Neurocomputing 496 (2022) 192–207

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep Learning for Weakly-Supervised Object Detection and Localization: A Survey

Feifei Shao [a], Long Chen [b,*], Jian Shao [a], Wei Ji [c], Shaoning Xiao [a], Lu Ye [d], Yueting Zhuang [a], Jun Xiao [a]

[a] College of Computer Science, Zhejiang University, Hangzhou 310027, China
[b] Department of Electrical Engineering, Columbia University, New York 10027, USA
[c] School of Computing, National University of Singapore, 117417, Singapore
[d] School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China

ARTICLE INFO

Article history:
Received 30 September 2021
Revised 4 January 2022
Accepted 23 January 2022
Available online 29 January 2022

Keywords:
Weakly-supervised learning
Object detection and localization
Basic framework
Techniques
Future directions

ABSTRACT

Weakly-Supervised Object Detection (WSOD) and Localization (WSOL), *i.e.*, detecting multiple and single instances with bounding boxes in an image using image-level labels, are long-standing and challenging tasks in object detection. Hundreds of WSOD and WSOL methods and numerous techniques have been proposed in the deep learning era. To this end, in this paper, we consider WSOL as a sub-task of WSOD and provide a comprehensive survey of the recent achievements of WSOD. Specifically, we firstly describe the formulation and setting of the WSOD, including the background, challenges, basic framework. Meanwhile, we summarize and analyze all advanced techniques and training and test tricks for improving detection performance. Then, we introduce the widely-used datasets and evaluation metrics of WSOD. Lastly, we discuss the future directions of WSOD. We believe that these summaries can help pave a way for future research on WSOD and WSOL.

1. Introduction

As an important part of multimodal learning, cross-media reasoning has received unprecedented attention, such as Visual Question Answering (VQA) [1–3,4,5], visual captioning [6–10], cross-media retrieval [11–19] and so on. Object detection [20–23]—locating and classifying object instances in an image—which is not only a traditional computer vision task but also a foundation technique for these cross-media reasoning tasks connecting vision and other modalities.

With the development of convolutional neural networks (CNNs) in visual recognition [24–26] and release of large scale dateset [27,28], today's state-of-the-art object detector can achieve near-perfect performance under fully-supervised setting, *i.e.*, Fully-Supervised Object Detection (FSOD) [29–34]. Unfortunately, the fully-supervised object detection task suffers from two inevitable limitations: 1) The large-scale instance annotations are difficult to obtain and labor-intensive. 2) When labeling instance-level data, they may inadvertently introduce imprecise and ambiguous manual annotations, which are hurtful for FSOD.

To avoid the mentioned problems, the community starts to solve object detection in a weakly-supervised setting, *i.e.*, Weakly-Supervised Object Detection (WSOD) [35–37]. Different from the fully supervised setting (cf. Fig. 1 (a)), WSOD aims to detect instances with only image-level labels (*e.g.*, categories of instances in the whole images). Meanwhile, WSOD can benefit from large-scale datasets on the web, such as Facebook and Twitter. Another task similar to WSOD is Weakly-Supervised Object Localization (WSOL) [38,39,40], which only detects one instance in an image. Since WSOD and WSOL detect multiple and single instances respectively, we consider WSOL as a sub-task of WSOD and use WSOD to represent both WSOD and WSOL in the following paper.

In this paper, we go over all typical WSOD methods and give a comprehensive survey (cf. Fig. 2) of recent advances in WSOD. Since the number of papers on WSOD is breathtaking, we sincerely apologize to those authors whose research on WSOD and other related fields are not included in this survey. In Section 2, we introduce the background, main challenges, and basic framework. In Section 3, according to the development timeline of WSOD, we introduce several modern classical methods in detail. Then, in-depth analyses are provided towards the all advanced techniques

* Corresponding author.
E-mail addresses: sff@zju.edu.cn (F. Shao), zjuchenlong@gmail.com (L. Chen), jshao@zju.edu.cn (J. Shao), weiji0523@gmail.com (W. Ji), shaoningx@zju.edu.cn (S. Xiao), yelue@zust.edu.cn (L. Ye), yzhuang@zju.edu.cn (Y. Zhuang), junx@zju.edu.cn (J. Xiao).

https://doi.org/10.1016/j.neucom.2022.01.095
0925-2312/

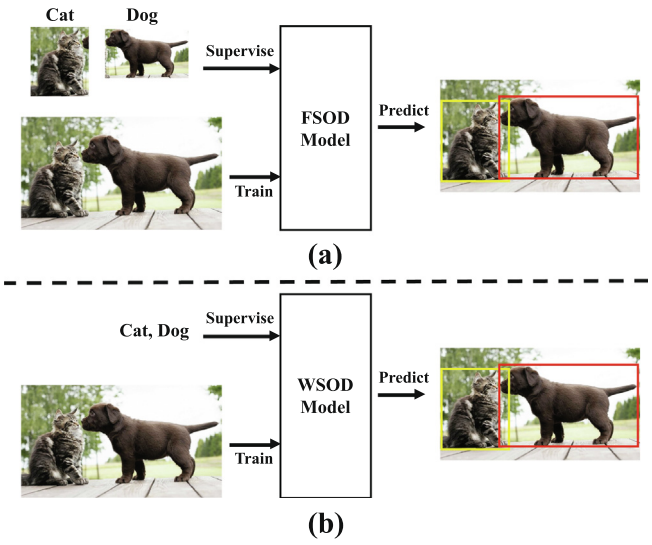


Fig. 1. (a) Fully-Supervised Object Detection (FSOD) uses the *instance-level* annotations as supervision. (b) Weakly-Supervised Object Detection (WSOD) uses the *image-level* annotations as supervision.

and tricks for the main challenges. In Section 7, we demonstrate all prevailing benchmarks and standard evaluation metrics for WSOD. In Section 8, we briefly discuss the future directions.

2. WSOD

2.1. A problem definition

WSOD aims to classify and locate object instances using only image-level labels in the training phase. As shown in Fig. 1 (b), given an image with cat and dog, WSOD not only classifies the cat and dog but also locates their location using bounding boxes. Different from FSOD that uses instance-level annotations in the training phase shown in Fig. 1 (a), WSOD only accesses image-level labels. Because of this restriction, though hundreds of WSOD methods have been proposed, the performance gap between WSOD and FSOD is still large. For example, the mAP of state-of-the-art FSOD approach [41] and WSOD approach [42] are 86.9% and 56.8% on PASCAL VOC 2007 dataset [43], respectively. Therefore, there are still many challenges in the task of WSOD for researchers to solve, especially in the direction of improving the detection performance.

2.2. Main challenges

The main challenges of WSOD come from two aspects: localization accuracy and speed. For localization accuracy, it consists of a discriminative region problem and multiple instances with the same category problem. Speed is an important characteristic of real applications. In Table 1, we summarize most of WSOD methods and their contributions to these challenges.

Discriminative Region Problem. It is that detectors [44,45] tend to focus on the most discriminative parts of the object. During training, there may exist more than one proposal around an object, and the most discriminative part region of the object is likely to have the highest score (*e.g.*, the region A is the most discriminative region in Fig. 3 (left) and it has a higher score than that of other regions). If the model selects positive proposals only using scores, it is easy to focus on the most discriminative part of the object rather than the whole object extent.

Multiple-instance Problem. It is a huge challenge to accurately detect all instances when there may exist several objects with the same category in an image since detectors [44,49] tend to select the highest score proposal of each category as the positive proposal and ignore other possible instance proposals.

Speed Problem. At present, proposal generators (*e.g.*, Selective Search (SS) [88], Edge Boxes (EB) [89] and Sliding Window (SW)) that are widely used in WSOD are too time-consuming. Specifically, SS and EB require 10 and 0.25 s respectively to generate 5,000 high-quality proposals [89]. A typical sliding window detector requires $\sim 10^6$ times classification per image [90], which significantly increases the computational cost of the detector. For example, the typical sliding window detector OverFeat [91] requires 2 s per image.

2.3. Basic WSOD framework

The basic framework of WSOD methods can be categorized into MIL-based networks and CAM-based networks.

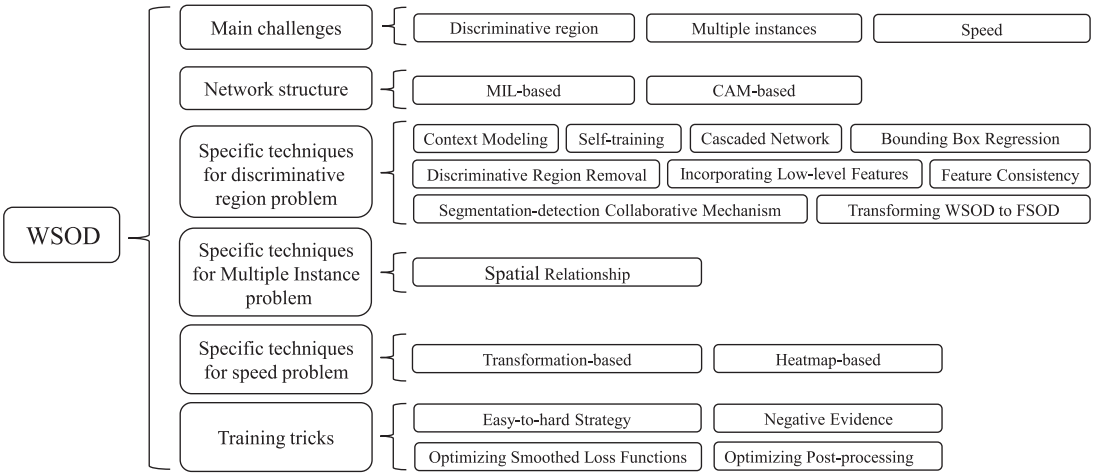


Fig. 2. The main content of this paper. Firstly, main challenges and network structures are introduced in Section 2.2 and Section 2.3, respectively. Secondly, in-depth analyses are provided towards specific techniques for discriminative region, multiple-instance and speed problem in Section 4, Section 5.1, and Section 5.2, respectively. Thirdly, in Section 6, we categorize the training and test tricks into four groups.

Table 1
A summary of the state-of-the-art WSOD methods. For the proposals, SS represents selective search, EB represents edge boxes, and SW represents a sliding window. The Challenges denote the main contributions of corresponding papers.

Approach	Proposals	Network		Challenges			Code on Github
		MIL-based	CAM-based	Discriminative Region	Multiple Instances	Speed	
WSDDN [44]$_{CVPR2016}$	EB	√					hbilen/WSDDN
CAM [45]$_{CVPR2016}$	Heatmap		√			√	zhoubolei/CAM
WSLPDA [46]$_{CVPR2016}$	EB	√		√			jbhuang0604/WSL
WELDON [47]$_{CVPR2016}$	SW	√		√		√	
ContextLocNet [48]$_{ECCV2016}$	SS	√		√			vadimkantorov/contextlocnet
OICR [49]$_{CVPR2017}$	SS	√		√			ppengtang/oicr
WCCN [50]$_{CVPR2017}$	EB	√		√			
ST-WSL [51]$_{CVPR2017}$	EB	√		√	√		
WILDCAT [52]$_{CVPR2017}$	Heatmap		√	√		√	durandtibo/wildcat.pytorch
Grad-CAM [53]$_{ICCV2017}$	Heatmap		√	√		√	ramprs/grad-cam
SPN [54]$_{ICCV2017}$	SW	√				√	ZhouYanzhao/SPN
TP-WSL [55]$_{ICCV2017}$	Heatmap		√	√		√	
PCL [56]$_{TPAMI2018}$	SS	√		√	√		ppengtang/pcl.pytorch
GAL-fWSD [57]$_{CVPR2018}$	EB	√				√	
W2F [58]$_{CVPR2018}$	SS	√		√	√	√	
ACoL [59]$_{CVPR2018}$	Heatmap		√	√		√	xiaomengyc/ACoL
ZLDN [60]$_{CVPR2018}$	EB	√		√			
TS^{2}C [61]$_{ECCV2018}$	SS	√		√			
SPG [62]$_{ECCV2018}$	Heatmap		√			√	xiaomengyc/SPG
WSRPN [63]$_{ECCV2018}$	EB	√					
C-MIL [64]$_{CVPR2019}$	SS	√					WanFang13/C-MIL
WS-JDS [65]$_{CVPR2019}$	EB	√		√			shenyunhang/WS-JDS
ADL [66]$_{CVPR2019}$	Heatmap		√			√	junsukchoe/ADL
Pred NET [67]$_{CVPR2019}$	SS	√					
WSOD2 [68]$_{ICCV2019}$	SS	√		√			researchmm/WSOD2
OAILWSD [69]$_{ICCV2019}$	SS	√		√			
TPWSD [70]$_{ICCV2019}$	SS	√		√			
SDCN [71]$_{ICCV2019}$	SS	√		√			
C-MIDN [72]$_{ICCV2019}$	SS	√		√			
DANet [73]$_{ICCV2019}$	Heatmap		√			√	xuehaolan/DANet
NL-CCAM [74]$_{WACV2020}$	Heatmap		√	√		√	Yangseung/NL-CCAM
ICMWSD [75]$_{CVPR2020}$	SS	√		√			
EIL [76]$_{CVPR2020}$	Heatmap		√	√		√	Wayne-Mai/EIL
SLV [77]$_{CVPR2020}$	SS	√		√			
RethinkingCAM [78]$_{ECCV2020}$	Heatmap		√	√		√	won-bae/rethinkingCAM
I^{2}C [79]$_{ECCV2020}$	Heatmap		√	√		√	xiaomengyc/I2C
UWSOD [80]$_{NeurIPS2020}$	SW	√		√			
CASD [42]$_{NeurIPS2020}$	SS	√		√			DeLightCMU/CASD
MCIR [81]$_{WACV2021}$	Heatmap		√	√		√	
CI-CAM [82]$_{ACMMM2021}$	Heatmap		√			√	shaofeifei11/CI-CAM
SLT-Net [83]$_{CVPR2021}$	Heatmap		√	√		√	gyguo/SLT-Net
SPA [84]$_{CVPR2021}$	Heatmap		√	√		√	
SPOL [85]$_{CVPR2021}$	Heatmap		√	√		√	weijun88/SPOL
IVR [86]$_{ICCV2021}$	Heatmap		√	√		√	
ORNet [87]$_{ICCV2021}$	Heatmap		√	√		√	Sierkinhane/ORNet

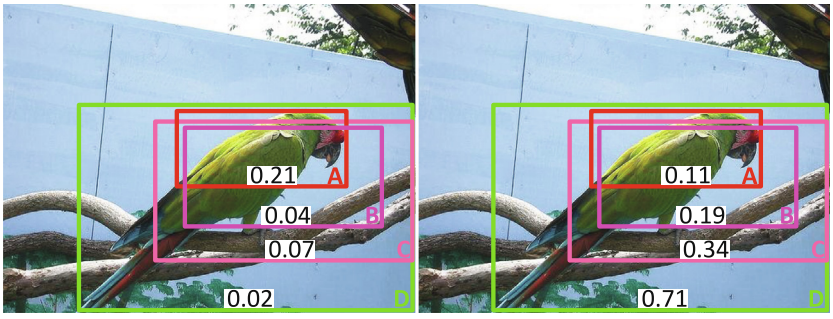


Fig. 3. Detection results between model without classifier refinement (left) and model with classifier refinement (right). The figure comes from [49].

2.3.1. MIL-based network

When the detection network predicts instances based on Multiple Instance Learning (MIL) problem [92], it is considered a MIL-based network. Taking Fig. 1 (b) for example, an image is interpreted as a bag of proposals in the MIL problem. If the image is labeled cat, it means that at least one of the proposals tightly contains the cat instance. Otherwise, all of the regions do not contain the cat instance (likewise for dogs). The MIL-based network is based on the structure of WSDDN [44] that consists of a proposal generator, backbone, and detection head.

Proposal Generator. Some proposal generators are usually used in MIL-based networks. 1) *Selective search (SS)* [88]: it lever-

ages the advantages of both exhaustive search and segmentation to generate initial proposals. 2) *Edge boxes (EB)* [89]: it uses object edges to generate proposals and is widely used in many approaches [44,46,50,51,60,51,65]. 3) *Sliding window (SW)*: it uses multiple boxes with different scales to continuously slide in an image and generates a proposal after each sliding. However, SW suffers from one inevitable problem. If the scales of the sliding window are too much, it will produce a large number of invalid proposals that increase the computational complexity and reduce the detection speed. If the scales are too few, it is easy to miss odd-shaped objects.

Backbone. With the development of CNNs and large scale datasets (*e.g.*, ImageNet [27]), the pretrained AlexNet [93], VGG16 [24], GoogLeNet [25], InceptionV3 [94] and SENet [95] are prevailing feature representation networks for both classification and object detection. Besides, DRN-WSOD [96] tries to leverage deep residual networks (*e.g.*, ResNet [26]) to effectively improve object instance localization and discriminative feature learning.

Detection Head. It includes a classification stream and a localization stream. The classification stream predicts class scores for each proposal, and the localization stream predicts every proposal's existing probability score for each class. Then the two scores are aggregated to predict the confidence scores of an image as a whole, which are used to inject image-level supervision in learning.

Given an image, we first feed it into the proposal generator and backbone to generate proposals and feature maps, respectively. Then, the feature maps and proposals are forwarded into a spatial pyramid pooling (SPP) [97] layer to generate fixed-size regions. Finally, these regions are fed into the detection head to classify and locate object instances.

2.3.2. CAM-based network

CAM-based network targets locating instances in terms of segmenting the class activation maps. It is based on the structure of CAM [45], which consists of three components: backbone, classifier, and class activation maps.

Backbone. It is similar to the backbone of the MIL-based network introduced in Section 2.3.1, which is responsible for providing good feature maps. Besides, TS-CAM [98] attempts to replace the Convolutional Neural Network (CNN) backbone with Transformer [99,100,101] for capturing long-range feature dependency among pixels.

Classifier. It is designed to classify the classes of an image, which includes a global average pooling (GAP) layer and a fully connected layer.

Class Activation Maps. It is responsible for locating object instances by using a simple segmentation technique. Because the class activation maps are produced by matrix multiplying the weight of the fully connected layer to the feature maps of the last convolutional layer, it spotlights the class-specific discriminative regions in every activation map. Therefore, it is easy to generate bounding boxes of every class by segmenting the activation map of the class.

Given an image, we first feed it into the backbone to generate feature maps of this image. Then, the feature maps are forwarded into the classifier to classify the image's classes. Meanwhile, we matrix multiply the weight of the fully connected layer to the feature maps of the last convolutional layer to produce class activation maps. Finally, we segment the activation map of the highest probability class to yield bounding boxes for object localization.

2.3.3. Discussions

In this section, in-depth analyses are provided towards the rationality of WSOD networks from both MIL-based and CAM-based.

Firstly, MIL-based network leverages SS [88], EB [89] or sliding window to generate thousands of initial proposals, but CAM-based network segments the activation map to one proposal for each class. Therefore, a MIL-based network is better than a CAM-based network when detecting multiple instances with the same category in an image. However, the training and inference speed of MIL-based networks is slower than CAM-based, since SS, EB, and sliding window are too time-consuming and yield plenty of initial proposals that most of them are invalid proposals.

Secondly, because the size of the proposals produced by SS, or EB is not consistent, the MIL-based network leverages an SPP layer to generate fixed-size vectors followed by feeding these fixed-size vectors into the fully connected layers for later training. However, a CAM-based network leverages a GAP layer to generate a fixed-size vector on the feature maps. Then, it feeds the vector into a fully connected layer for classifying.

Finally, both MIL-based networks and CAM-based networks suffer from the discriminative region problem and multiple-instance problem.

3. Milestones of WSOD

Since 2016, there are some landmark methods (cf. Fig. 4) for the research of WSOD. In the following, we will briefly introduce these milestones.

3.1. MIL-based methods

WSDDN. The biggest contribution of WSDDN [44] is using two streams network, which aims to perform classification and localization respectively. WSDDN first uses a SPP [97] on the top of the feature maps and generates a feature vector after two fully connected layer procedures. Next, the feature vector is fed into the classification stream and localization steam. Specifically, the classification stream is responsible for computing the class scores of each region, and the localization stream is designed to compute every region's existing probability for each class. Then, the matrix product of the class scores of each region and the existing probability for each class is considered as the final prediction scores. However, because of only accessing image-level labels in the training phase, the most discriminative part of the object will be paid more attention than the whole object instance in training. Due to the above limitation, WSDDN suffers from the discriminative region problem.

OICR. To alleviate the discriminative region problem, OICR [49] uses WSDDN as its baseline and adds three instance classifier refinement procedures after the baseline. Every instance classifier refinement procedure, which consists of two fully connected layers, is designed to further predict the class scores for each proposal. Because the output of each instance classifier refinement procedure is the supervision of its latter refinement procedure, OICR can continue to learn so that a larger area can have higher scores than WSDDN. Although the prediction of WSDDN may only focus on the discriminative part of the object, it will be refined after several instance classifier refinement procedures.

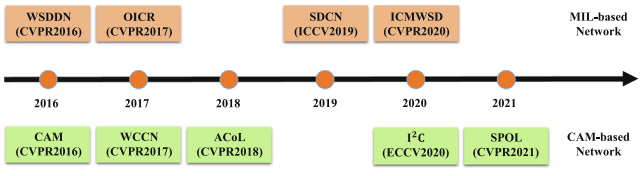


Fig. 4. The Milestones of WSOD since 2016.

SDCN. SDCN [71] introduces a segmentation-detection collaborative mechanism. It consists of a detection branch and segmentation branch, which are responsible for detecting bounding boxes and generating segmentation masks respectively. In SDCN, the detection results will be converted to a heatmap by setting a classification score to all pixels within each proposal as the supervision mask of the segmentation branch. Meanwhile, the proposals of the highest overlap with the connected regions from the segmentation masks will be the pseudo-ground-truth boxes of the detection branch. Both detection and segmentation branch are optimized alternatively and promoted each other, so SDCN achieves better detection performance than OICR.

ICMWSD. Different from SDCN which leverages both object detection and segmentation collaboration mechanism, ICMWSD [75] addresses the problem of focusing on the most discriminative part of an object by leveraging context information. Firstly, ICMWSD obtains dropped features by dropping the most discriminative parts. Then, maximizing the loss of the dropped features force ICMWSD to look at the surrounding context regions.

3.2. CAM-based methods

CAM. The biggest contribution of CAM [45] is using class activation maps to predict instances. CAM firstly leverages a GAP layer on the last convolutional feature maps to generate a feature vector. Then, the feature vector is fed into a classifier with a fully connected layer to generate prediction scores of an image. Finally, CAM generates bounding boxes of each class by using a simple thresholding technique to segment the activation map of every class. However, class activation maps of CAM spotlight the regions that are the most discriminative parts of the object, so CAM also suffers from the discriminative region problem as WSDDN.

WCCN. To alleviate the discriminative region problem, WCCN [50] uses a cascaded network that has three cascade stages trained in an end-to-end pipeline. The first stage is the CAM [45] network that aims to generate class activation maps and initial proposals. The second stage is a segmentation network that uses the class activation maps to train object segmentation for refining object localization. The final stage is a MIL network that detects multiple instances on proposals extracted in the second stage. Because the second and third stages refine object localization, WCCN alleviates the problem that the detection model tends to focus on the most discriminative part of the object rather than the whole object region.

ACoL. To alleviate the discriminative region problem, ACoL [59] introduces two parallel-classifiers for object localization using adversarial complementary learning. Specifically, it first leverages the first classifier to localize the most discriminative regions. Then, ACoL uses the masked feature maps by masking the most discriminative regions discovered in the first classifier as the input feature maps of the second classifier. This forces the second classifier to select the next discriminative regions. Finally, ACoL fuses the class activation maps of both classifiers to generate bounding boxes of every class by segmenting the activation map of the highest probability class.

I^{2}C. Different from the above approaches [50,59], I^{2}C [79] utilizes the consistency of object features within the same categories to highlight all the object regions and depress the background in the localization map. Specifically, I^{2}C proposes stochastic consistency and global consistency for learning more robust and reliable localization map. On the one hand, stochastic consistency is designed to ensure that the pixels of objects within the same categories in a batch of images have the same semantic features. On the other hand, global consistency is responsible for guaranteeing the feature consistency of the object pixels within the same categories of training set. Due to the consistency of object features, the pixels among one object have the same brightness in the localization map, and the detection model can detect the whole object.

SPOL. SPOL [85] argues that the previous works fail to make full use of the shallow features. Thus, it designs the multiplicative feature fusion network (MFF-Net) and gaussian prior pseudo label (GPPL) module for generating more accurate object localization. The MFF-Net aggregates both features of shallow and deep layers for suppressing the noise while making object boundaries sharper. The GPPL enhances the responses of the area inside the object gravity by using the mean and variance for all coordinates. Specifically, SPOL first generates the initial localization map using MFF-Net and calculates the object gravity upon the initial localization map. Then, SPOL enhances the responses of the area inside the object gravity. Next, SPOL uses two thresholds to segment foreground and background, respectively. Finally, the foreground and background are used as pseudo labels to train a class-agnostic segmentation network for generating the final localization map.

4. Specific techniques for discriminative region problem

In this section, we will introduce several advanced techniques for solving the discriminative region problem.

4.1. Context modeling

The context of one region is external information of this region, which can be obtained by masking the region of the feature maps with special numbers (*e.g.*, zero). Although contexts are not part of the object instance, they also contain a lot of useful information. If we make full use of the contextual information, the detectability of our model will be greatly improved. Among the context modeling, methods can be further grouped into score strategy and loss strategy.

Score strategy. It selects the regions that have a big gap between their scores and their contextual region's scores as positive proposals. For example, WSLPDA [46] first replaces the pixel values within one proposal with zero to obtain the contextual region. Then, WSLPDA compares the scores of proposals and their contextual region. If the gap between the two scores is large, it indicates that the proposal is likely positive. ContextLocNet [48] subtracts the localization score of one proposal from the localization score of the external rectangle region of the proposal. Then, the subtraction is considered as the final localization score of the proposal. Similar to WSLPDA and ContextLocNet, TS^{2}C [61] selects a positive proposal by comparing the mean objectness scores of the pixels in one proposal and its surrounding region. But to alleviate the impact of background pixels in the surrounding region, TS^{2}C computes the mean objectness scores only using pixels with large confidence values in the surrounding region.

Loss strategy. It selects positive proposals by leveraging the loss of context regions. For example, OAILWSD [69] believes that a proposal not tightly covers the object instance if the loss of the context feature maps of this proposal tends to decrease. Thus, OAILWSD first leverages the context classification loss to label regions. Then, it selects the top-scoring regions whose context class probabilities are low as positive proposals. ICMWSD [75] first drops the most discriminative parts of the feature maps to obtain contextual feature maps. Then, it maximizes the loss of the contextual feature maps to force it to focus on the context regions.

4.2. Self-training algorithm

In the self-training algorithm, the early prediction instances are then used in the detector for latter training as the pseudo-ground-truth instances. The key idea of self-training is that even if the early

top-scoring proposals may only focus on the discriminative part of the object, they will be refined after several refinement procedures. There are two types of self-training algorithms: inter-stream and inter-epoch. In inter-stream self-training, the instances of each stream supervise its later stream. In inter-epoch self-training, the instances of each epoch supervise its later epoch.

4.2.1. Inter-stream self-training

OICR [49] expects B, C, and D can inherit the class score of A to correctly localize objects in Fig. 3 (right). Therefore, OICR adds three refinement classifiers with two fully connected layers in WSDDN to address the issue shown in Fig. 3 (left). Specifically, the supervision of the first refinement classifier is the output of WSDDN. As for other refinement classifiers, the supervision of the current refinement classifier is the output of its previous refinement classifier. Inspired by OICR, WSOD2 [68] and CASD [42] also consist of numerous classifiers. ICMWSD [75] and UWSOD [80] insert refinement streams in WSDDN, however, every refinement stream includes a classifier and a regressor. The supervision data of the regressor is the prediction bounding boxes of previous stream. Besides, some approaches [58,72,69,70] use OICR as their baseline for refining the localization performance.

4.2.2. Inter-epoch self-training

ST-WSL [51] uses relative improvement (RI) scores of each proposal of two adjacent epochs as a criterion for selecting the positive sample. It chooses the proposals of the previous epoch whose intersection over union (IoU) $\geqslant 0.5$ with the maximal RI proposal as the positive samples of the current epoch.

4.3. Cascaded network

The cascaded network includes several different types of modules and the output of the previous module is the supervision of the latter. The multiple modules are concatenated to completely preserve the advantages of each module.

WCCN [50] and TS^{2}C [61] consist of three modules. The first module is a CAM [45] that is to detects the most discriminative regions and generates initial proposals by segmenting the class activation maps. The second is an object segmentation module designed to refine the size and position of the initial proposals by leveraging the capabilities of semantic segmentation models. The last one is a multiple instance learning module that is responsible for detecting accurate objects. It uses the refined proposals produced by the segmentation module as supervision data. At the localization branch of SLT-Net [83], a localizer is responsible for generating high-quality pseudo bounding boxes and a regressor is trained by these pseudo bounding boxes. SPOL [85] first generates class activation maps followed by obtaining discriminative region and object gravity area. Then, SPOL uses a double thresholding strategy to produce pseudo pixel-level labels, which consists of foreground, background, and conflict regions. Finally, these pseudo labels are used as supervision to train a class-agnostic segmentation module.

4.4. Bounding box regression

Bounding box regression can improve object localization performance using instance-level annotations in the training phase, but the WSOD task only accesses image-level labels. To solve this inevitable limitation, some approaches propose to yield high-quality pseudo bounding boxes for regressor.

Now, numerous approaches [67,68,70,75,77,83] include at least one of the bounding box regressors for adjusting the candidate bounding boxes from SS [88] or EB [89]. The supervision of the regressor is the output of previous classifiers. Different from adjusting the candidate bounding boxes, UWSOD [80] is one of the pioneers, which uses the final prediction bounding boxes as the supervision data of bounding box regression to go back and adjust the initial proposals.

4.5. Discriminative region removal

From Fig. 3 (left), some researchers find that the highest score region only covers the most discriminative part of the object. To localize the whole object extent, masking the most discriminative part of the object is designed to force the detector to find the next discriminative region.

TP-WSL [55] is a two-phase learning network that detects the next discriminative regions by masking the most discriminative region. In the first phase, it yields class activation maps followed by masking the most discriminative region using a threshold among the activation map of the highest probability class. In the second phase, it multiplies the masked activation map by the feature maps of the second network to refine the feature maps for detecting the next discriminative regions. ORNet [87] first chooses the maximum response point in the localization map followed by randomly dropping high-response pixels in a rectangle around the maximum response point. Finally, the dropped map and source image are fed into the classifier together during training for obtaining a robust classifier.

Different from TP-WSL and ORNet that have two backbones, ACoL [59] consists of one shared backbone and two parallel-classifiers. The masked feature maps from the first classifier are fed into the second classifier to generate class activation maps. Finally, ACoL locates object instances in the fused activation maps by fusing the two-class activation maps of both classifiers. EIL [76] proposes to share the weights of the two parallel-classifiers of ACoL, and it only segments the activation map of the highest probability class from the unmasked branch to yield object proposals. Different from masking the discriminative region in the feature maps, MCIR [81] randomly masks regions in the input image twice and obtains two complementary images. Based on this, MCIR mines information from complementary regions in an image while alleviating the discriminative region problem.

Comparing C-MIDN [72] with ACoL, there are three differences. First, the detection network of C-MIDN is WSDDN [44], but the detection network of ACoL is CAM [45]. Second, C-MIDN does not compute the loss of high overlap with the first detection module's top-scoring proposal in the second branch, but ACoL masks the first detection module's top-scoring proposal's region with zero in the second branch. Finally, C-MIDN chooses the top-scoring proposals of the second detection module and the top-scoring proposals of the first detection module with low overlap with selected proposals as positive proposals, but ACoL yields positive proposals by segmenting the fused class activation maps.

4.6. Incorporating low-level features

Low-level features usually retain richer object details, such as edges, corners, colors, pixels, and so on. We can obtain accurate object localization if making full use of these low-level features.

Grad-CAM [53] leverages high-resolution Guided Backpropagation [102] that highlights the image's details to create both high-resolution and class-discriminative visualizations. WSOD2 [68] first computes the score of a region proposal. Then, it selects the same region in low-level image features and computes its score. Finally, the product of the two scores is the final score of the region proposal. SPOL [85] first aggregates the low-level features embedded in shallow layers and the high-level features embedded in deep layers by multiplicative fusion. Then, SPOL produces class activation maps upon the fused feature maps. Besides, ORNet

[87] produces class activation maps upon the low-level feature maps in the classifier.

4.7. Feature consistency

Since an image classifier consists of fully connected layers with thousands of weight parameters, for a specific output score of a pixel, the input semantic features of the classifier may be various [79]. However, part of these different semantic features can generate high responses in the activation map, and others generate low responses. Inconsistent features of the pixels among an object are the root cause for different responses.

I^2C [79] achieves the consistency of object features within the same categories by optimizing the Euclidean distance between features of different pixels. Specifically, I^2C proposes stochastic consistency and global consistency. The former is responsible for the consistency of semantic features between each pixel in an object, while the latter is designed to prompt the feature consistency of the object pixels of different images in the same categories. CASD [42] pays attention to the feature consistency of proposals. Specifically, CASD conducts consistent representation learning over input images under multiple transformations, which guarantees the feature consistency of related proposals of the same image under different transformations. Besides, CASD also conducts consistent representation learning over convolutional blocks and further enhances the feature consistency within the same proposal.

4.8. Segmentation-detection collaborative mechanism

Segmentation-detection collaborative mechanism includes a segmentation branch and a detection branch. The primary reasons for the collaborative mechanism are the following: 1) MIL (detection) can correctly distinguish an area as an object, but it is not good at detecting whether the area contains the entire object. 2) Segmentation can cover the entire object instance, but it cannot distinguish whether the area is a real object or not [71]. So, some models use deep cooperation between detection and segmentation by supervising each other to achieve localization.

WS-JDS [65] first chooses the region proposals with top-scoring pixels generated by the semantic segmentation branch as the positive samples of the detection branch. Then, it sets the classification score to all pixels within each positive proposal of the detection branch as the supervision mask of the segmentation branch. Similar to WS-JDS, SDCN [71] also combines the detection branch with the segmentation branch which is introduced in Section 3.1.

4.9. Transforming WSOD to FSOD

Transforming WSOD to FSOD is another popular technique to achieve precise object detection, which is designed to train an FSOD model using the output of the WSOD model.

The primary problem of transformation is to yield good pseudo-ground-truth boxes from WSOD. There are several strategies to mine boxes as pseudo-ground-truth boxes. 1) *top score*: numerous approaches [49,56,61,65,71,72] select top score detection boxes of WSOD as the pseudo ground-truth boxes. 2) *relative improvement (RI)*: ST-WSL [51] selects the boxes with the maximal relative score improvement of two adjacent epochs as the pseudo-ground-truth boxes. 3) *mergence*: W2F [58] merges several small boxes into a big candidate box and uses these merged boxes as the pseudo-ground-truth boxes for later training. SLV [77] first merges the scores of several boxes to the pixels and then generates bounding boxes of each class by segmenting the map of every class using a threshold.

In addition, there are several FSOD models that have been used as follows: Fast R-CNN [103], Faster R-CNN [29], and SSD [30]. Numerous approaches [49,51,56,61,77,71,72,77] use prediction boxes of WSOD as the pseudo-ground-truth boxes to train Fast R-CNN. W2F [58] uses prediction boxes of WSOD to train Faster R-CNN. GAL-fWSD [57] uses prediction boxes of WSOD to train SSD.

4.10. Discussions

In the previous sections, we individually introduce several techniques that are commonly used to address discriminative region problems by detailed listing numerous approaches. In this section, in-depth analyses are provided towards the rationality of these techniques.

Firstly, context modeling and discriminative region removal are two similar techniques. Context modeling is to calculate the scores of the proposal and its context region respectively. Then it chooses the positive proposal derived from the two scores. On the other hand, the discriminative region removal is to directly erases top-scoring regions by setting zero value in the feature maps of the first branch followed by feeding the erased feature maps into the second branch.

Secondly, the self-training algorithm usually co-occurs with bounding box regression. Bounding box regression is responsible for refining the initial proposals from SS [88] or EB [89]. And self-training algorithm is designed to refine the prediction result of the baseline. The core problem of both the self-training algorithm and bounding box regression is yielding good pseudo-ground-truth.

Thirdly, the cascaded network and segmentation-detection collaborative mechanism are two very similar techniques. They leverage the segmentation module to improve the performance of the object detection module. Specifically, a cascaded network is a sequential structure that the previous module is responsible for training the latter modulem. The segmentation-detection collaborative mechanism is a circular structure that leverages deep cooperation between detection and segmentation supervising each other to achieve accurate localization.

Fourth, incorporating a low-level features technique leverages the advantage of the high-resolution characteristics of low-level features to improve object localization. Feature consistency encourages achieving the consistency of object features within the same categories, which generates similar responses of the pixels among an object in the activation map for precise object localization. The key idea of transforming WSOD to FSOD technique is to make full use of the advantages of the network structure of the FSOD model (*e.g.*, Fast R-CNN [103]).

Moreover, we further select several representative baselines and summarize the performance gains of each specific technique in Table 2, which is detailed introduced in Section 7.3.2.

5. Specific techniques for multiple-instance and speed problem

5.1. Multiple-instance problem

Since there may exist several proposals corresponding to the same object instance, we need to choose the proposal with the most accurate localization. FSOD models can easily solve the above problem by leveraging score-based Non-Maximum Suppression (NMS) [105] and the instance-level labels. Unfortunately, due to the lack of instance-level labels and avoiding redundant invalid prediction boxes, WSOD models select the highest score proposal of each category as the prediction box and ignore other possible instance proposals. In this section, we will introduce how to make full use of the *spatial relationship* of proposals and the NMS algorithm to solve the multiple-instance problem.

Table 2
Performance gains from the techniques (cf. Section 4) on some typical works on Pascal VOC 2007. 1) Context: Context modeling, 2) Self-t: Self-training algorithm, 3) Cascaded: Cascaded network, 4) BboxR: Bounding box regression, 5) DisRegRem: Discriminative region removal, 6) Low-level: Incorporating low-level features, 7) FeaCon: Feature consistency, 8) Seg-Det: Segmentation-detection collaborative mechanism, 9) Transform: Transforming WSOD to FSOD.

Techniques	Approaches	use technique	mAP (%)	CorLoc (%)	Techniques	Approaches	use technique	mAP (%)	CorLoc (%)
Context	ContextLocNet [48]	X	30.5	50	Seg-Det	WS-JDS [65]	X	37.3	–
		✓	36.3 (+5.8)	55.1 (+5.1)			✓	45.6 (+8.3)	64.5
Self-t	OICR [49]	X	39.3	58.0		SDCN [71]	X	41.2	–
		✓	42.0 (+2.7)	61.2 (+3.2)			✓	50.2 (+9.0)	68.6
	PCL [56]	X	39.3	58.0		OICR [49]	X	42.0	61.2
		✓	45.8 (+6.5)	63.0 (+5.0)			✓	47.0 (+5.0)	64.3 (+3.1)
	ST-WSL [51]	X	29.6	37.9		PCL [56]	X	45.8	63.0
		✓	40.8 (+11.2)	56.1 (+18.2)			✓	48.8 (+3.0)	66.6(+3.6)
	CASD [42]	X	55.3	–		C-WSL [104]	X	45.6	63.3
		✓	56.1 (+0.8)	–			✓	47.8 (+2.2)	65.6 (+2.3)
Cascaded	TS^2C [61]	X	42.0	61.2		WSRPN [63]	X	47.9	66.9
		✓	44.3 (+2.3)	61.0 (−0.2)			✓	50.4 (+2.5)	68.4 (+1.5)
BboxR	TPWSD [70]	X	43.3	61.4		WS-JDS [65]	X	45.6	64.5
		✓	48.6 (+5.3)	66.8 (+5.4)			✓	52.5 (+6.9)	68.6 (+4.1)
DisRegRem	C-MIDN [72]	X	49.0	–		SDCN [71]	X	50.2	68.6
		✓	52.6 (+3.6)	68.7			✓	53.7 (+3.5)	72.5 (+3.9)
Low-level	WSOD2 [68]	X	45.9	–		C-MIDN [72]	X	52.6	68.7
		✓	48.1 (+2.2)	–			✓	53.6 (+1.0)	71.9 (+3.2))
FeaCon	CASD [42]	X	48.9	–		SLV [77]	X	53.5	71.0
		✓	55.3 (+6.4)	–			✓	53.9 (+0.4)	72.0 (+1.0)

ST-WSL [51] advocates replacing the highest score with the greatest degree in the NMS algorithm, which not only leverages a graph network to detect multiple instances with the same category in an image but also alleviates getting stuck into discriminative region problem introduced in Section 2.2. It first chooses N top-scoring proposals of each positive class as the nodes of the graph. The edge between two nodes indicates a large overlap between them. Then it selects the greatest degree (number of connections to other nodes) nodes as positive proposals using the NMS algorithm. PCL [56] introduces the proposal cluster to replace the proposal bag that includes all of the proposals of each category. PCL assigns the same label and spatially adjacent proposals to the same proposal cluster. If proposals do not overlap each other, they will be assigned to different proposal clusters. Then, PCL selects the highest score proposal from each proposal cluster as the positive proposal. Finally, the number of instances detected is equal to the number of proposal clusters. W2F [58] iteratively merges the highly overlapping proposals with top-scoring into big proposals. Finally, these big proposals are considered positive proposals as the final prediction bounding boxes.

Although the above methods deal with the multiple-instance problem in different ways, they are all based on the spatial relationship (*e.g.* overlap) between proposals.

5.2. Speed problem

In this section, we will introduce several advanced techniques for solving the speed problem introduced in Section 2.2. The main reason for the slow speed is that the MIL-based method adopts SS [88], EB [89], and sliding-window are time-consuming and generate a large number of invalid proposals that consume a lot of computational cost of the model.

The methods for improving speed can be broadly categorized into two groups: 1) *Transformation-based* [58,57]: these approaches use their prediction boxes as the pseudo-ground-truth boxes to train the faster and fully supervised model (*e.g.* Faster R-CNN [29] and SSD [30]) and then use the faster model to infer images in inference phase. 2) *Heatmap-based* [47,53,52,55,59,62,66,73,62,76]: these approaches replace SS and EB with segmenting the heatmap using a threshold to generate fewer proposals for improving the speed of proposal generation.

These two techniques have respective strengths and weaknesses: 1) Transformation-based solution relies on training another faster model, which increases the complexity and time of training. 2) Heatmap-based solution can detect objects of any shape by segmenting the heatmap of an image, but it is not easy to accurately detect multiple objects with the same category.

6. Training and test tricks

Besides the techniques in the previous chapter, training and test tricks without changing network structure also can improve detection results. In this section, we will introduce several training and test tricks for improving detection performance.

6.1. Easy-to-hard Strategy

Previous approaches [44,45,47–38,55] use all of the images at once without a training sequence to train the detection model. The easy-to–hard strategy denotes that the model is trained by using the images with progressively increasing difficulty. In this way, the model can gain better detection results. For example, ZLDN [60] first computes the difficulty scores of images. Then, all of the images are ranked in an ascending order using the difficulty scores. Finally, ZLDN uses the images with increasing difficulty to progressively train themselves for more better performance.

6.2. Negative evidence

Negative evidence contains the low-scoring regions, activations, and activation maps. For example, WELDON [47] uses the classification scores of the k top-scoring proposals and the m low-scoring proposals to generate the classification scores of the image by simply summing. WILDCAT [52] leverages the k^+ highest probability activations and k^- lowest probability activations of the activation map to generate the prediction score. NL-CCAM [74] uses the lowest probability activation maps. Specifically, it first ranks all of the activation maps in a descending order based on the probability of every class. Then, it fuses these class activation maps using a combinational function into one map, which is segmented to predict object instances.

6.3. Optimizing smoothed loss functions

If the loss function of the model is non-convex, it tends to fall into sub-optimal and falsely localizes object parts while missing a full object extent during training [64]. So C-MIL [64] replaces the non-convex loss function with a series of smoothed loss functions to alleviate the problem that the model tends to get stuck into local minima. At the beginning of training, C-MIL first performs the image classification task. During the training process, the loss function of C-MIL is slowly transformed from the convex image classification loss to the non-convex object detection loss function.

6.4. Optimizing post-processing

The CAM-based network detects object instances by segmenting the localization map using a threshold in post-processing. However, the localization results are sensitive due to the ambiguity between foreground and background pixels across different activation maps [87].

RethinkingCAM [78] and IVR [86] are two normalization methods that are relatively unaffected from the value shift by original activation map maximum and minimum values while the range is calibrated according to the given values. SPA [84] utilizes the high-order self-correlation (HSC) to extract the inherent structural information and then aggregates HSC of multiple points into the original activation map. Finally, SPA segments the aggregated map to yield bounding boxes for precise object localization. SPOL [85] highlights the gravity area of object instances using a gaussian module in post-processing, which helps locate the area of object gravity and cover larger object region.

6.5. Discussions

In the previous sections, we individually introduce some training and test tricks that are independent of the model structure. In this section, we will compare and discuss these tricks.

Firstly, an easy-to–hard strategy is applied to the data processing phase, which is responsible for adjusting the order of the training images. Secondly, negative evidence acts on the training phase, which is designed to refine positive proposals or feature maps. Thirdly, optimizing smoothed loss functions act on the optimizing phase, which is responsible for avoiding getting stuck into local minima. Finally, optimizing post-processing is a simple yet effective trick for precise object localization, without extra parameters.

7. Datasets and performance evaluation

7.1. Datasets

Datasets play an important role in WSOD task. Most approaches of the WSOD use PASCAL VOC [110], MS COCO [28], ILSVRC [27], or CUB-200 [111] as training, validation and test datasets.

PASCAL VOC. It includes 20 categories and tens of thousands of images with instance annotations. PASCAL VOC has several versions: PASCAL VOC 2007, 2010, and 2012. PASCAL VOC 2007 consists of 2,501 training images, 2,510 validation images, and 4,092 test images. PASCAL VOC 2010 consists of 4,998 training images, 5,105 validation images, and 9,637 test images. PASCAL VOC 2012 consists of 5,717 training images, 5,823 validation images, and 10,991 test images.

MS COCO. It is large-scale object detection, segmentation, and captioning dataset. MS COCO has 80 object categories, 330 K images (>200 K labeled), 1.5 million object instances. In object detection, MS COCO is as popular as PASCAL VOC datasets, but the difficulty of training on the MS COCO dataset is higher than that of PASCAL VOC datasets due to MS COCO having more images and categories.

ILSVRC. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is a large-scale dataset. In ILSVRC, the model usually uses 200 fully labeled categories and 1,000 categories in object detection and object localization, respectively. ILSVRC has several versions, such as ILSVRC 2013, ILSVRC 2014, and ILSVRC 2016. Specifically, ILSVRC 2013 which is usually used in object detection has 12,125 images for training, 20,121 images for validation, and 40,152 images for testing. In addition, ILSVRC 2014 and 2016 inherit the ILSVRC 2012 dataset in object localization, which contains 1.2 million images of 1,000 categories in the training set. And ILSVRC 2012 dataset has 50,000 and 100,000 images with labels in the validation and test set, respectively.

CUB-200. Caltech-UCSD Birds 200 (CUB-200) contains 200 bird species which is a challenging image dataset. It focuses on the study of subordinate categorization. CUB-200–2011 [112] is an extended version of CUB-200, which adds many images for each category and labels new part localization annotations. CUB-200–2011 contains 5,994 images in the training set and 5,794 images in the test set.

7.2. Evaluation metrics

In the state-of-the-art WSOD approaches, there are three standard evaluation metrics: mAP, CorLoc, and top error.

mAP (mean Average Precision). Average Precision (AP) is usually used in image classification and object detection. It consists of precision and recall. If *tp* denotes the number of the correct prediction samples among all of the positive samples, *fp* denotes the number of the wrong prediction samples among all of the positive samples, and *fn* denotes the number of the wrong prediction samples among all of the negative samples, precision and recall can be computed as

$$\begin{aligned} \text{recall} &= tp/(tp+fn), \\ \text{precision} &= tp/(tp+fp), \end{aligned} \tag{1}$$

where the correct prediction sample denotes IoU of the positive sample and its corresponding ground-truth box $\geqslant 0.5$. Meanwhile, the IoU is defined as

$$\text{IoU}(b, b^g) = area(b \cap b^g)/area(b \cup b^g), \tag{2}$$

where b denotes a prediction sample, b^g denotes a corresponding ground-truth box, and *area* denotes the region size of the intersection or union. The mAP is the mean of all of the class average precisions and is a final evaluation metric of performance on the test dataset.

CorLoc (Correct Localization). CorLoc denotes the percentage of images that exist at least one instance of the prediction boxes whose IoU $\geqslant 50\%$ with ground-truth boxes for every class in these images. CorLoc is a final evaluation metric of localization accuracy on the trainval dataset.

Top Error. Top error consists of Top-1 classification error (1-err cls), Top-5 classification error (5-err cls), Top-1 localization error (1-err loc), and Top-5 localization error (5-err loc). Specifically, Top-1 classification error is equal to $1.0 - cls_1$, where cls_1 denotes the accuracy of the highest prediction score (likewise for Top-1 localization error). Top-5 classification error is equal to $1.0 - cls_5$, where cls_5 denotes that it counts as correct if one of the five predictions with the highest score is correct (likewise for Top-5 localization error). Numerous approaches [33,59,74,73,74] use top error to evaluate the performance of the model.

7.3. Experimental results

7.3.1. Comparison to state-of-the-arts

Results on Pascal VOC. The results of state-of-the-art WSOD methods on datasets Pascal VOC 2007, 2010, and 2012 are shown in Table 3. The WSOD methods with "+FR" denote that their initial predictions are fed into the Fast R-CNN [103] and serve as pseudo-ground-truth bounding box annotations to train Fast R-CNN, *i.e.*, these methods transform the WSOD into FSOD problems. From the results, we can observe the performance on all three Pascal VOC datasets have achieved unprecedented progress in recent few years (*e.g.*, mAP 53.6% in NeurIPS'20 vs. 29.1% in CVPR'16 on Pascal VOC 2012). Comparing the methods and their counterparts with Fast R-CNN (*e.g.*, OICR vs. OICR + FR), the detection performance is further improved by using this FSOD transforming strategy.

Results on MS COCO. The results of state-of-the-art WSOD methods on dataset MS COCO are shown in Table 4. We only report the AP metric, and the AP@0.5 denotes that the IoU threshold is equal to 0.5. Similarly, the performance on MS COCO also doubled in the last few years (*e.g.*, AP@0.5 11.5% vs. 24.8% in test set). Since MS COCO contains more object categories than PASCAL VOC datasets, the results on MS COCO are still far from satisfactory. However, the performance gains by transforming WSOD to FSOD are relatively marginal (*e.g.*, 0.7% gains in AP for PCL model).

Results on ILSVRC 2020 and CUB-200. Table 5 summaries the object localization performance of state-of-the-art WSOD methods on these two datasets. From Table 5, we can find the performance gains are also significant (1-err cls 35.6% vs. 21.9% and 1-err loc 57.8% vs. 40.9% in ILSVRC 2012).

7.3.2. Ablation study

To better understand the effectiveness of the techniques (cf. Section 4) on some typical works, we summarize several ablation studies on Pascal VOC 2007. The results of our ablation studies are illustrated in Table 2. We observe that context modeling brings an extra 5.8% and 5.1% performance gains on the mAP and CorLoc, respectively. Cascaded network using multiple different modules improves 2.3% mAP over baseline. Bounding box regression brings performance gain of 5.4% on the CorLoc. Incorporating low-level features and feature consistency is improved by 2.2% and 6.4% on the mAP, respectively. In addition, self-training, segmentation-detection collaborative mechanism, transforming WSOD to FSOD can give an average improvement of 5.3%, 8.7%, and 3.1% on the mAP, respectively. At the same time, self-training and transforming WSOD to FSOD have brought average improvements of 8.8% and 2.8% on the CorLoc, respectively.

8. Future directions and tasks

Although we have summarized many advanced techniques and tricks for improving detection results, there are still several research directions that can be further explored.

Table 3
The summary of detection results (mAP (%) and CorLoc (%)) of state-of-the-art WSOD methods on Pascal VOC 2007, 2010, and 2012 datasets. The FR means Fast R-CNN [103].

Approach	2007		2010		2012	
	mAP	CorLoc	mAP	CorLoc	mAP	CorLoc
WSDDN [44]$_{CVPR2016}$	39.3	58.0	36.2	59.7	–	–
WSLPDA [46]$_{CVPR2016}$	39.5	52.4	30.7	–	29.1	–
ContextLocNet [48]$_{ECCV2016}$	36.3	55.1	–	–	35.3	54.8
OICR [49]$_{CVPR2017}$	42.0	61.2	–	–	38.2	63.5
WCCN [50]$_{CVPR2017}$	42.8	56.9	39.5	–	37.9	–
ST-WSL [51]$_{CVPR2017}$	41.7	56.1	–	–	39.0	58.8
SPN [54]$_{ICCV2017}$	–	60.6	–	–	–	–
TST [106]$_{ICCV2017}$	34.5	60.8	–	–	–	–
PCL [56]$_{TPAMI2018}$	45.8	63.0	–	–	41.6	65.0
GAL-fWSD [57]$_{CVPR2018}$	47.0	68.1	**45.1**	**68.3**	43.1	67.2
W2F [58]$_{CVPR2018}$	52.4	70.3	–	–	47.8	69.4
ZLDN [60]$_{CVPR2018}$	47.6	61.2	–	–	42.9	61.5
MELM [107]$_{CVPR2018}$	47.3	61.4	–	–	42.4	–
TS^{2}C [61]$_{ECCV2018}$	44.3	61.0	–	–	40.0	64.4
C-WSL [104]$_{ECCV2018}$	45.6	63.3	–	–	43.0	64.9
WSRPN [63]$_{ECCV2018}$	47.9	66.9	–	–	43.4	67.2
C-MIL [64]$_{CVPR2019}$	40.7	59.5	–	–	46.7	67.4
WS-JDS [65]$_{CVPR2019}$	45.6	64.5	39.9	63.1	39.1	63.5
Pred NET [67]$_{CVPR2019}$	53.6	**71.4**	–	–	49.5	70.2
WSOD2 [68]$_{ICCV2019}$	53.6	69.5	–	–	47.2	71.9
OAILWSD [69]$_{ICCV2019}$	47.6	66.7	–	–	43.4	66.7
TPWSD [70]$_{ICCV2019}$	51.5	68.0	–	–	45.6	68.7
SDCN [71]$_{ICCV2019}$	50.2	68.6	–	–	43.5	67.9
C-MIDN [72]$_{ICCV2019}$	52.6	68.7	–	–	50.2	71.2
ICMWSD [75]$_{CVPR2020}$	54.9	68.8	–	–	52.1	70.9
SLV [77]$_{CVPR2020}$	53.5	71.0	–	–	49.2	69.2
DRN-WSOD [96]$_{ECCV2020}$	52.8	70.1	–	–	51.1	**73.2**
UWSOD [80]$_{NeurIPS2020}$	45.0	63.8	–	–	46.2	65.7
CASD [42]$_{NeurIPS2020}$	**56.8**	–	–	–	**53.6**	–
OICR [49]$_{CVPR2017}$+FR	47.0	64.3	–	–	42.5	65.6
PCL [56]$_{TPAMI2018}$+FR	48.8	66.6	–	–	44.2	68.0
MEFF [108]$_{CVPR2018}$+FR	51.2	–	–	–	–	–
C-WSL [104]$_{ECCV2018}$+FR	47.8	65.6	–	–	45.4	66.9
WSRPN [63]$_{ECCV2018}$+FR	50.4	68.4	–	–	45.7	69.3
WS–JDS [65]$_{CVPR2019}$+FR	52.5	68.6	**45.7**	**68.1**	46.1	69.5
SDCN [71]$_{ICCV2019}$+FR	53.7	**72.5**	–	–	46.7	69.5
C-MIDN [72]$_{ICCV2019}$+FR	53.6	71.9	–	–	**50.3**	**73.3**
SLV [77]$_{CVPR2020}$+FR	**53.9**	72.0	–	–	–	–

Table 4
Detection results on MS COCO dataset comes from [75]. These models use VGG16 as their convolutional neural network. There is no difference between mAP@[.5,.95] and mAP@0.5 under the MS COCO context.

Approach	Val		Test	
	@[.5,.95]	@0.5	@[.5,.95]	@0.5
WSDDN [44]$_{CVPR2016}$	–	–	–	11.5
WCCN [50]$_{CVPR2017}$	–	–	–	12.3
PCL [56]$_{TRAMI2018}$	8.5	19.4	–	–
C-MIDN [72]$_{ICCV2019}$	9.6	21.4	–	–
WSOD2 [68]$_{ICCV2019}$	10.8	22.7	–	–
ICMWSD [75]$_{CVPR2020}$	11.4	24.3	**12.1**	**24.8**
UWSOD [80]$_{NeurIPS2020}$	3.1	10.1	–	–
CASD [42]$_{NeurIPS2020}$	**13.9**	**27.8**	–	–
Diba et al. [109]$_{arXiv\ 2017}$+SSD [30]	–	–	–	13.6
OICR [49]$_{CVPR2017}$+FR [103]	7.7	17.4	–	–
MEFF [108]$_{CVPR2018}$+FR [103]	8.9	19.3	–	–
PCL [56]$_{TPAMI2018}$+FR [103]	9.2	19.6	–	–

Table 5
Object localization performance on ILSVRC 2012 and CUB-200–2011 datasets.

Approach	ILSVRC 2012 (top error %)				CUB-200-2011 (top error %)			
	1-err cls	5-err cls	1-err loc	5-err loc	1-err cls	5-err cls	1-err loc	5-err loc
CAM [45]$_{CVPR2016}$	35.6	13.9	57.8	45.3	–	–	–	–
ACoL [59]$_{CVPR2018}$	32.5	12.0	54.2	36.7	–	–	54.1	39.1
SPG [62]$_{ECCV2018}$	–	–	51.4	35.1	–	–	53.4	40.6
DANet [73]$_{ICCV2019}$	32.5	12.0	54.2	40.6	24.6	7.7	47.5	38.0
NL-CCAM [74]$_{WACV2020}$	27.7	–	49.8	39.3	26.6	–	47.6	35.0
EIL [76]$_{CVPR2020}$	29.7	–	53.2	–	25.2	–	42.5	–
RethinkingCAM [78]$_{ECCV2020}$	24.2	–	49.4	–	25.0	–	40.5	–
GC-Net [113]$_{ECCV2020}$	22.6	**6.4**	50.9	41.9	23.2	7.7	36.8	24.5
I^2C [79]$_{ECCV2020}$	23.3	6.9	45.2	35.4	–	–	44.0	31.7
MCIR [81]$_{WACV2021}$	28.8	–	47.6	–	**22.7**	–	35.3	–
CI-CAM [82]$_{ACMMM2021}$	27.4	–	51.3	–	24.4	–	41.6	–
SLT-Net [83]$_{CVPR2021}$	**21.9**	–	44.3	34.6	23.4	–	32.2	–
SPA [84]$_{CVPR2021}$	–	–	47.3	35.7	–	–	39.7	27.5
SPOL [85]$_{CVPR2021}$	–	–	**40.9**	32.9	–	–	**19.9**	**6.6**
FAM [114]$_{ICCV2021}$	–	–	44.8	**22.4**	–	–	29.3	18.8
ORNet [87]$_{ICCV2021}$	28.4	9.6	48.0	36.1	23.0	**7.0**	32.3	19.2
TS-CAM [98]$_{ICCV2021}$	–	–	46.6	35.7	–	–	28.7	16.2

8.1. Model directions

Better Initial Proposals. The main proposal generators of the existing methods are selective search [88], edge boxes [89], heatmap, and sliding window. Selective search and edge boxes are too time-consuming and yield plenty of initial proposals that most of which are invalid proposals. Segmenting heatmap tends to focus on the discriminative part of the object. The performance of the sliding window is strongly dependent on the size of objects. For example, if the size of the object instance is roughly fixed, then the sliding window works very well. Otherwise, it works badly. Because these generators have inherent disadvantages, we need to design a proposal generator that can yield fewer and more accurate initial proposals. The quality of the initial proposals directly affects the detection performance of the detector. So how to yield good initial proposals may be a new research direction.

Better Positive Proposals. Most WSOD methods select the proposals with the top score as positive proposals, which tend to focus on the most discriminative parts of the object rather than the whole object region. Because of the above problem, ST-WSL [51] selects positive proposals relying on the relative improvement (RI) of the scores of each proposal of two adjacent epochs and its surrounding proposals. Besides, the key of self-training and cascaded networks is to select accurate proposals as the pseudo-ground-truth boxes for later training. Thus, how can we design a better algorithm that can accurately select positive proposals may be an important research direction.

Lightweight Network. Today's state-of-the-art object detectors [26,33] leverage a very deep CNN to extract image feature maps and high-dimension fully connected layers to detect object instances that can achieve satisfactory detection performance. But the deep CNN and high-dimension fully connected layers rely on large memory and strong GPU computation power. Hence, a deep network is difficult to deploy on CPU devices (*e.g.*, mobile phones). With the popularity of mobile devices, lightweight network with few parameters has received more and more attention from researchers, such as Light-Head R-CNN [115]. Thus, a lightweight network in weakly supervised object detection may be a potential research direction.

8.2. Application directions

Medical Imaging. With the development of deep learning, it has evolved into cross-learning with multiple disciplines, especially the medical field. Because of lacking brain's Magnetic Resonance Imaging (MRI) and X-rays images with sufficient labels, weakly-supervised brain lesion detection [103,117] has received attention from researchers. The purpose of weakly-supervised brain lesion detection is to give the model the ability to accurately locate lesion regions and classify lesion categories that help the doctor complete the diagnosis of the disease. Weakly-supervised lesion detection is not only applied in brain disease, but also other organ diseases, such as the chest, abdomen, and pelvis. In addition to lesion detection, weakly-supervised learning is applied in dis-

ease prognosis [118]. During hospital visits, patients need to undergo a large number of tests to pinpoint their disease. These tests are generally presented to doctors and patients in the form of graphic reports. However, these numerous graphic reports lack correct labeling information. So, medical imaging may be another potential and meaningful research direction in a weakly supervised setting.

3D Object Detection. In recent years, with the continuous improvement of the accuracy of object detection of images 29–35,119–121], 3D object detection [122–125] has received unprecedented attention. The purpose of 3D object detection is to detect object instances in the point cloud using 3D bounding boxes. Comparing with 2D object detection, 3D object detection tends to cost more computation and its supervision is more difficult to obtain and labor-intensive. Therefore, how to train light and accurate 3D detection models in the point cloud using simple labels may be a big challenge. Fortunately, weakly-supervised object detection is successfully applied in 2D object detection. According to the above analysis, we think that 3D weakly-supervised object detection that uses weak supervision(*e.g.*, 2D bounding boxes and text labels) to train object detection models in the 3D scene may be a hot research direction.

Video Object Detection. Video object detection [126,127] is to classify and locate object instances in a piece of video. One of the solutions is to break the video into many frames and the detector achieves object detection in these frame images [128,129]. However, the detector will suffer from one big problem that the quality of these frame images has deteriorated. To improve the performance of video object detection, expanding the training dataset is a good approach. Unfortunately, tagging object location and cat-

Table 6
A list of current methods using these specific techniques for discriminative region problem as well as training and test tricks. 1) Context: Context modeling, 2) Self-t: Self-training algorithm, 3) Cascaded: Cascaded network, 4) BboxR: Bounding box regression, 5) DisRegRem: Discriminative region removal, 6) Low-level: Incorporating low-level features, 7) FeaCon: Feature consistency, 8) Seg-Det: Segmentation-detection collaborative mechanism, 9) Transform: Transforming WSOD to FSOD, 10) Easy-hard: Easy-to–hard strategy, 11) NegEvi: Negative evidence, 12) SmoLoss: Optimizing smoothed loss functions, 13) Post: Optimizing post-processing.

Approach	Specific techniques for discriminative region problem									Training and test tricks			
	Context	Self-t	Cascaded	BboxR	DisRegRem	Low-level	FeaCon	Seg-Det	Transform	Easy-hard	NegEvi	SmoLoss	Post
WSDDN [44]$_{CVPR2016}$													
CAM [45]$_{CVPR2016}$													
WSLPDA [46]$_{CVPR2016}$	√												
WELDON [47]$_{CVPR2016}$											√		
ContextLocNet [48]$_{ECCV2016}$	√												
OICR [49]$_{CVPR2017}$		√							√				
WCCN [50]$_{CVPR2017}$			√										
ST-WSL [51]$_{CVPR2017}$		√							√				
WILDCAT [52]$_{CVPR2017}$											√		
Grad-CAM [53]$_{ICCV2017}$						√							
SPN [54]$_{ICCV2017}$													
TP-WSL [55]$_{ICCV2017}$					√								
PCL [56]$_{TPAMI2018}$		√							√				
GAL-fWSD [57]$_{CVPR2018}$									√				
W2F [58]$_{CVPR2018}$		√							√				
ACoL [59]$_{CVPR2018}$					√								
ZLDN [60]$_{CVPR2018}$										√			
TS^{2}C [61]$_{ECCV2018}$	√		√						√				
SPG [62]$_{ECCV2018}$													
WSRPN [63]$_{ECCV2018}$													
C-MIL [64]$_{CVPR2019}$												√	
WS-JDS [65]$_{CVPR2019}$								√	√				
ADL [66]$_{CVPR2019}$													
Pred NET [67]$_{CVPR2019}$				√					√				
WSOD2 [68]$_{ICCV2019}$		√		√		√							
OAILWSD [69]$_{ICCV2019}$	√	√											
TPWSD [70]$_{ICCV2019}$		√		√									
SDCN [71]$_{ICCV2019}$								√	√				
C-MIDN [72]$_{ICCV2019}$		√			√				√				
DANet [73]$_{ICCV2019}$													
NL-CCAM [74]$_{WACV2020}$											√		
ICMWSD [75]$_{CVPR2020}$	√	√		√									
EIL [76]$_{CVPR2020}$					√								
SLV [77]$_{CVPR2020}$		√		√					√				
RethinkingCAM [78]$_{ECCV2020}$													√
GC-Net [113]$_{ECCV2020}$													
I^{2}C [79]$_{ECCV2020}$							√						
UWSOD [80]$_{NeurIPS2020}$		√		√									
CASD [42]$_{NeurIPS2020}$		√					√						
MCIR [81]$_{WACV2021}$					√								
CI-CAM [82]$_{ACMMM2021}$													
SLT-Net [83]$_{CVPR2021}$			√	√									
SPA [84]$_{CVPR2021}$													√
SPOL [85]$_{CVPR2021}$			√			√							√
IVR [86]$_{ICCV2021}$													√
ORNet [87]$_{ICCV2021}$					√	√							

egory in videos is much more difficult than in 2D images. Therefore, training video object detection in the weakly supervised setting is necessary.

Cross-domain WSOD. Cross-domain weakly supervised object detection methods [130–133] learn transferable knowledge from the source dataset and use it to speed up learning categories in the weakly supervised target dataset [130]. The source dataset is a fully supervised dataset, whose categories may overlap with the weakly supervised target dataset. At present, some researchers [130–132] consider knowledge transfer from extra data to improve the localization performance of WSOD. For example, Zhong et al. [131] train a generic object detector on a fully supervised source dataset and apply it to weakly supervised target dataset. However, they ignore the category information in the source dataset and degrade the accuracy of classification [129,130]. Thus, how to make full use of knowledge transfer from the fully supervised dataset to WSOD requires continuous explorations and attentions.

9. Conclusions

In this paper, we summarize plenty of the deep learning WSOD methods and give a lot of solutions to solve the above challenges. In summary, the main contents of this paper are listed as follows.

- We analyze the background, and main challenges, and basic framework of WSOD. Furthermore, we introduce several landmark methods in detail.
- For main challenges, we analyze almost all of the WSOD methods since 2016 and summarize numerous techniques and training and test tricks (cf. Table 6).
- We introduce currently popular datasets and important evaluation metrics in the WSOD task.
- We conclude and discuss valuable insights and guidelines for future progress in model and application directions.

CRediT authorship contribution statement

Feifei Shao: Writing – original draft. **Long Chen:** Conceptualization, Writing – review & editing, Supervision. **Jian Shao:** Writing – review & editing. **Wei Ji:** Writing – review & editing. **Shaoning Xiao:** Writing – review & editing. **Lu Ye:** Writing – review & editing. **Yueting Zhuang:** Funding acquisition. **Jun Xiao:** Funding acquisition, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

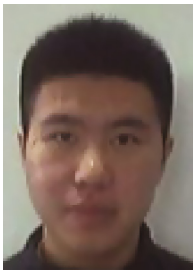
This work was supported by the National Key Research & Development Project of China (2021ZD0110700), the National Natural Science Foundation of China (U19B2043, 61976185), Zhejiang Natural Science Foundation (LR19F020002), Zhejiang Innovation Foundation(2019R52002), and the Fundamental Research Funds for the Central Universities.

References

[1] T. Le, N.T. Huy, N.M. Le, Multi visual and textual embedding on visual question answering for blind people, Neurocomputing..
[2] J. Hong, S. Park, H. Byun, Selective residual learning for visual question answering, Neurocomputing..
[3] J. Hong, J. Fu, Y. Uh, T. Mei, H. Byun, Exploiting hierarchical visual features for visual question answering, Neurocomputing..
[4] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, Y. Zhuang, Counterfactual samples synthesizing for robust visual question answering, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020).
[5] L. Chen, Y. Zheng, Y. Niu, H. Zhang, J. Xiao, Counterfactual samples synthesizing and training for robust visual question answering, arXiv preprint arXiv:2110.01013 (2021).
[6] J. Yang, Y. Sun, J. Liang, B. Ren, S.-H. Lai, Image captioning by incorporating affective concepts learned from both visual and textual components, Neurocomputing..
[7] S. Ding, S. Qu, Y. Xi, S. Wan, Stimulus-driven and concept-driven analysis for image caption generation, Neurocomputing..
[8] X. He, Y. Yang, B. Shi, X. Bai, Vd-san: Visual-densely semantic attention network for image caption generation, Neurocomputing..
[9] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T-S. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, Proceedings of the IEEE conference on computer vision and pattern recognition (2017).
[10] L. Chen, Z. Jiang, J. Xiao, W. Liu, Human-like Controllable Image Captioning with Verb-specific Semantic Roles, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021).
[11] L. Chen, C. Lu, S. Tang, J. Xiao, D. Zhang, C. Tan, X. Li, Rethinking the bottom-up framework for query-based video localization, Proceedings of the AAAI Conference on Artificial Intelligence (2020).
[12] C. Lu, L. Chen, C. Tan, X. Li, J. Xiao, DEBUG: A dense bottom-up grounding approach for natural language video localization, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019).
[13] S. Xiao, L. Chen, S. Zhang, W. Ji, J. Shao, L. Ye, J. Xiao, Boundary Proposal Network for Two-Stage Natural Language Video Localization, Proceedings of the AAAI Conference on Artificial Intelligence (2021).
[14] Y. Yuan, X. Lan, L. Chen, W. Liu, X. Wang, W. Zhu, A Closer Look at Temporal Sentence Grounding in Videos: Datasets and Metrics, arXiv preprint arXiv:2101.09028 (2021).
[15] L. Chen, W. Ma, J. Xiao, H. Zhang, S-F. Chang, Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding, Proceedings of the AAAI Conference on Artificial Intelligence (2021).
[16] M. Cao, L. Chen, M-Z. Shou, C. Zhang, Y. Zou, On pursuit of designing multi-modal transformer for video grounding, The 2021 Conference on Empirical Methods in Natural Language Processing (2021).
[17] S. Xiao, L. Chen, J. Shao, Y. Zhuang, J. Xiao, Natural language video localization with learnable moment proposals, The 2021 Conference on Empirical Methods in Natural Language Processing (2021).
[18] X. Yang, J. Dong, Y. Cao, X. Wang, M. Wang, T-S. Chua, Tree-augmented cross-modal encoding for complex-query video retrieval, Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020).
[19] X. Yang, F. Feng, W. Ji, M. Wang, T-S. Chua, Deconfounded Video Moment Retrieval with Causal Intervention, Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021).
[20] Z.-Q. Zhao, P. Zheng, S.-T. Xu, X. Wu, Object detection with deep learning: A review, IEEE transactions on neural networks and learning systems..
[21] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, Deep learning for generic object detection: A survey, International Journal of Computer Vision..
[22] T. Chen, X. Hu, J. Xiao, G. Zhang, S. Wang, Binet: Bidirectional interactive network for salient object detection, Neurocomputing..
[23] X. Fang, J. Zhu, R. Zhang, X. Shao, H. Wang, Ibnet: Interactive branch network for salient object detection, Neurocomputing..
[24] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556..
[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015.
[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.
[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision..
[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, 2014..
[29] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: IEEE transactions on pattern analysis and machine intelligence, 2015.
[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, 2016..
[31] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.
[32] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017..

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.
[34] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, European Conference on Computer Vision (2018).
[35] D. Zhang, J. Han, G. Cheng, M.-H. Yang, Weakly supervised object localization and detection: A survey, IEEE transactions on pattern analysis and machine intelligence..
[36] W. Jiang, Z. Zhao, F. Su, Y. Fang, Dynamic proposal sampling for weakly supervised object detection, Neurocomputing..
[37] D. Zhao, Z. Yuan, Z. Shi, F. Xie, Single-shot weakly-supervised object detection guided by empirical saliency model, Neurocomputing..
[38] L. Zhang, H. Yang, Adaptive attention augmentor for weakly supervised object localization, Neurocomputing..
[39] S.N. Benassou, W. Shi, F. Jiang, Entropy guided adversarial model for weakly supervised object localization, Neurocomputing..
[40] F. Shao, Y. Luo, L. Zhang, L. Ye, S. Tang, Y. Yang, J. Xiao, Improving Weakly-supervised Object Localization via Causal Intervention, ACM International Conference on Multimedia (2021).
[41] B. Singh, M. Najibi, L.S. Davis, Sniper: Efficient multi-scale training, in: arXiv preprint arXiv:1805.09300, 2018..
[42] Z. Huang, Y. Zou, V. Bhagavatula, D. Huang, Comprehensive attention self-distillation for weakly-supervised object detection, arXiv preprint arXiv:2010.12023..
[43] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge 2007 (voc2007) results..
[44] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016..
[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.
[46] D. Li, J.-B. Huang, Y. Li, S. Wang, M.-H. Yang, Weakly supervised object localization with progressive domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.
[47] T. Durand, N. Thome, M. Cord, Weldon: Weakly supervised learning of deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.
[48] V. Kantorov, M. Oquab, M. Cho, I. Laptev, Contextlocnet: Context-aware deep network models for weakly supervised localization, European Conference on Computer Vision (2016).
[49] P. Tang, X. Wang, X. Bai, W. Liu, Multiple instance detection network with online instance classifier refinement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.
[50] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, L. Van Gool, Weakly supervised cascaded convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.
[51] Z. Jie, Y. Wei, X. Jin, J. Feng, W. Liu, Deep self-taught learning for weakly supervised object localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.
[52] T. Durand, T. Mordan, N. Thome, M. Cord, Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.
[53] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.
[54] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, J. Jiao, Soft proposal networks for weakly supervised object localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.
[55] D. Kim, D. Cho, D. Yoo, I. So Kweon, Two-phase learning for weakly supervised object localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.
[56] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, A. Yuille, Pcl: Proposal cluster learning for weakly supervised object detection, IEEE transactions on pattern analysis and machine intelligence..
[57] Y. Shen, R. Ji, S. Zhang, W. Zuo, Y. Wang, Generative adversarial learning towards fast weakly supervised detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[58] Y. Zhang, Y. Bai, M. Ding, Y. Li, B. Ghanem, W2f: A weakly-supervised to fully-supervised framework for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[59] X. Zhang, Y. Wei, J. Feng, Y. Yang, T.S. Huang, Adversarial complementary learning for weakly supervised object localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[60] X. Zhang, J. Feng, H. Xiong, Q. Tian, Zigzag learning for weakly supervised object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[61] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, T. Huang, Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection, in: European Conference on Computer Vision, 2018..
[62] X. Zhang, Y. Wei, G. Kang, Y. Yang, T. Huang, Self-produced guidance for weakly-supervised object localization, European Conference on Computer Vision (2018).
[63] P. Tang, X. Wang, A. Wang, Y. Yan, W. Liu, J. Huang, A. Yuille, Weakly supervised region proposal network and object detection, European Conference on Computer Vision (2018).
[64] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, Q. Ye, C-mil: Continuation multiple instance learning for weakly supervised object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[65] Y. Shen, R. Ji, Y. Wang, Y. Wu, L. Cao, Cyclic guidance for weakly supervised joint detection and segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[66] J. Choe, H. Shim, Attention-based dropout layer for weakly supervised object localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[67] A. Arun, C. Jawahar, M.P. Kumar, Dissimilarity coefficient based weakly supervised object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[68] Z. Zeng, B. Liu, J. Fu, H. Chao, L. Zhang, Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
[69] S. Kosugi, T. Yamasaki, K. Aizawa, Object-aware instance labeling for weakly supervised object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
[70] K. Yang, D. Li, Y. Dou, Towards precise end-to-end weakly supervised object detection network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
[71] X. Li, M. Kan, S. Shan, X. Chen, Weakly supervised object detection with segmentation collaboration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
[72] Y. Gao, B. Liu, N. Guo, X. Ye, F. Wan, H. You, D. Fan, C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
[73] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, Q. Ye, Danet: Divergent activation for weakly supervised object localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
[74] S. Yang, Y. Kim, Y. Kim, C. Kim, Combinational class activation maps for weakly supervised object localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020.
[75] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, Y.J. Lee, A.G. Schwing, J. Kautz, Instance-aware, context-focused, and memory-efficient weakly supervised object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[76] J. Mai, M. Yang, W. Luo, Erasing integrated learning: A simple yet effective approach for weakly supervised object localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[77] Z. Chen, Z. Fu, R. Jiang, Y. Chen, X.-S. Hua, Slv: Spatial likelihood voting for weakly supervised object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[78] W. Bae, J. Noh, G. Kim, Rethinking class activation mapping for weakly supervised object localization, European Conference on Computer Vision (2020).
[79] X. Zhang, Y. Wei, Y. Yang, Inter-image communication for weakly supervised localization, European Conference on Computer Vision (2020).
[80] Y. Shen, R. Ji, Z. Chen, Y. Wu, F. Huang, Uwsod: Toward fully-supervised-level capacity weakly supervised object detection, Advances in Neural Information Processing Systems..
[81] S. Babar, S. Das, Where to look?: Mining complementary image regions for weakly supervised object localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021.
[82] F. Shao, Y. Luo, L. Zhang, L. Ye, S. Tang, Y. Yang, J. Xiao, Improving Weakly Supervised Object Localization via Causal Intervention (2021).
[83] G. Guo, J. Han, F. Wan, D. Zhang, Strengthen learning tolerance for weakly supervised object localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[84] X. Pan, Y. Gao, Z. Lin, F. Tang, W. Dong, H. Yuan, F. Huang, C. Xu, Unveiling the potential of structure preserving for weakly supervised object localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[85] J. Wei, Q. Wang, Z. Li, S. Wang, S.K. Zhou, S. Cui, Shallow feature matters for weakly supervised object localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[86] J. Kim, J. Choe, S. Yun, N. Kwak, Normalization matters in weakly supervised object localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[87] J. Xie, C. Luo, X. Zhu, Z. Jin, W. Lu, L. Shen, Online refinement of low-level feature based activation map for weakly supervised object localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[88] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, International Journal of Computer Vision..
[89] C.L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, European Conference on Computer Vision (2014).
[90] J. Hosang, R. Benenson, B. Schiele, How good are detection proposals, really?, arXiv preprint arXiv:1406.6962..

[91] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, arXiv preprint arXiv:1312.6229..
[92] T.G. Dietterich, R.H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artificial intelligence..
[93] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012..
[94] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.
[95] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[96] Y. Shen, R. Ji, Y. Wang, Z. Chen, F. Zheng, F. Huang, Y. Wu, Enabling deep residual networks for weakly supervised object detection, in: European Conference on Computer Vision, 2020..
[97] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE transactions on pattern analysis and machine intelligence..
[98] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, Q. Ye, Ts-cam: Token semantic coupled attention map for weakly supervised object localization, arXiv preprint arXiv:2103.14862..
[99] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929..
[100] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, 2021..
[101] W. Wang, L. Yao, L. Chen, B. Lin, D. Cai, X. He, W. Liu, CrossFormer: A Versatile Vision Transformer Hinging on Cross-scale Attention, The International Conference on Learning Representations (2022).
[102] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806..
[103] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2015..
[104] M. Gao, A. Li, R. Yu, V.I. Morariu, L.S. Davis, C-wsl: Count-guided weakly supervised localization, European Conference on Computer Vision (2018).
[105] A. Neubeck, L. Van Gool, Efficient non-maximum suppression, in: 18th International Conference on Pattern Recognition, 2006.
[106] M. Shi, H. Caesar, V. Ferrari, Weakly supervised object localization using things and stuff transfer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.
[107] F. Wan, P. Wei, J. Jiao, Z. Han, Q. Ye, Min-entropy latent model for weakly supervised object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[108] W. Ge, S. Yang, Y. Yu, Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[109] A. Diba, V. Sharma, R. Stiefelhagen, L. Van Gool, Object discovery by generative adversarial & ranking networks, arXiv preprint arXiv:1711.08174..
[110] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International Journal of Computer Vision..
[111] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, Caltech-ucsd birds 200..
[112] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset..
[113] W. Lu, X. Jia, W. Xie, L. Shen, Y. Zhou, J. Duan, Geometry constrained weakly supervised object localization, European Conference on Computer Vision (2020).
[114] M. Meng, T. Zhang, Q. Tian, Y. Zhang, F. Wu, Foreground activation maps for weakly supervised object localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[115] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, J. Sun, Light-head r-cnn: In defense of two-stage object detector, arXiv preprint arXiv:1711.07264..
[116] K. Wu, B. Du, M. Luo, H. Wen, Y. Shen, J. Feng, Weakly supervised brain lesion segmentation via attentional representation learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019.
[117] Z. Ji, Y. Shen, C. Ma, M. Gao, Scribble-based hierarchical weakly supervised learning for brain tumor segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019.
[118] M. Liu, J. Zhang, C. Lian, D. Shen, Weakly supervised deep learning for brain disease prognosis using mri and incomplete clinical scores, IEEE transactions on cybernetics..
[119] Q. Ren, S. Lu, J. Zhang, R. Hu, Salient object detection by fusing local and global contexts, IEEE Transactions on Multimedia..
[120] S. Wu, Y. Xu, B. Zhang, J. Yang, D. Zhang, Deformable template network (dtn) for object detection, IEEE Transactions on Multimedia..
[121] C. Deng, M. Wang, L. Liu, Y. Liu, Y. Jiang, Extended feature pyramid network for small object detection, IEEE Transactions on Multimedia..
[122] X. Chen, H. Ma, J. Wan, B. Li, T. Xia, Multi-view 3d object detection network for autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.
[123] Z. Yang, Y. Sun, S. Liu, X. Shen, J. Jia, in: Std: Sparse-to-dense 3d object detector for point cloud, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
[124] J. Chen, B. Lei, Q. Song, H. Ying, D.Z. Chen, J. Wu, A hierarchical graph network for 3d object detection on point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[125] W. Shi, R. Rajkumar, Point-gnn: Graph neural network for 3d object detection in a point cloud, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[126] F. Xiao, Y. Jae Lee, Video object detection with an aligned spatial-temporal memory, European Conference on Computer Vision (2018).
[127] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, T. Mei, Single shot video object detector, IEEE Transactions on Multimedia..
[128] H. Deng, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, H. Guan, Object guided external memory network for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
[129] Y. Chen, Y. Cao, H. Hu, L. Wang, Memory enhanced global-local aggregation for video object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[130] A. Rahimi, A. Shaban, T. Ajanthan, R. Hartley, B. Boots, Pairwise similarity knowledge transfer for weakly supervised object localization, in: European conference on computer vision, 2020.
[131] Y. Zhong, J. Wang, J. Peng, L. Zhang, Boosting weakly supervised object detection with progressive knowledge transfer, in: European conference on computer vision, 2020.
[132] T. Cao, L. Du, X. Zhang, S. Chen, Y. Zhang, Y.-F. Wang, Cat: Weakly supervised object detection with category transfer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[133] B. Dong, Z. Huang, Y. Guo, Q. Wang, Z. Niu, W. Zuo, Boosting weakly supervised object detection via learning bounding box adjusters, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.



Feifei Shao received a bachelor's degree in software engineering from Zhejiang University of Technology, China in 2016. He is currently working toward a master's degree at Zhejiang University, China. His research interests include computer vision and deep learning.

Long Chen is currently a postdoctoral research scientist at Columbia University, New York, USA. He was a senior researcher at Tencent AI Lab, Shenzhen, China. He received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2020, and the B.Eng degree in electrical information engineering from Dalian University of Technology, Dalian, China, in 2015. His research interests are computer vision, machine learning, and multimedia.

Jian Shao received the Ph.D. degree in signal and information processing from Institute of Acoustics, Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor with the College of Computer Science and Technology, Zhejiang University. His research interests include unstructured data management, cross media computing, and cognitive decision service.

Wei Ji received the B.S. degree from Nanjing University of Science and Technology, and the Ph.D. degree from Zhejiang University. He is currently a research fellow with the School of Computing, National University of Singapore. His current research interests include multimedia analysis, computer vision, and machine learning.

Shaoning Xiao received bachelor's degree in software engineering from Xidian University, China in 2017. He is currently working toward the Ph.D. degree at the College of Computer Science, Zhejiang University, China. His research interests include computer vision and multimedia retrieval.

Lu Ye is a professor of the School of Information and Electronic Engineering, Zhejiang University of Science and Technology. Her research interests mainly include computer science and multimedia technology.

Yueting Zhuang received his B.Sc., M.Sc. and Ph.D. degrees in Computer Science from Zhejiang University, China, in 1986, 1989 and 1998 respectively. From February 1997 to August 1998, Yueting Zhuang was a visiting scholar at Prof. Thomas Huang's group, University of Illinois at Urbana-Champaign. Currently, He is a full professor of the College of Computer Science, Zhejiang University. His research interests mainly include artificial intelligence, multimedia retrieval, computer animation and digital library.

Jun Xiao received the Ph.D. degree in computer science and technology from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2007. He is currently a professor with the College of Computer Science, Zhejiang University. His current research interests include computer animation, multimedia retrieval, and machine learning.