Weakly Supervised Object Localization (WSOL) Overview

CIS607 Computer Vision and Deep Learning Seminar

Tyler Newton

Weakly Supervised Object Localization (WSOL)

-training data are class labels (image-level labels)
-no information on location

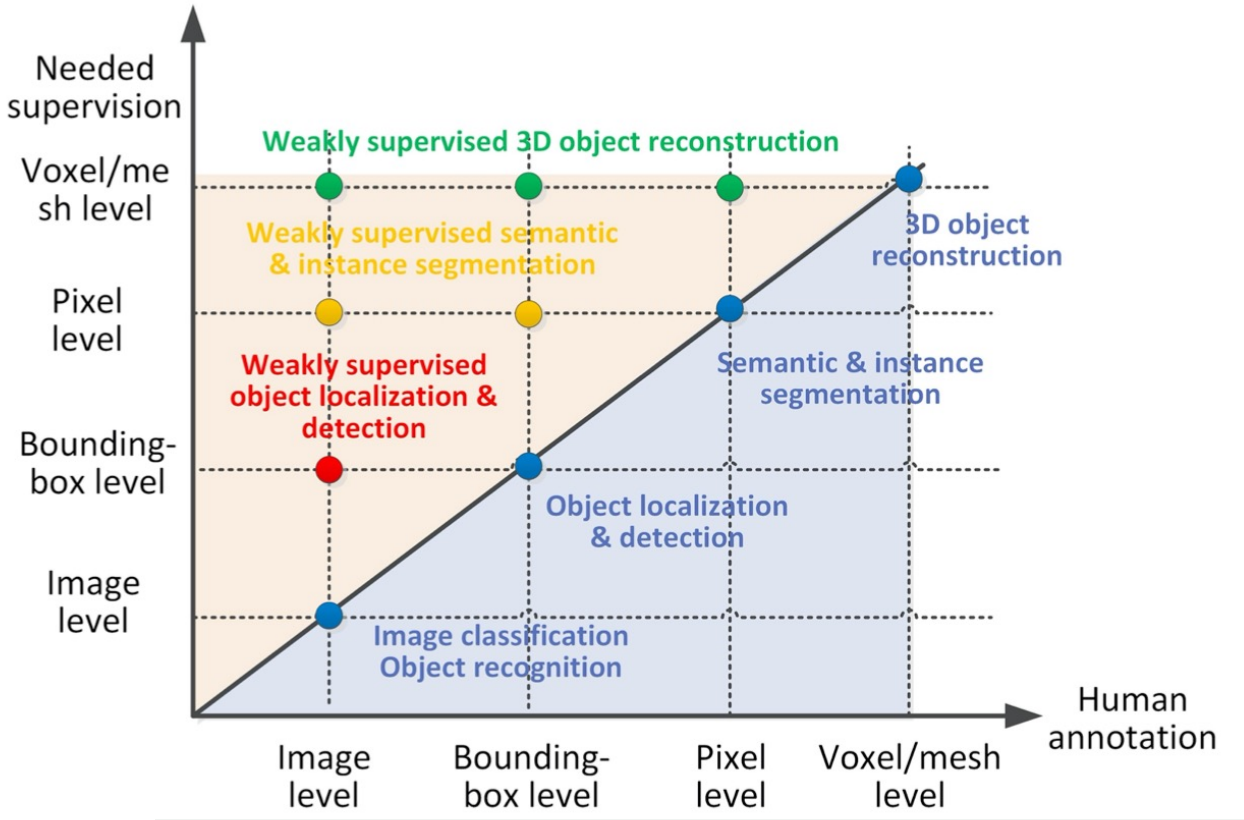


figure credit: Zhang et al. 2021

Weakly Supervised Object Localization (WSOL)

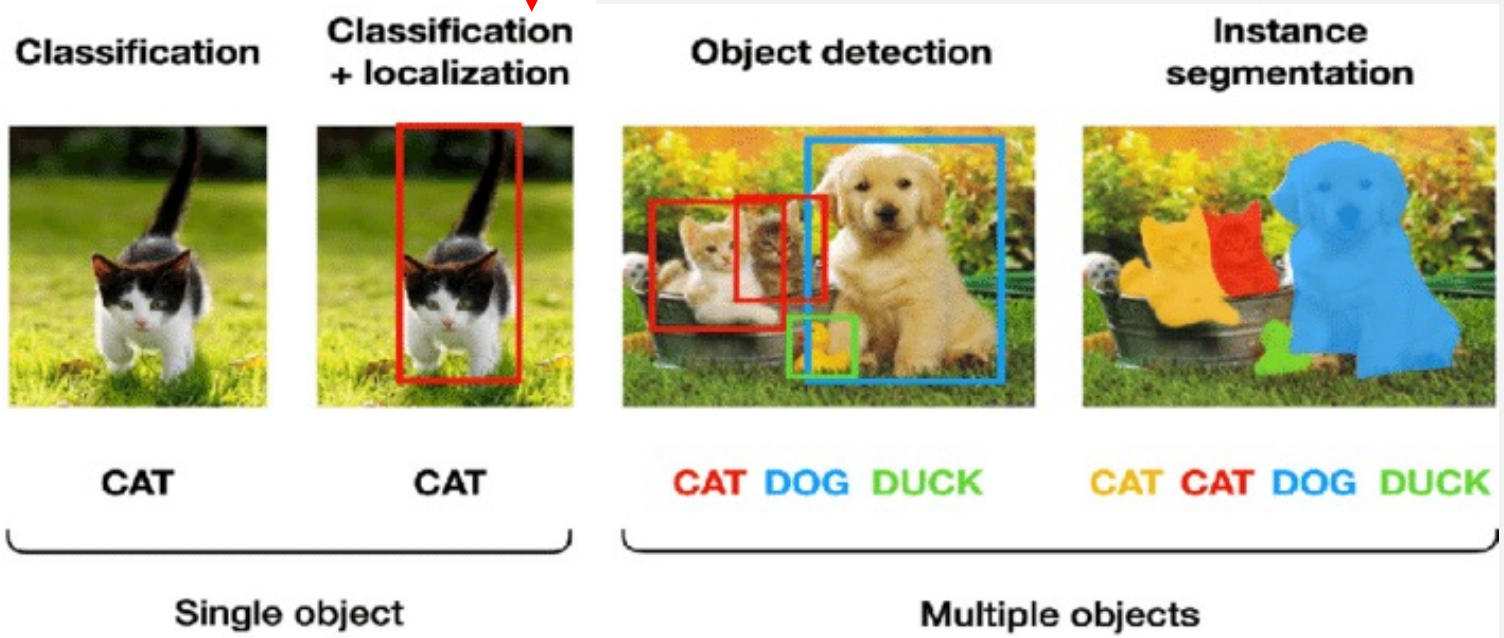


figure credit: Kang et al. 2020

Weakly Supervised Object Localization (WSOL)

ill-posed

input: image-level

output: instance-level

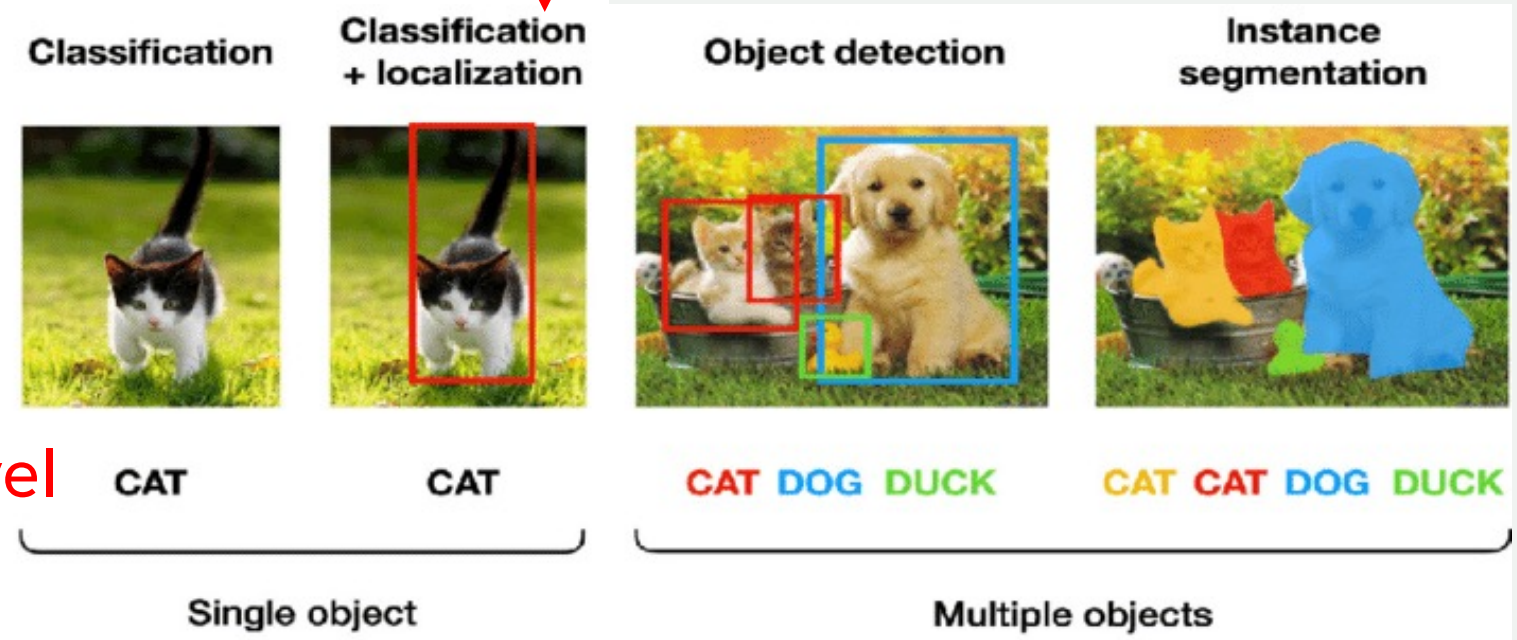


figure credit: Kang et al. 2020

Challenges of WSOL

The usual suspects:

- clutter
- intra-class variance
- occlusion

Mapping from class to location:

- part vs whole
- background vs object
- concurrence

+

Motivation/Cost of WSOL

Save labeling time at the expense of performance.

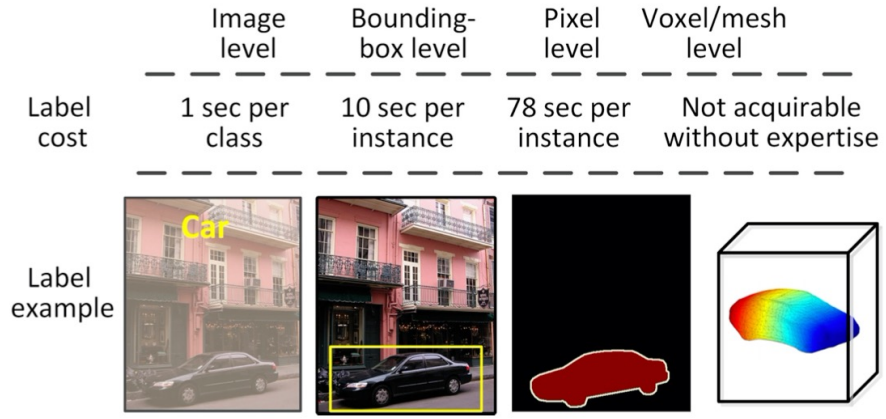


figure credit: Zhang et al. 2021

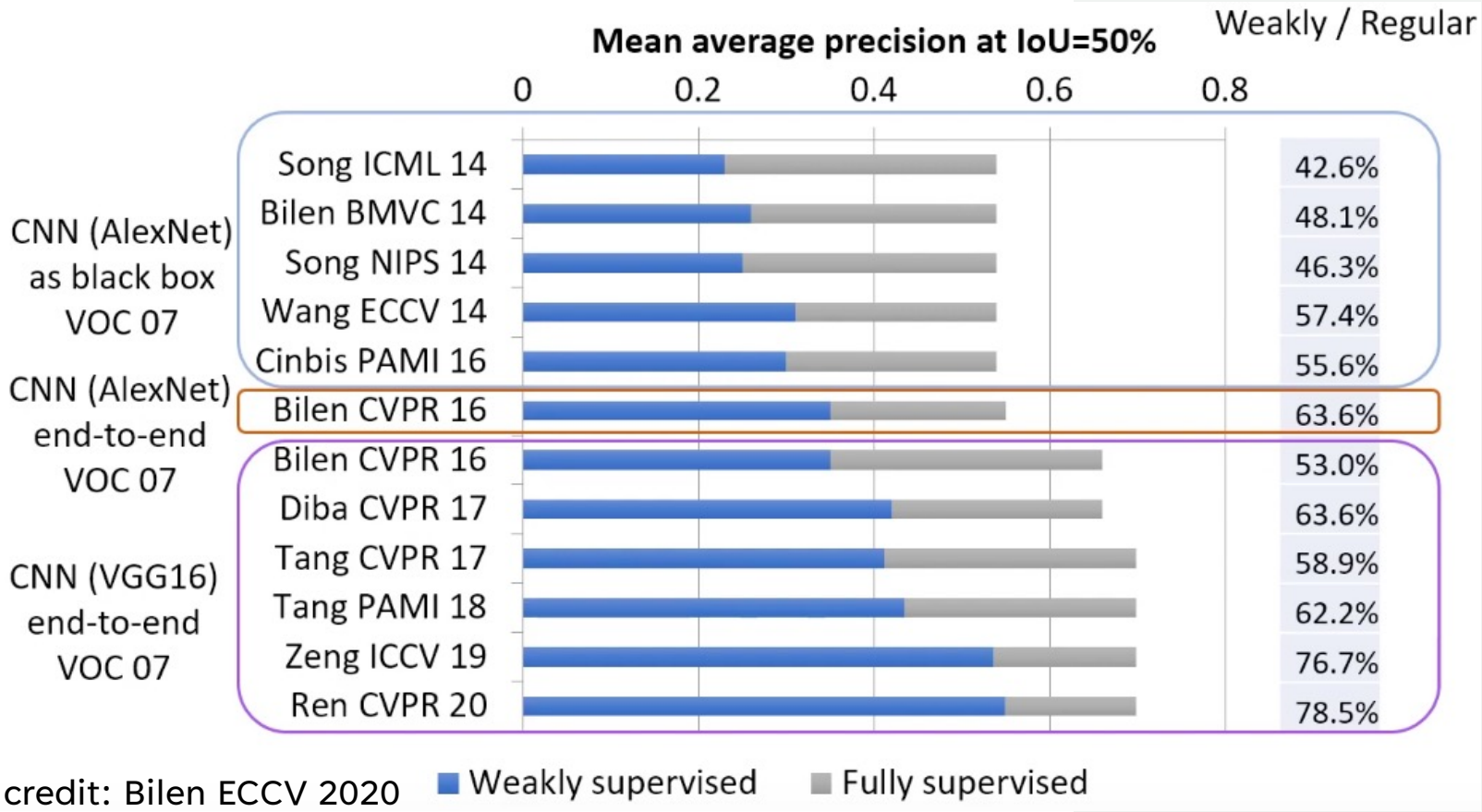


figure credit: Bilen ECCV 2020

Three main frameworks

1. classic approaches (feature engineering)
   - clever initialization methods
   - clever refinement methods
2. OTS feature representations
   - pre-trained deep features
   - inherent cues in deep models
3. deep learning frameworks
   - single-network
   - multi-network

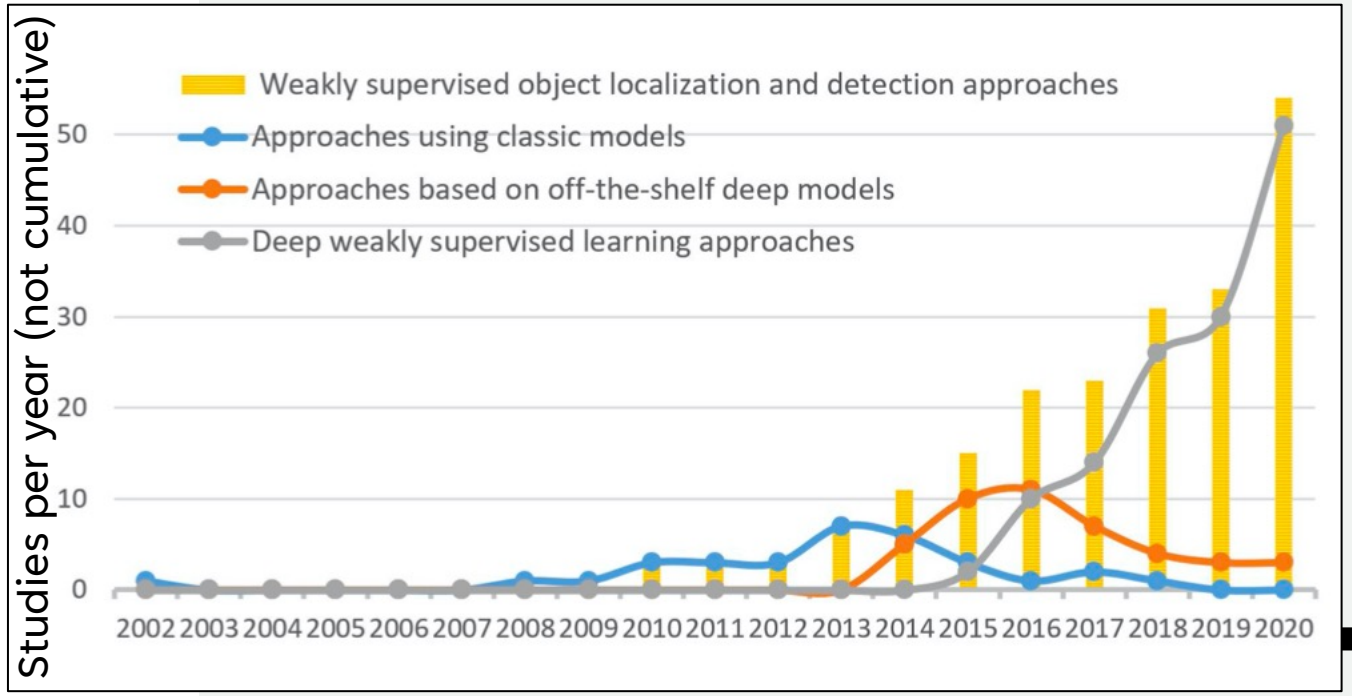


figure credit: Zhang et al. 2021

Classic approaches

1. Initialize 2. Refine 3. Detect

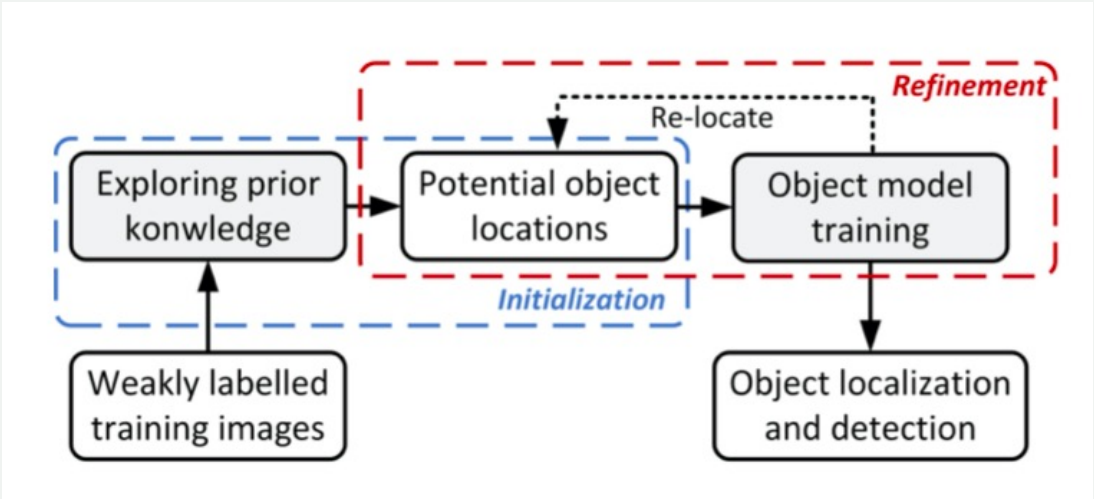
Initialization of region proposals:

- selective search (van de Sande, ICCV 2011)
- graphs (Song, ICML 2014)
- negative mining (Siva, CVPR 2013)
- etc. (some type of information cue on candidate object regions)

Refine:

- score region proposals in some way & iterate
- one positive instance per proposal (relax-max, nms)
- HOG, SIFT, etc.
- Deformable Parts Model, SVM

figure credit: Zhang et al. 2021



Classic approaches

1. Initialize 2. Refine 3. Detect

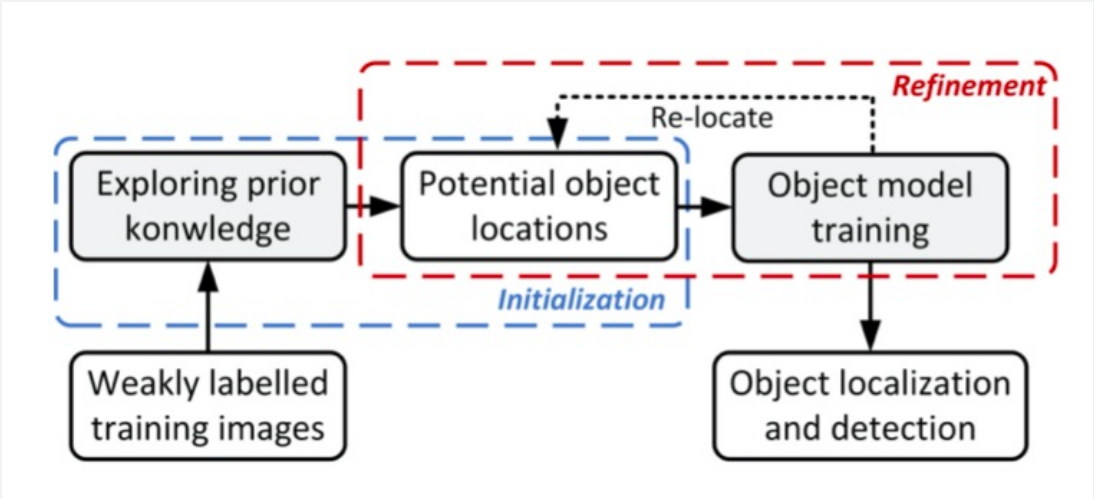
Initialization of region proposals:

- selective search (van de Sande, ICCV 2011)
- graphs (Song, ICML 2014)
- negative mining (Siva, CVPR 2013)
- etc. (some type of information cue on candidate object regions)

Refine:

- score region proposals in some way & iterate
- one positive instance per proposal (relax-max, nms)
- HOG, SIFT, etc.
- Deformable Parts Model, SVM

figure credit: Zhang et al. 2021



The building blocks of other models

Feature representations from off-the-shelf models

- Pre-trained deep features
  - use pre-trained models as feature extractors (DeCAF, MIL)
  - Gonthier 2018, Zadrija CVIU 2018, etc.
- Inherent cues in deep models
  - network layer activations and semantic scores
  - Li ISPRS 2018, Wilhelem DICTA 2017, etc.
- Fine-tuned deep models
  - tune OTS models during weakly supervised learning
  - Zhang IJCV 2019, Shi ICCV 2017, Singh CVPR 2016, etc.

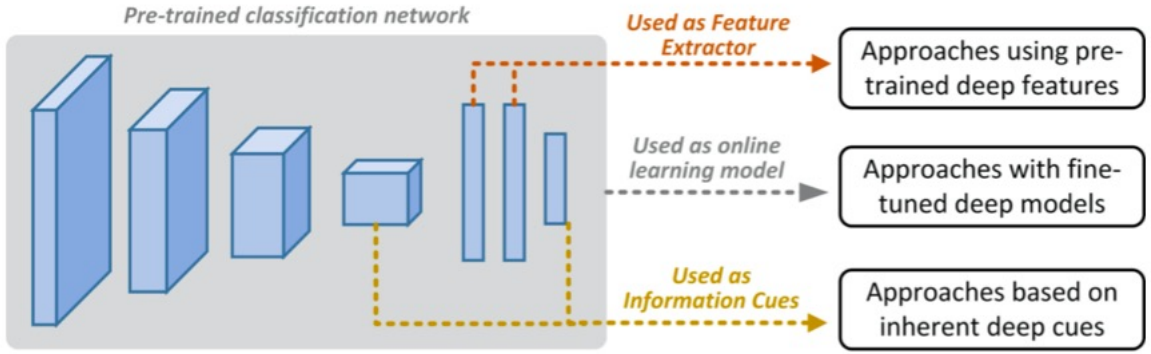


figure credit: Zhang et al. 2021

End-to-end deep weakly supervised learning

- Single-network training
  - object regions from end-to-end learning via layers for instance inference and propagation of image labels
  - outputs vary 🧐: semantic scores (Wu, Bilen), CAM maps (Zhou), bounding boxes, point
  - Huang NIPS 2020, Shen CVPR 2020, Mai CVPR 2020, Jiang ICCV 2019, ... , many others..., Wu CVPR 2015, ...
- Multi-network training
  - one network to mine initial object regions
  - another network for detection via MIL framework
  - sometimes another final object detection network (Fast[er] RCNN)
  - Zhang CVPR 2020, Zhong ECCV 2020m Kosugi ICCV 2019, ..., Zhang TGRS 2016, etc.
  - SOA despite lack of engineered features like bbox regression, negative mining, etc.

Evaluation metrics

via PASCAL VOC, ImageNet, and CUB datasets:

CorLoc metric – bounding box with highest class score has >= 50% overlap with ground-truth bounding box (requires a ground-truth, but remember this is WSOL)

mAP metric – mean IOU if IOU with ground-truth is > 50%

Other metrics: GT Loc, Top-1 Loc, Top-5 Loc

Good luck

Use cases

figure credits: Zhang et al. 2021

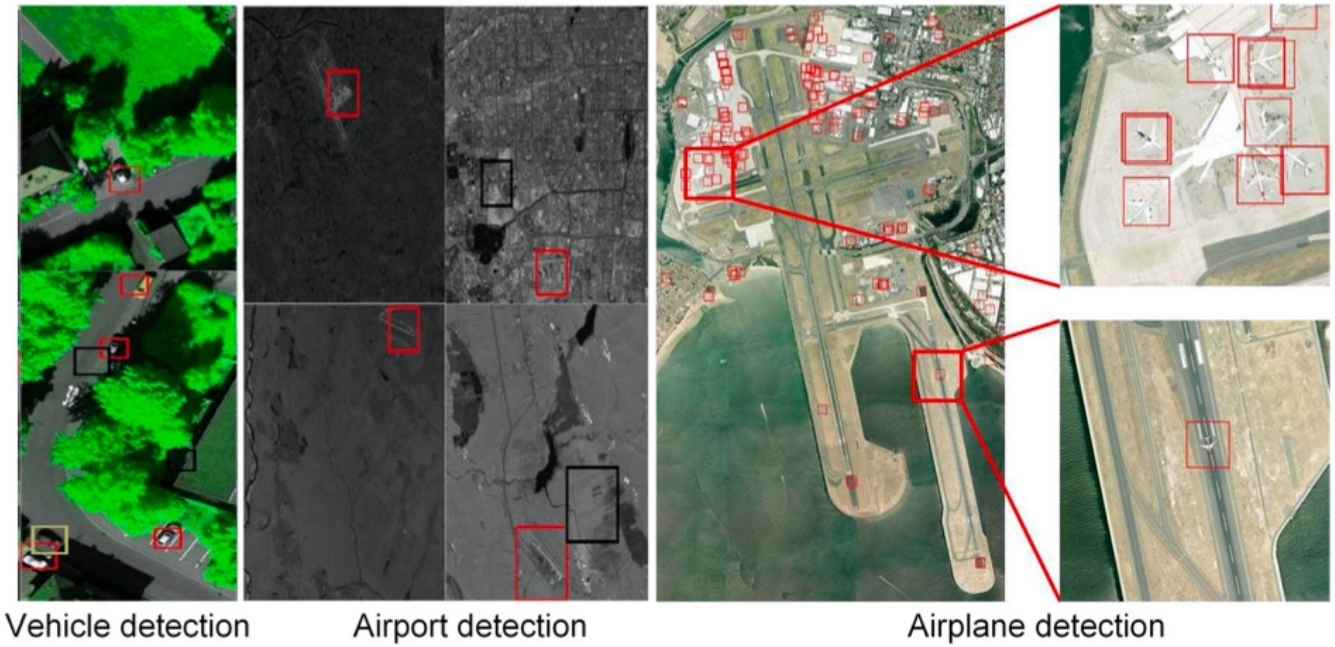


Fig. 10. Examples of the application of weakly supervised object localization or detection approaches in remote sensing imagery analysis. The examples are from [57], [208].

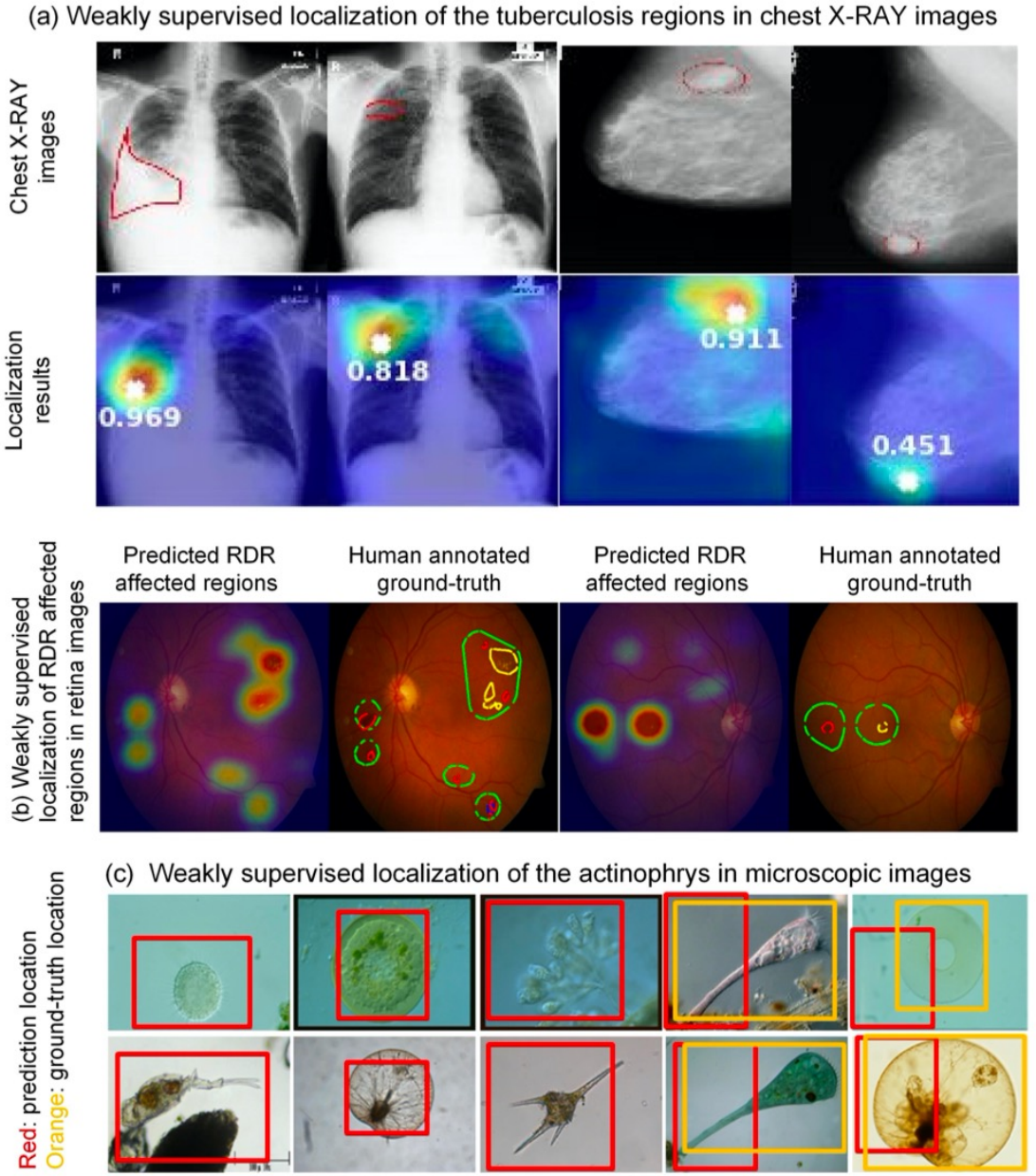


Fig. 9. Examples of the application of weakly supervised object localization or detection approaches in medical image analysis. The examples are from [53], [69], [90].

CAM methods are popular

CAM on OTS CNN models:

- struggle with global cues
- convolutions produce localized receptive fields leading to partial activation
- Layer-CAM (Jiang et al. 2021) tries to solve this by sampling shallow layers

Transformers to the rescue:

- TS-CAM (Gao et al. 2021) token semantic coupled attention maps via visual transformer

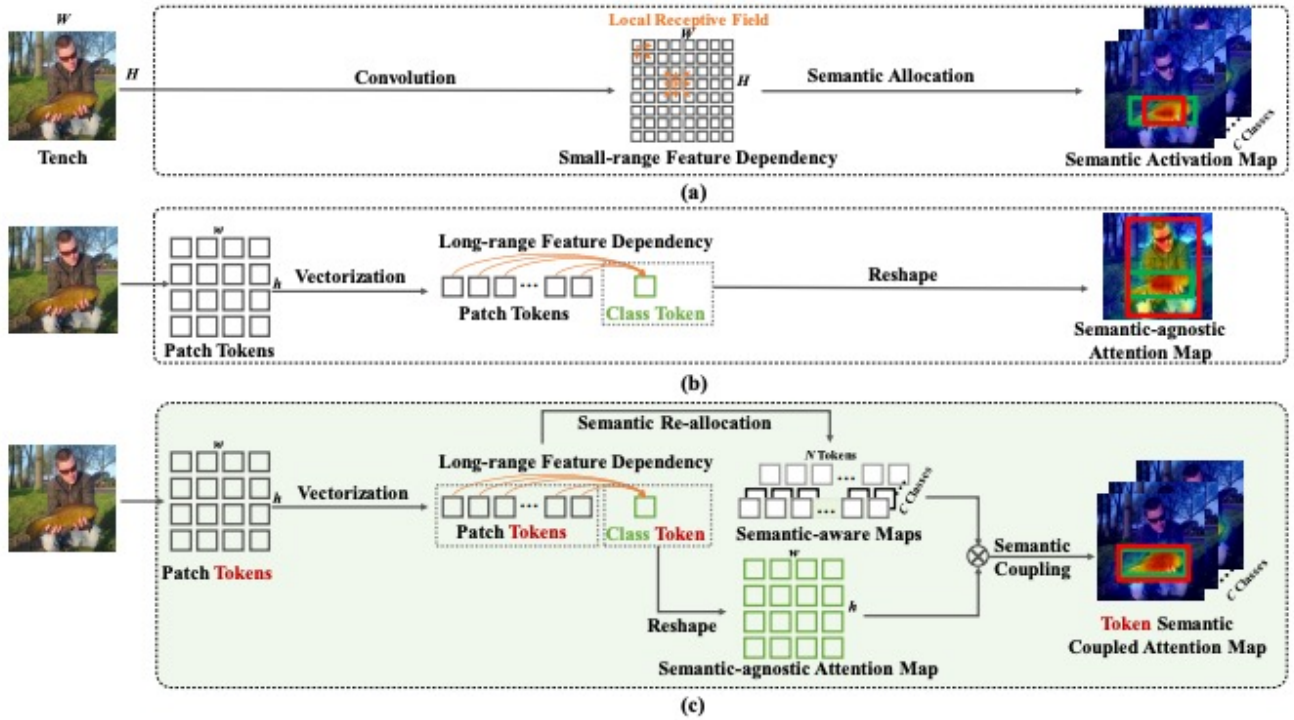
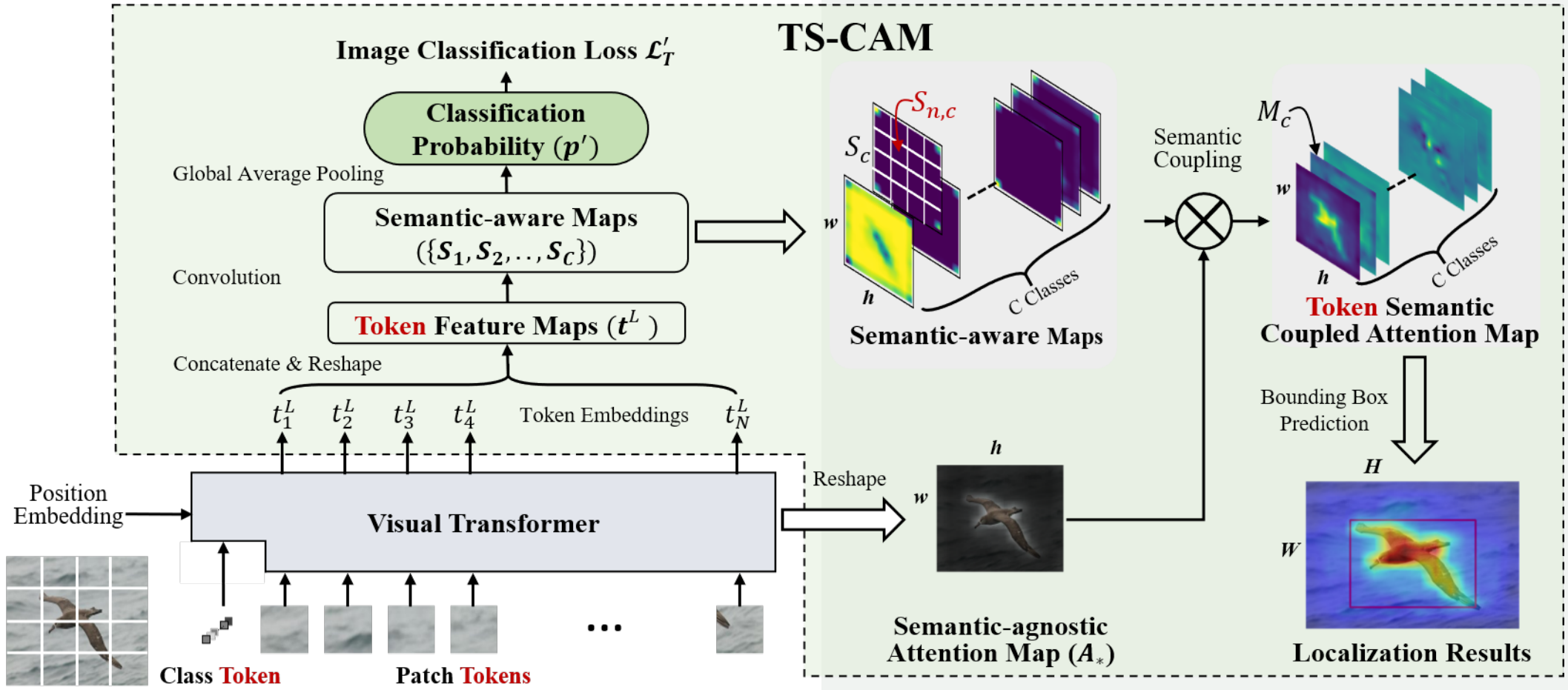


Figure 2. Comparison of the mechanisms of (a) CNN-based CAM, (b) Transformer-based Attention and (c) the proposed TS-CAM. The CNN-based CAM method is limited by the small-range feature dependency and the transformer-based attention is limited by the semantic-anostic issue. TS-CAM is able to produce semantic coupled attention maps for complete object localization. (Best viewed in color)

TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization



https://github.com/vasgaowei/TS-CAM

TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization

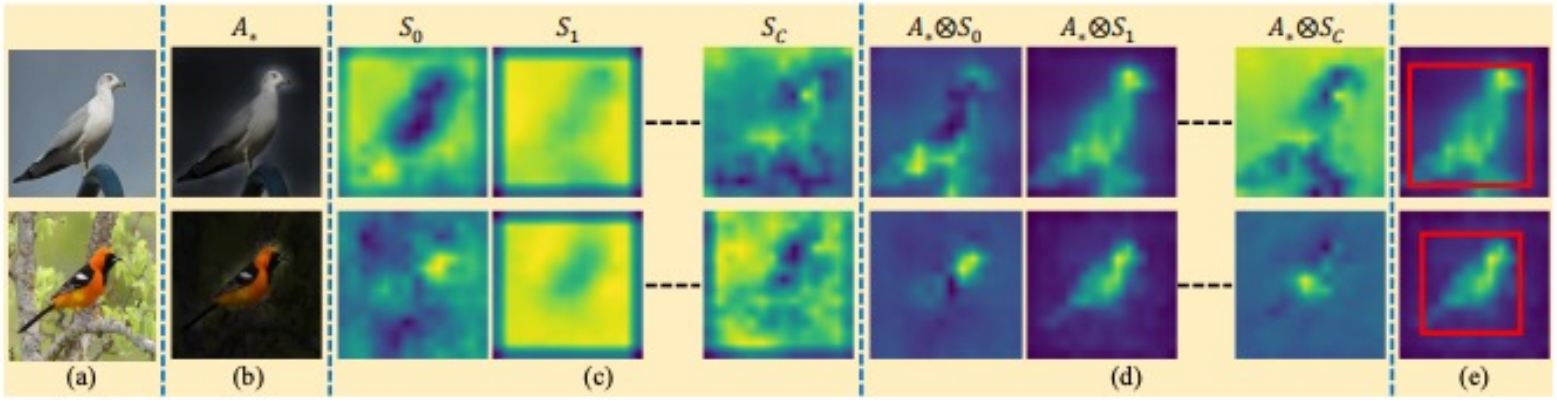


Figure 8. Visualization of semantic-attention coupling. (a) Input image. (b) Semantic-agnostic attention Map. (c) Token semantic-aware maps. (d) Token semantic coupled attention maps. (e) Localization results. (Best viewed in color)

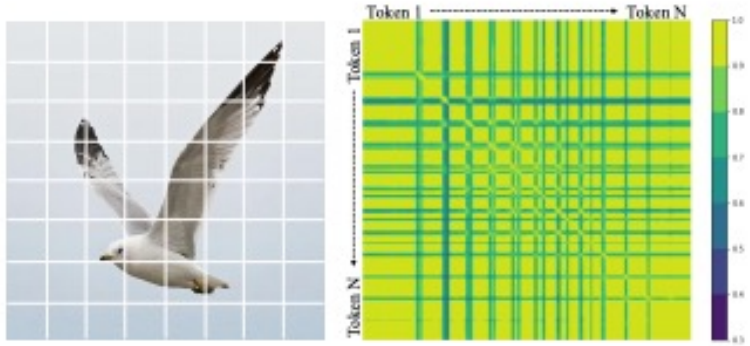


Figure 9. **Left**:Input patch tokens. **Right**: Visualization of the similarity matrix for patch token embeddings. Each row/column represents the cosine similarities between a patch token embedding with all patch token embeddings.

figure credits: Gao et al. 2021

TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization

CUB-200-2011 dataset

Backbone	Loc.Acc@1	Loc.Acc@5	Loc.Gt-Known	Cls.Acc@1	Cls.Acc@5	Baidu Drive	Google Drive
Deit-T	64.5	80.9	86.4	72.9	91.9	model	model
Deit-S	71.3	83.8	87.7	80.3	94.8	model	model
Deit-B-384	75.8	84.1	86.6	86.8	96.7	model	model
Conformer-S	77.2	90.9	94.1	81.0	95.8	model	model

ILSVRC2012 dataset

Backbone	Loc.Acc@1	Loc.Acc@5	Loc.Gt-Known	Cls.Acc@1	Cls.Acc@5	Baidu Drive	Google Drive
Deit-S	53.4	64.3	67.6	74.3	92.1	model	model

https://github.com/vasgaowei/TS-CAM

A great resource:

ECCV 2020 Tutorial on Weakly-Supervised Learning in Computer Vision

https://hbilen.github.io/wsl-eccv20.github.io/

A great review paper:

Weakly Supervised Object Localization and Detection: A Survey, Zhang et al. 2021

https://doi.org/10.1109/TPAMI.2021.3074313

Takeaways:

- methods that require initialization are sensitive to that choice and prone to overfitting
- end-to-end networks are SOA
- performance compared to standard supervision is ~70-80% in best cases

Let's demo 

- https://github.com/vasgaowei/TS-CAM/blob/master/tools_cam/visualization_attention_map_imaget.ipynb
- https://github.com/yeezhu/SPN.pytorch/blob/master/demo/EvaluationDemo.ipynb

Fake HW:

https://github.com/g41903/ObjectLocalization-WeaklySupervised

Extra slides

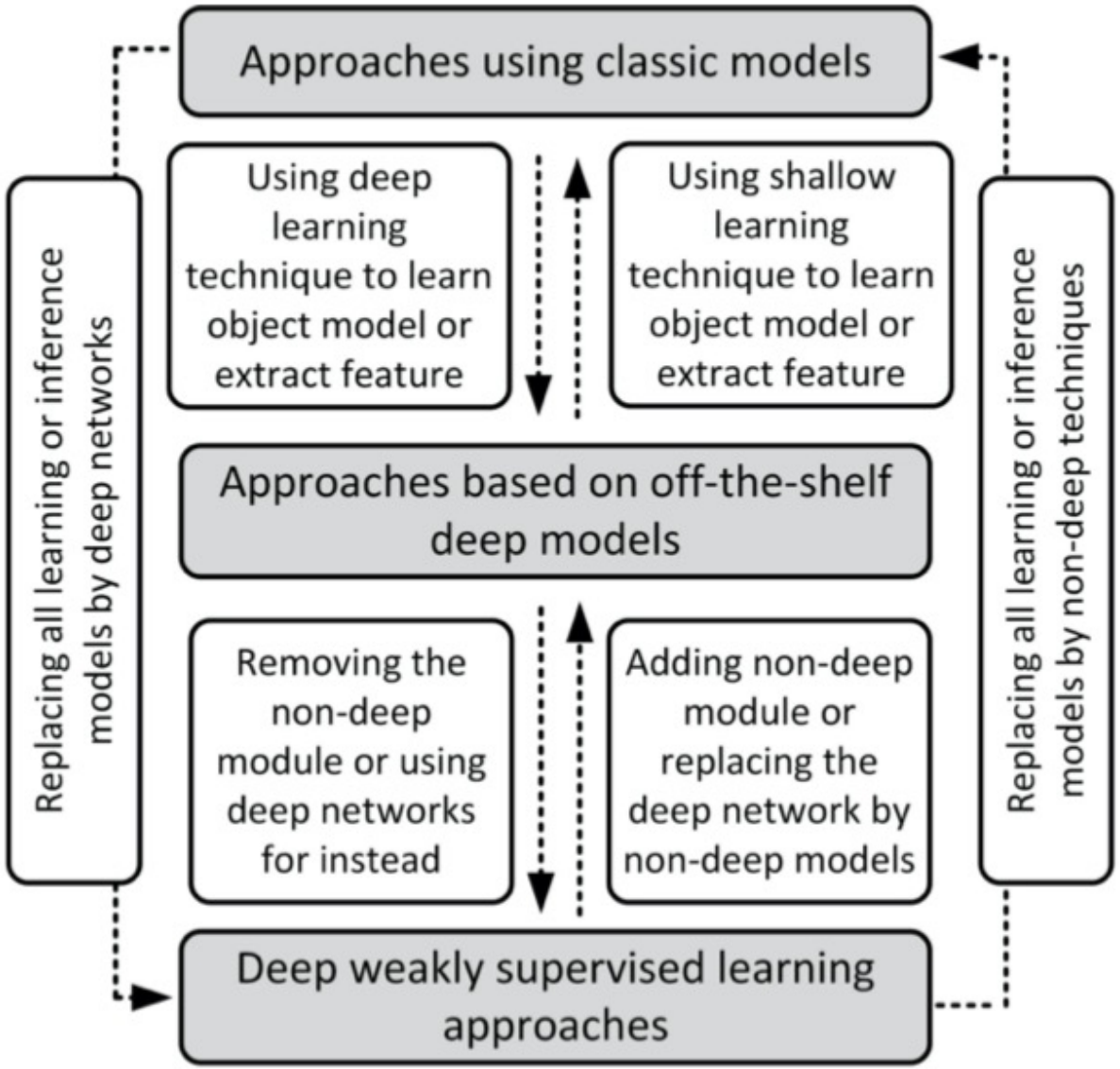


figure credit: Zhang et al. 2021

figure credit: Zhang et al. 202

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 [129]	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	IHOF	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
Cholakkal-CVPR-2016 [22]	None	SIFT	Saliency	None	SVM*	ScSPM-based top down saliency	Salient objects
Zadrija-GCPR-2015 [196]	None	SIFT+FV	None	None	GMM + linear classifier	Patch-level spatial layout learning	Traffic sign
Krapac-ICCVTA-2015 [84]	None	SIFT+FV	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	Saliency	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, pedestrain
Kanezaki-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
Galleguillos-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 deriviate filters	Exampler prior	Training exampler (box annotation)	GMM	Learning from exampler 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 trainsferring genearic 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

figure credit: Zhang et al. 2021

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)
Zadrija-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	None	ImageNet(tag label)	SVM*	MIL+bag splitting	General objects
Wan-ICIP-2016 [158]	SVM*	FC7	None	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-ICCV-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
Mathe-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

figure credit: Zhang et al. 2021

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
Wilhelem-DICTA-2017 [177]	None	CNN	None	ImageNEt(tage 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(tage 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(tage label)	SVM	Easy-to-hard	General objects
Uijlings-CVPR-2018 [154]	None	CNN	Semantic objectness(SSD*)	ImageNet(tage 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(tage 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(tage label), PASCAL Context (full annotation)	FCN,Fast RCNN	Localizing objects based on things and stuff prior and training Fast TCNN iteratively	General objects
Singh-CVPR-2016 [88]	Fast RCNN	CNN	Tracking prior	ImageNEt(tage label), Youtube-objects (unlabelled)	Fast RCNN	Discriminative region minning+transferring tracking object pattern + learn object detector	General objects
Li-CVPR-2016 [91]	VGG*	CNN	Mask out prior(AlexNet)	ImageNEt(tage label)	VGG*, SVM	Progressive Domain Adaptation	General objects
Liang-ICCV-2015 [97]	CNN	CNN	Instance example, motion prior	ImageNEt(tage 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

figure credit: Zhang et al. 2021

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 dat 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 hidding 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 calssification+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

figure credit: Zhang et al. 2021

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 Fast RCNN	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)	MIDN+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)	MIDN+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)	MIDN,TripNet, GoogleNet, FCN, Fast RCNN	Multi evidence fusion+ outlier filtering +pixel label preidction +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