Directed Variational Cross-encoder Network for Few-shot Multi-image Co-segmentation

Sayan Banerjee*, S Divakar Bhat*, Subhasis Chaudhuri, Rajbabu Velmurugan

Indian Institute of Technology Bombay

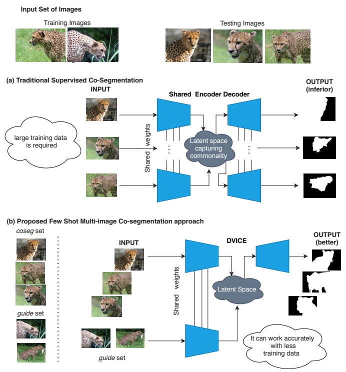
(* indicates equal contribution)

Problem Definition and Contribution

Goal: Multi-image co-segmentation using limited supervisory samples.

Motivations:

- Traditional supervised co-segmentation approaches require a large amount of annotated datasets.



Key Contributions:

- A novel multi-image co-segmentation framework capable of handling the **small sample size problem**.
- A novel encoder-decoder network to do explicit few-shot learning in co-segmentation task.

Problem Formulation

Main idea: Utilize a novel few-shot learning strategy to improve co-segmentation performance on the smaller target dataset $\mathcal{D}_{target}$.

- We develop an episodic training scheme, to handle the co-segmentation task with few training samples without overfitting.
- Each episode consists of a *guide* set and a *co-seg* set such that the set $\mathcal{G}$ provides the information of the common object to the *co-seg* set $\mathcal{C}$ over which co-segmentation is performed.
- Following operation removes the influence of outliers and determines robust features $\mathcal{O}^g$ of the common object.

$$\mathcal{O}^g = \frac{1}{|\mathcal{G}|}\sum_{i=1}^{k} \text{ChAM}(z_i^g). \tag{1}$$

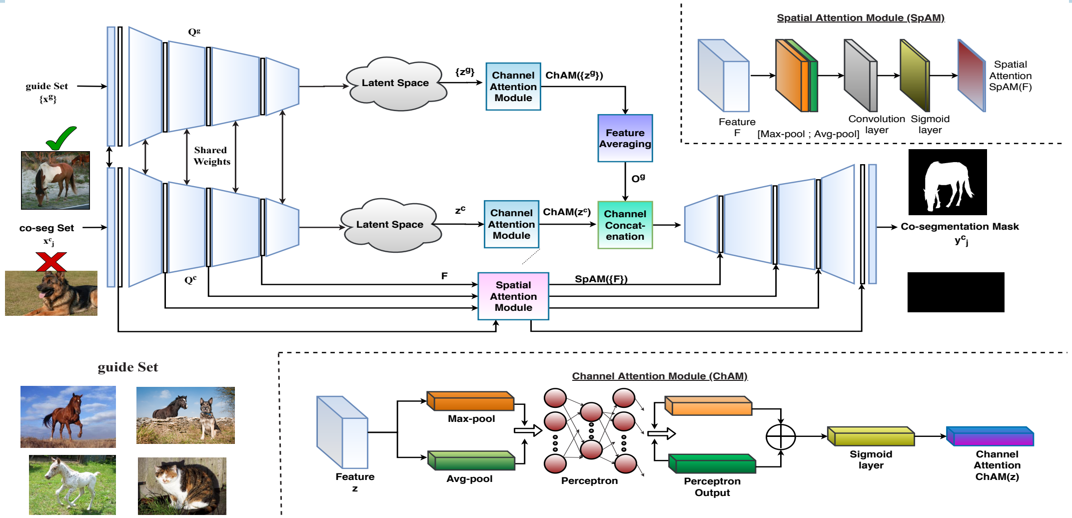
$\mathcal{O}^g$ is the common object prototype while z_j^c and z_j^g are the features obtained from the encoder for j^{th} image of the *co-seg* set and *guide* set, respectively.

- Feature of individual samples $x_j^c \in \mathcal{C}, j = 1 \ldots m$ is obtained as,

$$z_j^c = \text{ChAM}(E(x_j^c)). \tag{2}$$

- The proposed decoder implicitly checks the similarity between the $\mathcal{O}^g$ and z_j^c, and estimates co-segmentation

Method



Experiments & Results

Dataset:

- We consider Pascal-VOC as our $\mathcal{D}_{base}$.
- iCoseg dataset is a relatively smaller dataset which has 38 classes with 643 images. Some classes have less than 5 samples. Since, the number of labeled samples are small, we consider this dataset as one of our $\mathcal{D}_{target}$ dataset.
- Note that we are able to achieve fine control over the foreground extraction by varying the composition of majority samples as seen in the *guide* set 1 (where *pyramid* is the majority) and *guide* set 2 (where *horse* is the majority), the corresponding outputs obtained over the *coseg* set.

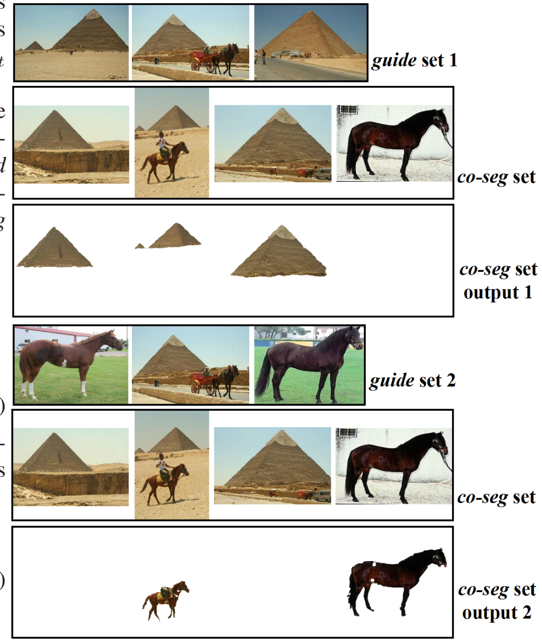
Loss function:

$$\begin{aligned}\log P(y^c, x^c) \geq\ & \mathbb{E}_{(\mathcal{O}^g,z^c)\sim Q(\mathcal{O}^g,z^c)}\left[\log P(y^c|\mathcal{O}^g, z^c)\right] \\ & - KL\left[Q(\mathcal{O}^g|\mathcal{G})||P(\mathcal{O}^g|\mathcal{G})\right] \\ & - KL\left[Q(z^c|x^c)||P(z^c|x^c)\right]\end{aligned} \tag{3}$$

- We derive an empirical loss ($\mathcal{L}$) from equation (3), calculated over the *co-seg* set, to train our model which is shown here,

$$\begin{aligned}\mathcal{L} = & -\sum_{j=1}^{m}\sum_{(a,b)} \log P(y_j^c(a,b)|\mathcal{O}^g, z_j^c) \\ & + KL\left[Q(\mathcal{O}^g|\mathcal{G})||P(\mathcal{O}^g|\mathcal{G})\right] \\ & + KL\left[Q(z^c|x^c)||P(z^c|x^c)\right]\end{aligned} \tag{4}$$

Visual results of the proposed method over one *co-seg* set but with different *guide* sets



Qualitative results on iCoseg dataset:

The first two rows depict the set of images used for co-segmentation (*co-seg* set) with their corresponding results to the immediate right of each image. The last row denotes the *guide* sets used to guide the network towards the desired foreground. The first three images correspond to the *guide* set for the first row, while the last three images from the last row correspond to the *guide* set of the second row of images. Note that the model is robust to the presence of outliers/noise in the *guide* sets as can be seen in the *guide* set corresponding to the *Panda*