OBJECT CO-SEGMENTATION WITH REGRESSION SIAMESE NETWORK

Wei-Yu Chen (陳威宇) Chiou-Shann Fuh (傅楸善)

Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
r04921038@ntu.edu.tw fuh@csie.ntu.edu.tw

ABSTRACT

Object co-segmentation is a task to segment similar objects from a pair of images, while the objects may vary from view angles and poses. A recent approach [1] of this task can be roughly divided into three stages: *segmentation proposals retrieval*, *feature extraction*, and *similarity measurement*. While it focuses on how to achieve better segmentation and select proper feature to describe it, i.e, the first two stages, the similarity measurement in the third stage is rather naive. In this paper, we propose ***Regression Siamese Network***, a modified Siamese structure to measure the similarity between object segmentation proposals. The experiment result shows that our framework gives more accurate result than single image segmentation without the need of extracting extra features including both images.

Keywords *Object Co-Segmentation; Siamese Network;*

1. INTRODUCTION

Object co-segmentation problem is a task of segmenting similar objects from two images. Take Fig.1 as example. A pair of bear images is given, and our goal is to pixel-wisely extract the similar object, i.e., the bear, from these two images. The difficulty lies in the large variation of object poses and view angles, such as one bear staying in the pool with only head left on the water, while the other bear standing on the ground and we can clearly see its whole body. This task is beneficial to several application areas, such as image retrieval.

Most existing methods consider to apply a Markov Random Field (MRF) segmentation on both images, and to add a penalization term based on the dissimilarity between the histogram of foreground image region [2, 3, 4]. However, our goal is to extract "object" from image, and even though the image region is similar, it may contain unrelated "stuffs" (such as leaves or sky). It would be more decent to first consider regions likely to be object first, or say, with high ***objectness***.

A state-of-the-art approach [1] of this task considers a learning based approach to first select region with high objectness. The steps of the work can be divided into 3 stages: segmentation proposals retrieval, feature extraction, and similarity measurement. Once again taking Fig.1 as example, segmentation proposals retrieval stage aims at segment possible foreground objects from background, such as bear, pool, or rock. Each possible segment is known as ***proposal***, and only proposals with high objectness would be used**.** Once proposals have been extracted from both images, feature extraction stage would turn these proposals into a fixed size feature, and in similarity measurement stage, the most similar proposal pair would be selected as their co-segments.

Fig. 1: Co-segmentation between images. Identical parts between two images should be segmented together. Top row: raw image pairs. Bottom row: object segments, where darker area indicates background

However, the ways to extract feature and measure similarity is rather naïve in [1], where it use histograms as features and consider Random forest regressor to measure segmentation similarity. In contrast, in the recent study field of learning based approaches, the development of neural network and deep learning has achieved great performance in numerous applications which are known for combining feature learning and classification or regression power together with various and flexible structures. Thus, integrating structure of neural network with original co-segmentation framework has great potential to achieve better performance.

Among all kinds of neural network structures, Siamese network [5] is particularly used to verify the similarity between two feature pairs, since it's a structure projecting similar instances closer onto the feature space. And it has been demonstrated its capability in tasks such as face verification [6, 7] (determine two faces are from identical person or not) and one-shot learning task [8] (only observe an example of each possible class to predict test instance). Despite its usefulness, it has not been use on co-segmentation tasks before.

In this work, we aims at exploiting the capability of Siamese network to further improve the similarity measurement between object segments. While standard Siamese network is only used to classify whether two feature are belongs to same category or not, our proposed ***Regression Siamese Network*** aims at regressing their similarity.

In the experiment section, we demonstrate our framework indeed improving the similarity measurement in comparison with baseline single object segment approach without co-segmentation, even though cross–image features does not directly extracted, which is helpful for performance but quite inefficient for retrieval tasks.

To conclude, we list our contribution as follow:

- We are a pioneer work combines the task of object co-segmentation task with deep neural network based structure, and it could enhance similarity measurement without cross–image features be extracted.
- We extend the Siamese Network into a regression version, which relaxes its capability onto more general problems.

2. RELATED WORK

2.1 object segmentation

Object segmentation focus on dividing foreground object from background. This technique has been studied for more than 40 years since [9]. However, since a single image may contain multiple objects, lots of researches focus on partitioning images into multiple regions, where each region belongs to different object or background. Sometimes it is also called "image segmentation".

Recent approaches, such as the one we use in our work [10], consider a slightly different task: they give several binary segmentations, and each segmentation specifies one object while left other parts as background. Each segmentation, or say proposal, is ranked according to their object likeness, or objectness.

The objectness is one of the crucial consideration in our work. Other works related to objectness including [11], which is a sliding-window based object detection approach and computes object scores in bounding box to reduce search space.

It is worth-noting that, since we first extract object proposals and choose the best matched ones, the accuracy of our framework is bounded by the best proposal. Thus, choosing better object segmentation algorithm is helpful for our overall performance.

2.2 object co-segmentation

As mentioned in introduction, most existing methods consider to apply MRF segmentation on both images, and they minimize an energy function equivalent to MAP estimation [2, 3] or consider it as a max-flow min-cut optimization problem [4].

Recently, different techniques were applied in this field. Region correspondence based approaches [12, 13] first divide images into small parts such as super-pixels, and match similar part between two images. An interesting approach [14] would first partition super-pixel and extract features from both images, and use SVM classifier to determine they are either foreground object or background. However, these methods do not explicitly consider the objectness of the segment as a whole.

Existing methods can also be categorized by methods for features extracted to measure similarity. Earlier method simply consider color histograms [2], while recent approaches take various features such as SIFT [15] or Gabor [4] into consideration. And for us, our feature to extract and our similarity measurement are rely on Siamese network.

2.3 neural network for image segmentation

Here we address about neural network approaches related to segmentation methods. SegNet [16] is a convolutional neural network based approach for semantic segmentation, i.e., segment images into several regions with sematic meanings. The structure is an encoder-decoder network, which would first encode the input image into small size high-level feature, and thereafter decode it into label map. Training data it used is the ground truth segmentation mask.

On the other hand, neural decision forest (NDF) [17] could be applied on semantic labeling as well. NDF is a special network cast original random forest structure into a neural network one. Image area with random size is consider as input data to predict pixel-wise segmentation label.

With proper handling, these works have potential to replace methods we use to generate segmentation proposals. In that case, the whole framework would become an end-to end neural network structure and could be optimized simultaneously. However, this is only a possible development direction.

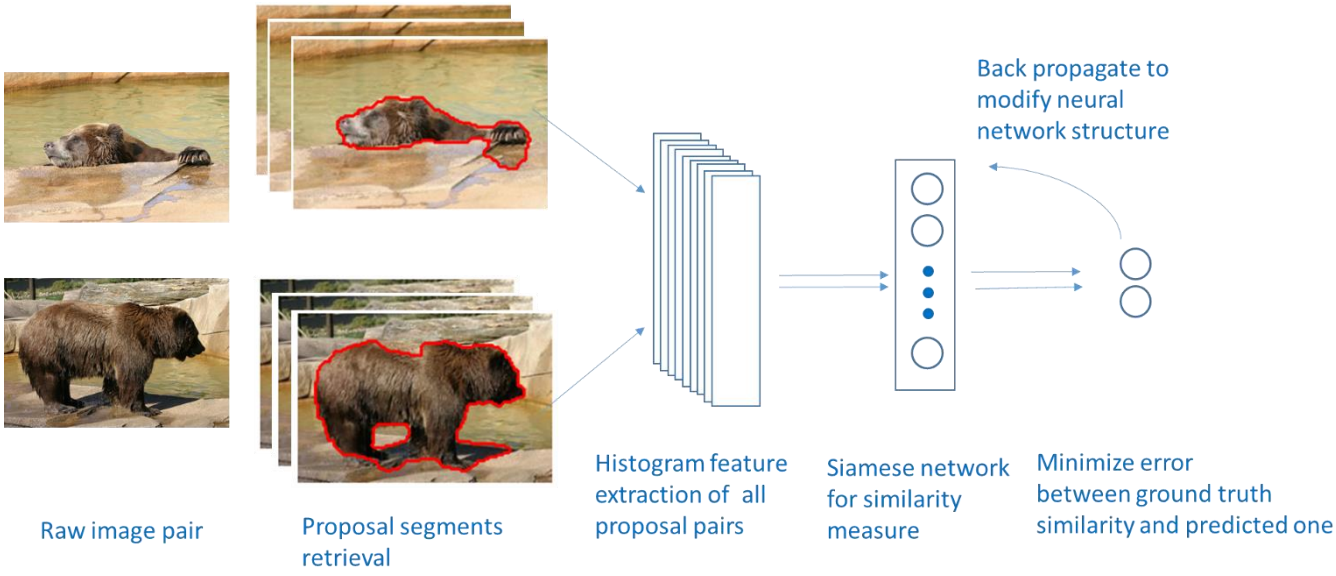


Fig. 2: Our proposed framework.

3. OBJECT COSEGMENTATION FRAMEWORK

Our framework is extended from the one proposed in [1]. As shown in Fig. 2, the whole framework can be roughly divided into three stages: ***proposals segmentation retrieval, feature extraction,*** and ***similarity measurement***.

To co-segment identical parts between two images, first we need to segment possible objects from background. These possible objects are known as ***proposals***. To compare whether these proposals share similar parts, we then extract a fixed-sized feature from each proposal. Thereafter, we train a Siamese network model to measure the similarity between these proposals with ground-truth co-segments. After the model is trained, we can use it to evaluate whether proposals of two images contain the same objects. Finally, we choose the proposal pairs most likely contain the same object as the co-segments between to images.

3.1 Proposals segmentation retrieval

As in [1], we start by extracting proposal segmentations based on the work of Carreira et al. [10], as shown in Fig. 3. Numerous object segmentation proposals would be generated by min-cut algorithm.

Thereafter, features would be extracted from each proposal to judge the ***objectness score*** *O* of these proposals, i.e., the score describing whether proposals cover objects accurately. Properties for these features including Graph partition properties (i.e., sum of affinity along the segment boundary), Region properties (e.g., area, perimeter, region centroid), and Gestalt properties (e.g., inter/intra-region texton similarity, brightness similarity and contour energy) Further details of these properties are stated in [10].

Since similar proposals would have similar ranks, top ranking proposals would have large overlapping area to each other, which would make selected proposals very similar. Thus, Maximal Marginal Relevance (MMR) measure is applied to diversify the top ranking proposals.

3.2 Feature extraction

To evaluate whether proposal segments from two images are similar, we need to generate feature describing proposal pairs. Since one of our goal is to integrate co-segmentation task with neural network structure, in the ideal case, the feature representation should be automatically generated during the training process of neural network model, without any manual feature involved in. The most likely solution is to train a Convolution Neural Network (CNN) by raw images and black-white masks of each proposal. However, training a CNN often requires a huge amount of training data. Due to the lack of datasets of object co-segmentation tasks, we still first extract histogram features to describe each proposal.

Fig. 3: Each proposal (other than first image, which is raw image) corresponds to a binary segmentation to segment an object from an image.

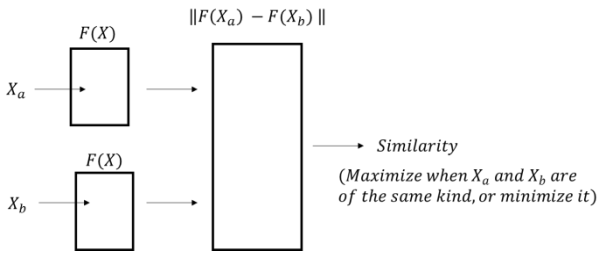


Fig. 4 Structure of standard Siamese network.

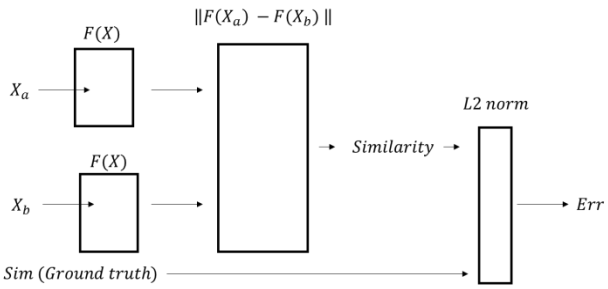


Fig. 5 Structure of Regression Siamese network.

For simplicity, we directly use the histogram feature used mentioned in section 3.1 in to evaluate objectness score of proposals, since its property can also represent the traits of each segmentation.

It is worth-noting that, in [1], beside histogram features retrieved to describe each proposal independently, it also extracts features including both images. However, since every different pairs of image would derive distinct features, it is quite inefficient if we want to extract co-segments from a bunch of images. For example, the image retrieval tasks is to extract the image in dataset that are the most similar to query image, and to extract feature between query and all image in dataset is unrealistic.

On the other hand, we rely on the capability of neural network structure to correctly measure similarity between independent histogram features, without additionally extract common features.

3.3 Similarity measurement

First we describe the definition of ground truth similarity between proposals of two images. For images belonging to different categories, i.e., one contains a bear while one is for a car, they have 0 similarity. On the other hand, the similarity between proposals of images of the same category is based on its overlap with ground truth segmentation. To be more formal, the overlap ratio *Ol* of a proposal P is given by:

$$Ol(P) = (P \cap GT)/(P \cup GT) \tag{1}$$

where *GT* indicates the area of ground truth.

And the similarity of two proposals of same category is the **average overlap ratio:**

$$Y = \begin{matrix} (Ol(P_a) + Ol(P_b))/2 & if\ c_a = c_b, \\ 0 & else \end{matrix} \tag{2}$$

The denoted *Y* is the ground truth similarity between two proposals, a,b are two images, and c is the category of an image. Our goal is to seek a method to correctly measure this similarity without provided ground truth segmentation.

In the next section, we'll further introduce the concept of standard Siamese network, and how do we modify it into regression Siamese network to fulfill the task of object co-segmentation.

4. REGRESSION SIAMESE NETWORK

4.1 Neural network

For the sake of completeness, we first address about the general concept of neural network. Neural network is a machine learning structure for classification and regression tasks. Given a sufficient amount of input instance vectors *X* and ground truth corresponded value *Y*, it could updates its parameter to predict corresponded value of unseen input instances.

The structure of neural network is composed of several layers, and a layer typically contains a weight matrix and an activation function, e.g., hyperbolic tangent function or rectified Linear unit function. When passed through a layer, input instance vector of a layer would be multiplied by weight matrix and non-linearly transformed by activation function. The output would become the input of next layer.

Sequentially passing forward, the output of final layer is $\tilde{Y}$, i.e., the prediction of corresponded value *Y*. The difference between *Y* and $\tilde{Y}$ is the error, and would be back-propagated to update weight matrixes in previous layers.

4.2 Standard Siamese network

Next, we briefly introduce the concept of Siamese network. Proposed in [6], Siamese network is a symmetric architecture for verification tasks such as Face Verification [7, 8].

The structure of a Siamese network is illustrated in Fig. 4 below. The inputs of a Siamese network are a pair of instances: X_a and X_b. Both instances would pass the identical neural network layers *F*. The goal is to minimize distance D_W between $F(X_a)$ and $F(X_b)$ if them are from the same category, while maximize it otherwise. Thus, the loss function to minimize is defined as:

$$\mathrm{L}(\mathrm{W}, \mathrm{Y}, X_a, X_b) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{\max(0, m - D_W)\}, \quad (3)$$

where Y is the ground truth similarity. In this case, it is a binary value (0 or 1) indicating whether X_a and X_b are of the same kind. Besides, m is a threshold, and distance D_W is the L1-loss between two instances, i.e.,

$$D_W = \|F_W(X_a) - F_W(X_b)\|. \quad (4)$$

However, this structure only support the verification task, i.e., either $Y = 0$ or 1. In our object co-segmentation tasks, two proposals has a non-binary similarity score between then, i.e., Y is a similarity score ranges from 0 to 1. If we simply set a threshold to judge two proposals are of same or not same kind, it would loss lots of information and can hardly tell the most similar proposal pairs.

4.3 Regression Siamese network

To address this problem, we change the formulation of loss function into

$$L = \frac{1}{2}Y * \left(Y - \tilde{Y}\right)^2, \quad (5)$$

where

$$\tilde{Y} = 1 - \tanh(\|F_W(X_a) - F_W(X_b)\|). \quad (6)$$

where Y is the ground truth similarity between X_a and X_b; $\tilde{Y}$ is the similarity measure between F(X_a) and F(X_b), and its value is bounded between 0 and 1. The new loss function aims at minimizing difference between Y and $\tilde{Y}$, i.e., a ***regression*** function. Thus, we call our performed structure as ***regression Siamese network***. Besides, since our goal is to predict the most similar proposals, we enhance the importance of instance pairs with higher similarity by multiplying the loss function by Y.

Next, we address about the neural network layers F. As mention in section 4.1, F is composed of a series of layers with activation function, and here we use reLu (rectified Linear unit) as activate function. To limit output of reLu in a proper value range, we normalized the result of reLu to 0~1. This could feasibly avoid the situation that the output response of reLu being so large that making the tanh function in (6) saturate.

4.4 Optimization

To minimize the loss function, we apply backpropagation technique to update the weights in Siamese network. We rely on TensorFlow to automatically perform the backpropagation process, and thus would not state details here.

(a) (b) (c) (d)

Fig. 6: Co-segments with high similarity but low objectness. (a)(b) raw image pairs (c)(d) failed co-segmentation

4.5 Combined with objectness score

The Siamese network can effectively measure the similarity between two proposals. However, we found that even we have already select top-ranking proposals with highest objectness scores, we still need to refer them when decide the best matched proposals. It is because meaningless partition of background sometimes still has moderate objectness score, and these parts are likely similar to each other. This phenomena is demonstrated in Fig. 6, where (a)(b) are our raw image pair, and (c)(d) are the most similar segmentation proposal. We must admit that they share the same physical meaning: they are both the partition of ground between legs of baseball player. While the baseball player itself is our target segment.

Thus, we combined predicted similarity $\tilde{Y}$ and objectness score O together to evaluate the best proposal as our co-segment. The combination is simply $O + \beta\tilde{Y}$, where β is a trade-off ratio smaller then 1. The reason why objectness score is more important is that half of proposals are actually meaningless background, i.e., if we retrieve 20 proposals from image, there would be about 10 proposals are purely segments of background. We would see the experiment in section 5.3. Another way to see this combination is that the best co-segment should take two models into consideration: one is used in [10] for measure objectness, and one is proposed here to measure similarity.

5. EXPERIMENT

5.1 Dataset and setting

The dataset we use is the CMU-Cornell iCoseg dataset dataset [18], which contains 37 categories of objects with ground-truth segmentation, and each category has 5~25 images. Examples of the dataset are shown in Fig. 7.

For experiment, we randomly consider four categories including baseball (25 images), bear (19 images), Ferarri (11 images), and gymnastics (16 images) to extract feature. We use the 5 images per

category for testing and use the rest for training, and for each image, we select top-20 rank proposals of it.

While the amount of data seems to be scarce at first glance, note that data input our regression Siamese network are a pair of proposal features. Take baseball category as example, 20 image are used for training, and thus it have 190 possible image pairs. Each image would be extracted 20 proposals, and thus each image pair contains 400 proposal pairs. That is, for baseball category, we have 190x400 proposal pairs for training, which is sufficient to train a neural network structure.

The feature extracted as mentioned in section 2.2 has dimension 4096. The source code of [10] has been released on the internet, and thus we could guarantee the quality of our generated object proposals.

For the *F* in Siamese network structure, we consider 5-layer reLu network. The dimensions of each layer is (input dim) – 1024 – 256 – 64 – 16. For implementation, we use TensorFlow, an open source python library for neural network.

5.2 Experiment result

First, we show some successful co-segmentation result. As shown in Fig. 8, even though the objects varies from view angle and posing, our overall framework can still achieve a descent performance. In fact, the performance is bounded by how good the retrieval of proposals segmentation is. If our framework works with more precise proposal retrieval algorithm, the result can be even better.

To quantitatively evaluate the performance of our framework, we consider the average overlap ratio as defined in (2), which is also the similarity between the best matched proposals from the pair of images. As mentioned in Sec. 5.1, we have 5 images per category. For object co-segmentation, we want to extract the most similar object proposal from a pair of image belongs to the same category. In other words, for a pair of image, we want to find out their most similar proposals. Since we have 20 proposals from each image, a pair of image

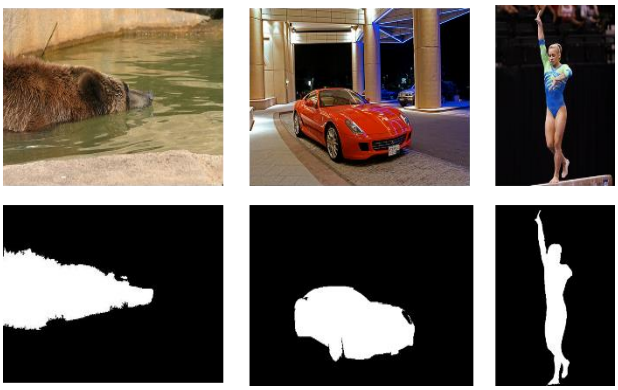


Fig. 7: Examples of iCoseg dataset.
Top Row: raw images.
Bottom row: ground truth segmentation masks

have totally 400 proposal pairs. Among all these proposal pairs, we select the best matched one and evaluate their average overlap ratio.

For comparison, we consider the average overlap ratio with single image segmentation only as baseline. After retrieving proposals segmentation as in Sec. 3.1, we could obtain the objectness score of each proposals. Select the proposal with highest objectness score and we can get the single image segmentation, and we can also calculate average overlap ratio of two segmentations.

As shown in table 1, the ratio is improved in all category about 5% in average. This demonstrate our proposed framework is indeed helpful for co-segmentation task. The effect is illustrated in Fig. 9, and take the baseball category as example: with single image segmentation, the court is considered as the common object, while the baseball player is corrected segmented with our co-segmentation algorithm. Note that we didn't compare our result to [10] because our setting for dataset is different, and we do not use feature including both image, which would likely raise performance but is not efficient as mentioned in section 3.2.

Fig. 8: Some good co-segmentation result of our proposed framework. Number is their average overlap ratio.

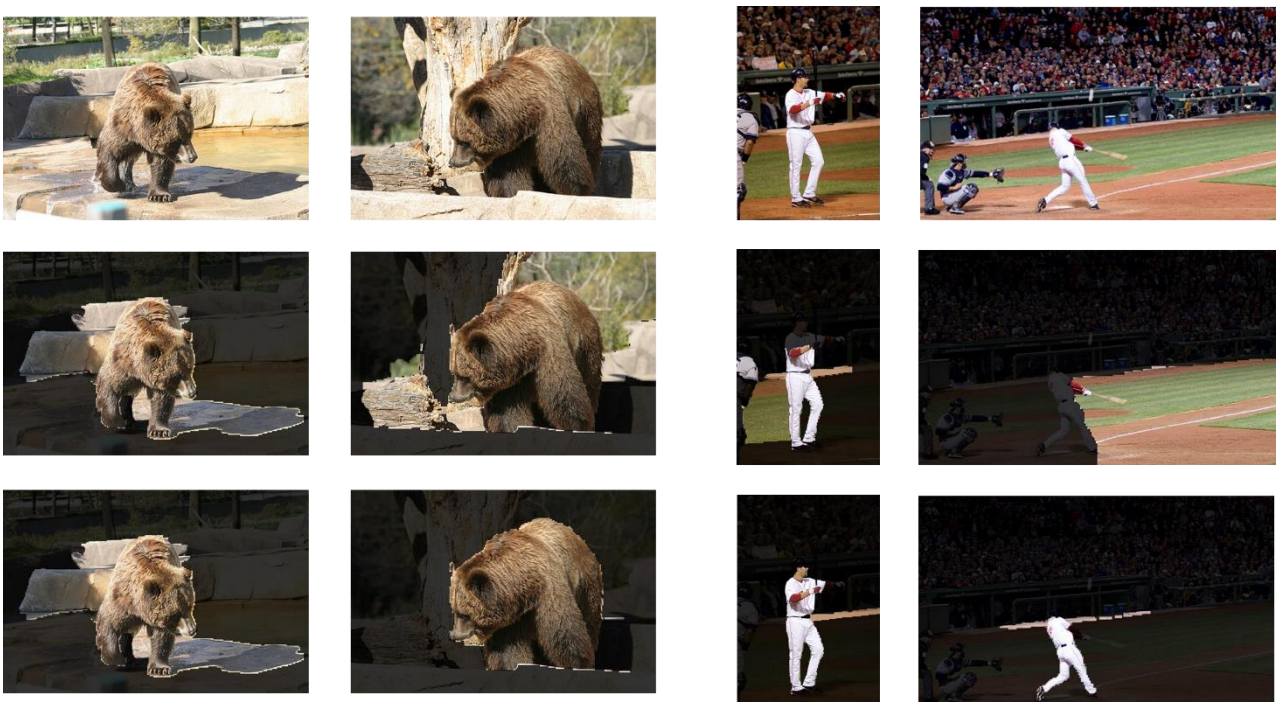


Fig. 9: The comparison between segmentation result between single image segmentation and proposed co-segmentation algorithm. Top row: raw image pairs. Middle row: result of single image segmentation. Bottom row: result of co-segmentation

Table 1. The average overlap ratio (%) of test image pairs from different categories: the higher the better.

	(a) single image segmentation	(b) proposed object co-segmentation
Baseball	46.11	53.89
Bear	51.57	60.20
Ferarri	65.12	66.19
Gymnastic	70.94	73.04
Average	58.43	63.33

Furthermore, we show the spatial relationship of proposals in high dimensional feature space in Fig. 10. Brighter color indicates higher overlapping ratio of the proposal, and similarity of two proposal is the average of their overlapping ratio. Thus, two points with lighter colors should be close to each other on the feature space. It is shown that, in the raw histogram feature space, lighter proposal points are scattered everywhere. After passing our regression Siamese network, they are projected onto the center of feature space.

5.3 Parameter analysis

Here we discuss about the effect of beta, the trade of ratio between predicted average overlap ratio $\tilde{Y}$ and objectness score O, as mentioned in section 4.5. In Fig. 11, it is show that we achieve the best performance when β is roughly 0.1. However, as β increase, the average overlap ratio tends to decrease and become even worse than single image segmentation. As

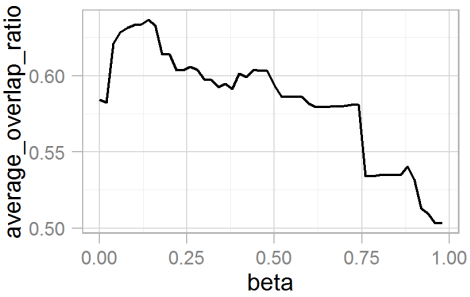
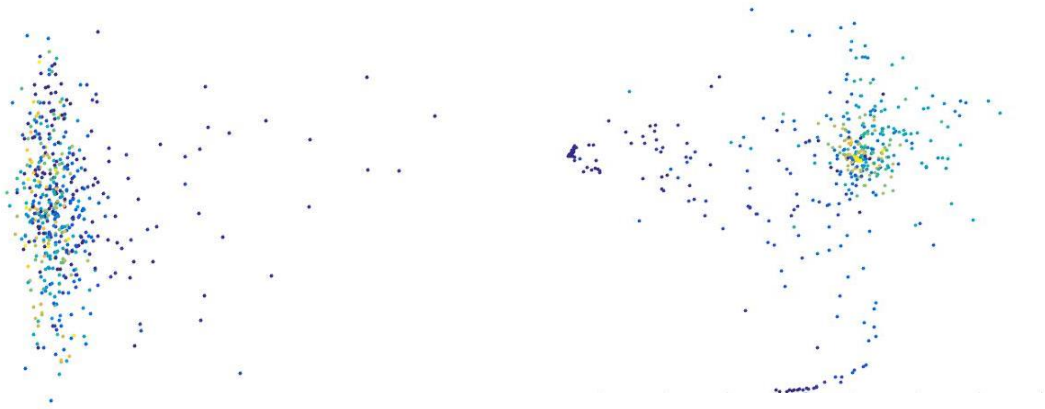


Fig. 11: Average overlap ratio v.s. β.

discussed in section 4.5, it is because the effect of background segments with small objectness take a large proportion in proposals. This, β should be smaller to avoid ruin the overall performance.

6. CONCLUSION

In this work, we proposed Regression Siamese Network to deal with object co-segmentation tasks. Unlike previous work as [1], our work does not extract features including both image, which is very ineffective when retrieving images with similar segmentation from a database. Experiment results of our work is stated in section 5, though the improvement is limited, it is still helpful on object co-segmentation task. Our next goal is to directly extract proposal segments with a CNN structure and integrate it with our Siamese network, and generate an end-to-end co-segmentation network.



(a) (b)

Fig. 10: The visualization of proposals of images from the gymnastic category (a) with raw histogram features (b) After passing Siamese network. Two axis are the first two dimensions of principle components. Brighter color indicates higher overlapping ratio of the proposal.

REFERENCES

[1] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov, "Object cosegmentation," in CVPR, 2011.

[2] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs," in CVPR, 2006.

[3] Lopamudra Mukherjee, Vikas Singh, and Charles R Dyer, "Half-integrality based algorithms for cosegmentation of images," in CVPR, 2009.

[4] Dorit S Hochbaum and Vikas Singh, "An efficient algorithm for co-segmentation," in ICCV, 2009.

[5] Jane Bromley, James W Bentz, L´eon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard S¨ackinger, and Roopak Shah, "Signature verification using a siamese time delay neural network," 1993.

[6] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in CVPR, 2005.

[7] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in CVPR, 2014.

[8] Gregory Koch, Siamese neural networks for one-shot image recognition, Ph.D. thesis, University of Toronto, 2015.

[9] John L Muerle, Donald C Allen, et al., "Experimental evaluation of techniques for automatic segmentation of objects in a complex scene," Pictorial pattern recognition, vol. 1, pp. 3–13, 1968.

[10] Joao Carreira and Cristian Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in CVPR, 2010.

[11] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari, "What is an object?," in CVPR, 2010.

[12] Edward Kim, Hongsheng Li, and Xiaolei Huang, "A hierarchical image clustering cosegmentation framework," in CVPR, 2012.

[13] Jose C Rubio, Joan Serrat, Antonio L´opez, and Nikos Paragios, "Unsupervised co-segmentation through region matching," in CVPR, 2012.

[14] Yuning Chai, Victor S Lempitsky, and Andrew Zisserman, "Bicos: A bi-level co-segmentation method for image classification.," in ICCV, 2011.

[15] Lopamudra Mukherjee, Vikas Singh, and Jiming Peng, "Scale invariant cosegmentation for image groups," in CVPR, 2011.

[16] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," in arXiv preprint arXiv:1505.07293 (2015).

[17] Samuel Rota Bulo and Peter Kontschieder, "Neural decision forests for semantic image labelling," in CVPR, 2014.

[18] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in CVPR, 2010.