

Weakly Supervised Object Localization and Detection: A Survey

Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang

Abstract—As an emerging and challenging problem in the computer vision community, weakly supervised object localization and detection plays an important role for developing new generation computer vision systems and has received significant attention in the past decade. As methods have been proposed, a comprehensive survey of these topics is of great importance. In this work, we review (1) classic models, (2) approaches with feature representations from off-the-shelf deep networks, (3) approaches solely based on deep learning, and (4) publicly available datasets and standard evaluation metrics that are widely used in this field. We also discuss the key challenges in this field, development history of this field, advantages/disadvantages of the methods in each category, the relationships between methods in different categories, applications of the weakly supervised object localization and detection methods, and potential future directions to further promote the development of this research field.

Index Terms—Weakly supervised learning, Object localization, Object detection.

INTRODUCTION

W EAKLY supervised learning (WSL) has recently received much attention in computer vision community. A plethora of methods on this topic have been proposed in the past decade to address the challenging computer vision tasks including semantic segmentation [16], object detection [111], and 3D reconstruction [38], to name a few. As shown in Fig. 1, a WSL problem is defined as the learning process when some partial information regarding the task (e.g., class label or object location) on a small subset of the data points is at our disposal. Compared to the conventional learning framework, e.g., fully supervised learning approaches, the WSL framework needs to operate on the small amount of weakly-labelled training data to learn the target model, which alleviates a huge amount of human labor to annotate training samples. It can also facilitate the learning process when the fine-grained annotation is extremely labor intensive and time consuming to even obtain the whole labeled data required by the fully-supervised approaches.

While a plethora of WSL-based vision methods have been developed, this survey mainly focuses on the task of weakly supervised object localization and detection, which is shown as the red dot in Fig. 1. It is well-known that object localization and detection is a fundamental research problem in computer vision. Learning object localization and detection models under weak supervision has attracted much attention in the past decades. While existing methods treat weakly supervised object localization (WSOL) and weakly supervised object detection (WSOD) as two different tasks¹, we consider these as a common task due to several reasons: 1) these tasks learn with the same image-level human annotation; 2) these two tasks need certain supervision as input and usually aim to localize objects on the bounding-

box level as output; 3) WSOD task can be accomplished by directly training off-the-shelf fully supervised object detectors on the object locations obtained from WSOL.

During the last decade, considerable efforts have been made to develop various approaches for learning object detectors with weak supervision. Some of the existing algorithms only learn weakly supervised object detectors for one or several certain object categories, such as vehicles [12], traffic signs [83], [196], pedestrians [9], [169], faces [43], [60], tuberculosis bacilli [69], aircrafts [57], [199], [222], [223], and human actions [60], [104]. While other approaches, e.g., [8], [112], [161], focus on developing weakly supervised learning frameworks for unconstrained object categories, i.e., learning frameworks can be extended to learn object detector for the given category-specific weakly-labelled training images. As enormous methods have been developed for these important tasks, a comprehensive review of the literature concerning weakly supervised object localization and detection is of great importance.

As weakly supervised object localization and detection methods mainly exploit the image-level manual annotation, the learning frameworks not only need to address the typical issues, such as the intra-class variations in appearance, transformation, scale and aspect ratio, encountered in conventional fully supervised object localization and detection task, but also the **learning under uncertainty** challenges caused by the inconsistency between human annotations and real supervisory signals. In weakly supervised object localization and detection, the accuracy of object locations and learning processes are closely related. The key is to propagate the image-level supervisory signals to the instance-level (bounding-box-level) training data for the learning processes. As each training image can be labeled by numerous bounding boxes of different accuracy, propagating such weak supervision inevitably involves a large amount of ambiguous and noisy information as each training instance. More specifically, the **learning under uncertainty** issue would cause the following challenges that make the weakly supervised learning process challenging:

- **Learning with inaccurate instance locations:** This issue is mainly caused by the definition ambiguity in object parts and context. Without precise annotation or definition, it is difficult for a learner to decide whether an object category

D. Zhang, J. Han, and G. Cheng are with Brain and Artificial Intelligence Laboratory, School of Automation, Northwestern Polytechnical University, Xi'an, China. Webpage: https://nwpu-brainlab.gitee.io/index_en.html.

M.-H. Yang is with EECS, University of California at Merced, Merced, California United States 95344. E-mail: mhyang@ucmerced.edu.

This work is supported in part by the National Key R&D Program of China under Grant 2017YFB1002201, the National Science Foundation of China under Grants 61876140 and 61773301. M.-H. Yang is supported in part by NSF CAREER Grant 1149783. (Corresponding author: Junwei Han.)

1. The difference between WSOL and WSOD mainly lies in that WSOL mainly aims at localizing a single known (or unknown) object from each given image scene. The goal of WSOD is to instead detect every possible object instance from the given image scene. This makes WSOD a little more difficult than WSOL.

label associates with a discriminative object part, the whole object region, or the object with a certain context region. As a result, the bounding-box instance locations inferred by the learner may contain many inaccurate samples including the ones with local object parts or undesired contextual regions. These samples would negatively affect the performance of WSL-based detectors.

- **Learning with noisy samples:** Even when the bounding-box locations can be precisely labeled, the training examples enclosed by bounding-boxes may still be noisy as background pixels are usually included. As there is no additional information to separate foreground objects from the background, the learner may tag a “background” label to an object region when it fails to recognize the object category. In addition, the learner may mistakenly label a bounding-box that contains a bicycle as a motorcycle, as these two object categories share many similar features.
- **Learning with domain shifts:** For a certain object category, the image regions localized during the learning process may only contain samples with limited diversity in object shape, appearance, scale, and view angle. This makes the subsequent learning process biased to limited knowledge of the object category and does not generalize well for test samples. For instance, a learner can hardly localize or detect a flying swan when all the training samples contain the swimming ones on lakes. This issue happens frequently among the weakly supervised learning process when there is a large gap between the training and testing domains.
- **Learning with insufficient instance samples:** Similar to the issues in conventional learning methods, it is difficult to train effective object detectors under the weakly-supervised setting when the amount of training samples is limited. In addition, the number of positive samples is usually much smaller than that of negative samples for binary classes. Furthermore, the data distributions for a large number of categories is usually long-tailed. This issue is significantly exaggerated for the WSL-based methods using deep learning.

To address the above-mentioned issues in learning weakly supervised object detectors, existing methods are usually constructed based on two steps: initialization and refinement. The initialization stage is used to leverage certain prior knowledge to propagate image-level annotation into instance level, and thus can generate instance-level annotation (but with label noise, sample bias, and limited quality in location accuracy) for the learning process. The refinement stage is used to leverage new instance samples obtained from the first stage to mine truthful knowledge about the objects of interest gradually and finally obtain the desired object models for localization and detection. These two learning stages need to collaborate to address the aforementioned five-fold challenges. In initialization stage, efforts should be made to improve the annotation quality as much as possible to generate training instances with proper locations, accurate labels, high diversity, and high recall rate. As the annotation quality obtained in the learning stage cannot be perfect, in the refinement stage, further efforts should be made to improve the learner’s robustness to cope with the inaccurate instance location, noisy examples, biased instance sample, insufficient instance sample issues as well as the capacity to take advantage of the unlabelled instance samples. When properly addressing the problems in each learning stage, good weakly supervised object detectors can be learned.

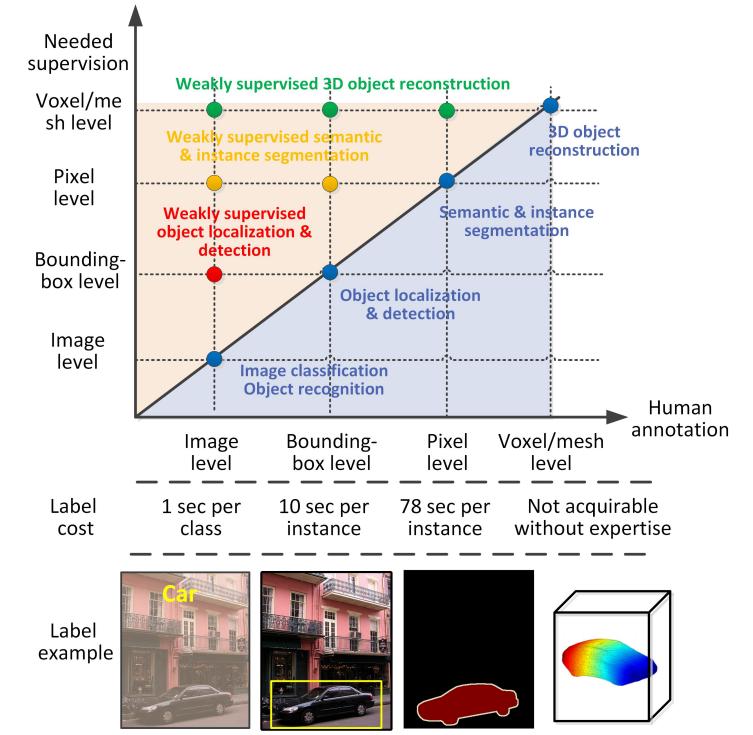


Fig. 1. Illustration of the weakly supervised learning tasks in computer vision community. The blue area in the top block indicates the conventional fully-supervised learning tasks, while the red area in the top block indicates the weakly supervised learning tasks. The coordinate axes shows different levels of human annotation or supervision requirement, from low cost to high cost. Notice that the high cost annotation can be transformed to low cost annotation easily, e.g., from bounding-box level to image level, whereas the low cost annotation is hard to be transformed to high cost annotation. In the bottom block, we also show the label cost, in terms of annotation time, and the examples of different type of annotations. In this survey, we mainly focus on reviewing the research progress in weakly supervised object localization and detection, i.e., the red dot in the top block.

In this work, we review the existing weakly supervised object localization and detection approaches², which are divided into three main categories and eight subcategories. These three main categories are based on classic approaches, feature representations from off-the-shelf deep models, and deep learning frameworks. The eight subcategories include approaches for initialization, refinement, initialization and refinement, pre-trained deep features, inherent cues in deep models, fine-tuned deep models, single-network training, and multi-network training. We further discuss the relationship between the approaches in different categories. In addition, we also discuss open problems and challenges of current studies and propose several promising research directions in the future for constructing more effective weakly supervised object localization and detection frameworks.

2 TAXONOMY

In the last decade, a plethora of methods have been developed for weakly supervised object localization and detection. We can generally categorize existing methods based on classic formulations, feature representations from off-the-shelf deep models, and deep weakly supervised learning algorithms. While inside each main category, we further divide the approaches into two or three subcategories. Fig. 2 shows our taxonomy of the studies in the research field of weakly supervised object localization and

2. Some early methods, such as [24], [42], learn to localize category-wise key points under the weak supervision, while this survey mainly focuses on the methods for localizing instances with bounding-boxes.

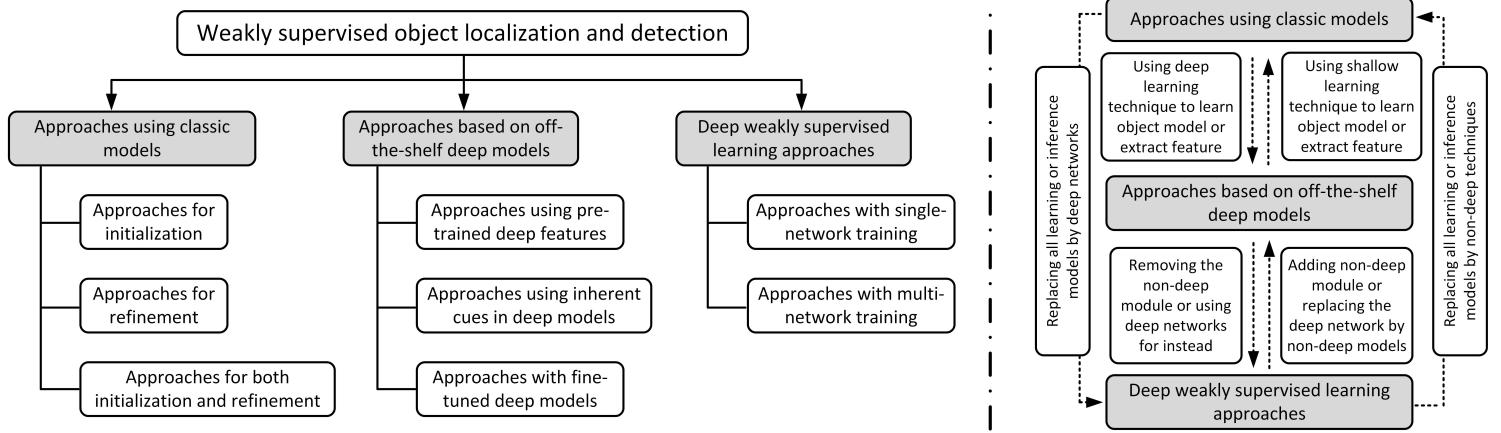


Fig. 2. In the left block, taxonomy of the existing approaches for weakly supervised object localization and detection, which includes three main categories and eight subcategories. In the right block, the relationships between the approaches in different categories are shown.

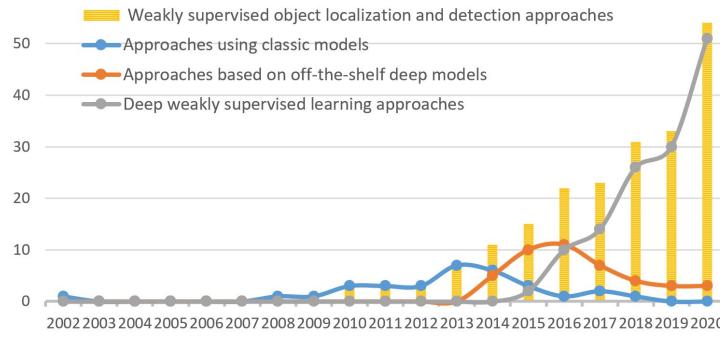


Fig. 3. Developments of weakly supervised localization and detection methods. The yellow histogram shows the number of publications in this research field in each year, and the curves show the number of proposed methods each year for a particular category of approach.

detection. In addition, Fig. 3 reviews the development history of each of the main category as well as that of the whole research field. A few approaches based on classic formulations appeared around 2002. From 2002 to 2009, the research in this field went through a very slow pace. Since 2014, numerous approaches based on both classic formulations and learned feature representations from deep models have been developed and received much attention. While in the last few years, more approaches solely based on deep learning have become the main stream to address the problems of weakly supervised object localization and detection. While a plethora of methods have been developed to address different aspects of these problems in the past decades, this field is gaining increasing attention.

As shown in the right block of Fig. 2, methods in main categories are related in several aspects. Numerous methods are developed based on the classic formulations with the advances of feature representations from deep models. Similarly, a number of methods based solely on deep models are end-to-end trainable by considering classic formulations and feature extraction schemes.

3 CLASSIC MODELS

In this section, we review the classic approaches that learn weakly supervised object localizer or detector without using deep features. These methods typically consist of one initialization module followed by one refinement process as shown in Fig. 4. In [26], [27], [78], [128], [138], the detector is based

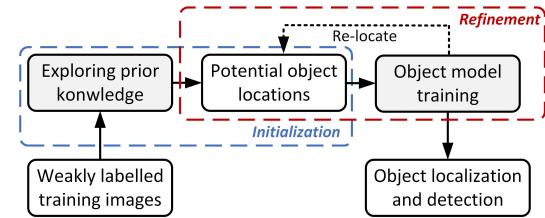


Fig. 4. Flowchart of the weakly supervised object localization and detection approaches using classic models.

on the deformable part model (DPM) [39]. In other approaches [13], [52], [167], [216], the detector is based on the support vector machine (SVM) classifier. The features used by these approaches are hand-crafted feature descriptors, such as HOG in [13], [150], [171], [216], SIFT in [134], [137], [138], [165], and Lab color in [129], [138], [165], which are sometimes used to build higher level representation such as bag-of-words (BOW) in [129], [134], [131], [137], [60], Fisher vector representation in [52], and subspace-based representation in [13]. In the following, we divide these approaches for initialization and refinement process.

3.1 Initialization

Numerous methods have been developed to mine reliable instance samples, using prior knowledge, as weak supervision for the following processes. A brief summary of these approaches are shown in Table 1.

Zhang et al. [216] leverage the prior-knowledge of object co-occurrence to identify translation and scale invariant high order features for weakly supervised object localization. In [130] Shi et al. propose a transfer learning paradigm to first use a RankSVM to learn the mapping relationship between the box overlap and appearance similarity from an auxiliary training data (with bounding-box level annotation) and then transfer the learned prior-knowledge for localizing objects of interest in the given weakly labeled images. A simple yet effective approach, named as negative mining, is developed by Siva et al. to explore the inter-class variance among the object regions in weakly labeled training images. The final object locations are obtained by using a linear combination of the inter-class variance and saliency prior. Similarly, Tang et al. [150] and Xie et al. [181] use the saliency prior and intra-class consistency to mine the initial object locations, respectively. In [128], [129], Shi et al. explore the appearance prior and geometry prior in their topic

TABLE 1

Summary of the approaches for initialization, which is a subcategory in the weakly supervised object localization and detection approaches that learn by classic models. An approach is considered for general object category when it is tested for detecting more than five object categories in the corresponding literature. The approaches with None detector indicate the weakly supervised object localization approaches.

Methods	Detector	Descriptor	Prior knowledge	Extra training data	Learning model	Learning strategy	Object category
Cao-PR-2017 [13]	SVM	HOG+PCA	Road map prior + density prior	None	SVM	MIL with density estimation	Vehicle in satellite imagery
Shi-TPAMI-2015 [29]	DPM	SIFT+Lab+LBP, BOW	Appearance prior + geometry prior	None	Topic model	Bayesian inference	General objects
Tang-ICIP-2014 [150]	DPM	HOG	Saliency (objectness score)	None	DPM	Select initial boxes + DPM training	General objects
Xie-VCIP-2013 [181]	None	SIFT	Intra-class consistency	None	None	Low-rank and sparse coding	General objects
Siva-CVPR-2013 [138]	DPM	Lab+SIFT	Saliency	Labelme + PASCAL07, 12 (unlabelled)	None	None	General objects
Shi-ICCV-2013 [128]	DPM	SIFT	Appearance prior (spatial distribution, size, saliency)	None	Topic model	Bayesian inference	General objects
Sikka-FG-2013 [134]	None	SIFT, LLC, BOW	None	None	MilBoost	Generating multiple segments for initialization and use Milboost for learning	Pain (on face)
Siva-ECCV-2012 [137]	None	SIFT+BOW	Iter-class variance + saliency	None	None	Negative mining	General objects
Shi-BMVC-2012 [130]	None	BOW	Mapping relationship between the box overlap and appearance similarity	Part of PASCAL 07 (box annotation)	RankSVM	Transfer learning by ranking	General objects
Khan-AAPRW-2011 [78]	DPM	Phog/phow	None	Internet image (weakly annotated)	MIL	Learning from internet image	Pascal@8
Zhang-BMVC-2010 [216]	SVM	IHOE	Co-occurrence	None	SVM	High order feature learning by exploring co-occurrence	General objects

TABLE 2

Summary of the approaches for refinement, which is a subcategory in the weakly supervised object localization and detection approaches that learn by classic models. Here, * indicates a certain variation of the corresponding model. An approach is considered for general object category when it is tested for detecting more than five object categories in the corresponding literature. The approaches with None detector indicate the weakly supervised object localization approaches.

Methods	Detector	Descriptor	Prior knowledge	Extra training data	Learning model	Learning strategy	Object category
Wang-TMI-2018 [168]	None	Color	None	None	Low-rank model	Low-rank Factorization	General lesion
Wang-TC-2017 [165]	None	SIFT+LAB	None	None	Probability model	BOW learning+instance labeling	General objects
Cholakalal-CVPR-2016 [22]	None	SIFT	Salient	None	SVM*	ScSVM-based top down saliency	Salient objects
Zadrida-GCP-2015 [196]	None	SIFT+PV	None	None	GMM + linear classifier	Patch-level spatial layout learning	Traffic sign
Krapac-ICCVTA-2015 [84]	None	SIFT+PV	None	None	Sparse logistic regression	Sparse classification	Traffic sign
Cinbis-CVPR-2014 [52]	SVM	FV	Center prior	None	SVM	Multi-fold MIL	General objects
Wang-ICIP-2014 [167]	SVM + graph model	SIFT	None	None	SVM + graph model	Maximal entropy random walk	Car, dog
Wang-ICIP-2014 [162]	SVM	SIFT	None	None	SVM	Clustering for window mining	General objects
Tang-CVPR-2014 [144]	None	SIFT	Salient	None	Boolean constrained quadratic program	Mine similarity and discriminativeness both for image and box	General objects
Hoai-PR-2014 [60]	SVM	SIFT,BOW	None	None	SVM*	Localization-classification SVM	Face,car, human motion
Wang-WACV-2013 [171]	Task-specific detectors	HOG/SC	Background saliency	None	MIL+AdaBoost*	Soft-label Boosting after MIL	Vehicle, pedestrian
Kaneko-MM-2013 [76]	Linear classifiers	3D voxel feature (color+C3HLAC +Intensity, texture, GRSD)	None	None	Linear classifiers	Multi-class MIL	Balls, tools
Pandey-ICCV-2011 [109]	DPM*	HOG	None	None	DPM*	Learning DPM with fully latent variable	General objects
Blaschko-NIPS-2010 [9]	None	BOW/HOG	None	None	SVM*	Learning SVM with structured output ranking objective	Cat, pedestrian
Hoai-ICCV-2009 [106]	SVM	SIFT,BOW	None	None	SVM*	Localization-classification SVM	Face,car, human motion
Galliguillou-ECCV-2008 [43]	None	SIFT+BOW	None	None	MilBoost	Train MIL classifier for localization	Landmarks, faces, airplanes, leopard, motorbike, car
Rosenberg-BMVC-2002	GMM	Orientation derivate filters	Exampler prior	Training exemplar (box annotation)	GMM	Learning from exemplar training data to weakly labelled training data	Telephone

TABLE 3

Approaches for both initialization and refinement, which is a subcategory in the weakly supervised object localization and detection approaches that learn by classic models. An approach is considered for general object category when it is tested for detecting more than five object categories in the corresponding literature. The approaches with None detector indicate the weakly supervised object localization approaches.

Methods	Detector	Descriptor	Prior knowledge	Extra training data	Learning model	Learning strategy	Object category
Wang-cvpr-2013 [169]	None	Color+SIFT	None	None	HST+SVM	Joint parsing and attribute localization	Scene attributes
Deselaers-IJCV-2012 [27]	DPM	GIST+CH+BOW+HOG	Generic knowledge	Meta-training data with box annotation	CRF+DPM	Learning appearance model by transferring generic knowledge	General objects
Siva-ICCV-2011 [139]	DPM	SIFT+BOW+HOG	Inter-class prior + intra-class prior	None	DPM	Model drift learning	General objects
Deselaers-ECCV-2010 [26]	DPM	GIST+CH+BOW+HOG	Generic knowledge	Meta-training data with box annotation	CRF+DPM	Learning appearance model by transferring generic knowledge	General objects

model to build a Bayesian joint modeling framework for weakly supervised object localization. On the other hand, Cao et al. [13] exploit the road map prior and density prior to mine the initial vehicle locations from the weakly labeled satellite images and then trained the vehicle detector under a modified multiple-instance learning (MIL) model.

3.2 Refinement

After potential object instances are obtained, these hypotheses are verified in the following refinement processes. The goals of these approaches are to design learning objective functions, optimization strategies, or learning mechanisms to gradually determine objects of interest from the extracted initial instance training samples. A brief summary of these approaches is shown in Table 2.

Hoai et al. [60], [106] propose an approach which localizes the instances of the positive class and learns a sub-window classifier to recognize the corresponding object class. Blaschko et al. [9] use a structured output SVM to learn a regressor from the weakly labeled training images to object locations that are parameterized by the coordinates of the bounding boxes. The object locations were treated as latent variables, while the image-level annotation was used to constrain the set of values the latent variable can take. Similarly, Pandey et al. [109] learn weakly supervised object detectors by using DPMs with latent SVM training. In [171], a soft-label boosting approach is developed to exploit the soft labels that are estimated during the MIL process to train object detectors based on Boosting algorithm. In [144], Tang et al. treat the weakly supervised object localization problem as an object co-localization task, and present a joint image-box formulation to mine reliable object locations via a Boolean constrained quadratic

TABLE 4

Summary of the approaches using pre-trained feature representations, which is a subcategory in the weakly supervised object localization and detection approaches based on the off-the-shelf deep models. * indicates a certain variation of the corresponding model. An approach is considered for general object category when it is tested for detecting more than five object categories in the corresponding literature. The approaches with None detector indicate the weakly supervised object localization approaches.

Methods	Detector	Descriptor	Prior knowledge	Extra training data	Learning model	Learning strategy	Object category
Gonthier-arxiv-2018 [54]	None	CNN	Supervised objectness score (Fast R-CNN(Resnet))	ImageNet(tag label)	SVM	MIL	Objects in art (watercolor2K, people-art)
Zadrjia-CVIU-2018 [195]	None	VGG19 Conv 5_4, SIFT, Fisher vector	None	ImageNet(tag label)	Sparse model	None	Zebra crossings, traffic signs
Cinbis-TPAMI-2017 [23]	SVM*	FV+CNN	Center prior	ImageNet(tag label)	SVM*	Multi-fold MIL	General objects
Wei-IJCAI-2017 [174]	None	CNN	None	ImageNet(tag label)	None	Deep descriptor transforming	General objects
Zhang-IJCAI-2016 [207]	SVM	CNN	Saliency prior	ImageNet(tag label)	SVM	Easy-to-hard(SPL+CL)	General objects
Li-ECCV-2016	None	FC6	Strong detector prior (sparsity)	ImageNet(tag label)	SVM	Regularizing score distribution	General objects
Ren-TPAMI-2016 [112]	SVM*	FC6		ImageNet(tag label)	SVM*	MIL+bag splitting	General objects
Wan-ICIP-2016 [158]	SVM*	FC7		ImageNet(tag label)	SVM*	Correlation suppression+part suppression	General objects
Rochan-IVC-2016 [114]	None	Color histogram +CNN	Saliency	PASCAL (edge box), ImageNet	SVM	None	General objects
Shi-ECCV-2016 [127]	SVM	FC7(Alexnet)	Size prior	ImageNet(tag label), PASCAL2012 (object size)	SVM	Easy-to-hard(curriculum)	General objects
Wang-TIP-2015 [161], [167]	SVM	FC6	None	ImageNet(tag label)	pLSA, SVM	Online latent category learning	General objects
Rochan-CVPR-2015 [115]	SVM	CNN	Objectness score, word embedding prior	YouTube-Objects (for parameter validation), Familiar object categories(detector)	SVM, Sparse reconstruction	Appearance transfer from text representation	General objects
Bilen-CVPR-2015 [7]	LSVM	FC7+spatial features	None	ImageNet(tag label)	LSVM	Convex clustering	General objects
Wang-IJCV-2015 [173]	None	FC6	None	PASCAL(edge box), ImageNet	SVM*	Relaxed multiple-Instance SVM	General objects
Zhou-ICMBD-2015 [223]	SVM	FC7	Saliency prior	ImageNet(tag label)	SVM	Negative Bootstrapping	Airplanes in remote sensing
Han-TGRS-2015 [57]	SVM	DBM	Saliency, intra-class compactness, inter-class separability	None	DBM + SVM	Bayesian framework for initialization + refinement detector training	Objects in remote sensing
Matthe-Arxiv-2014 [105]	Sequential detector	FC6	Human fixation	ImageNet(tag label)	MIL*+RL	Constrained multiple instance SVM learning + reinforcement learning of detector	Human actions
Wang-ECCV-2014 [167]	SVM	FC6	None	ImageNet(tag label)	pLSA, SVM	Online latent category learning	General objects
Bilen-BMVC-2014 [6]	LSVM	DeCAF	None	ImageNet(tag label)	LSVM*	LSVM with posterior regularization on symmetry and mutual exclusion	General objects
Song-NIPS-2014 [141]	DPM	FC7	Objectness score	ImageNet(tag label)	LSVM	Frequent configuration mining+detector training	General objects
Song-ICML-2014 [140]	LSVM	DeCAF	None	ImageNet(tag label)	Graph model+LSVM*	Initialization via discriminative submodular cover+smoothed LSVM learning	General objects

program. This approach can handle noisy labels in the image-level annotations. To address the property that the MIL process may converge to poor local optima after the initialization, Cinbis et al. [52] design a multi-fold MIL training paradigm. This method divides the whole weakly labelled training images into multiple folds and implements the detector training process and object re-localization process in different folds, thereby alleviating the issue with convergence of poor local optima.

3.3 Initialization and Refinement

A number of iterative approaches have been developed that take both initialization and refinement into account. In [139], Siva et al. propose an intra-class metric and an inter-class metric to initialize the potential object locations. After obtaining the initial object locations, this method iteratively trains a DPM object detector and uses a model drift detection approach to identify the termination refinement dynamically. Deselaers et al. [26], [27] present a conditional random field (CRF) model, which is used to learn generic prior knowledge of the objects from meta-training data firstly to localize the potential objects of interest in the weakly labelled training images. This algorithm updates the CRF model to learn the appearance and shape models for the target object category and localizes the objects of interest in the refinement stage. The alternation of localization and learning processes progressively transforms the CRF model from class-generic prior knowledge into the specialized knowledge for a certain target class. For learning weakly supervised attribute localizer, Wang et al. [169] initialize the learning process by building a Hierarchical Space Tiling (HST) scene configuration model [170] and the corresponding appearance models are trained based on HST. A joint inference and learning process is designed to update the scene attributes and the correlations between the scene parts and attributes gradually.

3.4 Discussion

Although the classic weakly supervised learning models are studied in early age, the two-stage learning frameworks, i.e., the learning initialization stage and refinement stage, built by these methods have been widely applied in future works. In the

learning initialization stage, these methods provide two kinds of information cues to infer the candidate object regions. The first one is the bottom-up cues, including the region saliency, objectness, intra-class consistency, inter-class discriminability, et al. The other one is the top-down cues, which usually provide the appearance prior for the learning process. Notice that as such top-down cues are hard to obtain from the weakly labeled data, auxiliary training data (with instance-level manual annotation) are usually leveraged to explore the top-down cues which are then transferred to the weakly labeled target data. In the refinement stage, classic machine learning models, such as SVM and CRF, are adopted to gradually refine both the appearance model and locations of the objects of interest.

The advantage of the classic weakly supervised learning methods is that the learning processes can be implemented on small-scale training data and the whole frameworks are quick both in the training phase and the testing phase. The disadvantage is that their performance is not satisfactory, which is due to the limitation in feature representation and model complexity.

4 OFF-THE-SHELF DEEP MODELS

In this section, we review the approaches that learn weakly supervised object localizer or detector based on classic formulations and feature representations based on the deep neural networks, either pre-trained from the ImageNet dataset [116] (with image tag annotation) or further fine-tuned on the weakly supervised training images in the target domain. The feature representations are based on the widely used deep models for image classification, such as AlexNet [85] and VGG [135]. The detectors are constructed based on classic formulations such as DPM and SVM [3], [57], [61], [112], [141], [149], [222], or recent models such as RCNN [50] and fast RCNN [49] [17], [75], [88], [126], [154]. We further divide these approaches into three subcategories using pre-trained deep features, inherent cues in deep models, and fine-tuned deep models as shown in Fig. 5.

4.1 Pre-trained Deep Features

The methods of this category replace the hand-crafted feature representations with the pre-trained deep features typically from

TABLE 5

Summary of the approaches using visual cues, which is a subcategory in the weakly supervised object localization and detection approaches based on the off-the-shelf deep models. * indicates a certain variation of the corresponding model. An approach is considered for general object category when it is tested for detecting more than five object categories in the corresponding literature. The approaches with None detector indicate the weakly supervised object localization approaches.

Methods	Detector	Descriptor	Prior knowledge	Extra training data	Learning model	Learning strategy	Object category
Li-ISPRS-2018 [95]	CAM* (VGG-F)	CNN	None	ImageNet(tag label)	VGG-F*	Learning learning + CAM learning (patch level)	Remote sensing objects
Wilhellem-DICTA-2017 [177]	None	CNN	None	ImageNet(tag label)	CAM	CAM+KDE refine	General objects
Tang-TMM-2017 [149]	DPM	CNN	Saliency + objectness score	ImageNet(tag label)	DPM+ CNN	Region initialization+DPM and feature learning+bounding box modification	General objects
Kolesnikov -BMVC-2016 [81]	None	CNN	Human feedback annotation	ImageNet(tag label)	CAM	Active learning for identifying object cluster	General objects
Bency-ECCV-2016 [4]	None	CNN	None	ImageNet(tag label)	VGG16	Beam-search based on CNN classifier	General objects
Zhou-MSSP-2016 [222]	SVM	FC7	Saliency prior	ImageNet(tag label), remote sensing data(unlabelled)	CNN (AlexNet), SVM	Deep feature transfer +MIL	Remote sensing objects (airplane, car, airport)
Bergamo-WACV-2016 [3]	SVM	CNN	None	ImageNet(tag label)	CNN, SVM	Mask out initialization + SVM detector training	General objects
Hoffman-CVPR-2015 [61]	SVM	FC7	Detector prior+ representation prior	ImageNet(tag label), ILSVRC13 validation subset(box annotation)	CNN, Latent SVM	Transferring detectors and representation from auxiliary data	General objects

TABLE 6

Summary of the approaches with fine-tuned deep models, which is a subcategory in the weakly supervised object localization and detection approaches based on the off-the-shelf deep models. * indicates a certain variation of the corresponding model. An approach is considered for general object category when it is tested for detecting more than five object categories in the corresponding literature.

Methods	Detector	Descriptor	Prior knowledge	Extra training data	Learning model	Learning strategy	Object category
Zhang-IJCV-2019 [204]	Fast RCNN (VGG16)	Pre-trained FC7	Tag number + mask out prior(AlexNet)	ImageNet(tag label)	SVM	Easy-to-hard	General objects
Uijlings-CVPR-2018 [154]	None	CNN	Semantic objectness(SSD*)	ImageNet(tag label), ILSVRC(full annotation)	SSD*+SVM+ Fast RCNN	MIL+knowledge transfer	General objects
Jie-CVPR-2017 [75]	Fast RCNN (VGG16)	CNN	Image-to-object transfer prior	ImageNet(tag label)	Fast RCNN (VGG16)	Initialization based on classification network and subgraph discovery + iterative Fast RCNN learning	General objects
Shi-ICCV-2017 [126]	Fast RCNN	CNN	Things and stuff prior	ImageNet(tag label), PASCAL Context (full annotation)	FCN, Fast RCNN	Localizing objects based on things and stuff prior and training Fast RCNN iteratively	General objects
Singh-CVPR-2016 [88]	Fast RCNN	CNN	Tracking prior	ImageNet(tag label), Youtobe-objects (unlabelled)	Fast RCNN	Discriminative region mining+transferring tracking object pattern + learn object detector	General objects
Li-CVPR-2016 [91]	VGG*	CNN	Mask out prior(AlexNet)	ImageNet(tag label)	VGG*, SVM	Progressive Domain Adaptation	General objects
Liang-ICCV-2015 [97]	CNN	CNN	Instance example, motion prior	ImageNet(tag label)	CNN, R-CNN	Seed selection based on instance example and instance tracking	General objects
Chen-ICCV-2015 [17]	RCNN	CNN	Online data type	Web data (weak label)	BLVC net +E-LDA + RCNN	Simple image initialization + graph-based representation adaptation on hard image	General objects
Zhou-CVPR-2015 [220]	RCNN	FC7	None	ImageNet(tag label)	SVM, R-CNN	Max-margin visual concept discovery + Domain-specific detector selection	General objects

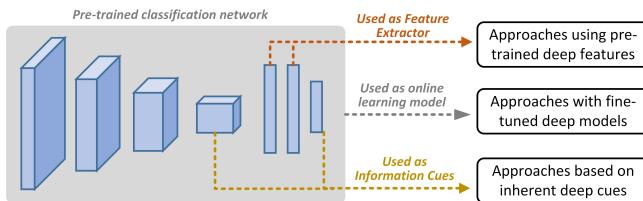


Fig. 5. Illustration of different usages of the off-the-shelf deep neural networks by the weakly supervised object localization and detection approaches based on the off-the-shelf deep models.

as AlexNet and VGG. A brief summary of these approaches is shown in Table 4.

Song et al. [140], [141] determine discriminative feature configurations of an object class via graph modeling, and train object detectors within the multiple-instance learning paradigm. The deep features of this work are extracted based on the DeCAF scheme [30] and AlexNet [85]. By using the deep features and spatial features to represent each proposal region, Bilen et al. [7] propose a convex clustering process for learning the object models under the weak supervision. The learning objective is able to enforce the similarity among the selected proposal windows. In [127], Shi and Ferrari develop a curriculum learning strategy to feed training images into the MIL loop in a pre-defined order, where images containing larger objects are learned at the early stages while images containing smaller objects are learned at later stages. Ren et al. [112] present a bag-splitting-based MIL mechanism that iteratively generated new negative bags from the positive ones. This algorithm can gradually reduce the ambiguity in positive images and thus facilitate the learning of more reliable training instance samples. In [174], [175], Wei

et al. leverage the pre-trained CNN model to implement a Deep Descriptor Transforming process, which can obtain the category-consistent image regions via evaluating the correlations of the descriptors in the convolutional activations of the CNN model.

4.2 Inherent Cues in Deep Models

Instead of using the pre-trained deep models as feature extractor, the methods of this category obtain useful information cues (such as the activations in the intermediate network layers and the semantic scores in the output network layer) from the pre-trained deep neural networks to facilitate the weakly supervised learning process. The focus of these approaches mainly lies in the initialization stage of the weakly supervised learning process. A brief summary of these approaches is shown in Table 5.

Bergamo et al. [3] propose a self-taught deep learning approach for localizing objects of interest under weak supervision. In the initialization stage, they design a mask-out strategy based on the deep semantic cues from a pre-trained classification network. Specifically, this method first calculates the degeneration of the image-level classification scores when masking out a certain object proposal region and then selects those with large differences as the interested object regions. After the initialization stage, this method trains an SVM-based object detector in the subsequent refinement stage for final object localization. Similar to [3], Bency et al. [4] propose a beam search algorithm to leverage the activation maps of a pre-trained classification network to localize the objects of interest. This method is based on the observation that when image regions centered around objects of interest are classified by a pre-trained DNN, they obtain higher semantic scores than other image regions. Hoffman et al. [61] develop a transfer learning-based algorithm, where the deep neural network is first trained on both the weakly labeled

TABLE 7

A brief summary of the approaches using single-network training scheme, which is a subcategory in the weakly supervised object localization and detection approaches with deep weakly supervised learning algorithms. * indicates a certain variation of the corresponding model. An approach is considered for general object category when it is tested for detecting more than five object categories in the corresponding literature. The approaches with None detector indicate the weakly supervised object localization approaches.

Methods	Detector	Descriptor	Prior knowledge	Extra training data	Learning model	Learning strategy	Object category
Huang-NIPS-2020 [67]	Faster RCNN	CNN	None	ImageNet(tag label)	Faster RCNN*(VGG16/ResNet50)	Proposal attention aggregation and distillation	General objects
Shen-CVPR-2020 [121]	Faster RCNN*	CNN	None	ImageNet(tag label) + Flickr	Faster RCNN*(VGG16)	bagging-mixup + background noise decomposition + clean data modelling	General objects
Mai-CVPR-2020 [102]	None	CNN	None	ImageNet(tag label)	VGG/InceptionV3	Integrating discriminative region mining and adversarial erasing	General objects
Zhang-ECCV-2020 [213]	None	CNN	Cross-image consistency	ImageNet(tag label)	VGG/InceptionV3/ResNet50	Inter-image stochastic consistency and global consistency	General objects
Yang-WACV-2020 [187]	None	CNN	None	ImageNet(tag label)	VGG	Weighted classification activation map combination	General objects
Yang-ICCV-2019 [186]	Faster RCNN*(VGG16)	CNN	None	ImageNet(tag label)	Faster RCNN*(VGG16) + CAM	Online classifier learning with bounding box regression	General objects
Wan-CVPR-2019 [156]	Faster RCNN*(VGG16)	CNN	None	ImageNet(tag label)	Faster RCNN*(VGG16)	Continuation MIL	General objects
Shen-CVPR-2019 [123]	WSDDN*(VGG16)	CNN	None	ImageNet(tag label)	Two-stream CNN (WS-DDN+DeepLab)	Joint detection and segmentation with cyclic guidance	General objects
Wan-CVPR-2019 [157]	Fast RCNN	CNN	None	ImageNet(tag label)	Two-stream CNN	continuation instance selection and detector estimation	General objects
Choe-CVPR-2019 [21]	None	CNN	None	ImageNet(tag label)	CNN	Feature learning by attention-based dropout	General objects
Jiang-ICCV-2019 [72]	None	CNN	None	ImageNet(tag label)	VGG16/Resnet101	Online attention accumulation on CAM	General objects
Sangineto-TPAMI-2018 [117]	Fast RCNN (VGG16)	CNN	None	ImageNet(tag label)	Fast RCNN (VGG16)	Easy-to-hard	General objects
Inoue-CVPR-2018 [70]	SSD	CNN	None	PASCAL (full label as source domain),ImageNet	SSD	Domain transfer + pseudo-labeling	Cartoon objects
Wan-CVPR-2018 [159]	Faster RCNN*(VGG16)	CNN	None	ImageNet(tag label)	Faster RCNN*(VGG16)	Min-entropy latent modeling	General objects
Shen-TNNLS-2018 [122]	VGG16*	CNN	None	ImageNet (tag label)	vgg16*	Object-specific pixel gradient mapping+iterative component mining	General objects
Tang-TPAMI-2018 [145]	Fast RCNN*(model ensemble)	CNN	None	ImageNet(tag label)	Fast RCNN*(model ensemble)	MIL+oicr+multi-scale+proposal cluster learning	General objects
Zhang-CVPR-2018 [211]	None	CNN	None	ImageNet(tag label)	VGG16*	Adversarial complementary erasing	General objects (for ILSVRC)
Choe-BMVC-2018 [20]	ResNet	CNN	None	Tiny ImageNet(tag label)	ResNet	GoogLeNet Resize (GR) augmentation	General objects (for Tiny ImageNet)
Zhang-ECCV-2018 [212]	Inception-v3+CAM	CNN	None	ImageNet(tag label)	Inception-v3+CAM	Self-produced guidance learning	General objects (for ILSVRC and CUB)
Gao-ECCV-2018 [44]	Fast RCNN*	CNN	Count (human label)	ImageNet(tag label)	Fast RCNN*-Fast RCNN	WSL with count-based region selection	General objects
Singh-ICCV-2017 [136]	CAM* (GoogLeNet)	CNN	None	ImageNet (tag label)	CAM* (GoogLeNet)	Random hiding patches	General objects (for ILSVRC)
Zhu-ICCV-2017 [226]	None	CNN	None	ImageNet(tag label)	GoogLeNet*	Soft proposal layer+CAM	General objects
Wan-ICIP-2017 [160]	None	CNN	None	ImageNet(tag label)	CAM*(VGG)	CAM with spatial pyramid pooling layer	General objects
Durand-CVPR-2017 [33]	None	CNN	None	ImageNet(tag label)	CAM*(ResNet101)	CAM with multi-map transfer layer	General objects
Tang-CVPR-2017 [146]	Fast RCNN*(model ensemble)	CNN	None	ImageNet(tag label)	Fast RCNN*-Fast RCNN	MIL+oicr+multi-scale	General objects
Jiang-CVPR-2017 [73]	Fast RCNN*(AlexNet)	CNN	None	PASCAL (edge box),ImageNet	AlexNet+ ROIpool	Region classification+region selection+multi-scale	General objects
Diba-CVPR-2017 [28]	Faster RCNN*(VGG16)	CNN	None	ImageNet(tag label)	Multi-stream CNN	Cascading LocNet (CAM), SegNet, and MILNet+multi-scale	General objects
Selvaraju-ICCV-2017 [119]	None	CNN	None	ImageNet(tag label)	VGG*	Gradient-based class activation mapping	General objects
Tang-PR-2017 [147]	None	CNN	None	ImageNet(tag label)	Fast RCNN*(VGG-16)	SPP with discovery block and classification block	General objects
Gudi-BMVC-2017 [56]	None	CNN	None	ImageNet(tag label)	CAM* (VGG-16)	CAM with Spatial Pyramid Averaged Max (SPAM) Pooling	General objects
Bilen-CVPR-2016 [8]	Fast RCNN*(model ensemble)	CNN	None	PASCAL (edge box),ImageNet	Fast RCNN*	MIL+multi scale	General objects
Kantorov-ECCV-2016 [77]	Fast RCNN*(VGG-F)	CNN	Context	ImageNet(tag label)	Fast RCNN*	MIL+multi-scale	General objects
Teh-BMVC-2016 [152]	CNN	CNN	None	PASCAL (edge box),ImageNet	CNN	Proposal attention learning	General objects
Durand-CVPR-2016 [34]	None	CNN	None	ImageNet(tag label)	CNN	Feature extraction network+weakly supervised prediction module	General objects
Zhou-CVPR-2016 [221]	None	CNN	None	ImageNet(tag label)	GoogLeNet*	Class activation mapping	General objects
Oquab-CVPR-2015 [108]	None	CNN	None	ImageNet(tag label)	CNN	Fully convolutional CNN with global max pooling	General objects
Wu-CVPR-2015 [179]	None	CNN	None	PASCAL (BING),ImageNet	CNN	Deep multiple instance learning network	General objects

auxiliary training data and the strongly labeled training data to obtain the background detector. Then, an SVM-MIL model is adopted to learn the object detectors based on the potential foreground regions that are obtained by using the pre-trained DNN to screen the image background regions. To overcome the problem that the objects of interest would sometimes co-occur with the distracting image background, Kolesnikov et al. [81] propose a user-guided weakly supervised learning framework to improve the localization capacity. This approach first trains a classification network under the image level annotation. For each image, the intermediate feature maps of the pre-trained network are clustered, and the object clutter is identified by a user.

4.3 Fine-tuned Deep Models

The methods of this category fine-tune the off-the-shelf DNN models during the weakly supervised learning process to obtain strong object detectors [17], [126], [184]. A brief summary of these approaches is shown in Table 6.

Chen et al. [17] first train CNNs from the web image data via an easy-to-hard learning scheme. The learned deep features are used to mine object locations by using the exemplar-LDA detector [59]. The off-the-shelf RCNN detector is then adopted to learn object models based on their localization results. In [91], Li et al. first use the mask-out strategy based on the pre-trained classification network to obtain the class-specific object proposals. An SVM-based MIL process is used to localize object

instances and the classification network is further fine-tuned on the localized object instances for better performance. Shi et al. [126] propose to transfer the prior knowledge of *things* and *stuff* to help the weakly supervised learning process. A semantic segmentation network is trained from the source set (with available bounding-box annotations) to generate the stuff map and thing map for the weakly labeled images in the target set. These maps are used to obtain potential object locations, and the fast RCNN model is adopted in a Deep MIL scheme to train the object detectors. Recently, [154] revisits knowledge transfer for training the weakly supervised object detector. In this method, a DNN-based proposal extractor is learned from the source data firstly. The DNN is designed based on the SSD [99] architecture and trained with a semantic hierarchy. The network is then used to provide proposals and other prior knowledge for the weakly labeled images in the target set. An MIL process is used to determine the proposals that cover the objects of interest based on which the fast RCNN model is adopted to learn the final object detectors. In [204], Zhang et al. first learn to localize the objects of interest via a collaborative self-paced curriculum learning mechanism based on pre-trained deep features. The fast RCNN model is applied to learn object detectors.

4.4 Discussion

Introducing the off-the-shelf deep models into the weakly supervised object localization or detection framework is the most straightforward approach to integrate deep learning and weakly supervised learning. The methods in this category show that: 1) feature learning is an important factor to improve the weakly supervised learning process; 2) DCNN models can infer discriminative spatial locations when learned under the image-level supervision; 3) pre-training DNN models on large-scale auxiliary training data is a simple but effective way to encode useful cues for the weakly supervised learning process. Compared with the classic models, the methods of this category exploit large-scale auxiliary training data to learn powerful feature representations and top-down cues. By using DNN models as the object detector or localizer, a significant performance gain can be obtained. However, more effective feature learning models can be exploited in the weakly supervised learning process.

5 DEEP WEAKLY SUPERVISED LEARNING

In this section, we review the methods that learn weakly supervised object localizers or detectors by designing novel deep weakly supervised learning frameworks³. Different from the approaches discussed in previous sections, both the feature representations and the object detectors of the approaches in this category are learned by newly-designed deep neural networks. The whole weakly supervised learning framework may be designed in a compact network model, such as in [47], [77], [103], [108], [119], [185], [187], [209], [224], [226], or contain several function-distinct DNN components, such as in [28], [100], [143], [164], [176], [198], [214]. We categorize these approaches into two groups using single-network training and multi-network training, respectively.

3. Notice that here we mainly indicate that the core learning blocks used in the learning frameworks are based on deep learning, while some minor computational components in pre-processing or post-processing, such as proposal extraction and bounding-box modification, et al., are not necessarily be implemented by deep learning.

5.1 Single-Network Training

The methods of this category are designed with a single deep neural network using the training images (or together with the extracted object proposals) as inputs and the image-level classification labels as the outputs. These approaches do not usually rely on meticulously designed initialization processes to obtain the potential object regions. Instead, these methods discover the interested object regions solely based on the end-to-end learning process of the designed DNN models. The DNN models used in these approaches usually have similar feature learning layers as the conventional image classification network, e.g., AlexNet, VGG, GoogleNet, and ResNet, followed by the instance label inferring and image label propagation layers to inherently predict the labels of each proposal region and generate the final image labels from the predicted proposal labels, respectively. Some of the methods contain multiple network streams for online inferring multiple informative cues. A brief summary of these approaches are shown in Table 7.

In the DNN models proposed by Wu et al. [179] and Bilen et al. [8], the network inputs are the training images and extracted object proposal regions while the outputs are the image-level semantic scores. The first parts of these networks extract the features and infer the labels for each proposal region, and the second parts propagate the proposal scores to the image-level via the max pooling scheme [179] or the two-stream score regularization method [8]. Zhou et al. [221] present an end-to-end weakly supervised deep learning based on the class activation mapping (CAM). The weights of the feature maps in the intermediate layers are inferred based on the correspondence between the feature map and a certain object category. The feature maps are then combined to form the class activation maps based on the inferred weights, which highlight the locations of the objects of interest. Notably, this method is determined to be highly efficient in recent work [19]. Built on CAM [221], Durand et al. [33] introduce the multi-map transfer layer and the WILDCAT pooling layer to facilitate the more accurate deep MIL process.

Recently, a number of two-branch MIL models have been developed in which one is based on a typical deep network and the other one is introduced for weakly supervised learning. Based on the WSDDN [8], Tang et al. [147] integrate MIL branch and the instance classifier refinement branch into a unified deep learning framework such that more accurate online instance classifier learning is realized under the weak supervision. In [28], Diba et al. propose a weakly supervised cascaded convolutional network, which contains three branches. The first branch adopts the CAM module to generate the class activation maps. The second branch uses the generated class activation maps as the supervision signal to learn a segmentation module to generate the segmentation masks of the objects of interest. Using the candidate object proposals selected based on the obtained segmentation masks as supervision, the third network branch uses a MIL process to mine accurate object locations from the candidate object regions. In [211], Zhang et al. present a CAM-based network architecture which contains a classification branch and a counterpart classifier branch for object localization. Specifically, the classification branch is used to localize the discriminative object regions, which drives the counterpart classifier branch to discover new and complementary object regions by erasing its discovered regions from the feature maps. Wan et al. [159] propose a min-entropy latent model (MELM) for weakly supervised object detection based on the assumption that minimizing entropy results in minimum randomness of a system. The network architecture is similar to [147], but global

TABLE 8

A brief summary of the approaches with multi-network training, which is a subcategory in the weakly supervised object localization and detection approaches with deep weakly supervised learning algorithms. * indicates a certain variation of the corresponding model. An approach is considered for general object category when it is tested for detecting more than five object categories in the corresponding literature. The approaches with None detector indicate the weakly supervised object localization approaches.

Methods	Detector	Descriptor	Prior knowledge	Extra training data	Learning model	Learning strategy	Object category
Zhang-CVPR-2020 [197]	None	CNN	Common object co-localization	ImageNet(tag label)	VGG/InceptionV3/ResNet50/DenseNet161	Classification + pseudo supervised object localization	General objects
Zhong-ECCV-2020 [219]	Faster RCNN	CNN	Location prior	ImageNet(tag label) + COCO (box label)	Oneclass universal detector + MIL classifier (on ResNet50)	Progressive knowledge transfer	General objects
Kosugi-ICCV-2019 [82]	Fast RCNN*	CNN	Mask-out prior	ImageNet(tag label)	Mask-out net + OICR*	Mask-our prior-guided label refinement	General objects
Singh-CVPR-2019 [87]	Fast RCNN*	CNN	Motion prior	ImageNet(tag label), videos	RPN+WSDDN/OICR (VGG16)	Training RPN using weakly-labeled videos for WSOD	General objects
Arun-CVPR-2019 [1]	Fast RCNN	CNN	None	ImageNet(tag label)	Fast RCNN (VGG16) + Fast RCNN* (VGG16)	Employ dissimilarity coefficient for modeling uncertainty	General objects
Li-TPAMI-2019 [94]	Faster RCNN* (VGG16)	CNN	Objectness (classifier) prior	ImageNet(tag label), ILSVRC2013(box label for unseen categories)	Faster RCNN* (VGG16) *2	Objectness transfer+MIL +multi-scale	General objects
Zhang-ECCV-2018 [214]	fast RCNN* (VGG16)	CNN	None	PASCAL (edge box), ImageNet(tag label)	Multi-view WSDDN+multi view	Two phase multi-view learning	General objects
Shen-CVPR-2018 [125]	SSD	CNN	None	ImageNet(tag label)	SSD+Fast RCNN*	MIL+GAN	General objects
Zhang-CVPR-2018 [215]	Faster RCNN (VGG16)	CNN	None	ImageNet(tag label)	MDN+Faster RCNN	WSOD+Pseudo Ground-truth Mining+FOD	General objects
Zhang-CVPR-2018 [210]	Fast RCNN* (VGG16)	CNN	None	PASCAL (edge box), ImageNet	WSDDN + Fast RCNN*	WSDDN+easy-to-hard FOD	General objects
Tang-ECCV-2018 [148]	Fast RCNN* (VGG16)	CNN	None	ImageNet(tag label)	Fast RCNN* (VGG16)	Alternating training of WSRPN and WSOD+multi-scale	General objects
Tao-TMM-2018 [151]	Fast RCNN* (VGG16)	CNN	Web image	Web dataset(weak label),imageNet(tag label)	Midn	Easy-to-hard	General objects
Wang-IJCAI-2018 [163]	Faster RCNN (VGG16)	CNN	Model consistency	ImageNet(tag label)	Faster RCNN+Fast RCNN*	MIL+collaborative learning	General objects
Wei-ECCV-2018 [176]	Faster RCNN* (VGG16)	CNN	Shape prior+ context prior	ImageNet(tag label)	MDN+CAM+ DeepLab	Tight Box Mining+MIL +OICR+multi-scale	General objects
Ge-CVPR-2018 [48]	Faster RCNN (VGG16)	CNN	Local objectness and global attention	ImageNet(tag label)	MDN,TripNet, GoogleNet, FCN, Fast RCNN	Multi evidence fusion+ outlier filtering +pixel label prediction+box generation+multi-scale	General objects
Dong-MM-2017 [31]	Fast RCNN*	CNN	None	ImageNet(tag label)	Fast RCNN*+R-FCN	Easy-to-hard	General objects
Li-BMVC-2017 [93]	Fast RCNN*	CNN	Shape prior	ImageNet(tag label)	Fast RCNN*+CAM+DeepLab	Easy-to-hard	General objects
Wang-CVPR-2017 [164]	CNN	CNN	None	ImageNet(tag label)	CAM*	Image-level training+pixel-level fine tuning	Salient objects
Sun-CVPR-2016 [143]	None	CNN	None	ImageNet(tag label)	Multi-scale FCN+ CNN(vgg16)	Cascade localization and recognition	General objects
Zhang-TGRS-2016 [208]	CNN	CNN	None	ImageNet(tag label), auxiliary data(image label)	CPRNet+LocNet	Alternative training CPRNet and LocNet	Aircraft

min-entropy and local min-entropy losses are introduced to train a DNN model to select the proposal cliques of largest object probability and mine truthful object locations from the selected proposal cliques. Zhang et al. [212] develop a self-generated guidance method for weakly supervised object localization. In this work, a self-generated guidance map is derived from a CAM layer to help learning features and object location masks from the previous network layers. More recently, Gao et al. [46] propose a token semantic coupled attention mapping for WSOL, which models the long-range visual dependency of the image regions and thus avoid partial activation. Ren et al. [113] introduce the instance-associative spatial diversification constraints and build the parametric spatial dropout block to address the instance ambiguity and incomplete localization problems. Besides, they additionally adopt a sequential batch back-propagation algorithm, which enables their model to use a large ResNet as the backbone⁴. Although there are other methods that use ResNet as the backbone [2], [21], [193], there is very limited exploration of using more recent backbone architectures, e.g., DesNet [65] and Res2net [45], in both WSOL and WSOD frameworks.

5.2 Multi-Network Training

The methods in this school collaborate multiple networks, either in one training stage or in multiple training stages, to accomplish the weakly supervised object localization or detection task. The approaches of this category usually train a network to mine the initial object regions [48], [94], [176] and another network for the detection task under the MIL framework [94], [148], [163], [210]. An additional object detection network, e.g., Fast RCNN, may also be used to train the final object detectors [48], [200], [215]. By integrating multiple networks, these methods tend to achieve

better performance both in object localization and detection. A brief summary of these approaches is shown in Table 8.

Li et al. [93] propose a multiple instance curriculum learning method, where a network based on the WSDDN [8] model is used to mine candidate object proposals and another one based on the CAM [221] algorithm to generate saliency maps from the selected proposals. A curriculum is designed to select confident training examples based on the consistency between the regions outputted by the two networks. The object detectors are then trained by using the confident training examples iteratively. Dong et al. [31] present a dual-network progressive approach for weakly supervised object detection, where a positive instance selection network and a region refinement network are adopted to minimize the classification error and modify object localization, respectively. These two networks are worked under a co-training paradigm. In [48], Ge et al. first obtain intermediate object localization and pixel labeling results using a classification network. A triplet-loss network and an instance classification network are then constructed to detect outlier and filter object instances. Finally, the filtered object instances are used as the supervision to train another Fast RCNN-based detection network. In order to overcome the limitations brought by the imprecise of the extracted object proposals, Wei et al. [176] propose to mine object proposals with tight boxes to learn weakly supervised object detector. The assumption is that the proposals with tight boxes are more likely to contain the objects of interest thus mining such kind of proposals would help screen the cluttered background regions. In their approach, a semantic segmentation network is first learned using the object localization map generated by CAM as the pseudo ground-truth. The predicted segmentation masks are used to mine object proposals with tight boxes, and fed into the online instance classifier refinement (OICR) network to learn weakly supervised object detector. With the same motivation as [176], Tang et al. [148] propose to combine a two-stage region proposal network with an OICR network to learn the weakly supervised object detectors. Instead of mining proposals based

4. As discussed in [194], it is non-trivial to introduce the deeper network backbones, e.g. the deep residual network, into the weakly supervised object detection frameworks as it would encounter dramatic deterioration of detection accuracy and training non-convergence.

TABLE 9

Brief summarization of the characteristics of the datasets. The top three are the datasets usually used for the WSOD task, while the bottom two are the common datasets for the WSOL task. #Categories indicates the number of object categories. #Images indicates the number of images. “GO” is short for Generic Objects.

Dataset	Content	#Categories	#Images	Metrics
PASCAL VOC 07			9,962	
PASCAL VOC 10	GO	20	21,738	mAP, CorLoc
PASCAL VOC 12			22,531	
CUB-200-2011	Birds	200	11,788	Top-1/5 Loc
ILSVRC 2016	GO	1000	1.2 M	GT Loc

on the semantic segmentation cue, Tang et al. use both the low-level cues (feature maps in shallow layers) and the high-level cues (semantic scores in deep layers) to mine reliable proposals in the two stages, respectively. The parameters of the region proposal network are obtained based on the network trained by [147]. Zhang et al. [210] explore the reliability of each training image by evaluating the image difficulty and then feed the images into the learning procedure in an easy-to-hard order. Specifically, the image difficulty is evaluated by diagnosing the localization outputs of the pre-trained WSDDN model based on the concentrateness of the high-scored proposal locations. In [215], Zhang et al. use three networks to learn weakly supervised object detectors. First, an OICR network is trained to generate the initial object regions. Then, they train an RPN [111] based on the pseudo ground-truth boxes obtained after implementing a post-process on the initial object regions, and use the learned RPN to generate more accurate object locations. Finally, fully supervised object detectors [49], [111] are trained based on the obtained object locations.

5.3 Discussion

Compared with the off-the-shelf deep model-based weakly supervised object detection and localization methods, the deep weakly supervised learning methods exploits the merits of deep learning and weakly supervised learning approaches. Without complex design in the learning initialization stage, the end-to-end deep weakly supervised learning methods have been shown to perform well by introducing the MIL mechanism into the network design of the DCNN models. While the multi-network training methods can further improve the learning performance by combining multiple function-specific networks. On the other hand, the performance of these methods is limited by whether the information extracted from the weakly-supervised module is effective or not. As such, prior knowledge may be useful to guide the deep learning process without solely relying on the weakly-supervised network.

6 DATASETS AND EVALUATION METRICS

During the last decades, significant efforts have been made to develop various methods for learning weakly supervised object localizer or detector. For fair performance evaluation, it is of great importance to introduce some publicly available benchmark datasets and evaluation metrics.

Existing weakly supervised object detection methods are usually evaluated on the PASCAL VOC datasets, including the PASCAL VOC 2007, 2010, and 2012 sets. The PASCAL VOC 2007 [36], PASCAL VOC 2010 [37] and PASCAL VOC 2012 [35] contain 9,962, 21,738, and 22,531 images of 20 object classes. These three datasets are divided into train, val, and test sets, where the



Fig. 6. Illustration of examples from the PASCAL VOC (top block), CUB-200-2011 (bottom-left block), and ILSVRC 2016 (bottom-right block) datasets.

trainval set (5,011 images for PASCAL VOC 2007, 10,869 images for PASCAL VOC 2010, and 11,540 images for PASCAL VOC 2012) are used to train the weakly supervised object detector, and the rest for evaluation. The mean of AP (mAP) metric is used to measure the performance where one object is successfully detected if the intersection over union (IoU) between the ground-truth and predicted boxes is more than 50 percentage.

The weakly supervised object localization performance is usually evaluated on the PASCAL VOC, ILSVRC, and CUB datasets. On the PASCAL VOC datasets [35], [36], [37], weakly supervised object localization methods only use the trainval sets, which are different from the setting in weakly supervised object detection. That is, both the weakly supervised learning process and localization process are implemented on the same image data. To evaluate the localization performance on the PASCAL VOC datasets, the correct localization metric (CorLoc) is adopted, where the bounding-box with the highest class-specific score from each image is examined to be whether correct (with more than 50% overlap with the ground-truth box) or not. In addition to the PASCAL VOC datasets, the ILSVRC 2016 dataset [116] (i.e., the ImageNet) and CUB-200-2011 dataset [155] are also widely used for performance evaluation. The ILSVRC 2016 dataset contains more than 1.2 million images of 1,000 classes for training, while the validation set, which contains 50,000 images, is used for testing. The CUB-200-2011 dataset contains 11,788 images of 200 categories with 5,994 images for training and 5,794 for testing. For these two datasets, The commonly-used evaluation metrics are GT-known localization accuracy (GT Loc), Top-1 localization accuracy (Top-1 Loc), and Top-5 localization accuracy (Top-5 Loc). Specifically, GT Loc judges the localization results as correct when the intersection over union (IoU) between the ground-truth bounding box and the estimated box is no less than 50%, while Top-1 Loc considers the localization results as correct when the class predicted with the highest score is equal to the ground-truth class and the estimated bounding box has no less than 50% IoU with the ground-truth bounding box [21]. Top-5 Loc differs from Top-1 Loc in that it checks if the target label is one of the top 5 predictions. As can be seen, the Top-1 Loc is a harder metric than the GT Loc as it needs to additionally

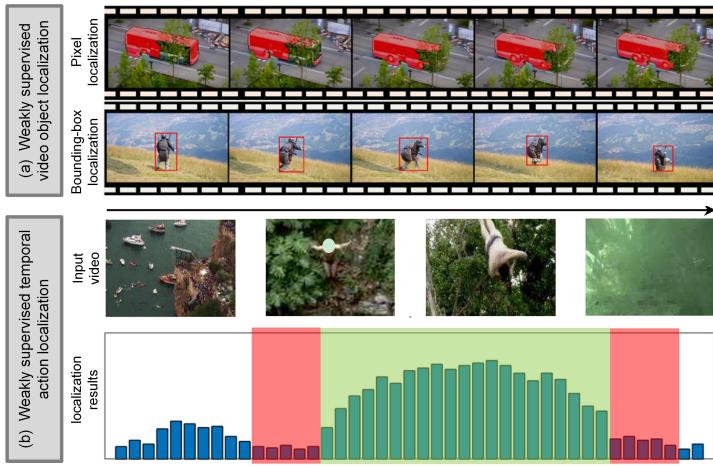


Fig. 7. Weakly supervised object localization or detection methods for video understanding. The examples are from [132], [202].

predict the class label correctly. This will dramatically increase the task difficulty when performing under very large or fine-grained semantic spaces. The difficulty of Top-5 Loc is between Top-1 Loc and GT Loc as it requires the model to predict the class label but does not restrict the prediction to be perfectly correct.

We provide a brief summarization of the characteristics of the aforementioned datasets in Table 9. We additionally show some examples from each dataset to illustrate the bias of image content in different datasets. As displayed in Fig. 6, the PASCAL datasets contain relatively more complex image content, where multiple object instances and categories may appear in a single image and different images would contain objects with significant scale variations. Although it only contains 20 object categories, its category diversity is higher than the CUB-200-2011 dataset as all the 200 categories in the CUB-200-2011 dataset are related to birds. ILSVRC 2016 contains far more images and categories than the PASCAL datasets. However, the content of the images from the ILSVRC 2016 dataset tends to be simpler than that of the PASCAL datasets—each of the images from the ILSVRC 2016 dataset typically contains only one single object and the objects have more consistent sizes and are placed in clearer background context relative to those from the PASCAL datasets.

7 APPLICATIONS

In recent years, weakly supervised object localization and detection techniques have been used in numerous vision problems, especially when it is difficult to collect ground-truth labels.

7.1 Video Understanding

As it is time-consuming to obtain object-level annotations for each video frame, weakly supervised object localization and detection methods have also been applied in the field of video understanding [14], [68], [101], [118], [132], [191], [202] (see Fig. 7). For example, Chanda et al. [14] build a two-stream learning framework, which adapts the information from the labeled images (source domain) to the weakly labeled videos (target domain). In [202] Zhang et al. propose a self-paced fine-tuning network for learning two network heads to localize and segment the object of interest from the weakly labeled training videos. The network is equipped with the multi-task self-paced learning function which can integrate confident knowledge from

each single task (localization or segmentation) and use it to build stronger deep feature representation for both tasks. On the other hand, [107], [132], [136], [166] develop methods to localize temporal actions in the given untrimmed videos, where the main goal is to predict the temporal boundary of each action instance contained in the weakly labeled training videos. Essentially, such a weakly supervised action localization (WSAL) task is an emerging, yet rapidly developing topic in recent years, and the methods for solving this task are highly related to the weakly supervised object detection and localization methods. The additional challenges are: (i) the duration of the interesting action has very large variation, i.e., from a few seconds to thousands of seconds; and (ii) the features extracted to represent the interesting action would be entangled with those of the complex scenes of the video frame. Notice that when applying to video understanding, there are strong correlations among adjacent video frames. So, additional informative constraints can be introduced to facilitate the weakly supervised object detection or localization under this scenario.

7.2 Art Image Analysis

One interesting application of the weakly supervised object localization and detection techniques is the analysis of art images (see Fig. 8). Inoue et al. [70] propose a cross-domain weakly-supervised object detection framework for learning the object detectors from weakly labeled watercolor images. A progressive domain adaptation method to transfer the style of the fully-labeled data from the source domain (the normal RGB domain) to the target domain (the watercolor domain) is developed. In [54] Gonthier et al. propose a weakly supervised learning algorithm for detecting objects in paintings. The IconArt database which contains object classes that are absent from the photographs in daily life is developed for performance evaluation. In addition, Crowley and Zisserman [25] adopt a weakly supervised object localization scheme for automatically annotating images of gods and animals in decorations on classical Greek vases. When applying to art image analysis, a key challenge arises due to the distinctiveness of the content domain—even the same semantics and image scenes would be presented differently to those in the natural environment. Under this scenario, models with stronger self-domain adaptation capacity would be required for the task.

7.3 Medical Imaging

As shown in Fig. 9, medical image analysis is one area where the weakly supervised object localization and detection methods are of critical importance as only few annotations of target objects by trained experts in bio-image (e.g., organ or tissues). To alleviate this problem, Hwang and Kim [69] develop a two-stream DNN model to localize the tuberculosis regions from the chest X-ray images. Considering that the medical image-based applications usually do not have the pre-trained networks, this work proposes a weakly supervised learning scheme without requiring any pre-trained network parameters. The proposed network contains a fully connected layer-based classification branch and a CAM-based localization branch with shared convolutional layers for feature extraction. Both of the two branches are supervised by image label annotation, where a weighting parameter is introduced to dynamically control the relative importance between them to gradually switch the focus of the learning process from the classification branch to the localization branch. The authors demonstrate that the features

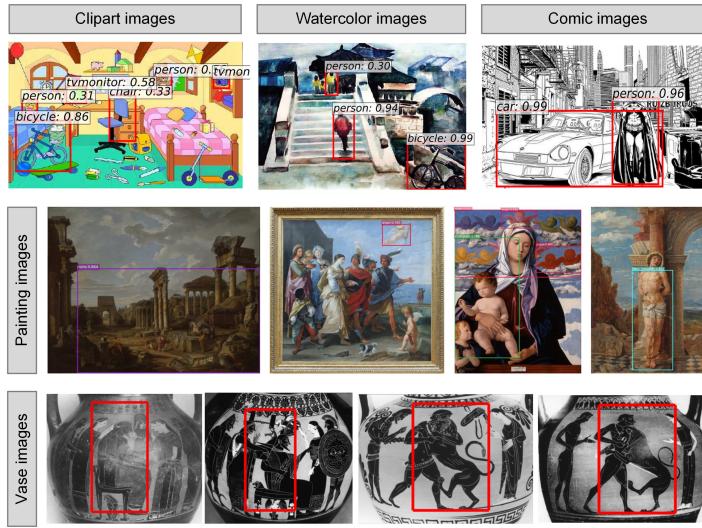


Fig. 8. Examples of the application of weakly supervised object localization or detection approaches for the analysis of art images. The examples are from [25], [54], [70], where detection results in different colors in the painting images indicate different types of objects.

learned from the classification layer at the early stage can provide informative cues to learn the localization branch at the late stage. For detecting a general type of lesions, Wang et al. [168] model the normal image as the combination of background and noise, while modeling the abnormal images as the combination of background, blood vessels, and noise. With the assumption that the noise for the normal image and abnormal image is the unified distribution, the image data can then be decomposed by the low-rank subspace learning technique to obtain the vessel areas. In [53], Gondal et al. apply the weakly supervised object detection network on the retina images and achieve few false positives with high sensitivity on the lesion-level prediction. Li et al. [90] apply a sparse coding-based weakly supervised learning method for localizing actinophrys in microscopic images. Dubost et al. [32] propose weakly supervised regression neural networks for detecting brain lesions. Besides, some recent works also show great research interests in weakly supervised learning-based brain image analysis, such as brain disease prognosis [98], brain tumor or lesion segmentation [71], [180], brain structure estimation [10], etc. Notice that compared to common images, medical imaging data usually suffers from issues of low contrast and limited texture. Fortunately, some spatial priors could be obtained for different organs or lesions. These priors can be used to guide the weakly supervised learning process on medical imaging data.

7.4 Remote Sensing Imagery Analysis

Remote sensing imagery analysis is one of the most widely studied applications based on weakly supervised object localization and detection, where the input images are usually of large scale and the annotation process tends to be very time-consuming [18], [40], [41], [189] (see Fig. 10). Zhang et al. [199] propose a saliency-based weakly supervised detector learning method to learn the detectors of the airplane, vehicle, and airport from the remote sensing images collected from different sensors. In this work, a weakly supervised object detection benchmark dataset for remote sensing imagery analysis is developed. Han et al. [57] and Zhou et al. [223] introduce the Bayesian inference and negative bootstrapping methods, respectively, for effective weakly supervised detectors for remote sensing images. In [208],

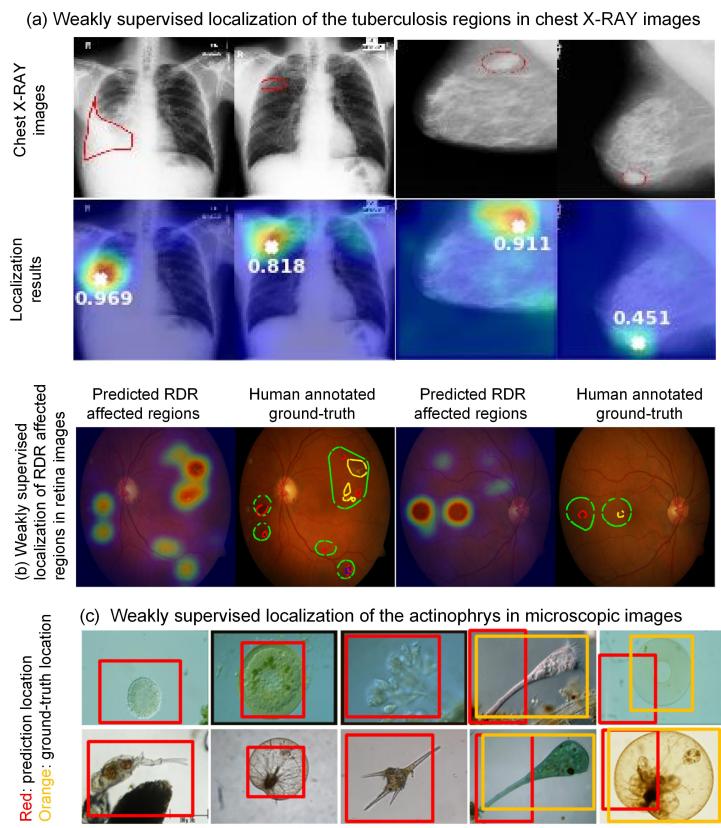


Fig. 9. Examples of the application of weakly supervised object localization or detection approaches in medical image analysis. The examples are from [53], [69], [90].

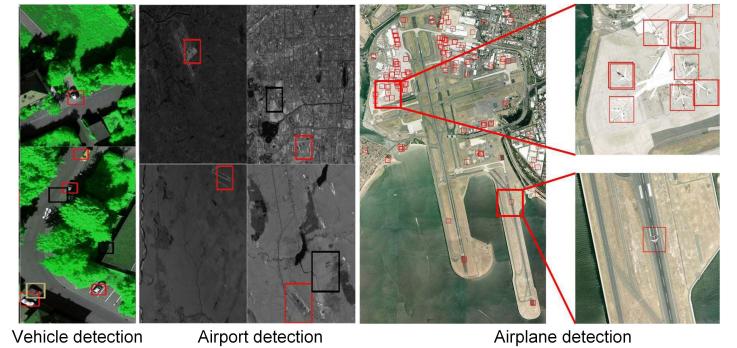


Fig. 10. Examples of the application of weakly supervised object localization or detection approaches in remote sensing imagery analysis. The examples are from [57], [208].

Zhang et al. propose a coupled CNN method which combines a candidate region proposal network and a localization network to detect aircrafts in images. When applying to remote sensing imagery analysis, the target objects are usually very small in size, which would dramatically increase the localization and detection difficulty given only weak supervision.

8 FUTURE DIRECTIONS

We discuss the issues to be addressed in this field for future research.

8.1 Multiple Instance Learning

Weakly supervised object localization or detection methods can be easily formulated within the MIL framework. Early methods in this field usually add prior knowledge [140] or post

regularization [6] on the classic MIL models, such as LSVM [190], while the current research obtains the breakthrough by building deep MIL models [172], [179], [221]. To further improve the weakly supervised learning performance, efforts should be made to introduce the most advanced ideas and techniques in the research field of MIL, such as the set-level problem [182] the key instance shift issue [217] and the scalable issue [66] in MIL. Further research towards more advanced MIL techniques would also bring helpful insights for the WSOL and WSOD in the future.

8.2 Multi-Task Learning

Another future direction is to combine multiple weakly supervised learning tasks into a unified learning framework. These tasks may include object detection [49], semantic segmentation [16], instance segmentation [79], 3D shape reconstruction [38], and depth estimation [51]. Essentially, efforts for simultaneously accomplishing multiple aforementioned tasks have been made in the conventional fully supervised learning scenario [58], [79], [178], [218], [225], which have demonstrated that such learning mechanism can bring helpful information from one task to the other ones. The methods proposed by Zhang et al. [203] is an early attempt to implement such a weakly supervised multi-task learning mechanism and the experimental results show that training object segmentation and 3D shape reconstruction models jointly indeed benefits the both weakly supervised learning tasks. With similar spirits to [203], Zhang et al. [202] and Shen et al. [124] establish a self-paced fine-tuning network and a cyclic guidance network to jointly learn object localization and segmentation models under the weak supervision, respectively. Under the multi-task weakly supervised learning scenario, one key problem is that the learning ambiguity of each individual task might be aggregated and the imprecise prediction on one task might affect the learning on other tasks. To deal with this problem, one needs to disentangle the complex multi-task learning, separately learning each individual task first, and then leveraging the confidence knowledge from each task to provide informative priors to guide the learning processes of the other tasks.

8.3 Robust Learning Theory

To address the learning under uncertainty issue that is inherently existed in the weakly supervised learning process, robust learning strategy will become one of the key techniques in the future. The goal is to alleviate the influence of the noisy samples during the learning process. In implementation, such learning strategy is usually achieved by selecting easy and confident training samples in the early learning stages while using hard and more ambiguous training samples in the late learning stages. Essentially, a number of recent methods have already introduced the robust learning strategies into their learning frameworks. For example, Shi and Ferrari [127] propose a curriculum learning strategy to feed training images into the WSOL learning loop in order from images containing bigger objects down to smaller ones. The training order is determined by the size of the object, which is estimated based on a regression model. Similarly, Zhang et al. [210] design a zigzag learning strategy, where they first develop a criterion to automatically rank the localization difficulty of an image, and then learn the detector progressively by feeding examples with increasing difficulty. As can be seen, these methods are just intuitive ways to introduce robust learning strategy into the weakly supervised object localization and detection frameworks, while they have already

achieved obvious performance gains when compared with the conventional learning strategy. Along this line, Zhang et al. [204] propose a self-paced curriculum learning framework for weakly supervised object detection. By integrating the curriculum learning [5] with the self-paced learning [86], the established learning framework provides a more theoretical-sounded way to improve the learning robustness. However, the solid robust learning theory is still lack in this research field.

8.4 Reinforcement and Adversarial Learning

Besides the conventional CNN models, it is also worth trying to apply some more advanced learning models into the learning process of the weakly supervised object detector. Here we give two examples. The first one is the deep reinforcement learning. According to [89], biological vision systems are believed to have a sequential process with changing retinal fixations that gradually accumulate evidence of certainty when searching or localizing objects. Several existing methods [11], [64], [74], [96], [192] have also demonstrated that designing deep reinforcement learning frameworks to model such a sequential searching process can indeed help to address the object localization, detection, and tracking problems in the computer vision community. Thus, it is highly desirable, both biologically and computationally, to explore deep reinforcement learning models that facilitate the weakly supervised object localization and detection systems in such a sequential searching process [205]. The second one is the generative adversary learning. As we know, generative adversary learning has been demonstrated to have advantages in unsupervised and semi-supervised learning scenarios [55], [133], [142], [153]. It can generate the desired data distribution based on very weak supervision, i.e., “real” or “fake”. Such capacity endows generative adversary learning very large potential in solving the weakly supervised object localization and detection problems. Although existing methods, such as [29], [125], [211], have already made efforts to introduce such a learning mechanism into the weakly supervised object localization and detection, there is still much room for improvement along this research direction.

8.5 Prior-guided Deep MIL

From Table 7 and Table 8, we can observe that most of the current deep weakly supervised object detection methods have not introduced any prior knowledge into their learning frameworks. However, from our review on classic models (see Sec. 3), prior knowledges actually play important roles in avoiding the weakly supervised learning process from drifting to trivial solutions. Considering this issue, some recent works utilize prior knowledges of saliency [92], objectness [110], [219], shape [93], count [44], [63], human action [188], human object interaction [80], mask-out scoring [183] in their frameworks. However, research towards building effective deep MIL frameworks (such as the one with prior knowledge distillation [15] or cross domain adaptation [62]) to embed helpful prior knowledge into the weakly supervised learning process needs to be further explored in the future. In addition, the co-occurring patterns mined in co-saliency detection [201], [206] and object co-localization [120], [175] approaches can also be used as informative priors to guide the deep multiple instance learning process in weakly supervised object localization and detection.

9 CONCLUSIONS

In this paper, we provide a comprehensive survey of existing literatures in the research field of weakly supervised object

localization and detection. We start with the introduction of the definition of the task and the key challenges that make the weakly supervised learning process hard to implement. Then, we introduce the development history of this field, the taxonomy of methods for weakly supervised object localization and detection, and the relationship between different categories. After reviewing existing literatures in each category of methodology, we introduce the benchmark datasets and evaluation metrics that are widely used in this field, which are followed by the reviewing of the applications of the existing weakly supervised object localization and detection algorithms. Finally, we point out several future directions that may further promote the development of this research field.

REFERENCES

- [1] A. Arun, C. Jawahar, and M. Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 9
- [2] W. Bae, J. Noh, and G. Kim. Rethinking class activation mapping for weakly supervised object localization. 9
- [3] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani. Self-taught object localization with deep networks. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 5, 6
- [4] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. Manjunath. Weakly supervised localization using deep feature maps. In *European Conference on Computer Vision*, pages 714–731. Springer, 2016. 6
- [5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009. 13
- [6] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference*, volume 3, 2014. 5, 13
- [7] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1081–1089, 2015. 5, 6
- [8] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 1, 7, 8, 9
- [9] M. B. Blaschko, A. Vedaldi, and A. Zisserman. Simultaneous object detection and ranking with weak supervision. In *International Conference on Neural Information Processing Systems*, 2010. 1, 4
- [10] D. Bontempi, S. Benini, A. Signoroni, M. Svanera, and L. Muckli. Cerebrum: a fast and fully-volumetric convolutional encoder-decoder for weakly-supervised segmentation of brain structures from out-of-the-scanner mri. *Medical image analysis*, 62:10168, 2020. 12
- [11] J. C. Caicedo and S. Lazebnik. Active object localization with deep reinforcement learning. In *ICCV*, 2015. 13
- [12] L. Cao, L. Feng, C. Li, Y. Sheng, H. Wang, W. Cheng, and R. Ji. Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning. *Pattern Recognition*, 64:417–424, 2016. 1
- [13] L. Cao, F. Luo, L. Chen, Y. Sheng, H. Wang, C. Wang, and R. Ji. Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning. *Pattern Recognition*, 64:417–424, 2017. 3, 4
- [14] O. Chanda, E. W. Teh, M. Rochan, Z. Guo, and Y. Wang. Adapting object detectors from images to weakly labeled videos. In *MBVC*, 2017. 11
- [15] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 742–751, 2017. 13
- [16] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 1, 13
- [17] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1431–1439, 2015. 5, 6, 7
- [18] G. Cheng, J. Yang, D. Gao, L. Guo, and J. Han. High-quality proposals for weakly supervised object detection. *IEEE Transactions on Image Processing*, 29:5794–5804, 2020. 12
- [19] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020. 8
- [20] J. Choe, J. H. Park, and H. Shim. Improved techniques for weakly supervised object localization. 2018. 7
- [21] J. Choe and H. Shim. Attention-based dropout layer for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7, 9, 10
- [22] H. Cholakkal, J. Johnson, and D. Rajan. Backtracking scspm image classifier for weakly supervised top-down saliency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5278–5287, 2016. 4
- [23] R. G. Cibnis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2017. 5
- [24] D. J. Crandall and D. P. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *European Conference on Computer Vision*, 2006. 2
- [25] E. J. Crowley and A. Zisserman. Of gods and goats: Weakly supervised learning of figurative art. *learning*, 8:14, 2013. 11, 12
- [26] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *European conference on computer vision*, pages 452–466. Springer, 2010. 3, 4, 5
- [27] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision*, 100(3):275–293, 2012. 3, 4, 5
- [28] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 914–922, 2017. 7, 8
- [29] A. Diba, V. Sharma, R. Stiefelhagen, and L. Van Gool. Weakly supervised object discovery by generative adversarial & ranking networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 13
- [30] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013. 6
- [31] X. Dong, D. Meng, F. Ma, and Y. Yang. A dual-network progressive approach to weakly supervised object detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 279–287. ACM, 2017. 9
- [32] F. Dubost, H. Adams, P. Yilmaz, G. Bortsova, G. van Tulder, M. A. Ikram, W. Niessen, M. Vernooy, and M. de Bruijne. Weakly supervised object detection with 2d and 3d regression neural networks. *arXiv preprint arXiv:1906.01891*, 2019. 12
- [33] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 642–651, 2017. 7, 8
- [34] T. Durand, N. Thome, and M. Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4743–4752, 2016. 7
- [35] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 10
- [36] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc2007) results. 2007. 10
- [37] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 10
- [38] H. Fan, S. Hao, and L. Guibas. A point set generation network for 3d object reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 13
- [39] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. 3
- [40] X. Feng, J. Han, X. Yao, and G. Cheng. Progressive contextual instance refinement for weakly supervised object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 12
- [41] X. Feng, J. Han, X. Yao, and G. Cheng. Tcanet: Triple context-aware network for weakly supervised object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 12
- [42] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision*, 71(3):273–303, 2007. 2
- [43] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie. Weakly supervised object localization with stable segmentations. In *European Conference on Computer Vision*, pages 193–207. Springer, 2008. 1, 4
- [44] M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis. C-wsl: Count-guided weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 152–168, 2018. 7, 13
- [45] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 9
- [46] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. *arXiv preprint arXiv:2103.14862*, 2021. 9
- [47] Y. Gao, B. Liu, N. Guo, X. Ye, F. Wan, H. You, and 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 (ICCV)*, October 2019. 8
- [48] W. Ge, S. Yang, and 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 Conference on Computer Vision and Pattern Recognition*, pages 1277–1286, 2018. 9
- [49] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 5, 10, 13
- [50] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 5
- [51] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 13
- [52] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2409–2416, 2014. 3, 4, 5
- [53] W. M. Gondal, J. M. Köhler, R. Grzeszick, G. A. Fink, and M. Hirsch. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2069–2073. IEEE, 2017. 12
- [54] N. Gonthier, Y. Gousseau, S. Ladjal, and O. Bonfait. Weakly supervised object detection in artworks. In *European Conference on Computer Vision*, pages 692–709. Springer, 2018. 5, 11, 12
- [55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 13
- [56] A. Gudi, N. van Rosmalen, M. Loog, and J. van Gemert. Object-extent pooling for weakly supervised single-shot localization. *British Machine Vision Conference*, 2017. 7
- [57] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3325–3337, 2015. 1, 5, 12
- [58] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014. 13
- [59] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *European Conference on Computer Vision*, pages 459–472. Springer, 2012. 7
- [60] M. Hoai, L. Torresani, F. De la Torre, and C. Rother. Learning discriminative localization from weakly labeled data. *Pattern Recognition*, 47(3):1523–1534, 2014. 1, 3, 4
- [61] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 2883–2891, 2015. 5, 6
- [62] W. Hong, Z. Wang, M. Yang, and J. Yuan. Conditional generative adversarial network for structured domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 13
- [63] C.-Y. Hsu and W. Li. Learning from counting: Leveraging temporal classification for weakly supervised object localization and detection. 13
- [64] C. Huang, S. Lucey, and D. Ramanan. Learning policies for adaptive tracking with deep feature cascades. In *ICCV*, 2017. 13
- [65] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger. Condensenet: An efficient densenet using learned group convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2761, 2018. 9
- [66] S.-J. Huang, W. Gao, and Z.-H. Zhou. Fast multi-instance multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 13
- [67] Z. Huang, Y. Zou, B. Kumar, and D. Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *Advances in Neural Information Processing Systems*, 33, 2020. 7
- [68] D. Huh, T. Kim, and J. Kim. Patch-wise weakly supervised learning for object localization in video. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAICC)*, pages 263–266. IEEE, 2019. 11
- [69] S. Hwang and H.-E. Kim. Self-transfer learning for weakly supervised lesion localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–246. Springer, 2016. 1, 11, 12
- [70] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5001–5009, 2018. 7, 11, 12
- [71] Z. Ji, Y. Shen, C. Ma, and M. Gao. Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 175–183. Springer, 2019. 12
- [72] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, and H.-K. Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2079, 2019. 7
- [73] W. Jiang, T. Ngo, B. Manjunath, Z. Zhao, and F. Su. Optimizing region selection for weakly supervised object detection. *arXiv preprint arXiv:1708.01723*, 2017. 7
- [74] Z. Jie, X. Liang, J. Feng, X. Jin, W. Lu, and S. Yan. Tree-structured reinforcement learning for sequential object localization. In *NIPS*, 2016. 13
- [75] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1377–1385, 2017. 5, 6
- [76] A. Kanezaki, Y. Kuniyoshi, and T. Harada. Weakly-supervised multi-class object detection using multi-type 3d features. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 605–608. ACM, 2013. 4
- [77] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *European Conference on Computer Vision*, pages 350–365. Springer, 2016. 7, 8
- [78] I. Khan, P. M. Roth, and H. Bischof. Learning object detectors from weakly-labeled internet images. In *35th OAGM/AAPR Workshop*, volume 326, 2011. 3, 4
- [79] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 13
- [80] D. Kim, G. Lee, J. Jeong, and N. Kwak. Tell me what they're holding: Weakly-supervised object detection with transferable knowledge from human-object interaction. *arXiv preprint arXiv:1911.08141*, 2019. 13
- [81] A. Kolesníkov and C. H. Lampert. Improving weakly-supervised object localization by micro-annotation. *arXiv preprint arXiv:1605.05538*, 2016. 6, 7
- [82] S. Kosugi, T. Yamasaki, and K. Aizawa. Object-aware instance labeling for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6064–6072, 2019. 9
- [83] J. Krapac and S. Segvic. Weakly supervised object localization with large fisher vectors. In *International Conference on Computer Vision Theory and Applications*, 2015. 1
- [84] J. Krapac and S. Šegvić. Weakly supervised object localization with large fisher vectors. In *10th International Conference on Computer Vision Theory and Applications*, 2015. 4
- [85] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 5, 6
- [86] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010. 13
- [87] K. Kumar Singh and Y. Jae Lee. You reap what you sow: Using videos to generate high precision object proposals for weakly-supervised object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 9
- [88] K. Kumar Singh, F. Xiao, and Y. Jae Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3548–3556, 2016. 5, 6
- [89] H. Larochelle and G. E. Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *NIPS*, 2010. 13
- [90] C. Li, K. Shirahama, and M. Grzegorzek. Environmental microorganism classification using sparse coding and weakly supervised learning. In *Proceedings of the 2nd International Workshop on Environmental Multimedia Retrieval*, pages 9–14. ACM, 2015. 12
- [91] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016. 6, 7
- [92] G. Li, Y. Xie, and L. Lin. Weakly supervised salient object detection using image labels. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 13
- [93] S. Li, X. Zhu, Q. Huang, H. Xu, and C.-C. J. Kuo. Multiple instance curriculum learning for weakly supervised object detection. *BMVC*, 2017. 9, 13
- [94] Y. Li, J. Zhang, K. Huang, and J. Zhang. Mixed supervised object detection with robust objectness transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 9
- [95] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:182–196, 2018. 6

- [96] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, 2017. 13
- [97] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 999–1007, 2015. 6
- [98] M. Liu, J. Zhang, C. Lian, and D. Shen. Weakly supervised deep learning for brain disease prognosis using mri and incomplete clinical scores. *IEEE transactions on cybernetics*, 50(7):3381–3392, 2019. 12
- [99] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 8
- [100] W. Lu, X. Jia, W. Xie, L. Shen, Y. Zhou, and J. Duan. Geometry constrained weakly supervised object localization. *arXiv preprint arXiv:2007.09727*, 2020. 8
- [101] F. Ma, L. Zhu, Y. Yang, S. Zha, G. Kundu, M. Feiszli, and Z. Shou. Sf-net: Single-frame supervision for temporal action localization. In *European Conference on Computer Vision*, pages 420–437. Springer, 2020. 11
- [102] J. Mai, M. Yang, and W. Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7
- [103] J. Mai, M. Yang, and 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*, pages 8766–8775, 2020. 8
- [104] S. Mathe and C. Sminchisescu. Multiple instance reinforcement learning for efficient weakly-supervised detection in images. In *Arxiv*, 2014. 1
- [105] S. Mathe and C. Sminchisescu. Multiple instance reinforcement learning for efficient weakly-supervised detection in images. *arXiv preprint arXiv:1412.0100*, 2014. 5
- [106] M. H. Nguyen, L. Torresani, F. De La Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1925–1932. IEEE, 2009. 4
- [107] P. Nguyen, T. Liu, G. Prasad, and B. Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761, 2018. 11
- [108] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?–weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015. 7, 8
- [109] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *2011 International Conference on Computer Vision*, pages 1307–1314. IEEE, 2011. 4
- [110] A. Rahimi, A. Shaban, T. Ajanthan, R. Hartley, and B. Boots. Pairwise similarity knowledge transfer for weakly supervised object localization. 13
- [111] S. Ren, R. Girshick, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 1, 10
- [112] W. Ren, K. Huang, D. Tao, and T. Tan. Weakly supervised large scale object localization with multiple instance learning and bag splitting. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):405–416, 2016. 1, 5, 6
- [113] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, Y. J. Lee, A. G. Schwing, and 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*, pages 10598–10607, 2020. 9
- [114] M. Rochan, S. Rahman, N. D. Bruce, and Y. Wang. Weakly supervised object localization and segmentation in videos. *Image and Vision Computing*, 56:1–12, 2016. 5
- [115] M. Rochan and Y. Wang. Weakly supervised localization of novel objects using appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4315–4324, 2015. 5
- [116] 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*, 115(3):211–252, 2015. 5, 10
- [117] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe. Self paced deep learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):712–725, 2019. 7
- [118] J. Schroeter, K. Sidorov, and D. Marshall. Weakly-supervised temporal localization via occurrence count learning. *Proceedings of International Conference on Machine Learning*, 2019. 11
- [119] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 7, 8
- [120] A. Shaban, A. Rahimi, S. Bansal, S. Gould, B. Boots, and R. Hartley. Learning to find common objects across few image collections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5117–5126, 2019. 13
- [121] Y. Shen, R. Ji, Z. Chen, X. Hong, F. Zheng, J. Liu, M. Xu, and Q. Tian. Noise-aware fully weakly supervised object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7
- [122] Y. Shen, R. Ji, C. Wang, X. Li, and X. Li. Weakly supervised object detection via object-specific pixel gradient. *IEEE transactions on neural networks and learning systems*, (99):1–11, 2018. 7
- [123] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7
- [124] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 697–707, 2019. 13
- [125] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang. Generative adversarial learning towards fast weakly supervised detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5764–5773, 2018. 9, 13
- [126] M. Shi, H. Caesar, and V. Ferrari. Weakly supervised object localization using things and stuff transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3381–3390, 2017. 5, 6, 7, 8
- [127] M. Shi and V. Ferrari. Weakly supervised object localization using size estimates. In *European Conference on Computer Vision*, pages 105–121. Springer, 2016. 5, 6, 13
- [128] Z. Shi, T. M. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2984–2991, 2013. 3, 4
- [129] Z. Shi, T. M. Hospedales, and T. Xiang. Bayesian joint modelling for object localisation in weakly labelled images. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):1959–1972, 2015. 3, 4
- [130] Z. Shi, P. Siva, and T. Xiang. Transfer learning by ranking for weakly supervised object annotation. *BMVC*, 2012. 3, 4
- [131] Z. Shi, P. Siva, and T. Xiang. Transfer learning by ranking for weakly supervised object annotation. *arXiv preprint arXiv:1705.00873*, 2017. 3
- [132] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang. Autoloc: weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018. 11
- [133] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2107–2116, 2017. 13
- [134] K. Sikka, A. Dhall, and M. Bartlett. Weakly supervised pain localization using multiple instance learning. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013. 3, 4
- [135] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5
- [136] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017. 7, 11
- [137] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *European Conference on Computer Vision*, pages 594–608. Springer, 2012. 3, 4
- [138] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3238–3245, 2013. 3, 4
- [139] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. 2011. 4, 5
- [140] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. *arXiv preprint arXiv:1403.1024*, 2014. 5, 6, 12
- [141] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *Advances in Neural Information Processing Systems*, pages 1637–1645, 2014. 5, 6
- [142] J. T. Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015. 13
- [143] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. Bourdev. Pronet: Learning to propose object-specific boxes for cascaded neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3493, 2016. 8, 9
- [144] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1464–1471, 2014. 4
- [145] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. L. Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 7
- [146] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection

- network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2843–2851, 2017. 7
- [147] P. Tang, X. Wang, Z. Huang, X. Bai, and W. Liu. Deep patch learning for weakly supervised object classification and discovery. *Pattern Recognition*, 71:446–459, 2017. 7, 8, 10
- [148] P. Tang, X. Wang, A. Wang, Y. Yan, W. Liu, J. Huang, and A. Yuille. Weakly supervised region proposal network and object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–368, 2018. 9
- [149] Y. Tang, X. Wang, E. Dellandrea, and L. Chen. Weakly supervised learning of deformable part-based models for object detection via region proposals. *IEEE Transactions on Multimedia*, 19(2):393–407, 2017. 5, 6
- [150] Y. Tang, X. Wang, E. Dellandrea, S. Masnou, and L. Chen. Fusing generic objectness and deformable part-based models for weakly supervised object detection. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4072–4076. IEEE, 2014. 3, 4
- [151] Q. Tao, H. Yang, and J. Cai. Exploiting web images for weakly supervised object detection. *IEEE Transactions on Multimedia*, 2018. 9
- [152] E. W. Teh, M. Rochan, and Y. Wang. Attention networks for weakly supervised object localization. In *BMVC*, pages 1–11, 2016. 7
- [153] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 13
- [154] J. Uijlings, S. Popov, and V. Ferrari. Revisiting knowledge transfer for training object class detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1101–1110, 2018. 5, 6, 8
- [155] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 10
- [156] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019. 7
- [157] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7
- [158] F. Wan, P. Wei, Z. Han, K. Fu, and Q. Ye. Weakly supervised object detection with correlation and part suppression. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3638–3642. IEEE, 2016. 5
- [159] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1306, 2018. 7, 8
- [160] Z. Wan and H. He. Weakly supervised object localization with deep convolutional neural network based on spatial pyramid saliency map. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4177–4181. IEEE, 2017. 7
- [161] C. Wang, K. Huang, W. Ren, J. Zhang, and S. Maybank. Large-scale weakly supervised object localization via latent category learning. *IEEE Transactions on Image Processing*, 24(4):1371–1385, 2015. 1, 5
- [162] C. Wang, W. Ren, and K. Huang. Window mining by clustering mid-level representation for weakly supervised object detection. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4067–4071. IEEE, 2014. 4
- [163] J. Wang, J. Yao, Y. Zhang, and R. Zhang. Collaborative learning for weakly supervised object detection. *arXiv preprint arXiv:1802.03531*, 2018. 9
- [164] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017. 8, 9
- [165] L. Wang, D. Meng, X. Hu, J. Lu, and J. Zhao. Instance annotation via optimal bow for weakly supervised object localization. volume 47, pages 1313–1324. IEEE, 2017. 3, 4
- [166] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017. 11
- [167] L. Wang, J. Zhao, X. Hu, and J. Lu. Weakly supervised object localization via maximal entropy random walk. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1614–1617. IEEE, 2014. 3, 4, 5
- [168] R. Wang, B. Chen, D. Meng, and L. Wang. Weakly-supervised lesion detection from fundus images. *IEEE transactions on medical imaging*, 2018. 4, 12
- [169] S. Wang, J. Joo, Y. Wang, and S.-C. Zhu. Weakly supervised learning for attribute localization in outdoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3111–3118, 2013. 1, 4, 5
- [170] S. Wang, Y. Wang, and S. C. Zhu. Hierarchical space tiling for scene modeling. *ACCV*, 2012. 5
- [171] W. Wang, Y. Wang, F. Chen, and A. Sowmya. A weakly supervised approach for object detection based on soft-label boosting. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 331–338. IEEE, 2013. 3, 4
- [172] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 13
- [173] X. Wang, Z. Zhu, C. Yao, and X. Bai. Relaxed multiple-instance svm with application to object discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1224–1232, 2015. 5
- [174] X.-S. Wei, C.-L. Zhang, Y. Li, C.-W. Xie, J. Wu, C. Shen, and Z.-H. Zhou. Deep descriptor transforming for image co-localization. *arXiv preprint arXiv:1705.02758*, 2017. 5, 6
- [175] X.-S. Wei, C.-L. Zhang, J. Wu, C. Shen, and Z.-H. Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 88:113–126, 2019. 6, 13
- [176] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 434–450, 2018. 8, 9
- [177] T. Wilhelm, R. Grzeszick, G. A. Fink, and C. Woehler. From weakly supervised object localization to semantic segmentation by probabilistic image modeling. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2017. 6
- [178] B. Wu and R. Nevatia. Simultaneous object detection and segmentation by boosting local shape feature based classifier. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 13
- [179] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, 2015. 7, 8, 13
- [180] K. Wu, B. Du, M. Luo, H. Wen, Y. Shen, and J. Feng. Weakly supervised brain lesion segmentation via attentional representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–219. Springer, 2019. 12
- [181] Y. Xie, C. Huang, T. Song, J. Ma, and J. Jing. Object co-detection via low-rank and sparse representation dictionary learning. In *2013 Visual Communications and Image Processing (VCIP)*, pages 1–6. IEEE, 2013. 3, 4
- [182] B.-C. Xu, K. M. Ting, and Z.-H. Zhou. Isolation set-kernel and its application to multi-instance learning. In *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2019. 13
- [183] W. Xu, Y. Wu, W. Ma, and G. Wang. Adaptively denoising proposal collection for weakly supervised object localization. *Neural Processing Letters*, pages 1–14, 2019. 13
- [184] W. Xu, Y. Wu, W. Ma, and G. Wang. Adaptively denoising proposal collection for weakly supervised object localization. *Neural Processing Letters*, 51(1):993–1006, 2020. 7
- [185] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, and Q. Ye. Danet: Divergent activation for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 8
- [186] K. Yang, D. Li, and Y. Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8372–8381, 2019. 7
- [187] S. Yang, Y. Kim, Y. Kim, and C. Kim. Combinational class activation maps for weakly supervised object localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 7, 8
- [188] Z. Yang, D. Mahajan, D. Ghadiyaram, R. Nevatia, and V. Ramanathan. Activity driven weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2917–2926, 2019. 13
- [189] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo. Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 12
- [190] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, volume 2, page 5, 2009. 13
- [191] Y. Yuan, Y. Lyu, X. Shen, I. W. Tsang, and D.-Y. Yeung. Marginalized average attentional network for weakly-supervised learning. *arXiv preprint arXiv:1905.08586*, 2019. 11
- [192] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *CVPR*, 2017. 13
- [193] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019. 9
- [194] Y. W. Z. C. F. Z. F. H. Yunhang Shen, Rongrong Ji and Y. Wu. Enabling deep residual networks for weakly supervised object detection. 9
- [195] V. Zadrija, J. Krapac, S. Šegvić, and J. Verbeek. Sparse weakly supervised models for object localization in road environment. *Computer*

- Vision and Image Understanding*, 176:9–21, 2018. 5
- [196] V. Zadrija, J. Krapac, J. Verbeek, and S. Segvic. Patch-level spatial layout for classification and weakly supervised localization. In *German Conference on Pattern Recognition*, pages 492–503. Springer, 2015. 1, 4
- [197] C.-L. Zhang, Y.-H. Cao, and J. Wu. Rethinking the route towards weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 9
- [198] C.-L. Zhang, Y.-H. Cao, and J. Wu. Rethinking the route towards weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13460–13469, 2020. 8
- [199] D. Zhang, J. Han, C. Gong, Z. Liu, S. Bu, and G. Lei. Weakly supervised learning for target detection in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 12(4):701–705, 2014. 1, 12
- [200] D. Zhang, J. Han, G. Guo, and L. Zhao. Learning object detectors with semi-annotated weak labels. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3622–3635, 2018. 9
- [201] D. Zhang, J. Han, C. Li, J. Wang, and X. Li. Detection of co-salient objects by looking deep and wide. *International Journal of Computer Vision*, 120(2):215–232, 2016. 13
- [202] D. Zhang, J. Han, L. Yang, and D. Xu. Spftn: A joint learning framework for localizing and segmenting objects in weakly labeled videos. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 11, 13
- [203] D. Zhang, J. Han, Y. Yang, and D. Huang. Learning category-specific 3d shape models from weakly labeled 2d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4573–4581, 2017. 13
- [204] D. Zhang, J. Han, L. Zhao, and D. Meng. Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. *International Journal of Computer Vision*, pages 1–18, 2018. 6, 8, 13
- [205] D. Zhang, J. Han, L. Zhao, and T. Zhao. From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 13
- [206] D. Zhang, D. Meng, and J. Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):865–878, 2016. 13
- [207] D. Zhang, D. Meng, L. Zhao, and J. Han. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. *arXiv preprint arXiv:1703.01290*, 2017. 5
- [208] F. Zhang, B. Du, L. Zhang, and M. Xu. Weakly supervised learning based on coupled convolutional neural networks for aircraft detection. *IEEE Transactions on Geoscience and Remote Sensing*, 54(9):5553–5563, 2016. 9, 12
- [209] M. Zhang and B. Zeng. A progressive learning framework based on single-instance annotation for weakly supervised object detection. *Computer Vision and Image Understanding*, 193:102903, 2020. 8
- [210] X. Zhang, J. Feng, H. Xiong, and Q. Tian. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4262–4270, 2018. 9, 10, 13
- [211] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 7, 8, 13
- [212] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 597–613, 2018. 7, 9
- [213] X. Zhang, Y. Wei, and Y. Yang. Inter-image communication for weakly supervised localization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 271–287, 2020. 7
- [214] X. Zhang, Y. Yang, and J. Feng. Mi-locnet: Improving object localization with multi-view learning network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 240–255, 2018. 8, 9
- [215] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 928–936, 2018. 9, 10
- [216] Y. Zhang and T. Chen. Weakly supervised object recognition and localization with invariant high order features. In *BMVC*, pages 1–11, 2010. 3, 4
- [217] Y.-L. Zhang and Z.-H. Zhou. Multi-instance learning with key instance shift. In *IJCAI*, pages 3441–3447, 2017. 13
- [218] L. Zhao and L. S. Davis. Closely coupled object detection and segmentation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 454–461. IEEE, 2005. 13
- [219] Y. Zhong, J. Wang, J. Peng, and L. Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In *Proceedings of European Conference on Computer Vision*, 2020. 9, 13
- [220] B. Zhou, V. Jagadeesh, and R. Piramuthu. Conceptlearner: Discovering visual concepts from weakly labeled image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2015. 6
- [221] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 7, 8, 9, 13
- [222] P. Zhou, G. Cheng, Z. Liu, S. Bu, and X. Hu. Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping. *Multidimensional Systems and Signal Processing*, 27(4):925–944, 2016. 1, 5, 6
- [223] P. Zhou, D. Zhang, G. Cheng, and J. Han. Negative bootstrapping for weakly supervised target detection in remote sensing images. In *2015 IEEE International Conference on Multimedia Big Data*, pages 318–323. IEEE, 2015. 1, 5, 12
- [224] Y. Zhou, Z. Chen, H. Shen, Q. Liu, R. Zhao, and Y. Liang. Dual-attention focused module for weakly supervised object localization. *arXiv preprint arXiv:1909.04813*, 2019. 8
- [225] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler. segdeepm: Exploiting segmentation and context in deep neural networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4703–4711, 2015. 13
- [226] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1850, 2017. 7, 8



Dingwen Zhang received his Ph.D. degree from the Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently a full professor in School of Automation, Northwestern Polytechnical University. From 2015 to 2017, he was a visiting scholar at the Robotic Institute, Carnegie Mellon University. His research interests include computer vision and multimedia processing, especially on saliency detection, video object segmentation, and weakly supervised learning.



Junwei Han (M'12-SM'15) is a Professor with Northwestern Polytechnical University, Xi'an, China. He received Ph.D. degree in Northwestern Polytechnical University in 2003. He was a Research Fellow in Nanyang Technological University, The Chinese University of Hong Kong, and University of Dundee. His research interests include computer vision and brain imaging analysis. He has published over 100 papers in IEEE TRANSACTIONS and top tier conferences. He is currently an Associate Editor of IEEE Trans. on Neural Networks and Learning Systems, IEEE Trans. on Circuits and Systems for Video Technology, IEEE Trans. Cognitive and Developmental Systems, IEEE Trans. on Human-Machine Systems, Neurocomputing, and Machine Vision and Applications.



Gong Cheng received the B.S. degree from Xidian University, Xi'an, China, in 2007, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, in 2010 and 2013, respectively. He is currently a Professor with Northwestern Polytechnical University. His main research interests are computer vision and pattern recognition.



Ming-Hsuan Yang received the PhD degree in computer science from the University of Illinois at Urbana-Champaign, in 2000. He is a professor in Electrical Engineering and Computer Science from the University of California, Merced. He served as an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence from 2007 to 2011, and is an associate editor of the International Journal of Computer Vision, the Computer Vision and Image Understanding, the Image and Vision Computing, and the Journal of Artificial Intelligence Research. He received the NSF CAREER award in 2012, and the Google Faculty Award in 2009. He is a Fellow of the IEEE and Senior Member of the ACM.