# CoSegNet: Image Co-segmentation Using a Conditional Siamese Convolutional Network

**Sayan Banerjee**[1*], **Avik Hati**[2], **Subhasis Chaudhuri**[1] and **Rajbabu Velmurugan**[1]
[1]Indian Institute of Technology Bombay, India
[2]Istituto Italiano di Tecnologia,Genova, Italy
sayan91.ban@gmail.com, avikhatiece@gmail.com, {sc, rajbabu}@ee.iitb.ac.in

## Abstract

The objective in image co-segmentation is to jointly segment unknown common objects from a given set of images. In this paper, we propose a novel deep convolution neural network based end-to-end co-segmentation model. It is composed of a metric learning and decision network leading to a novel conditional siamese encoder-decoder network for estimating a co-segmentation mask. The role of the metric learning network is to find an optimum latent feature space where objects of the same class are closer and that of different classes are separated by a certain margin. Depending on the extracted features, the decision network decides whether input images have common objects or not and the encoder-decoder network produces a co-segmentation mask accordingly. Key aspects of the architecture are as follows. First, it is completely class agnostic and does not require any semantic information. Second, in addition to producing masks, the decoder network also learns similarity across image pairs that improves co-segmentation significantly. Experimental results reflect an excellent performance of our method compared to state-of-the-art methods on challenging co-segmentation datasets.

## 1 Introduction

Image co-segmentation plays a significant role in computer vision since it identifies images that have common objects and jointly segments out those objects from them (Fig. 1). Further, co-segmentation has the potential to improve result of single image segmentation as it integrates information from a group of similar images [Rother *et al.*, 2006]. It can be used as a tool for image retrieval, annotation, object recognition and person re-identification.

Existing common challenges in the co-segmentation problem are to find suitable features when the appearance, shape and pose of foregrounds vary significantly, foreground has notable similarities with background, and images contain missing foregrounds. There has been a significant amount of work on unsupervised co-segmentation [Hati *et al.*, 2018; Joulin *et al.*, 2012; Vicente *et al.*, 2010]. But, in general, these methods are sensitive to the selected hand-crafted features and require exhaustive feature tuning. There also have been some semi-supervised approaches [Batra *et al.*, 2010; Wang and Shen, 2016] where users provide sparse foreground and background labels on training images as scribbles and the co-segmentation algorithm learns background and foreground appearance models from these images. To handle the aforementioned challenges, these methods have to iterate between learning appearance models and acquiring human inputs, which is difficult.


Figure 1: Illustration of co-segmentation of an image pair. Columns 1-4 show an input image pair, the corresponding ground truth mask pair, the estimated mask pair and the corresponding common foreground object obtained using the proposed model.

In order to solve those difficulties, we propose a CNN based simple end-to-end architecture for co-segmentation (Fig. 2). Our model utilizes a siamese convolution-deconvolution (encoder-decoder) network, which takes a pair of images as input, produces intermediate convolutional features using feature encoders and conditionally maps those features into corresponding co-segmentation masks using decoders. These masks are used to extract the common object (Fig. 1). We use a siamese metric learning network that learns an optimal latent feature space where objects belonging to the same class are closer and objects from different classes are well separated, without using semantic class labels. Furthermore, we use a decision network on top of the metric learning network to produce a binary label, indicating the presence or absence of a common object in the input image pair. Together these two networks condition the siamese encoder-decoder network to perform an accurate co-segmentation depending on the presence of a common object in the input image pair.

During training, the binary ground truth masks of the train-

*We acknowledge the support provided by Bharti Centre for Communications, Department of Electrical Engineering, IIT Bombay. Sayan Banerjee is the corresponding author.

ing images guide the encoder-decoder network to differentiate common objects from the background based on the learned features. The metric learning part guides the encoder-decoder network to reduce intra-class object distance and increase inter-class object distance. As identifying outlier images is very difficult during co-segmentation, we propose a novel training strategy. In the case of positive samples (an image pair with a common object), we train the whole network. However for negative samples (an image pair with no common object), we only train the complete metric learning, decision, encoder and certain part of decoder network (please see Sec. 3.5 for more detail). During testing, the decision network predicts the presence or absence of a common object in them and accordingly the siamese encoder-decoder network estimates the corresponding co-segmentation masks.

The main contributions of this paper are:

- We propose a twofold siamese network architecture that is class agnostic.
- One of them being a novel conditional siamese encoder-decoder network and the other being a siamese metric learning network followed by a decision network, which takes input from intermediate layers of decoder network, as opposed to conventional approaches that use the encoder output as input. This design helps to generate better co-segmentation masks through image similarity.
- We propose a novel training strategy which helps the network to discard outlier images and boost performance with low amount of training data.

We perform extensive experiments and ablation studies on various challenging co-segmentation datasets such as the PASCAL-VOC dataset [Everingham *et al.*, 2010], the Internet dataset [Rubinstein *et al.*, 2013], the MSRC dataset [Rubinstein *et al.*, 2013], and demonstrate significant improvement in performance over state-of-the-art methods.

## 2 Related Work

The unsupervised object co-segmentation problem was introduced by [Rother *et al.*, 2006] where they provide a solution for handling two images with a Markov random field (MRF) based generative. [Joulin *et al.*, 2010] solved co-segmentation using clustering by finding discriminative object features, whereas, the method in [Vicente *et al.*, 2010] uses Boykov-Jolly model and a dual decomposition technique based optimization. The method in [Wang *et al.*, 2013] represents each image in a linear functional space and learns a linear transformation between different functional space to perform co-segmentation. Many researchers have also been trying to solve the co-segmentation problem with different degrees of supervision. [Rubio *et al.*, 2012] trained SVM classifier on top of a Gaussian Mixture Model to find out correspondence between different regions of input images. The method in [Rubinstein *et al.*, 2013] uses dense correspondences and saliency. [Hsu *et al.*, 2018] also used saliency along with a pretrained convolution network in an unsupervised manner to solve co-segmentation. [Chen *et al.*, 2014] used segmentation prior (seed) with learned detector. The method in [Quan *et al.*, 2016] uses CNN features obtained from VGG net and handcrafted features for superpixels and subsequently implements two separate graph-cuts for background and foreground superpixels. The methods in [Batra *et al.*, 2011; Batra *et al.*, 2010; Dong *et al.*, 2015; Wang and Shen, 2016] propose semi-supervised solutions using sparse scribbles. [Yuan *et al.*, 2017] proposed a deep neural network based end-to-end co-segmentation model where they extracted a set of object proposals and applied deep conditional random field to find co-occurring objects. Subsequently, they segmented objects from these proposals independently. Therefore, their method is not robust against outliers. A very recent CNN based deep co-segmentation network presented in [Chen *et al.*, 2018] additionally used attention for solving the problem. The work in [Li *et al.*, 2018] also uses a siamese CNN to estimate co-segmentation masks. Different from their architecture, our proposed siamese encoder-decoder network is conditioned by a siamese metric learning network and decoder network that learn similarity across common objects (discussed in detail in the next section) which helps us in obtaining significantly better co-segmentation performance.

## 3 Proposed Network Architecture

Given a pair of input images, our aim is to segment the common objects from them. It should be noted that our problem setting is quite challenging, since we assume that the images do not always contain common objects, and when there is a common object, their poses and appearances may vary significantly. Humans have the ability to identify objects of the same class even if they significantly differ in pose, shape and appearance. To capture this aspect, we incorporate a metric learning approach to learn a latent feature space, which ensures that objects from the same class will be projected very close to each other and that of different classes will be projected far apart at least by a margin irrespective of their shape, pose and appearance variations. Furthermore, the object should be distinguishable from its background. Let $\{I_i\}_{i=1}^2 \in \mathbb{R}^{q\times q}$ be a pair of input images and $\{M_i\}_{i=1}^2 \in \mathbb{R}^{q\times q}$ be the corresponding ground truth binary masks highlighting the regions of common object. Our objective in the proposed network is to estimate such masks $\{\hat{M}_i\}_{i=1}^2$ using a conditional siamese encoder-decoder network. The proposed network architecture is shown in Fig. 2(a). It is composed of three major components: a conditional siamese encoder-decoder network, a fully-connected siamese metric learning network and a decision network.

### 3.1 Conditional Siamese Encoder-Decoder Network

The proposed encoder-decoder network has a convolution feature encoder and co-segmentation mask decoder. The siamese encoder consists of two identical feature extraction CNNs with shared parameters and is built upon the VGG 16 architecture. The input image pair $I_1$ and $I_2$ ($224 \times 224$) is passed through the encoder network, which is composed of 13 convolutional layers (conv) and five max-pooling (MP) layers as shown in Fig. 2. The output of each encoders is a high level semantic feature map $f_1$ or $f_2$, having 512 channels with a spatial size of $7 \times 7$.

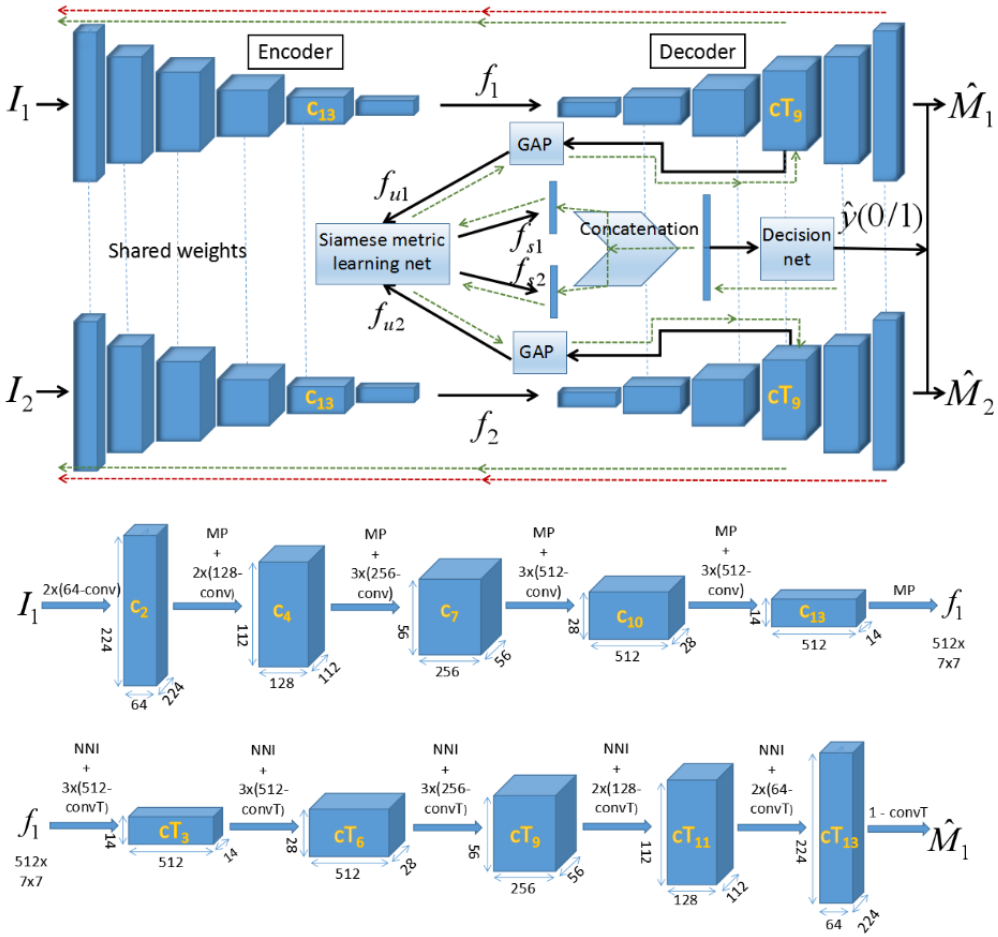


Figure 2: Illustration of the proposed deep convolution neural network architecture for co-segmentation. **[Top]**: An input image pair ($I_1$, $I_2$) is passed through a pair of encoder-decoder networks, with shared weights (indicated by vertical dotted lines). Output feature maps of the ninth decoder layer pair (cT$_9$) are vectorised ($f_{u1}, f_{u2}$) by Global Average Pooling (GAP) and fed to a siamese metric learning network. Its output vector pair ($f_{s1}, f_{s2}$) is then concatenated and fed to the final decision network, which is a multi-layer perceptron that produces a binary label ($\hat{y}$), one or zero, indicating presence or absence of any common object in the input image pair. **[Bottom]**: Details of the encoder-decoder network. 64-conv indicates convolution using 64 filters followed by ReLU. MP stands for max-pooling with a kernel of size $2 \times 2$. NNI stands for nearest neighbour interpolation. We perform deconvolution using convolution-transpose operation (convT). *Red* dotted line arrows show backpropagation for positive samples and *green* dotted line arrows show backpropagation for both positive and negative samples. The complete network is trained for positive samples. For negative samples, the decoder network after cT$_9$ is not trained.

The siamese decoder block that follows the encoder, performs the task of producing foreground masks of the common objects. It consists of two identical deconvolution networks. The input to this network is the semantic feature map pair $f_1$ and $f_2$ produced by the encoder. The decoder network is formed by five spatial interpolation layers with 13 transposed convolution layers (convT) as shown in Fig. 2 **[Bottom]**. In the encoder, we have used max-pooling as it makes the extracted features spatially invariant and contextual. Doing so, we loose the spatial resolution of the input images. The role of the decoder network is to transform these low resolution feature maps to co-segmentation masks with resolution equal to that of the input. To increase the size of the feature maps, we use nearest neighbour interpolation (NNI), which is fast compared to bi-linear or bi-cubic interpolation. However, this way of increasing size introduces blurring and spatial artifacts. Therefore after each NNI, transposed convolution has been performed to reduce these artifacts. All deconvolution or transposed convolution layers except the final layer are followed by a ReLU operation. The final deconvolution layer produces two single channel maps with size $224 \times 224$, which are converted to co-segmentation masks $\hat{M}_1$, $\hat{M}_2$ by sigmoid function. The output layer of this network (see Fig. 2) is gated by the binary output of the decision network (discussed in Sec. 3.3) to produce a conditional siamese convolutional network. The conditional parameter is used during the feed-forward and backpropagation stages, as detailed in Sec. 3.5.

### 3.2 Siamese Metric Learning Network

It consists of two fully connected layers with dimensions 128 and 64, respectively. The first layer has ReLU as non-linearity

and the second layer does not have any non-linearity. This network takes input $f_{u1}$ and $f_{u2}$ from the siamese decoder network and outputs a pair of feature vectors $f_{s1}$ and $f_{s2}$ that represent the objects in the learned latent space. Since the deconvolution layers at the middle of the decoder network capture sufficiently enough object information, we use the output of the ninth deconvolution layer ($56 \times 56$ and having 256 channels) as the input to this network. We use global average pooling (GAP) over each channel of the deconvolution layer to get two 256 dimensional vectors. To train the network, we use standard triplet loss Sec. 3.4. During backpropagation this also updates the nine decoder deconvolution layers and thirteen encoder convolution layers as shown in Fig. 2 which leads to better masks.

### 3.3 Decision Network

During testing, we are required to infer whether the input image pair contains common objects. We achieve this using a decision network that detects the presence or absence of common objects. This network is built upon the feature similarity network and has two fully connected layers with dimensions 32 and 1, respectively. The input to the network is a 128 dimensional vector obtained by concatenating $f_{s1}$ and $f_{s2}$. The first layer is associated with a ReLU non-linearity. The second layer is associated with a Sigmoid function that gives a probability signifying presence or absence of common object in the two input images. During testing, we threshold the probability and convert it to a binary label. For an input image pair if the decision network predicts a binary label one, we compute the corresponding binary masks at the output of the siamese decoder network.

### 3.4 Loss Function

The loss function used to train the proposed network is

$$L_{\text{final}} = w_1 L_1 + w_2 L_2 + w_3 L_3 \tag{1}$$

where $L_1$, $L_2$ and $L_3$ are the losses used for training the siamese encoder-decoder network, the siamese metric learning network and the decision network, respectively.

Given a set of positive and negative pair of images $\{(I_i^a, I_i^p), (I_i^a, I_i^n)\}$ (where $I_i^a$ is an anchor), corresponding pair of ground truth masks $\{(M_i^a, M_i^p), (M_i^a, M_i^n)\}$ and the predicted masks $\{(\hat{M}_i^a, \hat{M}_i^p), (\hat{M}_i^a, \hat{M}_i^n)\}$ obtained from the sigmoid layer of the decoder, we use pixel-wise binary cross entropy loss to train the encoder-decoder network as follows

$$L_1 = \sum_i \sum_{l \in [a,p,a,n]} \sum_{j=1}^{q} \sum_{k=1}^{q} M_i^l(j,k) \times \log(\hat{M}_i^l(j,k)) \tag{2}$$

where $\hat{M}(j,k)$ and $\hat{M}(j,k)$ is the value of the $(j,k)$-th pixel of the predicted and true mask, respectively. The loss of the metric learning network is the standard triplet loss given as

$$L_2 = \sum_i \max(0, \|f_u(I_i^a) - f_u(I_i^p)\|_2 - \|f_u(I_i^a) - f_u(I_i^n)\|_2 + \alpha) \tag{3}$$

$\alpha$ is a scalar valued margin. To train the decision network, we use binary cross entropy loss given as

$$L_3 = \sum_r y_r \log \hat{y}_r + (1 - y_r) \log(1 - \hat{y}_r) \tag{4}$$

$y_r = 1$ and 0 for a positive $(I_i^a, I_i^p)$ and negative pair $(I_i^a, I_i^n)$, respectively. And $\hat{y}_r$ is the predicted label obtained from it's final sigmoid layer.

### 3.5 Training Strategy

As mentioned earlier, at the time of training we backpropagate loss $L_1$ only for positive samples, and losses $L_2$ and $L_3$ for both positive and negative samples as shown in Fig. 2. An image belonging to a positive sample should produce an object mask. If the same image is part of a negative sample (outlier images), it is required to produce a null mask. Thus the decoder network is forced to produce two different masks for one image at different instances. For negative examples it is not required to produce any mask at all since the decision network notifies the presence of a common object. Hence, the role of the decision network is to reduce the overall difficulty level of the deconvolution layers by making them to produce object masks only for positive samples. It helps to train the network and also improves the performance as shown in Fig. 6. To summarize, we train the entire network for positive samples ($y_r = 1$) and a part of the network for negative samples ($y_r = 0$), thus making the mask estimation of the siamese network a conditioned one.

During testing, we obtain the co-segmentation masks by multiplying the output of the decoder network with the predcited label ($\hat{y}_r$) of the decision network because of the conditional network. Ideally, the output of the decision network for any positive example is one and that for any negative example is zero. Hence, for positive samples we obtain common object masks and for negative samples we obtain null masks, as desired. We show our experimental results in Fig. 6.

## 4 Experimental Results

We evaluate co-segmentation performance, using two common metrics. The first one is Precision, which is the percentage of correctly segmented pixels of both the foreground and the background. The second one is Jaccard Index, which is the intersection over union of the co-segmentation result and the ground truth common foreground segment. To evaluate the performance of the proposed co-segmentation algorithm on a set of $k$ input images $I_1, I_2, ..., I_k$, we perform co-segmentation on all such pairs and then report the average Precision and Jaccard index computed over all such pairs as the final co-segmentation accuracy for the given set. We evaluate the proposed method on three challenging co-segmentation datasets Pascal-VOC, Internet and MSRC and compare with state-of-the-art methods.

### 4.1 Implementation Details

We initialize our network with weights trained on Imagenet dataset. We use stochastic gradient descent as our optimizer and fix the learning rate and momentum at 0.00001 and 0.9, respectively for all three datasets. For Pascal-VOC and MSRC datasets, we set the weight decay to 0.0004 and for Internet we set it to 0.0005. At the time of training, we follow [Schroff *et al.*, 2015] for generating samples. For the case of positive samples, we set the weights $w_1 = w_2 = w_3 = 0.33$ in (1) and for negative samples we set $w_2 = w_3 = 0.5$

Figure 3: Visual results on the PASCAL-VOC dataset. Columns 1, 3 show input image pairs and Columns 2, 4 show the corresponding co-segmented objects obtained using the proposed method.

| Method | Precision (P) | Jaccard Index (J) |
|---|---|---|
| [Faktor and Irani, 2013] | 84.0 | 0.46 |
| *[Quan *et al.*, 2016] | 89.0 | 0.52 |
| [Wang *et al.*, 2017] | 84.3 | 0.52 |
| *[Hsu *et al.*, 2018] | 91.0 | 0.60 |
| Ours | **95.4** | **0.68** |

Table 1: Comparison of Precision (P) and Jaccard index (J) of the proposed method with state-of-the art methods on the PASCAL-VOC dataset (* denotes deep learning based methods).

and $w_1 = 0$ since $L_1$ is not backpropagated, as explained in Sec. 3.5. Due to memory constraint, we use a batch size of 3. Note that each input sample of the batch is a pair of input images, either positive or negative. We resize all the input images to $224 \times 224$ and set the margin $\alpha$ to 1.

### 4.2 PASCAL-VOC Dataset

The PASCAL-VOC dataset [Everingham *et al.*, 2010] has 20 classes and is one of the most challenging datasets due to the significant intra-class variations and presence of background clutter. We randomly split the dataset in the ratio of 3:1:1 for training, validation and testing sets. Since there is no standard split available, we repeat this splitting process 100 times and report the average performance computed over 100 such different test sets.

**Analysis**. In Table 1, we show comparative results of the proposed method with state-of-the-art methods and the proposed method significantly outperforms existing methods. The performance is improved by at least eight percent in terms of Jaccard Index and four percent in terms of precision. This can be explained by the fact that our model involves convolution-deconvolution with pooling operation, which involves a high degree of context for feature computation. Furthermore, the ground-truth mask helps to localize the foreground objects and with that the metric learning network learns a latent feature space where common objects come closer irrespective of their pose and appearance variations. Hence, the method becomes robust to pose and appearance changes.

### 4.3 Internet Dataset

The Internet dataset [Rubinstein *et al.*, 2013] has three classes: *Airplane*, *Car* and *Horse*. Each class also contains some outlier images. We use a subset of 100 images per class for our experiment as it is a common setting [Quan *et al.*, 2016; Li *et al.*, 2018; Hsu *et al.*, 2018; Yuan *et al.*, 2017].

**Analysis**. Comparative results of the proposed method with other state-of-the-art methods are shown in Table 2. It can be seen that the proposed method outperforms other deep learning based methods and unsupervised methods. The method in [Yuan *et al.*, 2017] is heavily dependent on extracted object proposals. Furthermore, after obtaining proposals that contain common objects, the segmentation of those proposals are done independently. Moreover, as observed in [Yuan *et al.*, 2017] itself the dense-CRF based model may worsen the co-segmentation result. Our method has some similarities with the method in [Li *et al.*, 2018], but our use of metric learning with a decision network as opposed to a mutual correlator as proposed in their work makes our model faster by 6 times per epoch, and this along with conditional siamese encoder-decoder increases co-segmentation performance by at least 6 to 8 percent from [Li *et al.*, 2018]. We show visual results of the proposed method in Fig. 4.

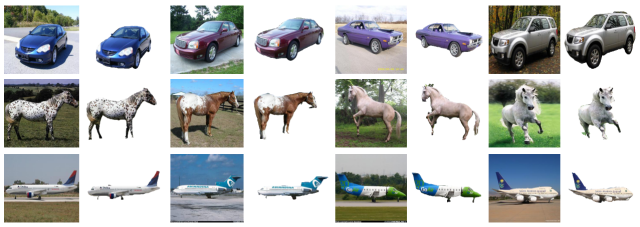

Figure 4: Visual results on Internet dataset. Columns 1,3 and 5,7 show two input image pairs and Columns 2,4 and 6,8 show the corresponding co-segmented objects.

### 4.4 MSRC Dataset

We evaluate our method on a subset of the MSRC dataset, which has been widely used by previous methods to evaluate co-segmentation performance. We select the same classes, which are *cow*, *plane*, *car*, *sheep*, *bird*, *cat* and *dog*, as reported by [Vicente *et al.*, 2010; Rubinstein *et al.*, 2013; Wang *et al.*, 2013]. Each class has 10 images and there is a single object in each image. The objects in each class have color, pose and scale differences. The experimental protocol and parameters are same as that of the Pascal-VOC dataset.

**Analysis**. Comparative results of the proposed method with other state-of-the-art methods are shown in Table 4. The model was trained on the Pascal-VOC dataset since the MSRC dataset does not have sufficient number of samples for training. Yet the proposed method outperforms other methods by at least five percent in terms of Jaccard index. In Fig. 5, we show visual results of the proposed method.

Figure 5: Visual results on MSRC dataset. Columns 1,3 and 5,7 show two input image pairs and Columns 2,4 and 6,8 show the corresponding co-segmented objects.

| Method | C (P) | C (J) | H (P) | H (J) | A (P) | A (J) | M (P) | M (J) |
|---|---|---|---|---|---|---|---|---|
| [Joulin *et al.*, 2010] | 59.2 | 0.37 | 64.2 | 0.30 | 47.5 | 0.15 | 57.0 | 0.28 |
| [Rubinstein *et al.*, 2013] | 85.4 | 0.64 | 82.8 | 0.32 | 88.0 | 0.56 | 82.7 | 0.51 |
| [Chen *et al.*, 2014] | 87.6 | 0.65 | 89.3 | 0.33 | 90.0 | 0.40 | 89.0 | 0.46 |
| *[Quan *et al.*, 2016] | 88.5 | 0.67 | 85.3 | 0.58 | 91.0 | 0.56 | 89.6 | 0.60 |
| *[Hsu *et al.*, 2018] | 93.0 | 0.82 | 89.7 | 0.67 | 94.2 | 0.61 | 92.3 | 0.70 |
| *[Yuan *et al.*, 2017] | 90.4 | 0.72 | 90.2 | 0.65 | 92.6 | 0.66 | 91.0 | 0.68 |
| *[Li *et al.*, 2018] | 94.0 | 0.83 | 91.4 | 0.65 | 94.6 | 0.64 | 93.3 | 0.70 |
| *[Chen *et al.*, 2018] | - | 0.80 | - | 0.71 | - | 0.71 | - | 0.73 |
| Ours | **95.2** | **0.87** | **96.2** | **0.72** | **96.7** | **0.71** | **96.1** | **0.77** |

Table 2: Comparison of Precision (P) and Jaccard index (J) of the proposed method with state-of-the art methods on the Internet dataset. C, H and A stands for *Car*, *Horse* and *Airplane* classes. M denotes Mean value.

| Method | C (P) | C (J) | H (P) | H (J) | A (P) | A (J) | M (P) | M (J) |
|---|---|---|---|---|---|---|---|---|
| Ours | 94.6 | 0.85 | 93.0 | 0.67 | 94.3 | 0.65 | 94.0 | 0.72 |

Table 3: Precision (P) and Jaccard index (J) of our proposed model trained with the Pascal-VOC, evaluated on the Internet dataset.

| Method | Precision | Jaccard Index |
|---|---|---|
| [Vicente *et al.*, 2010] | 90.0 | 0.71 |
| [Rubinstein *et al.*, 2013] | 92.2 | 0.75 |
| [Faktor and Irani, 2013] | 92.0 | 0.77 |
| [Wang *et al.*, 2013] | 92.2 | - |
| *[Li *et al.*, 2018] | 94.4 | 0.80 |
| *[Chen *et al.*, 2018] | 95.3 | 0.77 |
| Ours | **96.3** | **0.85** |

Table 4: Comparison of Precision and Jaccard index of the proposed method with state-of-the-art methods on the MSRC dataset.

### 4.5 Ablation Study

To show the role of our fully connected siamese metric learning network and decision network, we create a baseline model CoSegNet-base by removing the metric learning and decision networks from the proposed architecture. We concatenate features $f_1$ and $f_2$ along their channels to make a feature map with 1024 channels for image $I_1$ and feed it to the corresponding decoder network. The same is done for image $I_2$. It should be noted that, for our baseline model, we train the whole siamese encoder-decoder network for negative samples also using null mask. In Table 5, we compare the baseline model with our proposed model on different co-segmentation datasets. It can be seen from Table 5 that the proposed conditional siamese encoder-decoder network performs significantly better which justifies the inclusion of the fully connected siamese metric learning network and decision network with a novel training strategy. The advantage of the proposed CoSegNet architecture over the CoSegNet-base architecture is visually illustrated in Fig. 6. Different class objects in the image pair are incorrectly detected as common objects by CoSegNet-base, whereas CoSegNet correctly detects that there is no common object in the image pair. In Table 6, we compare the performance of our proposed model by feeding input features, obtained from different layers of the siamese deconvolution network, to the siamese metric learning network. In Table 6, $f_1$ and $f_2$ are the output feature maps of the siamese encoder network and $\text{cT}_3$, $\text{cT}_6$, $\text{cT}_9$ and $\text{cT}_{11}$ are the output feature maps of the third, sixth, ninth and eleventh deconvolution layers, respectively. The model performs the best for $\text{cT}_9$ because (a) sufficient object information has been fed to the input of the metric learning module, (b) the number of deconvolutional layers ($\text{cT}_{11}$ and $\text{cT}_{13}$) dedicated for producing co-segmentation masks is optimal.

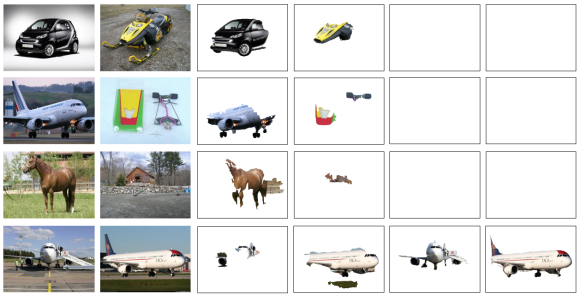

Figure 6: **[Ablation study]** Visual results on the Internet dataset. Columns 1, 2 show input image pairs, Columns 3, 4 show the objects obtained (incorrectly) using CoSegNet-base, and Columns 5, 6 show that the proposed method (CoSegNet) successfully performs co-segmentation on input image pairs even in the absence of common object (Rows 1-3), indicated by empty boxes.

| Dataset | Architecture | |
|---|---|---|
| | CoSegNet-base | CoSegNet |
| Pascal-VOC | 0.47 | 0.68 |
| Internet | 0.61 | 0.77 |
| MSRC | 0.63 | 0.85 |

Table 5: Comparison of Jaccard index (J) of the proposed model with the baseline model on different datasets.

| Dataset \ Layer | $f_1$,$f_2$ | $\text{cT}_3$ | $\text{cT}_6$ | $\text{cT}_9$ | $\text{cT}_{11}$ |
|---|---|---|---|---|---|
| Pascal-VOC | 0.63 | 0.65 | 0.66 | **0.68** | 0.64 |
| Internet | 0.67 | 0.68 | 0.68 | **0.72** | 0.65 |

Table 6: Comparison of Jaccard Index (J) of the proposed model for connecting the input of the metric learning network to different layers of the decoder network.

## 5 Conclusion

In this paper, we present a novel and efficient CNN-based architecture for solving image co-segmentation. Based on a conditional siamese encoder-decoder architecture, combined with a siamese metric learning and a decision network, we achieve better than state-of-the-art performances on various datasets, and demonstrate good generalization performance on segmenting objects of the same classes across different datasets, and robustness to outlier images.

## References


[Batra *et al.*, 2010] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3169–3176. IEEE, 2010.

[Batra *et al.*, 2011] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. Interactively co-segmentating topically related images with intelligent scribble guidance. *International Journal of Computer Vision*, 93(3):273–292, 2011.

[Chen *et al.*, 2014] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2035–2042. IEEE, 2014.

[Chen *et al.*, 2018] Hong Chen, Yifei Huang, and Hideki Nakayama. Semantic aware attention based deep object co-segmentation. *arXiv preprint arXiv:1810.06859*, 2018.

[Dong *et al.*, 2015] Xingping Dong, Jianbing Shen, Ling Shao, and Ming-Hsuan Yang. Interactive cosegmentation using global and local energy optimization. *IEEE Transactions on Image Processing*, 24(11):3966–3977, 2015.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[Faktor and Irani, 2013] Alon Faktor and Michal Irani. Co-segmentation by composition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1297–1304. IEEE, 2013.

[Hati *et al.*, 2018] Avik Hati, Subhasis Chaudhuri, and Rajbabu Velmurugan. Co-segmentation of non-homogeneous image sets. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 266–270, Oct 2018.

[Hsu *et al.*, 2018] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention CNNs for unsupervised object co-segmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI*, pages 748–756, 7 2018.

[Joulin *et al.*, 2010] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1943–1950. IEEE, 2010.

[Joulin *et al.*, 2012] Armand Joulin, Francis Bach, and Jean Ponce. Multi-class cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 542–549. IEEE, 2012.

[Li *et al.*, 2018] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. *arXiv preprint arXiv:1804.06423*, 2018.

[Quan *et al.*, 2016] Rong Quan, Junwei Han, Dingwen Zhang, and Feiping Nie. Object co-segmentation via graph optimized-flexible manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 687–695, 2016.

[Rother *et al.*, 2006] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 993–1000. IEEE, 2006.

[Rubinstein *et al.*, 2013] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1939–1946. IEEE, 2013.

[Rubio *et al.*, 2012] Jose C Rubio, Joan Serrat, Antonio López, and Nikos Paragios. Unsupervised co-segmentation through region matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 749–756. IEEE, 2012.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[Vicente *et al.*, 2010] Sara Vicente, Vladimir Kolmogorov, and Carsten Rother. Cosegmentation revisited: Models and optimization. In *European Conference on Computer Vision*, pages 465–479. Springer, 2010.

[Wang and Shen, 2016] Wenguan Wang and Jianbing Shen. Higher-order image co-segmentation. *IEEE Transactions on Multimedia*, 18(6):1011–1021, 2016.

[Wang *et al.*, 2013] Fan Wang, Qixing Huang, and Leonidas J Guibas. Image co-segmentation via consistent functional maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 849–856, 2013.

[Wang *et al.*, 2017] Chuan Wang, Hua Zhang, Liang Yang, Xiaochun Cao, and Hongkai Xiong. Multiple semantic matching on augmented $n$-partite graph for object co-segmentation. *IEEE Transactions on Image Processing*, 26(12):5825–5839, 2017.

[Yuan *et al.*, 2017] Zehuan Yuan, Tong Lu, and Yirui Wu. Deep-dense conditional random fields for object co-segmentation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3371–3377. AAAI Press, 2017.