SFCM: LEARN A POOLING KERNEL FOR WEAKLY SUPERVISED OBJECT LOCALIZATION

Zongxian Li[1], *Yemin Shi*[1], *Yonghong Tian*[1*], *Wei Zeng*[1], *Yaowei Wang*[2*]

[1] National Engineering Laboratory for Video Technology, School of EE&CS,
Peking University, Beijing, China
[2] School of Information and Electronics, Beijing Institute of Technology, Beijing, China

ABSTRACT

The weakly supervised object localization (WSOL) is to locate the objects in an image while only image-level labels are available during the training procedure. In this work, the **S**elective **F**eature **C**ategory **M**apping (SFCM) method is proposed, which introduces the **F**eature **C**ategory **M**apping (FCM) and the widely-used selective search method to solve the WSOL task. Our FCM replaces layers after the specific layer in the state-of-the-art CNNs with a set of kernels and learns the weighted pooling for previous feature maps. It is trained with only image-level labels and then map the feature maps to their corresponding categories in the test phase. Together with selective search method, the location of each object is finally obtained. Extensive experimental evaluation on ILSVRC2012 and PASCAL VOC2007 benchmarks shows that SFCM is simple but very effective, and it is able to achieve outstanding classification performance and outperform the state-of-the-art methods in the WSOL task.

Index Terms— Weakly Supervised Object Localization (WSOL), Selective Feature Category Mapping (SFCM), Global Learnable Pooling (GLP)

1. INTRODUCTION

Due to the development of supervised object detection methods, such as Faster RCNN [1], SSD [2] and YOLO [3], the precision of many public detection datasets have been pushed to a higher and higher record. However, these kinds of methods typically require a huge amount of manually annotated bounding boxes for each object, which is very time consuming and far more expensive than category labeling. In order to avoid the tedious bounding box annotation process, some previous works tried to localize the objects in a weakly supervised way [4, 5, 6, 7]. The weakly supervised object localization(WSOL) aims to localize the specific object in an image without any location annotation in the training procedure.

According to [8], both semantic and structure information can be retained in the feature maps of the last convolution layer but will be lost after going through fully-connected layers. To keep more localization information for each feature point, [9] proposed to apply the Global Average Pooling(GAP) [10] before classification. However, global average pooling treats all positions equally and the pooled value might not be able to well represent the whole feature map composed of hundreds of pixels. Motivated by this, we think of learning pooling weights so that the network can decide what to take care of and connect the discriminative area of feature maps to its corresponding category through the learned kernel.

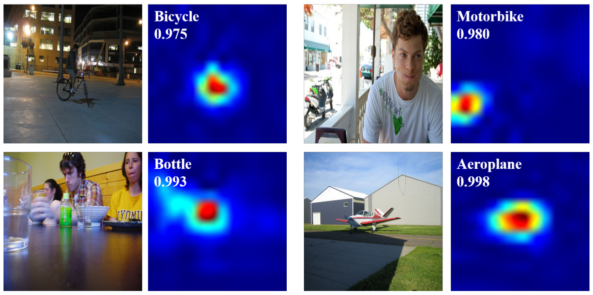

Fig. 1. Visualization of the output of feature category mapping(FCM), which applies the learnable pooling kernel to learn weighted sum in the training procedure and map the feature maps to their corresponding categories in the test procedure and locate the object accurately.

In this paper, we propose the **S**elective **F**eature **C**ategory **M**apping (SFCM) method to conduct the weakly supervised object localization. Instead of simply using the class activation mapping(CAM) [9], we build a bridge between the convolution feature map and the image level label through the learned pooling kernel and call it Feature Category Mapping (FCM). We replaced the last pooling and fully-connected layers with our FCM layer and train the network on classification datasets. As shown in Figure 1, we illustrated the heat maps obtained by FCM which are able to clearly highlight the corresponding object. Selective search (SS) [11] is a widely-used proposal generation method. Instead of using the SS as a proposal generator at the training phase as in many previous supervised works [12], the SS is employed to help FCM

Corresponding author: Yonghong Tian (email: yhtian@pku.edu.cn), Yaowei Wang(email:yaoweiwang@bit.edu.cn).

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

find a more precise localization box at the test phase. The final localization bounding boxes will be generated by fusing heat maps with the outputs of the SS. Despite the proposed method is apparently simple, we are able to achieve 10.65% top-5 error rate on ILSVRC2012 validation dataset for classification task, which is rather close to the 8.6% top-5 error rate achieved by VGG16 [13] . For WSOL task, our method has achieved 30.0% mean average precision(mAP) and 61.6% correct location rate(CorLoc) [14] and outperforms the state-of-the-art methods.

To summarize, the main contributions of our work are:

- A new Global Learnable Pooling (GLP) method is introduced in this work, which makes the pooling operation learnable and obtains a better representative value for each feature map.
- We propose the Selective Feature Category Mapping (SFCM) method for the weakly supervised object localization (WSOL) task. By mapping the feature maps pixel-wisely to their specific category and fusing with the selective search to get the final localization boxes, our method can achieve great performance on the PASCAL VOC2007 dataset in both mAP and CorLoc.

2. RELATED WORK

The WSOL aims to get the location and category of the object while the bounding box annotation is not available in the training procedure. Considering the existing evidence indicating that CNNs trained for image classification shows remarkable ability in object discover [8], Bilen *et al.* proposed WSDDN [15] by adding a detection stream in the pre-trained CNNs. Some recent works are also focusing on the object proposal stage. In [16], Zhu *et al.* designed a soft proposal(SP) component which can be easily plugged into any CNN architectures. Despite the remarkable representation ability of CNN features, several other works are trying to find out some more advanced cues or strategies, which are hard for CNN to learn, for the WSOL task, such as object size [6], objectness [4, 5, 17]. The Multiple Instance Learning(MIL) [18] is an important traditional method, which regards the images as a bag of instances. Based on the MIL, Shi *et al.* tried to solve the WSOL task by using a kind of transfer leaning [19], which transferred the knowledge learned from a source set to an unknown target set.

Due to the lack of accurate location information, the coordinates regression cannot be done during the training phase, which always leads to incomplete bounding boxes [15] or can only get a location point [20]. In order to avoid the semantic and structure information destroyed in the fully-connected layer and extract the discriminative feature area in the feature maps, [9] and [20] respectively use the global average pooling(GAP) and global max pooling(GMP) without any optimization in the pooling stage. After that, a heuristic search strategy is proposed by Bency *et al.* [21] to hypothesize location of feature maps in a multi-scale manner and grade the corresponding receptive fields by the classification layer. Besides the GAP and GMP, the pooling operation is also considered as a part that can be optimized in WSOL. Durand *et al.* proposed the WILDCAT framework [22] aiming to align image regions for gaining spatial invariance and learning strongly localized features. Inspired by Zhou's work [9], a global learnable pooling(GLP) is proposed in this paper, which makes the pooling stage learnable and results in a better representation of the feature map through the pooled value. For the WSOL task, the learned pooling kernel is acted as the bridge which connected feature maps and the specific category, highlighting the discriminative corresponding area in the convolution feature maps.

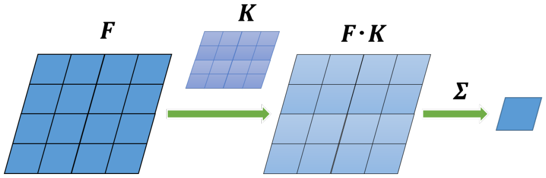

Fig. 2. The pooling kernel K will keep optimizing during the training phase to obtain an appropriate representation of the feature map F.

3. SFCM

In this section, we will introduce the proposed SFCM method for the WSOL task. In section 3.1, we first present the detail and training strategy of the Global Learnable Pooling(GLP). In section 3.2, we introduce the Feature Category Mapping(FCM) by applying the GLP, and we finally describe the steps to generate the localization bounding box.

3.1. Global Learnable Pooling

Some previous works have indicated that by removing the fully-connected layer except for the last one [23, 24] or replacing it with the GMP [20] or the GAP [10] layer will not result in an obvious drop in the performance. However, mapping a feature map composed of hundreds of pixels to a specific value via simply computing its max or mean might not obtain the best representation. In this paper, we introduced a new pooling method which makes the pooling operation learnable so that the network can decide what to take care of and tend to obtain an appropriate representation of the feature map through the training data.

Take the VGG16 as an example, we removed the following fully-connected layer $fc6$ and $fc7$ after the $conv5\_3$ layer. Instead, we added a convolution layer FCM_conv with 1024 kernels where the kernel size is 3×3. After that, we can get 1024 feature maps of size 14×14. Instead of mapping the

$196(14 \times 14)$ feature pixels to a single value by applying GAP or GMP, the **G**lobal **L**earnable **P**ooling(GLP) is introduced at this stage. For a given image I, let the F^i represents the i_{th} feature map of the FCM_conv layer. K^i refers to the learned pooling kernel which belongs to F^i. The pooled value P^i can be obtained as follows and the $\cdot$ here means the dot product among the feature map and the pooling kernel:

$$P^i = \sum_{m,n} relu(F^i \cdot K^i) \qquad (1)$$

where m, n is the index of the two dimensions of F^i and K^i and the $relu()$ is applied for nonlinear transform which can also inhibit the inactive area. We illustrate our pooling method in Figure 2.The optimization of our pooling kernel is similar with the convolution layer. In the back propagation stage, consider the specific value $k_{m,n}$ and $f_{m,n}$ in the pooling kernel K^i and feature map F^i, and J is the Softmax loss value. The optimization can be formulated as:

$$z_{m,n} = \textstyle\sum_{m,n} f_{m,n} k_{m,n} + b \qquad (2)$$

$$k^{'}_{m,n} = \begin{cases} k_{m,n} - \alpha \frac{\partial J}{\partial z_{m,n}} \frac{\partial z_{m,n}}{\partial k_{m,n}}, & z_{m,n} > 0 \\ k_{m,n}, & z_{m,n} \leq 0 \end{cases} \qquad (3)$$

where α refers to the learning rate and the b means the bias. Due to each feature map F^i owns a specific pooling kernel K^i and there is no sharing mechanism in our method, the kernels are able to focus on their corresponding feature map. Our GLP method are tending to learn a better representation of the given feature map which can retain both the semantic and spatial characteristic. Furthermore, our method build a bridge between the feature maps and the categories, which is the most important component in our Feature Category Mapping(FCM) method.

3.2. Selective Feature Category Mapping

As described in section 3.1, for each convolution feature map with size 14×14, the Global Learnable Pooling(GLP) outputs a specific value which has a good representation of the semantic and spatial information. As in section 3.1, the VGG16 is considered as our base network and the fully-connected layer $fc6$ and $fc7$ are replaced with the FCM_conv and GLP layer.After the learnable pooling operation, we obtained 1024 pooled units and then a simple fully-connected layer is applied as the classifier.

We defined the W_i^j refers to the weight between the pooled value P^i and the specific category j, which also implies the contribution of feature map to the category. Considering the learnable pooling kernel K, for a specific category j, a bridge has been built among the weights array and the feature map F by applying the learned kernel.We define the M_j as the feature category mapping for class j, by using the pooling kernel K learned from the GLP layer. The N refers to the number of the convolution feature maps. The heat map M_j can be calculated as follows:

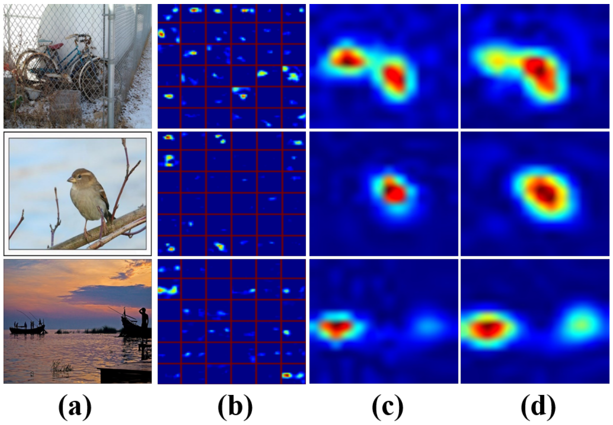

Fig. 3. The comparison between the feature maps by using different methods and the original convolution feature maps. Column (b) are the original feature maps obtained after the $conv5\_3$ layer in VGG16. Column (c) are generated by using CAM and the last column are the feature maps after applying our FCM method.

$$M_j = \sum_{i=0}^{N-1} relu(F^i \cdot K^i) W_j^i \qquad (4)$$

After obtaining the heat maps by using the FCM, an easy up-sampling is done for reverting it to its original size. We transfer the heat map to the binary form C^j by setting the threshold. And then, the localization box are generated according to the binary map. The $index()$ here are used for obtaining the coordinate of each pixel in the binary map. Specific steps are formulated as follows:

$$C^j = \begin{cases} 1, C^j_{m,n} \geq thresh \\ 0, C^j_{m,n} < thresh \end{cases} \qquad (5)$$

$$(\mathbf{x}, \mathbf{y}) = index(C^j_{m,n}) \qquad s.t. C^j = 1 \qquad (6)$$

$$box = \{min(\mathbf{x}), min(\mathbf{y}), max(\mathbf{x}), max(\mathbf{y}))\} \qquad (7)$$

Different from simply weighted linear sum in CAM [9], our GLP method learning pooling weights so that the network can decide what to take care of. As shown in Figure 3, it is obviously that some semantic information is retained on the original feature map, but it is not prominent enough for object localization. Comparing the columns (c) and (d), we find that FCM method proposed in this paper is very good at locating the object accurately and completely.

On the other branch, by applying Selective Search(SS) method, we got nearly 2,000 proposals for each image. For each proposal, the classification score is computed by forwarding the pre-trained CNNs and the ROI-Pooling [12] has

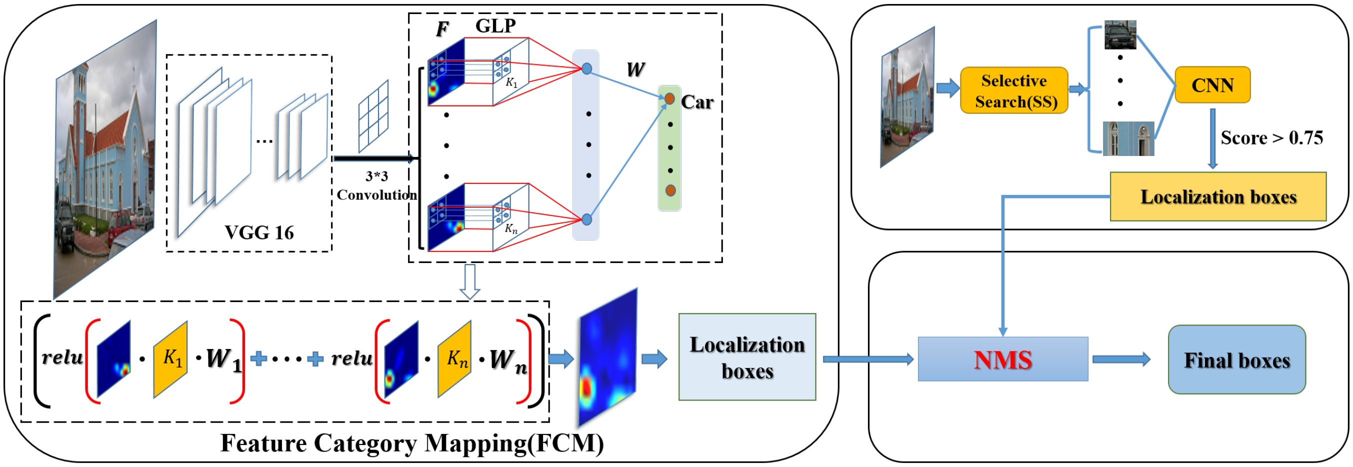

Fig. 4. The overview of our Selective Feature Category Mapping(SFCM) method. The standard VGG16 architecture was modified with our GLP layer and finally generating the localization boxes by merging our FCM method with the SS.

been applied for avoiding the time-consuming. After that, the proposals with score higher than 0.75 will be reserved. Finally, the Non-Maximum Suppression(NMS) is applied for obtaining the best localization bounding box from primary boxes generated from the FCM and Selective search proposals. We obtain the final localization boxes by merging the FCM and SS without any bounding box annotation during the training procedure. The overview of our Selective Feature Category Mapping(SFCM) method demonstrated in Figure 4.

4. EXPERIMENTS

In this section, we evaluate our method on the ILSVRC2012 and PASCAL VOC2007 datasets, as they are the most widely-used benchmark in image classification and weakly supervised object localization. For classification, we report the error rate in our method with other state-of-the-art CNNs architectures and the CAM modified version mentioned in [9]. For WSOL task, we use the training splits of the VOC2007 and evaluate the classification and localization performance on its validation splits. We use two performance measures. First, we assess CorLoc [14], a common-used weakly supervised object localization measure. Then, we report the localization mean average precision(mAP) on VOC2007 validation splits.

4.1. Classification on ILSVRC2012

Note that it is important for the network to perform well in classification tasks which will directly affect the localization performance in the next stage. In order to verify the performance of our method on classification task, we mainly evaluate the effect of modifying the state-of-the-art CNNs architecture such as AlexNet [25], VGGNet [13], GoogleNet [26], by applying the Global Learnable Pooling(GLP). Specifically, for each network mentioned above, the fully-connected layer or the GAP layers are removed except the final classification layer and we added a convolution layer FCM_conv of size 3×3 and applied the GLP layer after it. What is worth to mention that by removing the fully-connected layer, nearly 90% parameters have been decreased in VGG16.

For details, we remove the $pool5$, $fc6$ and $fc7$ layer and replaced it with the FCM_conv in AlexNet, and the GLP layer maps the 13×13 feature map to a specific value, the last fully-connected layer is retained as a classifier. The modification method of the VGG16 is generally the same as the AlexNet. For GoogleNet, the layers after the $Inception5b$ are replaced with the FCM_conv and GLP layer. Respectively, each network is trained on the ILSVRC2012 datasets from the scratch for 300,000 iterations.

For all evaluation, we report the top-5 error rate on ILSVRC2012 validation datasets. Table 1 summarizes the classification performance by using the origin AlexNet, VGG16, GoogleNet and its modified version by using our method. According to the Table 1, we find that by using the VGG16 or the AlexNet as our base network may cause 1-2% drop on top-5 error rate by modifying the networks in our method. For GoogleNet, there is almost no drop when comparing with the original architecture (0.7% drop). In addition, we find that our method does a better job than using the CAM modified CNNs with GAP in [9]. Overall, we can come to a conclusion that removing the fully-connected layer and applying the FCM_conv and GLP layer instead, will result in a comparable classification performance.

4.2. Weakly Supervised Object Localization

We evaluate the object localization performance of the SFCM on PASCAL VOC2007 validation splits, the CorLoc and mAP are reported as our measurement.

Table 1. Classification performances (error rate) on ILSVRC2012 validation sets.

Method	Top-5 val. error
VGG16 [13]	8.6
AlexNet [25]	16.4
GoogleNet [26]	8.4
VGG16+GLP+scratch	10.65
AlexNet+GLP+scratch	18.1
GoogleNet+GLP+scratch	9.1
VGG16+CAM_finetune [9]	12.7

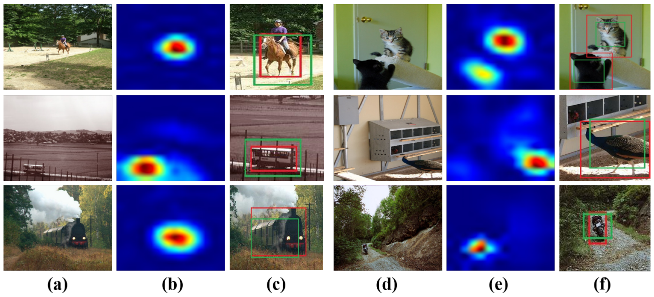

Fig. 5. Bounding boxes and localization results on VOC2007 validation set. The ground-truth boxes are in red and the predicted boxes by applying the SFCM are in green.

4.2.1. *Classification task*

As described in section 4.1, the good classification performance is important for robust object localization. We first evaluate the classification performance of our model on the VOC2007 validation splits. For the 20 classes (except Background), we summarized the classification results in Table 2. By modifying the VGG16 and GoogleNet in our way, we find that our model is able to outperform its original architecture and get 0.8% increase when using VGG16, and the GoogleNet_GLP performs comparable to GoogleNet.

4.2.2. *Object Localization*

SFCM localization performance is reported in Table 3 and our method significantly outperforms existing WSOL methods on CorLoc. As shown in Fig. 5, we demonstrate the predicted localization boxes by applying SFCM and the boxes visualization (c) and (f) are cropped for a better visual impression.

Table 2. Classification performance (Accuracy) on PASCAL VOC2007 validation splits.

Method	Accuracy
VGG16 [13]	89.3
GoogleNet [26]	89.8
VGG16+GLP_finetune	**90.1**
GoogleNet+GLP_finetune	88.9
VGG16+CAM_finetune [9]	87.6

Table 3. CorLoc and mAP on PASCAL VOC2007 validation splits.

Method	CorLoc	mAP
WSDDN [15]	54.2	34.5
Cinbis [4]	52.0	28.6
SP-VGG [16]	60.6	-
WSOL_Convex [7]	43.7	27.7
Shi *et al.* [19]	59.9	33.8
Ours	**61.6**	30.0

Comparison to the state-of-the-art. As shown in Table 3, we compare the results of our method to the state-of-the-art WSOL methods and our method is comparable to them in both CorLoc and mAP. From the Table 3, the CorLoc of our method outperforms the state-of-the-art works [4, 7, 15, 16], and the mAP is worse than the WSDDN [15] and Shi *et al.* [19]. However, WSDDN is designed to focus on obtaining precise bounding box while our method is trying to locate the specific objects, and the result also proves our motivation. SFCM can also be improved by using WSDDN's idea of considering the contribution of the discriminative area and optimizing in the training procedure. As for the [19], they transfer the model trained on another set to the WSOL task, which results in a better performance in mAP but is not very fair to SFCM. Comparing with the complex multi-instance model proposed by Cinbis [4] , our method outperforms by a large margin while being much simpler.

The analysis of the FCM. From highlighting heat maps shown in the Figure 5, the feature maps obtained through FCM shows the remarkable ability of reserving the semantic and spatial information and we can clearly draw the position of the specific objects. In order to prove the effectiveness of the Feature Category Mapping(FCM) in SFCM, we verify the the FCM and selective search performance respectively and results are reported in Table 4 detailly. By carefully observing the NMS procedure, we find that the SS is just acted as an assistant and most of the generated candidates has been dropped. From the ablation study, we can draw the conclusion that by using the SFCM combined with the FCM and the SS, the localization performance gained a significant improvement when compared tp using FCM and SS separately.

5. CONCLUSION

In this work, we develop a general Selective Feature Category Mapping (SFCM) method for the weakly supervised object localization (WSOL). By applying the CNNs modified with the GLP layer trained on the classification task, we build a bridge between convolution feature maps and final category units, which will highlight the discriminative area in feature maps. The final localization boxes are generated by merg-

Table 4. Comparison of using SFCM and separately using selective search and FCM

Method	mAP	CorLoc
Selective Search	17.8	38.8
FCM	20.7	49.5
Ours	**30.0**	**61.6**

ing the highlighted areas with the candidates generated from the selective search method. We report the classification and the WSOL performance of our method on the ILSCRC2012 benchmark and the PASCAL VOC2007 dataset, demonstrating that our method is able to achieve good classification performance and outperforms the state-of-the-art methods on the WSOL task. Since our technique provides the precise localization information for objects in the image, the weakly supervised semantic segmentation and the precise compact bounding boxes generation will be considered as our future work.

Acknowledgement. This work is partially supported by grants from the National Basic Research Program of China under grant 2015CB351806, the National Natural Science Foundation of China under contract No. U1611461, No. 61390515, No. 61425025, No. 61633002, and No. 61471042.

6. REFERENCES

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *ECCV*. Springer, 2016, pp. 21–37.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[4] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *TPAMI*, vol. 39, no. 1, pp. 189–203, 2017.

[5] Chong Wang, Kaiqi Huang, Weiqiang Ren, Junge Zhang, and Steve Maybank, "Large-scale weakly supervised object localization via latent category learning," *TMM*, vol. 24, no. 4, pp. 1371–1385, 2015.

[6] Miaojing Shi and Vittorio Ferrari, "Weakly supervised object localization using size estimates," in *ECCV*. Springer, 2016, pp. 105–121.

[7] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars, "Weakly supervised object detection with convex clustering," in *CVPR*, 2015, pp. 1081–1089.

[8] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Object detectors emerge in deep scene cnns," in *ICLR*, 2014.

[9] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

[10] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," in *ICLR*, 2014.

[11] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[12] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[14] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari, "Weakly supervised localization and learning with generic knowledge," *IJCV*, vol. 100, no. 3, pp. 275–293, 2012.

[15] Hakan Bilen and Andrea Vedaldi, "Weakly supervised deep detection networks," in *CVPR*, 2016, pp. 2846–2854.

[16] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao, "Soft proposal networks for weakly supervised object localization," in *ICCV*, 2017.

[17] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell, "On learning to localize objects with minimal supervision," *arXiv preprint arXiv:1403.1024*, 2014.

[18] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[19] Miaojing Shi, Holger Caesar, and Vittorio Ferrari, "Weakly supervised object localization using things and stuff transfer," in *ICCV*, 2017.

[20] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *CVPR*, 2015, pp. 685–694.

[21] Archith John Bency, Heesung Kwon, Hyungtae Lee, S Karthikeyan, and BS Manjunath, "Weakly supervised localization using deep feature maps," in *ECCV*. Springer, 2016, pp. 714–731.

[22] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *CVPR*, 2017.

[23] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin, "Thinet: A filter level pruning method for deep neural network compression," in *ICCV*, 2017.

[24] Yihui He, Xiangyu Zhang, and Jian Sun, "Channel pruning for accelerating very deep neural networks," in *ICCV*, 2017.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.