

Iterative Spectral Clustering for Unsupervised Object Localization

Aditya Vora* and Shanmuganathan Raman†

Electrical Engineering, Indian Institute of Technology Gandhinagar, Gujarat, India, 382355

Email: *aditya.vora@iitgn.ac.in, †shanmuga@iitgn.ac.in

Abstract—This paper addresses the problem of unsupervised object localization in an image. Unlike previous supervised and weakly supervised algorithms that require bounding box or image level annotations for training classifiers in order to learn features representing the object, we propose a simple yet effective technique for localization using iterative spectral clustering. This iterative spectral clustering approach along with appropriate cluster selection strategy in each iteration naturally helps in searching of object region in the image. In order to estimate the final localization window, we group the proposals obtained from the iterative spectral clustering step based on the perceptual similarity, and average the coordinates of the proposals from the top scoring groups. We benchmark our algorithm on challenging datasets like Object Discovery and PASCAL VOC 2007, achieving an average CorLoc percentage of 51% and 35% respectively which is comparable to various other weakly supervised algorithms despite being completely unsupervised.

Index Terms—Object Localization, Spectral Clustering, Unsupervised Localization.

I. INTRODUCTION

Object localization is an important computer vision problem where the task is to estimate precise bounding boxes around all categories of the objects present in the given image. Due to the intra-class variations, occlusion, and background clutter present in the real-world images, this becomes a challenging problem to solve. Compared to image classification, localization involves estimating precise location of an object in the image. Therefore, it proves to be more difficult problem to solve. Object localization is useful in several image understanding tasks like separating the foreground from the background, object recognition, and segmentation. Previous fully-supervised approaches relied on sliding window search in order to search for an object in the image. Because of their inefficiency in terms of speed, several efficient sub-window search algorithms were proposed which work quite well in localizing the object in an image ([1]). However, these techniques require strong supervision in the form of manually-annotated bounding boxes on locations of all the object categories in an image. Acquiring such human annotations for training accurate classifiers is a cumbersome task and is prone to human errors. As a result, supervised techniques for object localization do not prove to be useful in resource restricted settings. In order to overcome the huge manual efforts required in annotations of objects in the image in supervised learning algorithms, several weakly supervised approaches were proposed. Rather than bounding box annotations of target instances, weakly

supervised learning focuses on image level labelling which is based on the presence/absence of target object instances in an image ([2], [3], [4], [5], [6]). Though these techniques work well in terms of localization accuracy, they still require human annotation efforts especially when the training data is large.

In an effort to make the task of object localization completely unsupervised, various object co-localization algorithms were proposed which try to localize an object across multiple images ([7], [8], [9], [10], [11]) without any supervision. As co-localization algorithms assume that each image has the same target object instance that needs to be localized ([9], [10]), it imports some sort of supervision to the entire localization process thus making the entire task easier to solve using techniques like proposal matching ([11]) and clustering ([12]) across images. In contrast to these works, the work presented in this paper focuses on localizing a single object instance in an image in a completely unsupervised fashion. To the best of our knowledge there is no previous work that tries to solve this problem in an unsupervised way. In this work, we do not make any assumptions like that in co-localization algorithms, thus making the entire problem more practical and challenging one to be solved. Further, it is an important problem to be addressed because of the following reasons: (1) Proposed work is an unsupervised approach for object localization. As mentioned previously, all the manual labour required in annotating the data with accurate bounding box regions around the target object instances will not be required, which saves the resources as well as the training time. (2) Apart from being fully automatic and unsupervised, our technique is easy-to-fit in the current state-of-the-art object recognition pipelines like RCNN ([13]). Thus unlike current system, we do not need to classify each of the thousands of object proposals generated from an object proposal algorithm individually. Instead, we can localize the object directly in the input test image and then provide this localized object to the CNN pipeline that will classify the object appropriately. Such a type of functionality is not available with co-localization techniques. This restricts their applicability in real-world scenarios.

To solve this problem in an unsupervised manner, we start with extracting thousands of object proposals from the input image using an off-the-shelf object proposal algorithm ([14], [15]). We then try to filter out number of object proposals effectively in such a way that after the entire proposal filtering process, a good set of object proposals that contain the object are retained. In order to achieve this, we formulate the problem as an undirected graph problem and perform spectral clustering

on the constructed graph. This will split the set of proposals that can be discriminated based on the selected feature space. However, one iteration of spectral clustering would not be enough to filter the proposals by a significant amount. As a result, we repeat the process for a number of iterations after selecting appropriate cluster for subsequent partitioning. We compute a cluster score after each iteration and select the cluster that has higher score for further partitioning in the next iteration and discard all the proposals in the cluster having lower score. After this filtering step, we then estimate the final localized window by grouping the proposals based on the perceptual similarity among the proposals. We then pick top scoring groups and take the mean of coordinates of proposals present in that groups in order to get the final localized window.

The main contributions of this paper are summarized as follows: (1) A completely unsupervised object localization algorithm for an image containing a single object is presented and benchmarked on a challenging datasets like Object Discovery ([16]) and PASCAL VOC 2007 ([17]). (2) An iterative spectral clustering approach along with an appropriate cluster selection strategy is proposed which naturally helps in searching of object region in the image. This entire process takes place in a completely unsupervised fashion. (3) Proposal grouping technique is proposed which helps in estimating the final localized window in the image.

II. RELATED WORK

Previous works in object localization are described below based on the decreasing order of supervision.

Supervised approaches for object localization involves *sliding window* approaches that apply a classifier subsequently to subimages, thus obtaining a classification map. Indicator of the object region is obtained from the classification map as the region with the maximum score. As an average size image will have a lot of pixels, scanning all of them and deriving the classification map is a computationally expensive task. [1] and [18] tried to come up with a more efficient solution by proposing an *efficient subwindow search* for object localization which does not suffer from the above mentioned drawbacks. This scheme helps to optimize the quality function over all the possible subregions of the image with fewer number of classification evaluations and thus making the algorithm run in linear time or faster. [19] introduced the concept of *global and local context kernels* that tries to combine different context models into a single discriminative classifier. [20] proposed an integrated framework for image classification, localization and detection. They efficiently implemented a multiscale and sliding window approach within a convolutional neural network (CNN). They treat localization as a regression problem where the final layer is involved in predicting the coordinates of the bounding box. This entire system is trained end-to-end with bounding box annotations from the ImageNet dataset ([21]).

Weakly-supervised approaches for localization can be divided into 4 categories: (i) exhaustive search technique ([4], [5], [22], [23]), (ii) multiple instance learning ([24],

[25], [26], [27], [28], [29]), (iii) inter-intra-class modelling ([3], [30], [31], [32], [33]), and (iv) topic model ([34], [7]). Exhaustive search techniques try to learn discriminative sub-window classifiers from the weakly labelled data and then based on the scores of the most discriminative local regions of the image, they try to estimate the final localization window. Multiple-instance learning approaches try to learn various object categories from the bag of positive and negative labelled images. Different multiple-instance learning algorithms try to exploit various aspects associated with the image. For example, [29] tries to model the latent categories of the image like sky and grass in order to improve the overall localization accuracy. [28] proposed an alternative learning approach where they trained robust category models from images returned by keyword-based search engines. In order to improve the quality of the object regions, along with the inter-class models, researchers also model intra-class relations to improve the similarity of the regions within the same object class.

Co-localization approaches: [12] proposed an image-box formulation for solving object co-localization problem, where they simultaneously localize object of the same class across a set of images. [11] generalized the task of object localization by relaxing the condition that each image should contain the object from the same category. [10] proposed an iterative link analysis technique in order to estimate the region of interest in the image. [9] proposed an approach for colocalization which is based on the partial correspondences and the clustering of local features.

III. PROPOSED APPROACH

In this section, we describe the entire pipeline for the unsupervised object localization. The summary of the entire pipeline is shown in Fig. 1.

A. Object proposals extraction and scoring

We generate object proposals from the input image using off-the-shelf object proposal generation algorithm known as EdgeBoxes ([14]), which generates object proposals based on the edge information present in the image. We chose this technique for extracting object proposals as it is capable of generating proposals with high recall at a very fast rate. It extracts ~ 1000 object proposals per image in around 0.25 secs achieving an object recall of over 96% at an overlap of 0.5 on the PASCAL VOC dataset ([17]).

After the extraction of object proposals $B = \{b_1, b_2, \dots, b_N\}$ from the image I , we score each proposal based on the probability that the region contains an object. Here, we extract $N = 1000$ object proposals. EdgeBoxes algorithm computes objectness scores s_{obj} for each proposal in B which is based on the fact that the number of edge contours that are wholly contained by the proposal is indicative of the proposal containing an object. We combine appearance score of each proposal with the saliency score in order to compute the overall score of each proposal. In order to do this, we compute the saliency map S of the input image I using the saliency algorithm proposed by [35]. From the saliency map S , we compute the average saliency score for each object proposal which gives

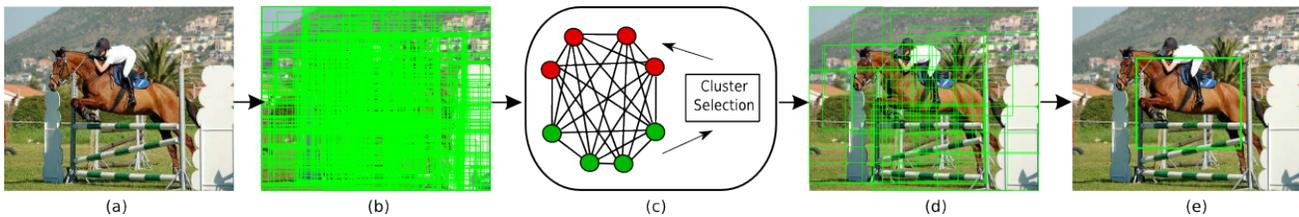


Fig. 1. **Overview of the proposed approach.** (a) Input image. (b) Extraction of object proposals (N). (c) Iterative spectral clustering stage. Construct a graph with nodes as HOG descriptor extracted from each proposal, red nodes = foreground region and green nodes = background nodes. Spectral clustering and cluster selection continues for a number of iterations with the graph being updated in each iteration based on the cluster selected. The process continues till the stopping criteria (T) is met. (d) Result after iterative spectral clustering after reaching stopping criteria (T). (e) Final estimated localized window b_{final} .

saliency score s_{sal} for that particular proposal. After this, the overall score s_i of all the object proposals in the set B are computed as $s_i = s_{obj} \times s_{sal}$, where $i = 1$ to N . As a result, proposals that have high objectness score and those that cover the salient region of the image will have high overall score.

B. Iterative spectral clustering for proposal filtering

Given a set of object proposals $B = \{b_1, b_2, \dots, b_N\}$ and proposal scores $s = \{s_1, s_2, \dots, s_N\}$ ($N = 1000$), we need to effectively select a subset of proposals that have a high probability of containing an object. We model the feature similarity among the object proposals using an undirected graph. For each proposal $b_i \in B$, we extract a HOG descriptor f_i on a 8×8 grid ([36]). We model graph W based on the HOG feature similarity among the computed proposals in the set B . Each node of the graph is the computed HOG descriptor f_i and each edge of the graph is weighted by w_{ij} , where w_{ij} is the gaussian similarity score computed as $w(f_i, f_j) = \exp(-\frac{\|f_i - f_j\|^2}{2\sigma^2})$. Here the parameter σ controls the width of the neighbourhood. For our experiments we select σ to be $0.05 \times \max(\|f_i - f_j\|)$. After constructing the graph W , we then compute the normalized Laplacian matrix of the graph W as $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where D is the diagonal matrix composed of the sum of rows of W . This choice is motivated by the work of [37] who showed that selecting the second smallest eigenvector of the normalized Laplacian graph L leads to bi-partitioning of the graph. As HOG features are able to discriminate between the foreground and background object proposals through the modelled graph, this bi-partitioning will try to partition these proposals into two separate clusters. However, one step of spectral clustering will not be able to select highly localized proposals from the huge set of object proposals. As a result, we perform a few iterations of spectral clustering until we are left with a few highly localized object proposals from which we can estimate the final detected window. In order to select a cluster for subsequent partitioning, we compute a cluster score by taking the average of all the scores s among all the proposals present in the cluster. We pick the cluster with higher score for further partitioning and discard the proposals in the cluster with lower score. We continue this iterative spectral clustering until the number of proposals are less than stopping criteria T (here, $T = 100$). As it can be seen in Fig. 2(b) that after iterative spectral clustering, highly localized object proposals

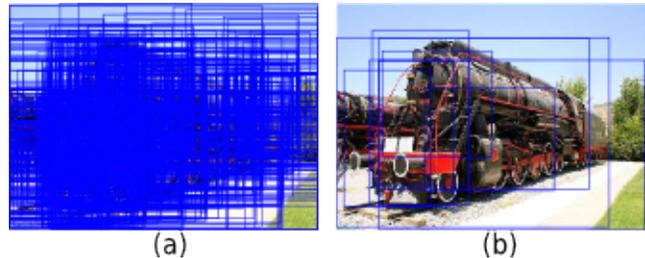


Fig. 2. **Results of iterative spectral clustering.** (a) Candidate regions extracted from the input image from the PASCAL VOC 2007 dataset ([17]). (b) Filtered object proposals of the image.

are retained and rest of the proposals are discarded from the original set of proposals (Fig. 2(a)).

C. Estimating the final localized window

After the iterative spectral clustering step, the proposals obtained from the final cluster localizes the object region in the image, as it can be seen in Fig. 2(b). This is due to the reason that the proposals in the final cluster have a high HOG feature similarity and they stand out based on the cluster score during the entire clustering process. We need to estimate a tight localization window around the object from these proposals in order to get good localization accuracy in a completely unsupervised fashion. One simple way to do this is to take the mean of all the proposals obtained after the iterative spectral clustering step. However, such a naive technique is prone to outliers (those object proposals that are larger in area will have less portion of object covered and have more background region) which will affect the overall mean of all the windows and thus affect the final localization accuracy. As a result, we need to come up with a better strategy to estimate the final window location which is immune to such outliers and thus resulting in an accurate localization.

In visual perception, a high contrast difference exists between the foreground and the background regions and low contrast difference exists between the foreground-foreground and background-background regions. We exploit this idea in order to get a good localization window. We group the object proposals obtained after clustering by considering each proposal as a seed proposal and selecting other members of the group as the proposals that have a high feature similarity with the seed proposal. We score each group based on the feature similarity of group's member proposals with the seed

Algorithm 1: Pipeline for Object Localization

Input: Image I
Result: Final localization window, b_{final}

- 1 **Procedure:**
- 2 Set Number of proposals, N .
- 3 Set stop criteria for iterative spectral clustering, T .
- 4 Set k for NN search and C for number of groups to select.
- 5 Extract object proposals $B = \{b_1, b_2, \dots, b_N\}$ from I .
- 6 Compute saliency map S of I .
- 7 Compute saliency score s_{sal} for each proposal.
- 8 Compute overall score $s_i = s_{obj} * s_{sal}$, $i = 1$ to N .
- 9 Compute HOG features for $b_i \in B$, i.e.,
 $F = \{f_1, f_2, \dots, f_N\}$.
- 10 **while** $N \leq T$ **do**
- 11 Construct similarity graph W on F feature set.
- 12 Compute normalized laplacian L and perform bi-partitioning.
- 13 Compute cluster scores and select appropriate cluster.
- 14 Update proposals B .
- 15 **end**
- 16 Compute Dense SIFT features $\hat{F} = \{\hat{f}_1, \hat{f}_2, \dots, \hat{f}_K\}$, for $\hat{b}_i \in \hat{B}$, where $\hat{B} = \{\hat{b}_1, \hat{b}_2, \dots, \hat{b}_K\}$, where $K < T$.
- 17 k -NN for $\hat{b}_i \in \hat{B}$ resulting in $G = \{g_1, g_2, \dots, g_K\}$ groups.
- 18 Compute scores for $g_i \in G$ using Eq. 1 and pick top- C groups.
- 19 Compute mean of all coordinates of proposals to get the final localized window b_{final} .

proposal and on the score of each proposal in the group. The main difference between iterative spectral clustering step and proposal grouping strategy is that, iterative spectral clustering helps in effectively filtering huge set of proposals at a very fast rate by reducing the number of proposals by about half in each iteration. However, each proposal remaining after the iterative spectral clustering step will have some implication on the final localization accuracy as it can be seen in Fig. 2(b). As a result, continuing iterative spectral clustering on these proposals will discard many such proposals that highly contribute towards the localization accuracy. Because of this, we come up with a proposal grouping technique that considers each proposals implication on localization accuracy through the contrast differences between the foreground and background regions, thus helping in better localization window estimation.

After the iterative spectral clustering step, we are left with a set of proposals $\hat{B} = \{\hat{b}_1, \hat{b}_2, \dots, \hat{b}_K\}$, where $K < T$. For each object proposal $\hat{b}_i \in \hat{B}$, we extract dense SIFT features as local discriminative features every 4 pixels and vector quantize each descriptor into a 1,000 word codebook. For each proposal, we pool the SIFT features within the proposal using 1×1 and 3×3 spatial pyramid matching (SPM) ([38]) pooling regions to generate a $d = 10,000$ dimensional feature descriptor for each box, similar to the one used by [12] for co-localization. We then normalize each feature descriptor using L_2 norm. As a result of this, we get a set of features $\hat{F} = \{\hat{f}_1, \hat{f}_2, \dots, \hat{f}_K\}$ for each proposal in \hat{B} .

We then group the set of object proposals that are perceptually similar. For this, we consider each proposal from the set \hat{B} as the seed proposal and find the k -nearest neighbours of this seed proposal using the \hat{F} feature set. Here we select $k = 10$. We do this k -nearest neighbour search for all proposals. After this, we get $G = \{g_1, g_2, \dots, g_K\}$ groups of object proposals with each group containing one member as a seed proposal and remaining members of the group with proposals that have higher similarity with the seed proposal. In order to get the final localization window, we need to select best groups from this set G . We compute the score for each group based on the feature similarity among the proposals in that group and the proposal score of the member object proposals in that group. The score for each group $g_i \in G$ is given as:

$$s_{group}(i) = \sum_{j=2}^k s(\hat{b}_j)(\hat{f}_j^T \hat{f}_i) \quad (1)$$

Here i is the seed proposal of the group and j is the index of the group members and thus the range is from 2 to k . We pick top- C proposal groups (here, $C = 5$) that have maximum group scores and obtain a final set of proposals by taking the union of all proposals in the top- C groups (i.e., many proposals in the top- C groups would be common, thus those proposals that are common will be considered only once in the final set if we take the union). We then take the mean of all the coordinates of the bounding boxes in the final set of proposals in order to obtain the final detection window b_{final} as shown in Figure. 1(e).

IV. EVALUATION

In order to evaluate our algorithm, we conducted experiments on two realistic datasets, the Object Discovery dataset ([16]) and PASCAL VOC 2007 dataset ([17]) and compare the results with various state-of-the-art algorithms for weakly supervised localization of object in the image like [4], [5], [32], [26], [34], [2], [27], [29]. We set the parameters of the experiments as follows: (1) Number of proposals per image, $N = 1000$. (2) Stopping criteria of iterative spectral clustering, $T = 100$. (3) k for NN search, $k = 10$. (4) Number of groups to merge, $C = 5$. Selecting a higher value of k and C will result into more proposals contributing to the estimation of the final localization window, thus affecting the localization accuracy. We tried with different values of k and C and found that the above values give best results. We use these values for our experiments and keep it constant throughout.

A. Evaluation criteria and runtime

Following previous works on weakly supervised localization of images, we use the CorLoc (correct location) metric, defined as the percentage of images correctly localized in the whole dataset. Here a correct localization in an image is obtained if the intersection-over-union score of the estimated bounding box and the ground truth bounding box of the image is greater than 0.5 i.e., $\frac{area(b_p \cap b_{gt})}{area(b_p \cup b_{gt})} > 0.5$. Here b_p is the predicted bounding box and b_{gt} is the ground truth bounding box of the same object. This evaluation criteria was suggested by [39].

TABLE I
PERFORMANCE OF OUR ALGORITHM ON OBJECT DISCOVERY DATASET.

Class	Aeroplane	Car	Horse	Average (%)
CorLoc	43.9	65.17	45.16	51.41

C. PASCAL VOC 2007 dataset and results

In order to compare our algorithm with various other weakly supervised algorithms, we benchmark it on PASCAL VOC 2007 dataset ([17]) which is a very challenging dataset consisting of images captured in real-life scenarios with considerable clutter, occlusion, and diverse viewpoints. It consists of 20 object classes. For a large scale evaluation of the algorithm we take all train+val dataset images which counts to a total of 5011 images. Each of the 20 object instances are spread across all the 5011 images in a range from 215 (diningtable) to 4690 (person), having a total of 12608 object instances. However, as mentioned above we evaluate our algorithm on per image basis i.e., the image is correctly classified in the image if atleast one object instance in the image is correctly localized.

The results on PASCAL dataset is shown in Table II. The second column in the table describes about the amount of data used by these mentioned algorithms. Positive images of the training set is denoted by P and negative images are denoted by N. [29] uses additional data to train a CNN model thus this is represented as A. We evaluate our algorithm on 5011 images. Our algorithm achieves an average CorLoc measure of about 35.08%, which performs better than some of the weakly supervised techniques i.e. [4], [5], [32], [26] and comparably to [34], despite being completely unsupervised. The best performing algorithm is that by [29] which gives an average CorLoc of 48.5%. However, it trains a convolutional neural network (CNN) on ILSVRC 2011 dataset in order to extract a 4096 deep feature descriptor for each proposal of the image. However, we achieve better results than 3 weakly supervised by just using simple conventional features like HOG and dense SIFT and with the power of spectral clustering. Images of the results of localization on PASCAL VOC dataset is shown in Fig. 3.

TABLE II
PERFORMANCE OF OUR ALGORITHM ON PASCAL VOC 2007 DATASET.

Method	Data used	CorLoc (%)
[4]	P+N	22.4
[32]	P+N	30.2
[26]	P+N	30.4
[34]	P+N	36.2
[2]	P+N	38.8
[27]	P	47.3
[29]	P+N+A	48.5
Ours	-	35.08

V. CONCLUSION

We have presented a new simple and efficient approach for unsupervised object localization using iterative spectral clustering and proposal grouping. We have shown that our algorithm is able to perform well in challenging scenarios

by benchmarking our algorithm on challenging datasets like Object Discovery dataset and PASCAL VOC 2007. We have achieved comparable results in terms of CorLoc when compared to other weakly supervised algorithms. In future work, we plan to extend this work to localizing multiple object instances in an image thus making the algorithm more useful for real-life scenarios.

REFERENCES

- [1] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Efficient subwindow search: A branch and bound framework for object localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2129–2142, 2009.
- [2] 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*, 2014, pp. 2409–2416.
- [3] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects while learning their appearance," *Computer Vision—ECCV 2010*, pp. 452–466, 2010.
- [4] M. H. Nguyen, L. Torresani, F. De La Torre, and C. Rother, "Weakly supervised discriminative localization and classification: a joint learning process," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1925–1932.
- [5] M. Hoai, L. Torresani, F. De la Torre, and C. Rother, "Learning discriminative localization from weakly labeled data," *Pattern Recognition*, vol. 47, no. 3, pp. 1523–1534, 2014.
- [6] 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.
- [7] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," 2005.
- [8] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1605–1614.
- [9] K. Grauman and T. Darrell, "Unsupervised learning of categories from sets of partially matching image features," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 19–25.
- [10] G. Kim and A. Torralba, "Unsupervised detection of regions of interest using iterative link analysis," in *Advances in neural information processing systems*, 2009, pp. 961–969.
- [11] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1201–1210.
- [12] 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*, 2014, pp. 1464–1471.
- [13] 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*, 2014, pp. 580–587.
- [14] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*. Springer, 2014, pp. 391–405.
- [15] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.
- [16] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2013.
- [17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>.
- [18] M. Villamizar, J. Andrade-Cetto, A. Sanfeliu, and F. Moreno-Noguer, "Bootstrapping boosted random ferns for discriminative and efficient object classification," *Pattern Recognition*, vol. 45, no. 9, pp. 3141–3153, 2012.
- [19] M. B. Blaschko and C. H. Lampert, "Object localization with global and local context kernels," in *BMVC*, 2009, pp. 1–11.

- [20] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [22] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1307–1314.
- [23] Y. Zhang and T. Chen, "Weakly supervised object recognition and localization with invariant high order features." in *BMVC*, 2010, pp. 1–11.
- [24] X. Qi and Y. Han, "Incorporating multiple svms for automatic image annotation," *Pattern Recognition*, vol. 40, no. 2, pp. 728–741, 2007.
- [25] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie, "Weakly supervised object localization with stable segmentations," *Computer Vision–ECCV 2008*, pp. 193–207, 2008.
- [26] P. Siva, C. Russell, and T. Xiang, "In defence of negative mining for annotating weakly labelled data," in *European Conference on Computer Vision*. Springer, 2012, pp. 594–608.
- [27] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 189–203, 2017.
- [28] S. Vijayanarasimhan and K. Grauman, "Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.
- [29] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 431–445.
- [30] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *International journal of computer vision*, vol. 100, no. 3, pp. 275–293, 2012.
- [31] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," *Computer Vision–ECCV 2012*, pp. 1–15, 2012.
- [32] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 343–350.
- [33] 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*, 2013, pp. 3111–3118.
- [34] 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*, 2013, pp. 2984–2991.
- [35] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1139–1146.
- [36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.
- [37] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
- [38] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.
- [39] 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*, vol. 88, no. 2, pp. 303–338, 2010.
- [40] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.
- [41] X. Chen, A. Shrivastava, and A. Gupta, "Enriching Visual Knowledge Bases via Object Discovery and Segmentation," in *Computer Vision and Pattern Recognition (CVPR)*, 2014.