Unsupervised Multi-class Object Discovery by Spherical Clustering of Deep Features

Kazuhiko Murasaki †,††, Yukinobu Taniguchi (member)††, Tetsuya Kinebuchi †

Abstract In this paper, we propose a novel method for multi-class co-localization; it offers unsupervised localization of the main object in each image from an image set containing multiple kinds of main objects. Our method utilizes deep features to tackle the co-localization problem. Deep features, which can be extracted by pre-trained neural networks, are effective against unsupervised co-localization from multi-class image set. Based on spherical clustering, we classify deep features into several clusters, and choose one dominant cluster for each image or each image set. Experiments show that this very simple approach is significantly better than conventional state-of-the-art techniques in terms of localization accuracy. Moreover, multi-class co-localization experiments show that our method has the potential to classify the object in each image at the same time as achieving localization.

Key words: Co-localization, Object discovery, Unsupervised learning, Deep feature, CNN, Spherical clustering

1. Introduction

Recognizing the appearance of objects from just an image set without any supervision remains a fundamental problem in image recognition. Solving this problem means the system will be able to acquire unknown object classes one after another from the large amounts of images available on the web etc.1). Toward this ultimate goal, stepwise problems in unsupervised learning have been tackled. Simple co-localization is the task of detecting target objects without supervision while assuming that all images in the dataset hold objects of the same target category. As a more realistic problem, co-localization from an image set that includes a small number of noise images, which do not include the target object, has been proposed2). Cho et al.3) tried to solve the more difficult problem of co-localization from an image set containing objects of multiple categories. However, the localization accuracy reported was low because of background clutter or small objects. In this paper, we also try to solve the co-localization problem given single category image sets and multiple category image sets. By using the intermediate output of a trained CNN as a deep feature, we develop a very...
simple method that can localize objects with high accuracy. There are three novelities in our approach. First, we apply spherical clustering to deep features to extract co-occurring object regions. Second, the main cluster indicating the foreground object is decided based on the intensities of the deep features belonging to each cluster. Third, our approach is easily extended to multi-class problem by increasing the number of clusters and selecting the main cluster for each image. Although this is a quite simple and straightforward approach, to the best of our knowledge, nobody has tried to apply spherical clustering to deep features. Through extensive experiments on various datasets, we show that our simple approach based on spherical clustering on pre-trained deep features is able to extract features of co-occurring objects. Additionally, the proposed method can perform not only localization but also classification simultaneously for an image set containing objects of multiple categories in an unsupervised way.

Experiments show that proposed method outperforms existing state-of-the-art methods in both co-localization of single category objects and co-localization of multiple category objects. In addition to completely unsupervised methods, it is compared with methods that also use pre-trained deep features and those that use negative images for weakly supervised learning. The results show that our method outperforms all of them in terms of localization accuracy. Furthermore, we conduct experiments on multiple category image sets with goal being to demonstrate unsupervised classification with localization. The results show that our method is effective for simple datasets. Some results of unsupervised multi-class object discovery are shown in Fig.1. The color maps in the figure indicate the discovered clusters, and show that three kinds of objects were simultaneously localized and classified correctly.

2. Related Works

2.1 Image Co-Localization

Extracting object position from an image set in which objects of the same category appear is known as co-localization\(^3\) \(^5\) and co-segmentation\(^2\) \(^6\) \(^8\). Various approaches have been proposed so far including clustering of local features\(^7\) \(^9\) \(^10\) and correspondence of local features between images\(^3\)\(^5\)\(^9\)\(^10\). Many studies on co-localization and co-segmentation assume just one target object is present in each image set\(^4\)\(^5\)\(^8\), which means the problem is to extract a consistent object area from the image set. Recent studies have tackled the more difficult problem of co-localization from image sets that include multiple target objects\(^3\)\(^6\). Wang et al.\(^6\) segment the foreground area by the approach based on alternate optimization of local image segmentation and global similarity connection between partial images, and Cho et al.\(^7\) localize the foreground object by the approach based on matching local regions and arrangement of them between similar images. In particular, the method of Cho et al. is effective even in the case where the number of categories in the image set is large or the set contains noise images which do not contain any target object. Co-localization from a multiple category image set considers the apparent difference of each category, but does not explicitly acquire the category information itself. Category discovery as well as localization is a more difficult task. Wang et al.\(^6\) propose a co-segmentation method that performs segmentation while simultaneously identifying multiple classes, but the image set is assumed to contain extremely similar objects or just two classes of general objects. To the best of our knowledge, no experiment on simultaneous localization and classification has proceeded without supervision being provided by an image set containing many general categories (e.g. 20 categories in Pascal VOC\(^11\)).

2.2 Pre-trained Deep Feature

Deep learning has been widely used, and the intermediate output of the trained deep neural network has, an image feature, been applied to various image recognition problems. The extracted features, called deep features, are known to be effective even if the target problem differs from the problem to train the original CNN. For example, learning based on deep feature is known to be effective for object proposal extraction\(^12\) and object detection\(^13\).

Pre-trained deep features are also useful in co-localization. Wei et al.\(^3\) reported that co-localization based on deep features is highly accurate. In DDT by Wei et al., PCA is applied to deep features obtained from an image set, and a region with strong the first principal component is simply extracted. However the commonly used deep feature is trained based on the 1000 category classification task of ILSVRC using ImageNet\(^14\), it has been shown that deep features are effective not only for the 1000 categories pre-trained but also for categories not present in ImageNet. Deep features are expected to be a versatile and powerful feature expression for various objects.

In this paper, we tackle unsupervised object discovery
by the simple analysis of deep features like DDT. We show that co-localization with high accuracy is realized by extracting co-occurring features through spherical clustering. The method proposed herein can be applied to the co-localization problem with multiple categories, which is impossible with DDT, and it yields significantly better performance than the state-of-the-art method. Furthermore, by assigning a consistent label to the entire image set, we can not only discover the main object in each image but also the object classes in the image set.

3. Proposed Method

In DDT, PCA is used to select deep features belonging to the main object in the image set. This method is quite simple in that if the value of the first principal component calculated from the dataset is greater than zero, it is considered to be part of the main object. They extract deep features by pre-trained CNN pixel by pixel and subtract the mean vector of all features from them. If the first eigenvector calculated from a set, $X$, of the normalized deep feature $x_i$ is taken to be $\xi_1$, the first principal component of each feature is given by

$$p_i = \xi_1^T x_i.$$  \hspace{1cm} (1)

If $p_i$ exceeds zero, it is regarded as belonging to the main object. This calculation implies that feature $x_i$ belongs to the main object if the direction of the eigenvector $\xi_1$ closely approaches that of the feature, and the features of main object are spreading in the direction of the eigenvector. Our proposed method is based on the hypothesis that the direction of deep feature $x_i$ indicates the category it belongs to, while its intensity indicates the confidence of its assignment. Fig. 2 provides a graphical comparison of our method with DDT. Fig. 2(a) and (b) show how DDT and our method extract features that correspond to the main object, respectively. Especially, Fig. 2(b) shows the example when the number of clusters is 3. Although PCA captures the direction of the feature vectors of the main object, DDT assigns the category by the hyper plane. On the other hand, our method assigns category explicitly based on the direction of feature vectors as given by the cosine distance from cluster centers $c_k$. To validate this hypothesis, Fig. 3 shows deep features representing regions of three classes in Object Discovery dataset. Deep features output from VGG19 are projected onto a 2D plane spanned by first and second principal vectors. We can see that the features of the same category spread radially from the origin, and features of three categories intersect around the origin. Following this hypothesis, we employ spherical clustering, which clusters features according to the directions of the vectors identified by cosine distance instead of Euclidean distance. We propose an approach of using spherical clustering to extract the direction of the deep features of the main category in the image set where decision reliability is given by the intensity of the direction component. The processing flow of our method is shown in Algorithm 1. Each function in the algorithm is detailed below.

### 3.1 Feature Extraction

First, deep features are extracted by a pre-trained CNN. As the pre-trained model, we employ VGG19 because it has a simple CNN structure with convolution, activation and pooling layers, and shares its deep feature characteristics with other basic CNN models. The other reason is to compare our method with DDT.

The 512 dimension intermediate output vectors from the conv 5 layer, which is the final output of the convolution layers, are taken to be the feature vectors. In VGG19, since the vertical and horizontal sizes are reduced by the pooling layers four times, deep feature width becomes $1/16$ times the width of the input image.
Algorithm 1 Co-localization via spherical clustering

Input: \( I = \{I_j\} \): Images, \( K \): Number of clusters
Output: \( B = \{B_j\} \): Bounding box for each image \( j \)

1: \( \text{function } \text{co-localization}(I) \)
2: \( F \leftarrow \text{empty} \)
3: \( \text{for } j = 1, 2, \ldots, N_i \text{ do} \quad \triangleright \text{ } N_i: \text{Number of images} \)
4: \( f_{\text{map}} \leftarrow \text{CNN}(I_j) \quad \triangleright \text{Extract deep feature map} \)
5: \( f_{\text{orig}} \leftarrow \text{Resize}(f_{\text{map}}) \)
6: \( \triangleright \text{Resize feature map to original image size} \)
7: \( F \leftarrow F \cup f_{\text{orig}} \quad \triangleright \text{Store feature vectors} \)
8: \( \text{end for} \)
9: \( \bar{F} \leftarrow F - \text{mean}(F) \quad \triangleright \text{Subtract mean value} \)
10: \( C \leftarrow \text{SphericalClustering}(\bar{F}, K) \)
11: \( \triangleright \text{Output cluster ID for each feature vector} \)
12: \( l \leftarrow \text{MainClusterSelection}(\bar{F}, C) \)
13: \( \text{for } j = 1, 2, \ldots, N_i \text{ do} \)
14: \( M_j \leftarrow \text{MainClusterMap}(I_j, C, j) \)
15: \( \triangleright \text{Create binary map of main cluster} \)
16: \( B_j \leftarrow \text{BoundingBox}(M_j) \)
17: \( \text{end for} \)
18: \( \text{return } B \)
19: \( \text{end function} \)

In this paper, we integrate atrous convolution\(^13\) into the VGG19 model in order to increase resolution and perform localization with higher definition. With the use of atrous convolution instead of convolution layers and setting the stride of the pooling layer to 1 in block 5 of VGG19, deep feature width becomes 1/8 times the width of the input image. This is also true with regard to height. After that, the deep feature map output from CNN is resized to original image size by bilinear interpolation, which ensures the correspondence of feature vectors and pixels.

3.2 Spherical Clustering

Spherical clustering is applied to the deep features extracted from the image set. It yields effective clustering when the distribution of the vector elements has significance rather more than their intensity, as is true in document analysis. In spherical k-means clustering\(^15\), each data point is projected onto a hypersphere of radius 1, and k-means clustering based on Euclidean distance is applied to them. The number of clusters, \( K \), must be given beforehand, while the initial value at the cluster center is set based on k-means++.\(^18\) We set the center of the hypersphere to the average value of the entire data set by subtracting the total average vector from the input features before applying clustering. K cluster centers \( c_k (k = 1 \cdots K) \) are output by spherical clustering, and cluster number \( l_i \), which corresponds to neighbor cluster centers, is given for each data point \( x_i \). An example in which \( K = 7 \) clustering is applied to the Object Discovery dataset is shown in the 2nd row of the Fig. 1. As shown in the figure, it can be seen that the deep features obtained from the same object are grouped in the same cluster.

3.3 Main Cluster Selection

In order to localize the main object in the data set, it is necessary to select the cluster number that indicates the target object. To select the cluster number corresponding to the main object, we define the intensity \( S_k \) for each cluster \( k \) as:

\[
S_k = \frac{1}{N_k} \sum_{i \in F_k} c_k^T x_i,
\]

where \( N_k \) is the number of features belonging to cluster \( k \), \( F_k \) denotes the group of features belonging to cluster \( k \) and the cluster center \( c_k \) is normalized so that the absolute value is 1. Considering the assumption that the category the deep feature belongs to is expressed by the direction of the feature, and the confidence is expressed by the intensity of the feature, the cluster with the highest confidence in the dataset becomes the main object. The confidence of each feature \( x_i \) is the component of \( x_i \) along the direction of the corresponding cluster center \( c_k \).

We select the cluster with the highest intensity \( S_k \) for the dataset as the main object cluster. In the case of multi-class co-localization, cluster selection is applied to each image. The cluster with the highest intensity for each image becomes the main object cluster in that image. The 3rd row of Fig. 1 shows the main object extracted from each image. Since the size of the image labeled by assigned cluster number is reduced with respect to the input image, the cluster number assigned to each data point and the pixels in the image are linked by upsampling the label image to the input image size. The pixels belonging to the main object cluster are considered to be the region of the main object.

The minimum rectangle bounding box that contains the largest connected component of each main object region is returned as the co-localization prediction results. The bottom row of Fig. 1 shows the predicted bounding boxes.

4. Experiments

We conduct several experiments to show that the proposed method is a powerful means of unsupervised object discovery in spite of its simplicity.

4.1 Evaluation Metric and Datasets

Following previous image co-localization works\(^{14,19}\), we take the correct localization (CorLoc) as the evaluation metric. CorLoc is defined as the percentage of images correctly localized according to the Intersection
over Union (IoU): \[ \text{IoU} = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} \] > 0.5, where \( B_p \) is the predicted bounding box and \( B_{gt} \) is the ground-truth bounding box. All CorLoc results are reported in percentages. When there are multiple objects in the image and there are multiple ground-truth bounding boxes, we calculate IoU for the bounding box nearest to the predicted bounding box.

Evaluation is conducted in three experimental setups. In single-class co-localization setup, there is only one category that appears in the image set in common. The ground-truth bounding box is always defined by the location of the target object. In multi-class co-localization setup, there are multiple categories appearing in the image set as main objects. The ground-truth bounding box changes according to the target category of evaluation. Finally, in multi-class co-localization with classification setup, while the input data is the same as multi-class co-localization, prediction output contains both bounding boxes and class labels. For each evaluation category, CorLoc metric is computed only for the images that contain the target category object.

The following three datasets are used in the performance evaluation.

(1) Object Discovery Dataset

The Object Discovery dataset\(^{21}\) was created to evaluate co-localization accuracy and data include noise images. Following the evaluation in the previous works\(^{14}\), 100 subsets of each of the three classes of Airplane, Car and Horse are used for in the evaluation. The images of the three categories included 18, 11, and 7 noise images, respectively, where none of the noise images included the target object.

(2) ImageNet Subsets Dataset

Since deep features are obtained by pre-training using 1000 categories included in ImageNet\(^{14}\), the acquired expressions are considered to be useful only for the categories covered by these 1000 classes. In order to evaluate the robustness of deep features against categories other than the pre-trained categories, subsets of ImageNet not included in the 1000 categories of ILSVRC are selected and used in the evaluation. Following the previous works\(^{14}\), we use six categories: Chipmunk, Rhino, Stoat, Raccoon, Rake, and Wheelchair.

(3) PASCAL VOC 2007 Dataset

PASCAL VOC 2007\(^{11}\) contains realistic images covering 20 categories. It is significantly more challenging than the Object Discovery dataset due to considerable clutter, occlusion, and diverse viewpoints. Following previous works, the images in the train+val set of 20 categories are used in the evaluation. However, images that difficult or truncated is labeled to all annotated objects are excluded from the data set. It should be noted that since the VOC dataset contains many images containing multiple target objects, the main object to be evaluated can not be uniquely determined. For example, when evaluating the co-localization of person category for an image showing a person and a bicycle at the same time, the person area becomes the main object, and when evaluating the co-localization of bicycle category, the bicycle area becomes the main object.

4.2 Single-Class Co-localization

First we evaluate the standard single-class co-localization problem. In this setup, the whole dataset is divided into the object categories and localization is attempted for each of the main objects in each partial dataset. We compare the proposed method with that of Cho et al.\(^{20}\) which is a purely unsupervised approach without any pre-trained information, and those of Wei et al.\(^{1}\) and Li et al.\(^{19}\) which take the unsupervised approach with pre-trained deep features. Furthermore, we also compare it with Wang et al.\(^{20}\) which is a weakly supervised approach with supervision of negative images that do not include the target object in addition to positive images and pre-trained features.

The accuracy evaluation for the Object Discovery dataset is shown in Table 1. This dataset is characterized by containing noise images, but since the proposed method does not attempt to find an area common to all images, highly accurate detection is realized without being negatively affected by the noise images. Since the target object is conspicuous in the image and the background is simple, the best accuracy is achieved when K=2. With K=2, our method offers processing similar to that of DDT and similar accuracy is achieved. With regard to the Horse category, many images contain horses and people at the same time and the background is complex, so K=3 gives the maximum accuracy. For categories Airplane and Car and the overall average, our method outperforms the conventional methods. The row ‘Best K’ shows the best score when K was changed in the range 2 to 5. It shows upper bound scores when the most appropriate K is decided for each category.

The accuracy evaluation using the ImageNet Subsets dataset is shown in Table 2. In this dataset, there are no noise images, only relatively simple images with just the target object, but there are objects not present.
in the ILSVRC 1000 categories used for pre-training deep features. This evaluation tested whether deep features were useful even for objects different from the pre-trained categories. The evaluation results in the table show that our method again yielded better accuracy than the conventional methods. Especially of the Rake category, the improvement in accuracy was significant. Even in the case of K=2, there is a large difference in accuracy from that of DDT\(^4\). This is because a person as a characteristic object appeared simultaneously with the rake, and the principal component direction obtained by PCA is strongly influenced by human characteristics. On the other hand, our method appropriately separates the feature of rake from the feature of person because it is based on clustering.

The accuracy evaluation using the VOC dataset is shown in Table 3. In this dataset, since objects of the background can appear simultaneously in each image and the background is cluttered, the accuracy is often maximized when K=3 or 4. Spherical clustering is effective against the complexity of the background, Using K=3, 4 or the best K for each category, our method significantly outperforms the conventional methods, not just fully unsupervised approaches\(^{24,19}\), but also the weakly supervised approach\(^{20}\). The type of cue used by each approach is indicated in the table as follows: positive images (P), negative images (N) and deep features (D). Boat and Sofa can be mentioned as categories for which the proposed method yields significantly different results from the conventional methods. Because these image types have no distinctive texture, it was difficult for the conventional methods to capture the co-occurrence between the images using either local or deep features.

In addition to the comparison with conventional methods, the score by k-means clustering based on Euclidean distance and the score by deep feature extracted by another CNN model are shown in the row ‘Euclid’ and ‘ResNet’ in Table 3. To validate the effect of spherical clustering, we tried standard k-means clustering based on Euclidean distance. Because we cannot decide the main cluster when using Euclidean distance, the main cluster is identified as the one that maximizes the CorLoc score. Even if main cluster selection is ideal, the Euclidean distance approach yields lower average score than our method, which is based on spherical clustering. This result supports our hypothesis that object category is related to deep feature direction. Moreover, ResNet50\(^21\), another pre-trained CNN model, was examined to elucidate the difference from VGG19. Although co-localization accuracy is slightly lower than the accuracy based on VGG19 features, it is still better than conventional methods.
### 4.3 Multi-Class Co-Localization

Because the proposed method employs a simple clustering approach, it is easily applied to *multi-class co-localization*. The cluster number indicating the main object is determined not for the entire dataset but for each image according to cluster intensity (Eq. 2) for the features in each image. We compare our method with that of Cho et al.\(^3\) which demonstrated the highest accuracy on *multi-class co-localization*.

Evaluation results for the Object Discovery dataset, ImageNet Subsets dataset and VOC dataset are shown in Table 4, 5 and 6, respectively. The number of clusters, K, is optimized by the average CorLoc accuracy to lie in the ranges of \((2,3,\ldots,10),(2,3,\ldots,15)\) and \((2,3,\ldots,25)\), respectively. The tables show that the proposal attained much better accuracy than Cho et al.\(^3\). The result for the ImageNet Subsets dataset shows almost the same recognition accuracy as the single-class task. This indicates that our method is a very robust to the number of categories. Furthermore, the result for the *Horse* category in the Object Discovery dataset exceeds the accuracy achieved in the single-class setup. The addition of images that do not include horses allows clustering to better captures the features of background objects such as people. An example of localization results for the VOC dataset are shown in Fig. 4. The three categories at the top of the figure show examples of accurate localization and the three categories at the bottom show examples of localization failure. In spite of the less texture (left example of *cat*), small size (right examples of *bus*, *cat* and *motorbike*), and combination with another object (left example of *motorbike*), our method could accurately localize the target objects. Failure cases are discussed in Section 5.

### 4.4 Multi-Class Co-Localization with Classification

Because the proposed method assigns a consistent cluster number to an image set that includes multiple categories of objects, it is possible to identify the class of the detected object simultaneously with localization. Wang et al.\(^6\) dealt with a similar problem in the field of co-segmentation, but the applicable image set was assumed to contain exactly the same object or only two classes. Experiments into *multi-class co-localization with classification* on large datasets like the VOC dataset have never been attempted before. The results on the three datasets are shown in Tables 7, 8 and 9. The number of clusters, K, is optimized by the average CorLoc accuracy to lie in the ranges of \((4,5,\ldots,9),(6,7,\ldots,20)\) and \((20,21,\ldots,40)\), respectively. The evaluation metric is the percentage of images correctly localized for each category and does not consider false detection such that the bounding box of the object class does not appear in the image. Cluster number is greedily assigned to the most applicable category to maximize the score.

#### Table 4 CorLoc(%) of multi-class co-localization on Object Discovery dataset.

| Method        | Airplane | Car   | Horse | Avg. |
|---------------|----------|-------|-------|------|
| Ours (K=7)    | 87.8     | 96.6  | 83.9  | 89.4 |
| Cho et al.\(^3\) | 81.7     | 94.4  | 71.0  | 82.4 |

#### Table 5 CorLoc(%) of multi-class co-localization on ImageNet Subsets dataset.

| Method        | Chipmunk | Rhino | Stoat | Racoon | Rake | Wheelchair | Avg. |
|---------------|----------|-------|-------|--------|------|------------|------|
| Ours (K=11)   | 85.3     | 94.4  | 75.1  | 87.4   | 58.4 | 64.8       | 77.6 |
| Cho et al.\(^3\) | 40.4     | 32.8  | 28.8  | 22.7   | 2.8  | 48.4       | 58.7 |

#### Table 6 CorLoc(%) of multi-class co-localization on VOC dataset.

| Method        | aero     | bicy    | bird    | boa     | bot     | bus      | car     | cat     | cha     | cow      | dbab     | dog      | hors     | pers     | plnt     | she      | sofa     | trai     | tv       | Avg.    |
|---------------|----------|---------|---------|---------|---------|----------|---------|---------|---------|----------|----------|----------|---------|----------|----------|---------|---------|---------|---------|---------|
| Ours (K=6)    | 73.7     | 55.0    | 60.6    | 46.8    | 10.8    | 70.5     | 62.6    | 70.0    | 20.5    | 67.0     | 36.7     | 71.1     | 68.9    | 73.3     | 4.6      | 7.2     | 52.9    | 57.5    | 71.4    | 8.1     | 49.5    |
| Cho et al.\(^3\) | 40.4     | 32.8    | 28.8    | 22.7    | 2.8     | 48.4     | 58.7    | 41.9    | 9.8     | 32.0     | 10.2     | 41.9     | 51.9    | 43.3     | 13.0     | 10.6    | 32.4    | 30.2    | 52.7    | 21.8    | 31.3    |

![Fig. 4](image.png) Examples of predicted object co-localization bounding box and pixels of main component cluster on VOC dataset.
The results for the Object Discovery dataset are not significantly inferior to the results with the multi-class co-segmentation setup. The 2nd and 3rd row of Fig. 1 show the results of multi-class co-segmentation with classification setup and the results of the multi-class co-segmentation setup, respectively, with K=7.

The CorLoc scores of multi-class co-segmentation with classification setup because the appearances of the three categories are quite different. On the other hand, the accuracy achieved with the VOC dataset and the Imagenet Subsets dataset is lower than that with the multi-class co-segmentation setup. This is because multiple categories are assigned to the same cluster. For example, since Chipmunk and Raccoon have very similar appearance, it is difficult to discriminate them. Although the accuracies are relatively low, this trend is also found with the VOC dataset and ImageNet Subsets dataset, which suggests that objects not included in 1000 pre-trained categories can be classified without supervision. The categories of aeroplane, bicycle, bus and cat are accurately recognized in the VOC dataset, but they are difficult to distinguish from boat, motorbike, car and dog, respectively, so the accuracies of the latter categories are remarkably worsened. In order to distinguish similar classes, it is necessary to hierarchically cluster the features according to the categories to be considered.

### 4.5 Computational Complexity
Another advantage of our method is low computation complexity. Spherical clustering, used in our method, computes similarities between all features and cluster centers, so the computation cost is $O(D_f N_i N_p K)$, where $D_f$, $N_i$, $N_p$ are the number of dimensions of the feature vector, the number of images in the dataset and the number of pixels in each image, respectively. Because Cho et al.\(^3\) compute similarities between every pair of local regions. It takes time $O(D_f N_n N_r^2)$, where $N_n$, $N_r$ are the number of neighbor images used in computing similarities and the number of local regions in each image, respectively. Wei et al.\(^4\) is based of PCA with complexity of $O(D_f^2 N_i N_p + D_f^3)$. Our method is more efficient than these conventional methods as well as more accurate.

### 5. Limitation
Examples of detection failure by the proposed method are shown at the bottom of Fig. 5. The figure shows three categories (person, pottedplant, tvmonitor) with low CorLoc scores for multi-class co-localization on the VOC dataset. Almost all detection failures were due to three factors: false selection of main object, over
division of clusters, and small size of target object. The examples from the person category show that motorbike or bottle were incorrectly detected as main objects, it can be seen that the cluster including the feature of person is not chosen in main cluster selection. In addition, the examples form the potted plant category show that plants and pots were divided into different clusters, and the whole potted plant was not captured because of over division. Furthermore, examples form the TV monitor category show that although the monitor area in the image was assigned to a constant cluster and the cluster is chosen as the main object, background noise is incorrectly extracted as the dominant part of the main component because of the small size of the TV monitor region. Solving these problems is future work. Another remaining task is to automatically determine the number of clusters, K.

6. Conclusion

In this paper, we showed that highly accurate co-localization can be achieved by spherical clustering with deep features and the automatic selection of clusters that include the main object. Although our method is quite simple, it offers highly accurate co-localization and high robustness to noisy images. Experiments show its superiority over conventional state-of-the-art alternatives and its applicability to multi-class tasks. Moreover, our method assigns class to the bounding boxes at the same time as localization. Our results suggest the possibility of discovering various categories from large amounts of images without any supervision.

References

1) X. Chen, A. Shirivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In Proc. of ICCV, pages 1409–1416, 2013.
2) M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In Proc. of CVPR, pages 1939–1946, 2013.
3) 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 Proc. of CVPR, pages 1201–1210, 2015.
4) X. Wei, C. Zhang, Y. Li, C. Xie, J. Wu, C. Shen, and Z. Zhou. Deep descriptor transformer for image co-localization. In Proc. of IJCAI, pages 3048–3054, 2017.
5) K. Tang, A. Joulin, L. Li, and L. Fei-Fei. Co-localization in real-world images. In Proc. of CVPR, pages 1464–1471, 2014.
6) F. Wang, Q. Huang, M. Ovsjanikov, and L. Guibas. Unsupervised multi-class joint image segmentation. In Proc. of CVPR, pages 3142–3149, 2014.
7) A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In Proc. of CVPR, pages 1943–1950, 2010.
8) T. Daryanto, S. Arif, and S. Yang. Survey: Recent trends and techniques in image co-segmentation challenges, issues and its applications. Int. J. of Computer Science and Software Engineering, 6(5):59, 2017.
9) K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In Proc. of CVPR, pages 19–25, 2006.
10) G. Kim and A. Torralba. Unsupervised detection of regions of interest using iterative link analysis. In Proc. of NIPS, pages 961–969, 2009.
11) M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. Int. J. of Computer Vision, 111(3):98–136, January 2015.
12) A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool. Deeprespool: Hunting objects by cascading deep convolutional layers. In Proc. of ICCV, pages 2578–2586, 2015.
13) R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proc. of CVPR, pages 580–587, 2014.
14) O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. Int. J. of Computer Vision, 115(3):211–252, 2015.
15) I. Dhillon and D. Modha. Concept decompositions for large sparse text data using clustering. Machine learning, 42(1-2):143–175, 2001.
16) K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.
17) G. Papandreou, I. Kokkinos, and P. Savalle. Untangling local and global deformations in deep convolutional networks for image classification and sliding window detection. CoRR, abs/1412.0296, 2014.
18) D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In Proc. of ACM-SIAM symposium on Discrete algorithms, pages 1027–1035, 2007.
19) Y. Li, L. Liu, C. Shen, and A. Hengel. Image co-localization by mimicking a good detector ‘s confidence score distribution. In Proc. of ECCV, pages 19–34, 2016.
20) C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In Proc. of ECCV, pages 431–445, 2014.
21) K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proc. of CVPR, pages 770–778, 2016.

Kazuhiko Murasaki received his B.E. from the University of Tokyo in 2009 and Master of Information Science and Technology in 2011. He joined NTT Corporation in 2011 and studied image recognition, especially about semantic segmentation, boundary detection and degradation detection of infrastructure. He is currently a researcher of NTT Media Intelligence Laboratories. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Yukinobu Taniguchi received the B.E., M.E., and Dr.Eng. degrees in mathematical engineering from the University of Tokyo in 1990, 1992, and 2002, respectively. He was with NTT Corporation from 1992 to 2015. He is currently a professor of Tokyo University of Science, Japan. His research interests include image/video processing and multimedia applications. He is a member of ACM, IPSJ, and IEICE.

Tetsuya Kinebuchi received the M.S. degree in physics from Tohoku University, Sendai, Japan, in 1997. He joined NTT Corporation in 1997. He is currently a senior research engineer, supervisor, of NTT Service Innovation Laboratory Group. His research interests include pattern recognition, image processing. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.