ATTENTION-BASED OBJECT LOCALIZATION USING CONVOLUTIONAL BLOCK ATTENTION MODULE AND FREQUENT ITEMSET MINING

PROJECT REPORT

Submitted by

NIVIA JOSE

REG NO : TKM20CSCE10

In partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Under the guidance of
Prof. Shyna A

Thangal Kunju Musaliar College of Engineering
Kerala

SEPTEMBER 2022

Thangal Kunju Musaliar College of Engineering
Dept. of Computer Science & Engineering

C E R T I F I C A T E

This is to certify that this report titled ***Attention-Based Object Localization Using Convolutional Block Attention Module (CBAM) and Frequent Itemset Mining*** is a bonafide record of the **Project** presented by **NIVIA JOSE (TKM20CSCE10)**, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, **M.Tech in Computer Science & Engineering** in **APJ Abdul Kalam Technological University** .

Coordinator	Supervisor	Head of the Department
Dr. Ansamma John Professor Dept. of CSE TKMCE	Prof. Shyna A Assistant Professor Dept. of CSE & TKMCE	Dr. Dimple A Shajahan Associate Professor Dept. of CSE TKMCE

ACKNOWLEDGEMENT

A successful project is a fruitful culmination of efforts by many people, some directly involved and some others indirectly, by providing support and encouragement. Firstly I would like to thank the almighty for giving me the wisdom and grace for making my project a memorable one. I thank him for steering me to the shore of fulfillment under his protective wings.

I express my sincere gratitude to **Dr. T A Shahul Hameed**, Principal of T.K.M College of Engineering for giving me an opportunity to present my project. I would like to thank **Dr. Dimple A Shajahan**, Associate Professor and Head of the Department, CSE, TKMCE, for her constant support and encouragement throughout the work.

With a profound sense of gratitude, I would like to express my heartfelt thanks to my guide **Prof. Shyna A**, Assistant Professor, CSE, TKMCE, and project coordinator **Dr. Ansamma John**, Professor, CSE, TKMCE for their expert guidance, cooperation and immense encouragement. I also extend my thanks to the entire faculty members and staffs of the Department of Computer Science Engineering, TKMCE, who has encouraged me throughout this work.

I also express my thanks to my loving parents, brother and friends, for their support and encouragement in the successful completion of this project work.

NIVIA JOSE

Abstract

Object localization is one of the core tasks in computer vision, as they are applied in many real-world applications such as autonomous vehicles and robotics. It refers to the task of locating an object in an image using a bounding box. Most of the existing object localization methods require a huge amount of annotations for training and are highly time-consuming. Thus, it is worth developing object localization methods for unlabeled images. However, this is far more challenging than typical co-localization or weakly super- vised localization tasks. To tackle this problem, a novel attention-based method is proposed that takes advantage of CNN models, attention mechanisms, and data mining. Specifically, the proposed method first converts the feature maps from a new feature map extractor model,VggCBAM, into a set of transactions and then discovers frequent patterns from the transaction database through pattern mining techniques. From the experimental results, it is observed that the feature maps extracted contain meaningful activations that increase focus on the object of interest while suppressing background and the discovered patterns typically hold appearance and spatial consistency. Motivated by observation, this method can easily discover and localize possible objects by merging meaningful patterns. This approach does not need any annotations yet still shows promising localization ability, which provides a new perspective to solve the localization problem.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Object detection and Localization are two of the core tasks in Computer Vision, as they are applied in many real-world applications such as human face recognition, remote sensing object detection, retail checkout recognition, automated driving, automatic monitoring systems, and Robotics. In simple words, localization means locating an object in an image using a bounding box. Several deep learning methods have recently shown outstanding performance in object detection [1]. Algorithms like Single Shot Detectors(SSD) [2] and You-Only-Look-Once(YOLO) [3] take the approach of finding all objects within the image in a single forward pass. These techniques, however, require expensive and time-consuming object proposals. Later, weakly supervised learning techniques [4],[5],[6] were introduced that require less detailed annotations compared to the fully-supervised setting. Image co-localization [8],[9],[10] is another problem in computer vision that simultaneously localize objects of the same class across a group of different images. However, all these methods demands labeled image data. Our work deals with object localization from unlabeled data.

Most images in real life are usually without labels and thus, it is worth developing object localization in unlabeled images. However, object Localization from unlabeled data is more challenging because it does not depend on any auxiliary information other than the given unlabeled image. A selective convolutional descriptor aggregation method (SCDA) [11] reused the pre-trained model for fine-grained image retrieval. They employed a simple "mean-threshold" strategy to localize objects. Motivated by SCDA, a pattern mining-based method called Object Location Mining (OLM) is proposed in [12] for object localization that takes advantage of pre-trained CNN models and data mining. Specifically, OLM first converts the feature maps from a pre-trained VGG16 CNN model into a set of transactions and then discovers frequent patterns from the transaction database through pattern mining techniques. But it is observed that the localization results

provided by [12] are not very robust on images with complex backgrounds. Hence this work focuses on an attention-based method for object localization that includes an attention module, (CBAM) to each convolutional block of vgg16 that can learn focus information on object from channel and space and suppress complex background, followed by frequent itemset mining. The attention mechanism in the visual perception system employs a complex set of filters to produce a blurring effect, keeping the object of interest in focus while blurring or fading the background.

The input to the method is only a single unlabeled image. Given an unlabeled image, the proposed method first extracts feature maps from deep layers of a new feature map extractor model, VggCBAM. In this model, an attention module called Convolutional Block Attention Module(CBAM) is applied at the end of each convolutional block of VGG16 to get refined feature maps. The extracted feature maps are then converted to a set of transactions or a transactional database. Frequent itemset mining techniques are then applied to the transaction database to mine frequently occurring regions. Finally, the optimal patterns are selected and merged to generate a support map representing the object's location.

The main contributions of this method are summarized as follows :

- A novel method for object discovery and localization to discover objects and localize their region from a given single unlabeled image.
- The method exploits the advantages of the attention mechanism to increase focus on the object of interest while suppressing background.
- Experimental results showed that our method outperforms the previous state-of-the-art method by a large margin.

The remainder of this report is organized as follows. Chapter 2 recalls some related works used as references for completing this study. Chapter 3 discusses the methodology. Chapter 4 presents the experimental results and analysis. Finally, the conclusion is provided in Chapter 5.

Chapter 2

Related Works

This section reviews related works on object localization, pattern mining, and attention mechanisms in computer vision that motivate us to come up with a novel approach to the proposed method.

Over the past few years, deep learning-based object detection has made significant advancements in the development and application of cutting-edge object detection systems. The learning capabilities of deep CNN and extensive bounding box annotations have helped object localization attain a remarkable performance[1][2][3]. Earlier attempts at object detection used classifiers as detectors. An overview of the various deep learning-based object detection approaches is given in [1]. Several deep learning-based network architectures, like Single Shot MultiBox Detector (SSD) [2], You Only Look Once (YOLO) [3], etc., were introduced in order to appropriately extract the features. In [2], a single neural network is proposed that can predict bounding boxes and class probabilities from entire images in a single evaluation. They view the complete image during training and testing, thus it implicitly incorporates contextual information about classes in addition to their appearance. The accuracy of these technologies, however, is inferior than cutting-edge detection methods. They also have difficulty precisely localizing some small objects. In contrast to [2], [3] introduced the first deep network-based object detector that is faster and more accurate than [2]. This object detector does not resample pixels or features for bounding box hypotheses. To deal with objects of varied sizes naturally in this situation, the network aggregated predictions from many feature maps with variable resolutions. These techniques, however, call for expensive and time-consuming object proposals, such as large bounding box annotations. Weakly supervised object localization techniques, which solely use image-level annotations to localize objects, were developed as a consequence of this. The majority of currently used weakly supervised object localization techniques [4],[5],[6],] identify the objects by training a classification convolutional neural network using image-level

annotations. In spite of a lack of information on the object's location, Zhou et al [7] has showed that the convolutional units of different layers of convolutional neural networks (CNNs) can, in fact, behave as object detectors. To preserve this remarkable ability to locate objects in the convolutional layers, [4] made use of the global average pooling layer. However, labels are only provided at the image level in weakly supervised systems. As a result, image co-localization problems received increased focus in later efforts on object localization [8],[9],[10]. Co-localization is the problem of simultaneously localizing objects of the same class across a group of different images (with bounding boxes). [8] suggested a method for co-localization that generates a joint optimization problem out of an image model and a box model. [9] shown how a common object detector can be learned by making its detection confidence scores distributed like those of a strongly supervised detector, which is a basic but intriguing solution to the co-localization problem. It was first shown in [10] that convolutional activations/descriptors in pre-trained models can serve as a detector for the common object. However, the majority of these initiatives typically require producing massive object proposals first. This could result in significant time consumption, and the effectiveness greatly depends on the quality of the proposals. Our work deals with localizing objects in unlabeled data. Object Localization from unlabeled data is more challenging due to the fact that it does not depend on any auxiliary information rather than a given unlabeled image. Only a few approaches, such [11], [12], are analogous to our work, which focuses on object localization from unlabeled data. [11] uses a simple "mean-threshold" approach to find the principal objects in fine-grained images by extracting feature descriptors from a pre-trained VGG-16 model's last max-pooling layer [13]. But the localization results of [11] are not very resistant to the noisy background on complex images.

The data mining community has been developing pattern mining techniques for many years. Typically, a collection of patterns combines a number of components, and the distinguishing information is recorded. This feature has led many researchers to look into the issue of using pattern mining to handle computer vision tasks. By fusing CNN features with a pattern mining technique, [14] addresses mid-level visual element discovery. How to convert an image into transaction data—which are appropriate for pattern mining and keep the majority of the related discriminative information—is a crucial challenge in pattern mining techniques. Each

dimension index of a CNN activation is treated as an item in [14], which converts a local patch into a transaction. [15] described a cutting-edge method for automatically identifying local feature configurations that appear frequently on instances of a certain object class and infrequently on the background. They use a Frequent Itemset Mining technique, which quickly scans the vast collection of all neighbourhoods and produces spatial configurations of local characteristics that frequently recur over training images. A useful method for extracting mid-level features for image classification, based on appropriate pattern mining, was proposed by [16]. The first application of the pattern mining technique to the challenge of object localisation was described in [12]. To locate potential object regions, they sought to identify the commonly active location information of feature maps. However, the localization results of [12] are inaccurate with complicated background, similar to [11]. Motivated by [12], a novel fine-grained image retrieval method was in [17] that obtain representative local features through frequent pattern mining.

The introduction of the Attention Mechanism in deep learning has improved the success of various models in recent years. The attention process used in deep learning is comparable to the human visual attention system. In the NeurIPS 2017 paper by Google Brain, titled "Attention Is All You Need," the concept of Attention Mechanisms was first widely used in the field of Natural Language Processing (NLP) [18]. In the Self Attention Generative Adversarial Networks study (SAGAN)[19], the concept was first translated into computer vision. They added a self-attention method to convolutional GANs in [19]. Convolutions work best in conjunction with the newly developed self-attention module. The potential of using multi-layered attention, including channel attention and spatial attention, was shown by SCA-CNN [20]. Later, the idea was expanded to Convolutional Block Attention Module (CBAM)[21]. Given an intermediate feature map, [21] sequentially infers attention maps along two distinct dimensions, channel and spatial, and then multiplies the attention maps by the input feature map for adaptive feature refinement. A new network structure called VCG is suggested in [22] for the classification of clothing image, and it is inspired by [21]. [22] introduced CBAM to the second convolution block to emphasise the channel and spatial information, improve attention to the target area, and extract the detailed information from the image. Image super-resolution was later developed with CBAM. In order to convey information between feature maps quickly, [23] incorporated

CBAM inside a dense block. In this work, we explore how to effectively locate objects using deep convolutional neural networks, attention mechanisms, and frequent-mining techniques.

Chapter 3

Methodology

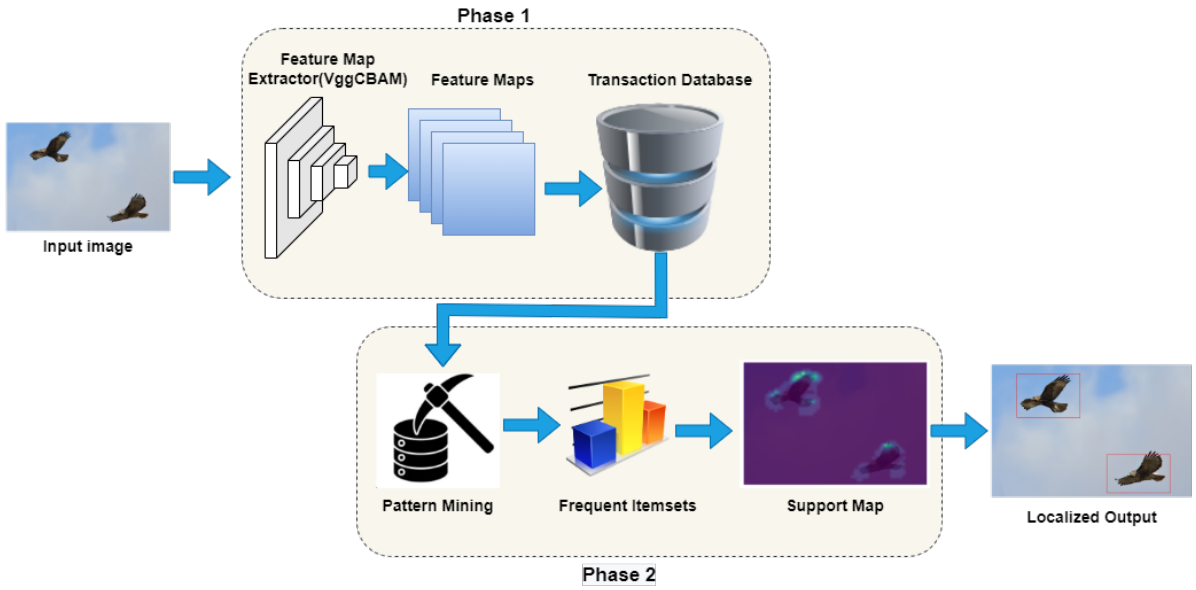

Figure 3.1: Architecture of the Proposed Method

Object discovery and localization is a basic computer vision technique that can be used to detect and locate objects in an image or video. Typical object localization techniques involve weakly supervised or image co-localization methods, which locate objects using image-level annotations. But in reality, the majority of images are often unlabeled. Therefore, it is worthwhile to develop object localization techniques for unlabeled images. However, object discovery from unlabeled data is difficult since it only uses the unlabeled image that is provided and does not rely on any auxiliary information. To tackle this problem, this work proposes an attention based architecture, that exploits the advantages of attention mechanism, convolutional neural networks and data mining. The architecture of the proposed method is shown in Figure 3.1. The whole architecture mainly includes two phases, a feature map extraction phase that utilizes attention mechanism, followed by a frequent itemset mining phase that helps to generate a support map. When an unlabeled image is input to the proposed

method, it will first extract feature maps from deep layers of a feature extraction module VggCBAM. These feature maps are then converted to transactions. Finally, a frequent itemset mining algorithm is applied on to these transactions and a support map is generated. This support map shows the location of an object within the image for bounding box localization.

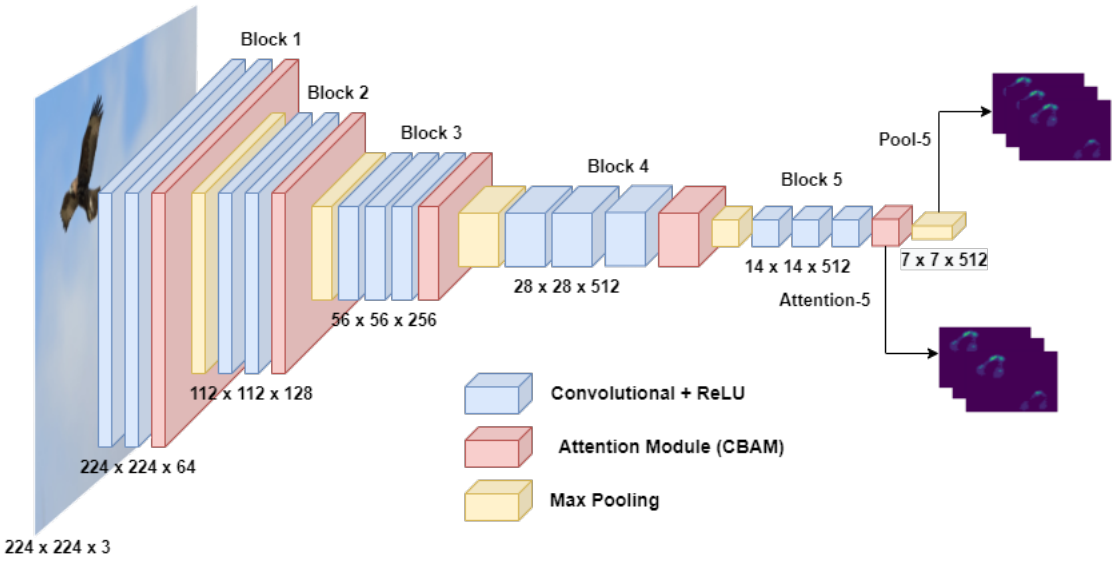

Figure 3.2: Proposed Feature Map Extractor VggCBAM

Recent studies have shown that various convolutional layers learn various level features. In particular, the discriminative power of the convolutional layer increases with depth. As a result, it is generally preferred to select deeper convolutional layers. These layers are able to maintain more co-occurring significant regions, which could be related parts of an object. However, certain areas in the background are also active. This may affect the localization results. Hence, a new feature map extractor model, named VggCBAM, is proposed, as shown in Figure 3.2. In this model, an attention module is applied at the end of each convolutional block of VGG16 to get refined feature maps. The attention module that is being used is Convolutional Block Attention Module (CBAM). By putting more emphasis on the channel and space information, CBAM can assist the network in effectively extracting features. CBAM consist of two consecutive sub-modules Channel Attention Module (CAM) and the Spatial Attention Module (SAM), that are deployed in that specific order. Channel attention identifies the feature maps that are crucial for enhancing learning. What must be learned from the feature map is communicated through spatial

attention. The input to the CBAM module are intermediate feature maps from previous convolutional layer, then the CBAM outputs refined feature maps. The working of each block of VggCBAM is shown in Figure 3.3.

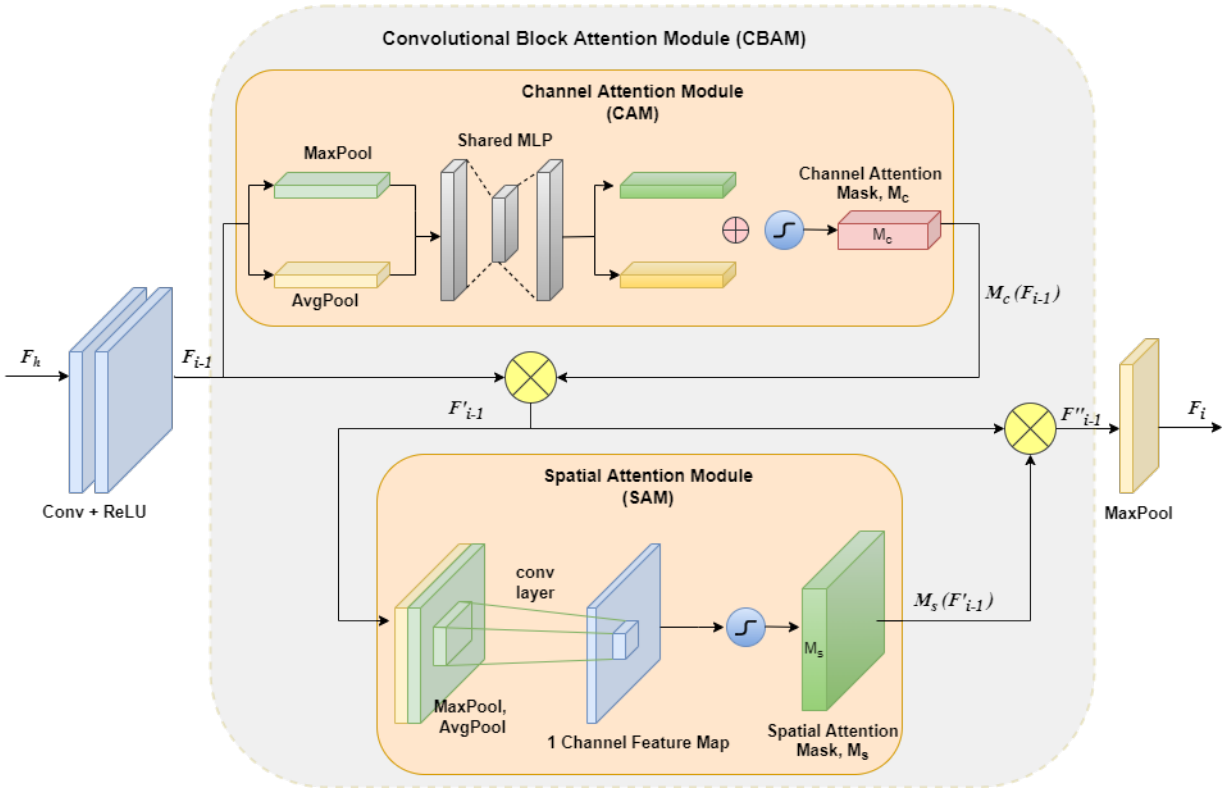

Figure 3.3: Block diagram of each VggCBAM block

Let F_h and F_i be the input and output of i^{th} block of VggCBAM. The input F_h is first fed in to the convolutional layers and a set of feature maps $F_{(i-1)}$ are extracted as output. The feature maps are then first passed through the Channel Attention Module (CAM) of CBAM. Since each channel of a feature map is viewed as a feature detector, channel attention focuses on "what" is significant given an input image. To efficiently compute the channel attention, we minimize the input feature map's spatial size. Average pooling and max pooling are applied on the feature maps to collect spatial and contextual information. The obtained spatial contextual descriptors: $F^c_{(i-1)max}$ and $F^c_{(i-1)avg}$, are then received by a shared network that is composed of a multi-layer perceptron (MLP) with a single hidden layer. To reduce parameter overhead, the hidden activation size is set to $R^{c/r \times 1 \times 1}$, where r is the reduction ratio. The resultant feature vectors are combined using element-wise summation after the shared network has been applied to each descriptor. A sigmoid activation is finally applied on to this output to produce a 1-dimensional channel attention

mask $M_c \in R^{c\times1\times1}$. This mask M_c is then multiplied to the input feature maps $F_{(i-1)}$ to obtain channel refined feature maps $F^{'}_{(i-1)}$.

The Spatial Attention Module (SAM) of CBAM then takes this channel refined feature maps $F^{'}_{(i-1)}$. Instead than focusing on "what," as channel attention does, spatial attention places more emphasis on "where". Prior to computing the spatial attention, we first execute average-pooling and max-pooling operations along the channel axis and concatenate them to provide an efficient feature descriptor. The concatenated feature descriptor is then fed in to a convolution layer to produce a one-channel feature map which is then subjected to a sigmoid activation function to generate a spatial attention map $M_s \in R^{1\times h\times w}$ that encodes where to emphasize or suppress. This map M_s is then multiplied to the input feature maps $F^{'}_{(i-1)}$ to obtain the final refined feature maps $F^{''}_{(i-1)}$. Finally, a maxpooling operation is applied on $F^{''}_{(i-1)}$ to obtain the result F_i. Every block of the proposed method is subjected to the same operations. The idea is to extract feature maps from last two layers of the model i.e., Attention-5 and Pool-5.

The mathematical representation of operations performed on each block of VggCBAM is illustrated as follows:

$$F_i = H_i(F_h) \tag{3.1}$$

where H_i denotes the overall operations of the i^{th} block of VggCBAM.

The output of the convolutional block layer can be represented as

$$F_(i-1) = ReLU(conv(X_h)) \tag{3.2}$$

where X_h = concat(F_{h1}, F_{h2},.....,F_{hm}). Now the output of convolutional operation becomes

$$O_{convh} = W_h * X_h + B_h \tag{3.3}$$

After applying ReLU activation function on O_{convh}, we get $F_{(i-1)}$ as :

$$F_{(i-1)} = max(0, O_{convh}) \tag{3.4}$$

By integrating all the functions, the overall output $F_(i-1)$ can be represented as:

$$F_{(i-1)} = max(0, W_h * concat(F_{h1}, F_{h2}, \ldots\ldots, F_{hm}) + B_h) \tag{3.5}$$

The overall attention process when $F_{(i-1)}$ is inputted in to CBAM module can be summarized as below, where $\otimes$ denotes element-wise multiplication.

$$F^{'}_{i-1} = M_c(F_{i-1}) \otimes F_{i-1} \tag{3.6}$$

$$F^{''}_{i-1} = M_s(F^{'}_{i-1}) \otimes F^{'}_{i-1} \tag{3.7}$$

The channel attention and spatial attention are computed as:

$$M_c(F_{i-1}) = \sigma(MLP(AvgPool(F_{i-1})) + MLP(MaxPool(F_{i-1})) \tag{3.8}$$

$$M_s(F^{'}_{i-1}) = \sigma(f^{7\times7}([(AvgPool(F^{'}_{i-1}); MaxPool(F^{'}_{i-1})])) \tag{3.9}$$

where σ denotes the sigmoid function and $f^{7\times7}$ represents a convolution operation with the filter size of 7x7 .

The final output F_i is obtained as:

$$F_i = MaxPool(F^{''}_{i-1}) \tag{3.10}$$

The extracted feature maps from Attention-5 and Pool-5 can preserve regions that may be corresponding parts of an object. However, not all activated regions in feature maps are useful. Thus, it is necessary to mine meaningful regions using frequent-itemset mining techniques. The most

important step when using frequent-itemset mining algorithms for computer vision applications is to transform input into transactions while keeping meaningful information. This method transforms each feature map into a transaction denoted by the letter T by treating each position index that is activated from a feature map as an item I. For instance, if there are five places fired on the j-th feature map, the accompanying transaction T would include five items, or $T_j = \{i_1, i_2, i_3, i_4, i_5\}$. The set of all transactions (i.e.all feature maps) is denoted by D. In practice, the proposed feature extraction method receives an image of a size of $h \times w \times 3$. Next, get 512 feature maps with sizes of $\frac{h}{32} \times \frac{w}{32}$ from Pool-5 and $\frac{h}{16} \times \frac{w}{16}$ from Attention-5, respectively. Resize the Pool-5 feature maps to the same size as Attention-5 using bilinear interpolation. Each feature map is therefore $\frac{h}{16} \times \frac{w}{16}$ in size. The following step is to choose useful descriptors that can be turned into items. Each feature map's tunable threshold should be determined as the mean value of activation responses greater than 0. When the location has a response magnitude that is higher than the mean, the index will be converted into an item.

Once the transaction is created, the next step is to apply frequent-itemset mining. In this case, we mine patterns using the Level-wise search technique. The Level-wise search algorithm uses a breadth-first, bottom-up approach to explore item sets given the transaction database D. Starting with an item, the algorithm estimates the frequency of a subset of items in the transactions with the same item set size at each iteration. Only items whose support values exceed a predefined minimum support threshold are then retained, and the item set size is subsequently increased by one. For a given minimum support threshold α, an itemset P is considered as frequent if $Supp(P) \geq \alpha$. In other words, which patterns are mined depends on the support threshold. Given an itemset $P \subseteq I$, we calculate the frequency K of the itemset P in the transaction database D by $K = |\{T \mid T \in D\}|$. Then, we define the support value of P as: $Supp(P) = \frac{K}{N}$, where N is the number of transactions in D. The frequent itemset P is the itemset whose support value exceeds a predetermined threshold. One target item is spatially continuous in a single image, thus the components of that object that are commonly represented by patterns should also be spatially continuous. Try to select the largest linked component based on the mined patterns in order to cover the entire target object as much as possible. The frequent patterns are then integrated to create a support map S for each image. The size of the support map is the same as the size of the feature

map. Using bilinear interpolation, we up sample the support map to make it the same size as the original image. Finally, take the bounding box surrounding the largest connected component shown on the support map.

Chapter 4

Experimental Results and Discussions

This section will evaluate the effectiveness of the proposed method on object localization task.

4.1 Implementation Details and datasets

4.1.1 Implementation details

In order to extract feature maps, a new feature extractor model called VggCBAM with VGG16 architecture as backbone is proposed in this work. The model is built by integrating Convolutional Block Attention Module(CBAM) to each convolutional block of VGG16. For obtaining frequent patterns, Apriori algorithm is used.

4.1.2 Datasets

We used CIFAR-10 dataset [24] for training the proposed VggCBAM model. The dataset consists of 60000, 32x32 colour images in 10 classes, with 6000 images per class. The different class labels of CIFAR-10 dataset are airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. For VggCBAM training purpose, we used 10000 images. 20% of the data were used as validation samples during training.

In order to evaluate the performance of the proposed approach, a set of qualitative and quantitative experiments is conducted on a variety of benchmark datasets. Specifically CUB-200-2011 [25], Stanford Dog [26], Standford Cars-196 [27], Object Discovery dataset [28] are used for object localization. We collected 150 images from CUB-200-2011 that contain images of "birds" category, 150 images from Object Discovery dataset (that comes with three categories: airplane, car, and horse), 100 images from Standford Cars-196 and 100 images from Stanford Dog.

4.1.3 Evaluation metrics

Following previous image localization works[29], [10], correct localization (CorLoc) is used as the evaluation metric to evaluate the performance of OLM. According the PASCAL-criterion, the CorLoc is defined as:

$$\frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} > 0.5 \tag{4.1}$$

where B_p is the predicted box and B_{gt} is the ground-truth box. If the CorLoc value between each of the candidate image and the respective ground truth box is greater than 0.5 then it is treated as correctly localized.

4.2 Feature Map Extractor Model Training

Proposed deep learning network is implemented using keras platform with tensorflow as the back-end. All the experiments were conducted on a PC with Intel® Core™ i7-5820k, 128G main memory, 3.30 GHz CPU and an Nvidia GeForce 980 Ti GPU. VggCBAM is trained with epoch of 50 and batch size of 16. For optimization, Adam optimizer is used and the learning rate set was set to 0.0001. To avoid overfitting of the deep learning model, a dropout of 0.3 and L2-regularization of 0.01 is used. In addition, the loss function is trained by categorical cross-entropy loss.

In order to prove the effectiveness of VggCBAM after adding CBAM to the VGG convolutional blocks, we randomly selected an image and visualized feature maps responses from ReLU-5 of original VGG16 and Attention-5 of our VggCBAM. As shown in Figure 4.1, we can clearly observe that the heatmaps of ReLU-5 and Attention-5 are not the same. The feature maps from Attention-5 gives more focus to the entire target object while suppressing the complex background. Whereas, the feature maps from ReLU-5 shows only smaller activations that can even include the background regions.

4.3 Experimental Results

Experiments were conducted on CUB-200-2011, Stanford Dog, Standford Cars-196 and Object Discovery datasets to evaluate the performance of the proposed method. The CorLoc % obtained for different datasets are

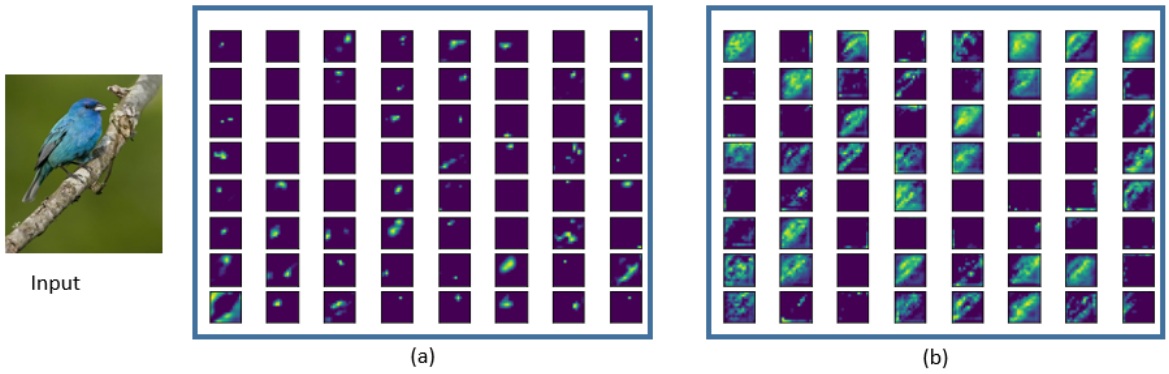

Figure 4.1: (a) Feature maps extracted from ReLU-5 layer of VGG16. (b) Feature maps extracted from Attention-5 layer of the proposed feature extractor

Methods	Supervision	**Full Set**
OLM	w/o	79
Ours	w/o	93

Table 4.1: Comparisons of CorLoc (%) on CUB-200-2011 Dataset.

represented in tabular form. It is observed that our method outperformed the previous work, OLM, by a large margin. The CorLoc % obtained by our method on CUB-200-2011 dataset is 93% ,whereas the CorLoc % of OLM is 79% only. Like wise, CorLoc % obtained by our method from other datasets are also high comparing to OLM. In following Figures, we can visualize the comparison of some of the localization results of our method and OLM. It is clear from the figures that our method successfully locates objects in images with complex backgrounds.

4.4 Limitations

Even though we can localize the objects in an image, it is hard to fix the value of minimum support. Thus, the minimum support value varies between images. Also, the apriori algorithm takes up large memory for certain images. Thus it is necessary to come up with a pattern mining algorithm that can efficiently mine the frequent items yet takes up comparatively small amount of memory.

Methods	Supervision	Airplane	Car	Horse	**Mean**
OLM	w/o	72	79	92	81
Ours	w/o	86	93	97	92

Table 4.2: Comparisons of CorLoc (%) on Object Discovery Dataset.

Methods	Supervision	**Full Set**
OLM	w/o	73
Ours	w/o	87

Table 4.3: Comparisons of CorLoc (%) on Stanford Dog Dataset.

Methods	Supervision	**Full Set**
OLM	w/o	68
Ours	w/o	81

Table 4.4: Comparisons of CorLoc (%) on Standford Cars-196 Dataset.

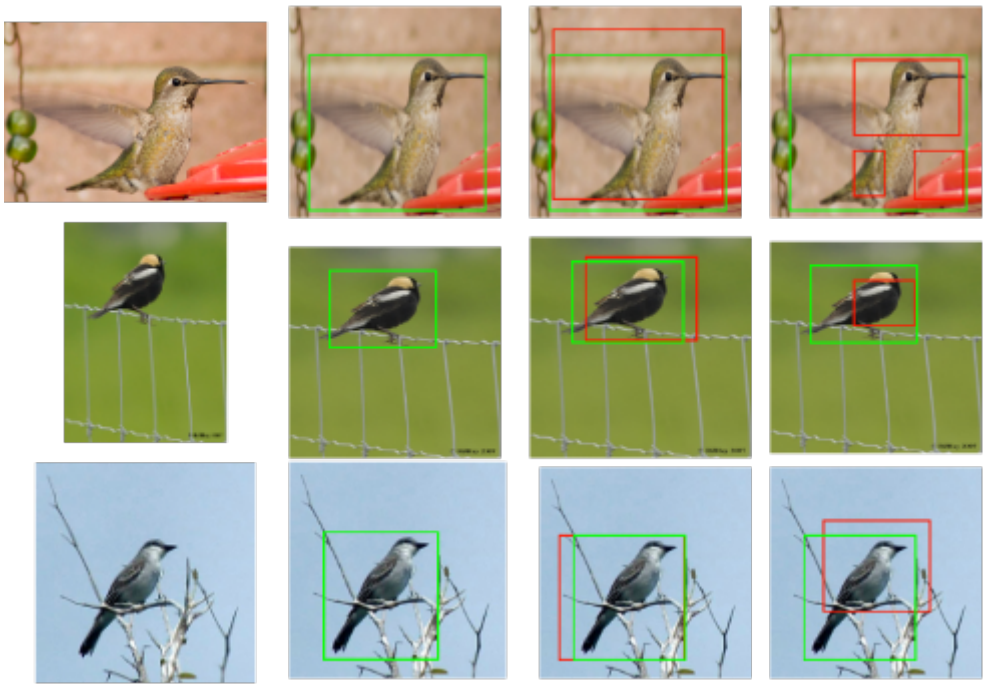

Figure 4.2: Localization results on CUB-200-2011. Figure shows input image, Ground Truth, Localized result from proposed method and Localized result from OLM method respectively

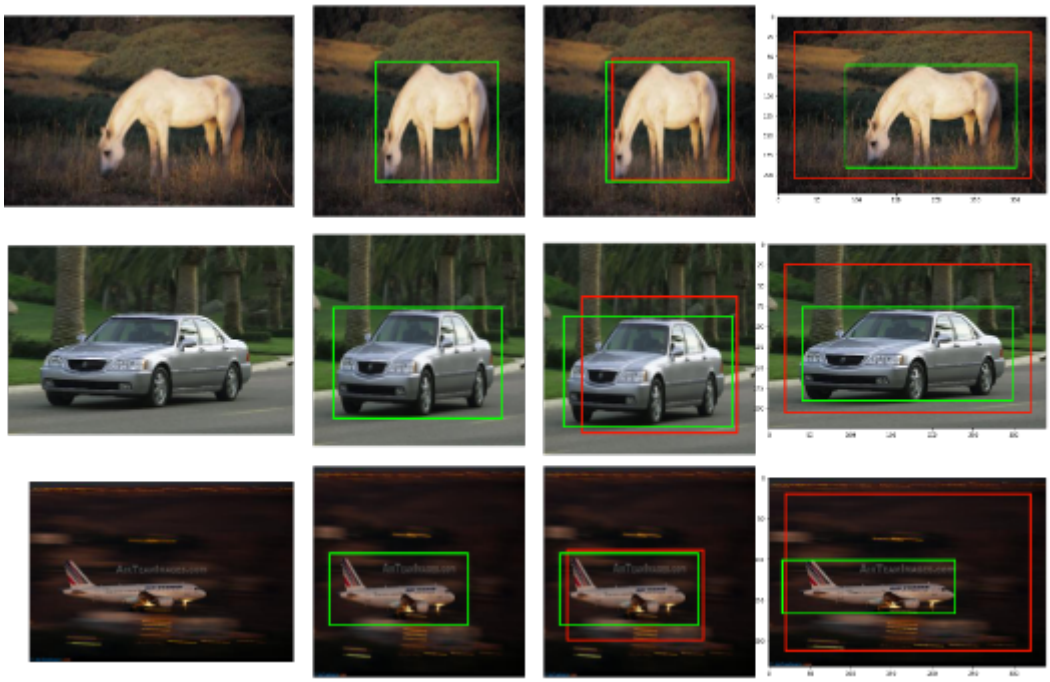

Figure 4.3: Localization results on Object Discovery Data. Figure shows input image, Ground Truth, Localized result from proposed method and Localized result from OLM method respectively

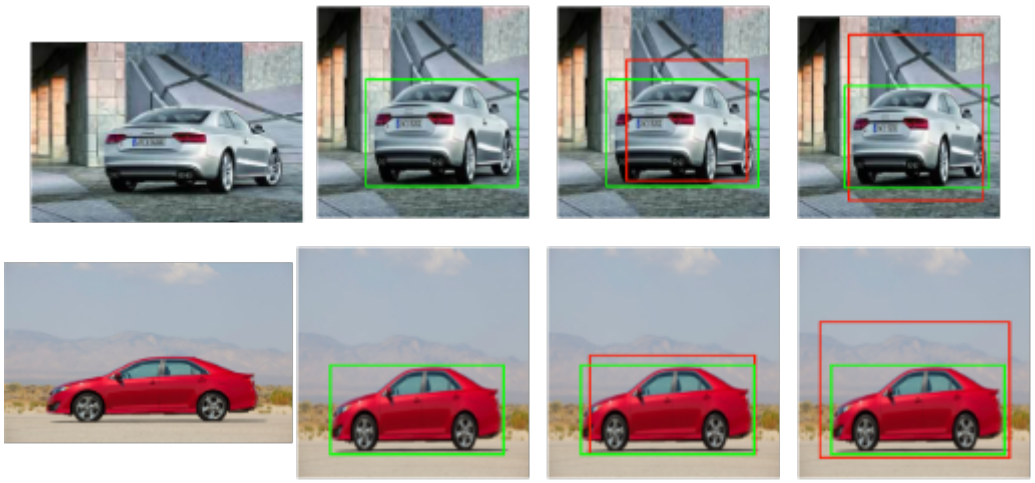

Figure 4.4: Localization results on Stanford Cars-196. Figure shows input image, Ground Truth, Localized result from proposed method and Localized result from OLM method respectively

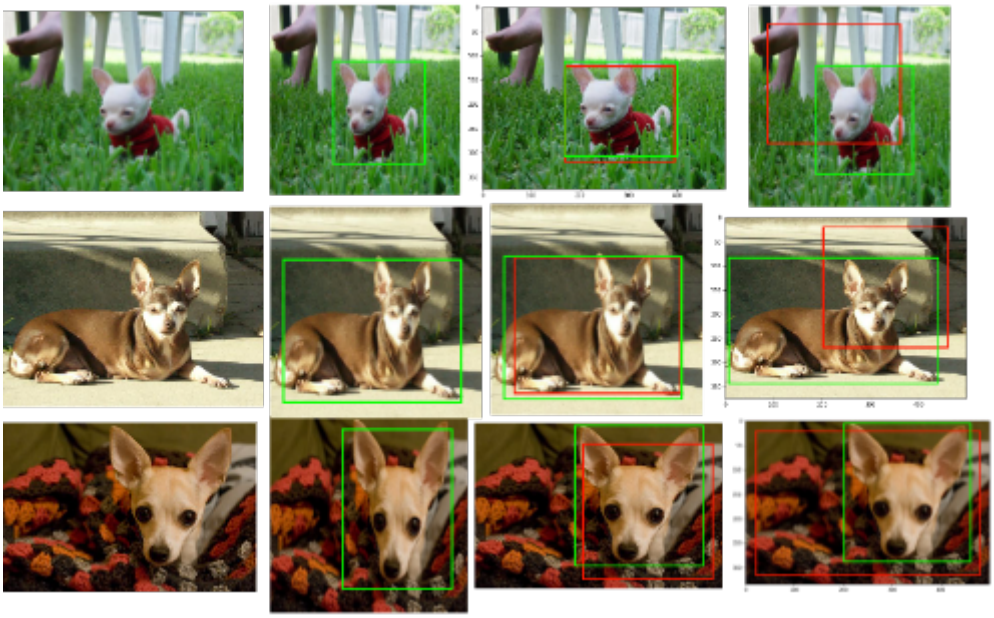

Figure 4.5: Localization results on Stanford Dogs. Figure shows input image, Ground Truth, Localized result from proposed method and Localized result from OLM method respectively

Chapter 5

Conclusion & Future Works

In this project, we propose a novel attention-based method for object discovery and localization from a single unlabeled image. The method exploits the advantage of attention mechanisms, deep learning CNN models, and data mining techniques. In order to extract feature maps, a new feature map extractor model, called VggCBAM, is proposed. The feature maps extracted from VggCBAM contain meaningful activations that increase focus on the object of interest while suppressing the background. Experimental results show that the proposed method outperformed the previous pattern mining-based work on object localization by accurately predicting the location of the object in an unlabeled image, thereby demonstrating the effectiveness of coupling pattern mining with the attention mechanism. The approach does not need any annotations yet still shows promising localization ability, which provides a new perspective to solve the localization problem. This method has the potential to be extended to other object detection applications such as localization on medical images, crowd density estimation, etc.

References

[1] C. Bhagya and A. Shyna, "An Overview of Deep Learning Based Object Detection Techniques," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-6, doi: 10.1109/ICIICT1.2019.8741359.

[2] Liu, W. et al. (2016). SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9905.

[3] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in CVPR, 2016, pp. 2921–2929.

[5] T. Durand, T. Mordan, N. Thome, and M. Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in CVPR, 2017, pp. 5957–5966.

[6] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in CVPR, 2018, pp. 1325–1334.

[7] B. Zhou, V. Jagadeesh, and R. Piramuthu. Conceptlearner: Discovering visual concepts from weakly labeled image collections. Proc. CVPR, 2015.

[8] K. Tang, A. Joulin, L. J. Li, and F. F. Li, "Co-localization in real-world images," in CVPR, 2014, pp. 1464–1471.

[9] Y. Li, L. Liu, C. Shen, and A. V. D. Hengel, "Image co-localization by mimicking a good detector's confidence score distribution," in ECCV, 2016, pp. 19–34.

[10] X. S. Wei, C. L. Zhang, Y. Li, C. W. Xie, J. Wu, C. Shen, and Z. H.Zhou, "Deep descriptor transforming for image co-localization," in IJCAI, 2017, pp. 3048–3054.

[11] X. S. Wei, J. H. Luo, J. Wu, and Z. H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval." IEEE TIP, vol. 26, no. 6, pp. 28682881, 2017.

[12] R. Zhang et al., "Object Discovery From a Single Unlabeled Image by Mining Frequent Itemsets With Multi-Scale Features," in IEEE Transactions on Image Processing, vol. 29, pp. 8606-8621, 2020.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in ICLR, 2014.

[14] Y. Li, L. Liu, C. Shen, and A. V. D. Hengel, "Mining mid-level visual patterns with deep cnn activations," IJCV, vol. 121, no. 3, pp. 1–21, 2016.

[15] T. Quack, V. Ferrari, B. Leibe, and L. J. V. Gool, "Efficient mining of frequent and distinctive feature configurations," in ICCV, 2007, pp. 1–8.

[16] B. Fernando, E. Fromont, and T. Tuytelaars, "Mining mid-level features for image classification," IJCV, vol. 108, no. 3, pp. 186–203, 2014.

[17] Zheng M, Geng Y, Li Q. Revisiting Local Descriptors via Frequent Pattern Mining for Fine-Grained Image Retrieval. Entropy. 2022; 24(2):156.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[19] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A. (2018). Self-Attention Generative Adversarial Networks. arXiv. https://doi.org/10.48550/arXiv.1805.08318.

[20] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T. (2016). SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. arXiv. https://doi.org/10.48550/arXiv.1611.05594.

[21] Woo, S., Park, J., Lee, J., Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. arXiv. https://doi.org/10.48550/arXiv.1807.06521.

[22] S. Yu, S. Jin, J. Peng, H. Liu and Y. He, "Application of a new deep learning method with CBAM in clothing image classification," 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), 2021, pp. 364-368, doi: 10.1109/ICESIT53460.2021.9696783.

[23] Lu, E., Hu, X. Image super-resolution via channel attention and spatial attention. Appl Intell 52, 2260–2268 (2022). https://doi.org/10.1007/s10489-021-02464-6.

[24] inproceedingsKrizhevsky2009LearningML, title=Learning Multiple Layers of Features from Tiny Images, author=Alex Krizhevsky,year=2009

[25] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, 2011.

[26] A. Khosla, N. Jayadevaprakash, B. Yao, and F. fei Li, "L.: Novel dataset for fine-grained image categorization," in CVPR Workshop on FGVC, 2011.

[27] J. Krause, M. Stark, D. Jia, and F. F. Li, "3d object representations for fine-grained categorization," in ICCV Workshops, 2013, pp. 554–561.

[28] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in CVPR, 2013, pp. 1939–1946.

[29] 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 CVPR, 2015, pp. 1201–1210.