from https://abs/1805.11973.

[7] Neelima Chavali, Harsh Agrawal, Aroma Mahendru, and Dhruv Batra. 2015. Object-proposal evaluation protocol is "gameable." Retrieved from http://arxiv.org/abs/1505.05836.

[8] Xinlei Chen and Abhinav Gupta. 2017. Spatial memory for context reasoning in object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. IEEE Computer Society, 4106–4116. DOI: https://doi.org/10.1109/ICCV.2017.440

[9] O. Chum and A. Zisserman. 2007. An exemplar model for learning object classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[10] R. G. Cinbis, J. Verbeek, and C. Schmid. 2017. Weakly supervised object localization with multi-fold multiple-instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1 (2017), 189–203.

[11] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. Retrieved from http://arxiv.org/abs/1606.09375.

[12] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. 2012. Weakly supervised localization and learning with generic knowledge. *Int. J. Comput. Vision* 100, 3 (Dec. 2012), 275–293. DOI: https://doi.org/10.1007/s11263-012-0538-3

[13] Ali Diba, Vivek Sharma, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. 2016. Weakly supervised cascaded convolutional networks. Retrieved from http://arxiv.org/abs/1611.08258.

[14] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple-instance problem with axis-parallel rectangles. *Artif. Intell.* 89, 1–2 (Jan. 1997), 31–71. DOI: https://doi.org/10.1016/S0004-3702(96)00034-3

[15] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. 2009. An empirical study of context in object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1271–1278.

[16] Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision* 111, 1 (Jan. 2015), 98–136. DOI: https://doi.org/10.1007/s11263-014-0733-5

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. [n.d.]. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. Retrieved from http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. [n.d.]. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. Retrieved from http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[19] Luis Felipe Zeni and Claudio R. Jung. 2020. Distilling knowledge from refinement in multiple-instance detection networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 768–769.

[20] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. 2017. DSSD: Deconvolutional single-shot detector. Retrieved from http://arxiv.org/abs/1701.06659.

[21] Carolina Galleguillos and Serge Belongie. 2010. Context-based object categorization: A critical survey. *Comput. Vis. Image Underst.* 114, 6 (June 2010), 712–722. DOI: https://doi.org/10.1016/j.cviu.2010.02.004

[22] C. Galleguillos, A. Rabinovich, and S. Belongie. 2008. Object categorization using co-occurrence, location and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[23] Mingfei Gao, Ang Li, Ruichi Yu, Vlad I. Morariu, and Larry S. Davis. 2017. C-WSL: Count-guided weakly supervised localization. Retrieved from http://arxiv.org/abs/1711.05282.

[24] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*. JMLR.org, 1263–1272.

[25] Ross Girshick. 2015. Fast R-CNN. Retrieved from http://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Girshick_Fast_R-CNN_ICCV_2015_paper.pdf.

[26] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. Retrieved from http://arxiv.org/abs/1311.2524.

[27] Michel Goossens, S. P. Rahtz, Ross Moore, and Robert S. Sutor. 1999. *The Latex Web Companion: Integrating TEX, HTML, and XML (1st ed.)*. Addison-Wesley Longman Publishing, Boston, MA.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 2980–2988.

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. Retrieved from http://arxiv.org/abs/1703.06870.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. Retrieved from http://arxiv.org/abs/1512.03385.

[31] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. 2014. How good are detection proposals, really? In *Proceedings of the British Machine Vision Conference*. BMVA Press. DOI: https://doi.org/10.5244/C.28.24

[32] Jan Hendrik Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. 2015. What makes for effective detection proposals? Retrieved from http://arxiv.org/abs/1502.05082.

[33] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2017. Relation networks for object detection. Retrieved from http://arxiv.org/abs/1711.11575.

[34] Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple-instance learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 2127–2136. Retrieved from http://proceedings.mlr.press/v80/ilse18a.html.

[35] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. 2015. Structural-RNN: Deep learning on spatio-temporal graphs. Retrieved from http://dblp.uni-trier.de/db/journals/corr/corr1511.html#JainZSS15.

[36] Ruyi Ji, Dawei Du, Libo Zhang, Longyin Wen, Yanjun Wu, Chen Zhao, Feiyue Huang, and Siwei Lyu. 2019. Learning semantic neural tree for human parsing. Retrieved from http://arxiv.org/abs/1912.09622.

[37] Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. 2020. Attention convolutional binary neural tree for fine-grained visual categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. IEEE, 10465–10474. DOI: https://doi.org/10.1109/CVPR42600.2020.01048

[38] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. 2017. Deep self-taught learning for weakly supervised object localization. Retrieved from http://arxiv.org/abs/1704.05188.

[39] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. 2016. ContextLocNet: Context-aware deep network models for weakly supervised localization. Retrieved from http://arxiv.org/abs/1609.04331.

[40] Thomas Kipf and Max Welling. 2016. Variational graph auto-encoders. Retrieved from https://abs/1611.07308.

[41] Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. Retrieved from http://arxiv.org/abs/1609.02907.

[42] Qiuqiang Kong, Changsong Yu, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D. Plumbley. 2019. Weakly labelled AudioSet classification with attention neural networks. Retrieved from http://arxiv.org/abs/1903.00765.

[43] Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. Association for Computing Machinery, New York, NY, 597–606. DOI: https://doi.org/10.1145/2783258.2783380

[44] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. Weakly supervised object detection with segmentation collaboration. Retrieved from http://arxiv.org/abs/1904.00551.

[45] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Graph convolutional recurrent neural network: Data-driven traffic forecasting. Retrieved from http://arxiv.org/abs/1707.01926.

[46] Chenhao Lin, Siwen Wang, Dongqi Xu, Yu Lu, and Wayne Zhang. 2020. Object instance mining for weakly supervised object detection. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20), the 32nd Innovative Applications of Artificial Intelligence Conference (IAAI'20), and the 10th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI'20)*. AAAI Press, 11482–11489. Retrieved from https://aaai.org/ojs/index.php/AAAI/article/view/6813.

[47] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2015. SSD: Single-shot MultiBox Detector. Retrieved from https://arxiv:1512.02325. DOI: https://doi.org/10.1007/978-3-319-46448-0_2

[48] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2015. SSD: Single-shot MultiBox detector. Retrieved from http://arxiv.org/abs/1512.02325.

[49] Xinhai Liu, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. 2019. Point2Sequence: Learning the shape representation of 3D point clouds with an attention-based sequence to sequence network. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19), the 31st Innovative Applications of Artificial Intelligence Conference (IAAI'19), and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI'19)*. AAAI Press, 8778–8785. DOI: https://doi.org/10.1609/aaai.v33i01.33018778

[50] Oded Maron and Tomás Lozano-Pérez. 1998. A framework for multiple-instance learning. In *Proceedings of the Conference on Advances in Neural Information Processing Systems 10 (NIPS'97)*. MIT Press, Cambridge, MA, 570–576.

[51] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. 2015. Is object localization for free? - Weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 685–694.

[52] Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. 2018. Adversarially regularized graph autoencoder for graph embedding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. AAAI Press, 2609–2615.

[53] M. Pandey and S. Lazebnik. 2011. Scene recognition and weakly supervised object localization with deformable part-based models. In *Proceedings of the International Conference on Computer Vision*. 1307–1314.

[54] Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Association for Computational Linguistics, Doha, Qatar, 455–466. DOI: https://doi.org/10.3115/v1/D14-1052

[55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, 8026–8037. Retrieved from http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[56] Minlong Peng and Qi Zhang. 2019. Address instance-level label prediction in multiple-instance learning. Retrieved from http://arxiv.org/abs/1905.12226.

[57] Pedro H. O. Pinheiro and Ronan Collobert. 2014. Weakly supervised semantic segmentation with convolutional networks. Retrieved from http://arxiv.org/abs/1411.6228.

[58] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2015. You only look once: Unified, real-time object detection. Retrieved from http://arxiv.org/abs/1506.02640.

[59] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15)*. MIT Press, Cambridge, MA, 91–99.

[60] W. Ren, K. Huang, D. Tao, and T. Tan. 2016. Weakly supervised large scale object localization with multiple-instance learning and bag splitting. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 2 (2016), 405–416.

[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2014. ImageNet large scale visual recognition challenge. Retrieved from http://arxiv.org/abs/1409.0575.

[62] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. 2009. The graph neural network model. *IEEE Trans. Neural Netw.* 20, 1 (2009), 61–80.

[63] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao. 2019. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 697–707.

[64] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang. 2018. Generative adversarial learning towards fast weakly supervised detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5764–5773.

[65] M. Shi, H. Caesar, and V. Ferrari. 2017. Weakly supervised object localization using things and stuff transfer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 3401–3410.

[66] Z. Shi, T. M. Hospedales, and T. Xiang. 2015. Bayesian joint modelling for object localisation in weakly labelled images. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 10 (2015), 1959–1972.

[67] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. 2006. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV'06)*, Aleš Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–15.

[68] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

[69] Krishna Kumar Singh and Yong Jae Lee. 2019. You reap what you sow: Using videos to generate high precision object proposals for weakly-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*.

[70] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. 2014. Weakly-supervised discovery of visual pattern configurations. Retrieved from http://arxiv.org/abs/1406.6507.

[71] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan L. Yuille. 2018. PCL: Proposal cluster learning for weakly supervised object detection. Retrieved from http://arxiv.org/abs/1807.03342.

[72] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. 2017. Multiple-instance detection network with online instance classifier refinement. Retrieved from http://arxiv.org/abs/1704.00138.

[73] Peng Tang, Xinggang Wang, Zilong Huang, Xiang Bai, and Wenyu Liu. 2017. Deep patch learning for weakly supervised object classification and discovery. Retrieved from http://arxiv.org/abs/1705.02429.

[74] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. 2018. Weakly supervised region proposal network and object detection. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 370–386.

[75] Torralba, Murphy, Freeman, and Rubin. 2003. Context-based vision system for place and object recognition. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, Vol. 1. 273–280.

[76] Z. Tu and X. Bai. 2010. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 10 (2010), 1744–1757.

[77] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. 2013. Selective search for object recognition. *Int. J. Comput. Vision* 104, 2 (2013), 154–171. https://ivi.fnwi.uva.nl/isis/publications/2013/UijlingsIJCV2013.

[78] Rianne van den Berg, Thomas Kipf, and Max Welling. 2017. Graph convolutional matrix completion. Retrieved from https://abs/1706.02263.

[79] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=rJXMpikCZ.

[80] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. 2019. C-MIL: Continuation multiple-instance learning for weakly supervised object detection. Retrieved from http://arxiv.org/abs/1904.05647.

[81] Fang Wan, Pengxu Wei, Zhenjun Han, Jianbin Jiao, and Qixiang Ye. 2019. Min-entropy latent model for weakly supervised object detection. Retrieved from http://arxiv.org/abs/1902.06057.

[82] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. 2014. Weakly supervised object localization with latent category learning. In *Proceedings of the European Conference on Computer Vision (ECCV'14)*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 431–445.

[83] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. 2019. Region proposal by guided anchoring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*.

[84] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. 2017. Non-local neural networks. Retrieved from http://arxiv.org/abs/1711.07971.

[85] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas S. Huang. 2018. TS2C: Tight box mining with surrounding segmentation context for weakly supervised object detection. Retrieved from http://arxiv.org/abs/1807.04897.

[86] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. 2015. Deep multiple-instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. IEEE Computer Society, 3460–3469. DOI: 10.1109/CVPR.2015.7298968

[87] Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, and Jure Leskovec. 2018. GraphRNN: Generating realistic graphs with deep auto-regressive models. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80, 5708–5717. Retrieved from http://proceedings.mlr.press/v80/you18a.html.

[88] Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit-Yan Yeung. 2018. GaAN: Gated attention networks for learning on large and spatiotemporal graphs. Retrieved from http://arxiv.org/abs/1803.07294.

[89] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. 2017. Single-shot refinement neural network for object detection. Retrieved from http://arxiv.org/abs/1711.06897.

[90] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. 2018. Zigzag learning for weakly supervised object detection. Retrieved from http://arxiv.org/abs/1804.09466.

[91] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem. 2018. W2F: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 928–936.

[92] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. 2008. Multi-instance learning by treating instances as non-I.I.D. samples. Retrieved from http://arxiv.org/abs/0807.1997.

[93] Zhi-Hua Zhou and Min-Ling Zhang. 2002. Neural networks for multi-instance learning. *Proceedings of the International Conference on Intelligent Information Technology*.

[94] Larry Zitnick and Piotr Dollar. 2014. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV'14)*. Retrieved from https://www.microsoft.com/en-us/research/publication/edge-boxes-locating-object-proposals-from-edges/.

Received July 2020; revised October 2020; accepted October 2020