One shot Joint Colocalization & Cosegmentation

Abhishek Sharma, Max Planck Institute for Informatics
asharma@mpi-inf.mpg.de

Abstract—This paper presents a novel framework in which image cosegmentation and colocalization are cast into a single optimization problem that integrates information from low level appearance cues with that of high level localization cues in a very weakly supervised manner. In contrast to multi-task learning paradigm that learns similar tasks using a shared representation, the proposed framework leverages two representations at different levels and simultaneously discriminates between foreground and background at the bounding box and superpixel level using discriminative clustering. We show empirically that constraining the two problems at different scales enables the transfer of semantic localization cues to improve cosegmentation output whereas local appearance based segmentation cues help colocalization. The unified framework outperforms strong baseline approaches, of learning the two problems separately, by a large margin on four benchmark datasets. Furthermore, it obtains competitive results compared to the state of the art for cosegmentation on two benchmark datasets and second best result for colocalization on Pascal VOC 2007.

Index Terms—Discriminative clustering, weak supervision, cosegmentation, colocalization, multi-task learning

1 INTRODUCTION

Localizing and segmenting objects in an image is a fundamental problem in computer vision since it facilitates many high level vision tasks such as object recognition, action recognition [39], natural language description of images [40] to name a few. Thus, any advancements in image segmentation and localization algorithm are automatically transferred to the performance of high level tasks [40].

With the recent success of deep networks, supervised top down segmentation methods obtain impressive performance [46] by learning on pixel level labelled datasets. The same is true for object detection [19]. However, the amount of annotations required to achieve pixel or bounding box level labelled datasets is tremendous [25]. Taking into account the cost of obtaining such annotations, recent work has explored the problem of weakly-supervised object discovery [10], [30], [36], [37]. The degree of supervision used in these problems varies from weak (positive and negative image-level labels for a target class [38]), very weak (image level labels e.g. colocalization [9], [21] and cosegmentation [2], [29], and null [30]. In this paper, we focus on colocalization and cosegmentation and use very weak supervision to imply that labels are given only at the image level.

Cosegmentation is the problem of segmenting common foreground regions out of a set of images whereas colocalization aims to localize the common object. Prior work in the supervised setting has used off-the-shelf object detectors to guide the segmentation process [4] and also used segmentation as an initial phase for detection. However, existing work for cosegmentation and colocalization either completely ignores these complimentary cues or use them in a two stage decision process, either as pre-processing step [28] or for post processing [31]. For example, Quan et al. [28] refines the coarse localization heat map obtained by a VGG network [32] to improve cosegmentation. However, it is difficult to recover from errors introduced in the initial stage and the post processing steps are prone to unwanted heuristics.

This paper advocates an alternative to the prevalent trends of either ignoring these complimentary cues or placing a clear separation between segmentation and localization. In the weakly supervised scenario, the goal of knowledge transfer between the two tasks becomes even more challenging. The key idea here is to avoid making hard decisions and instead, couple these two problems by linear constraints. We empirically show that constraining the two problems jointly improves the performance of both tasks significantly. Our work, although similar in spirit to the prior work that embeds pixels and parts in a graph [26], [27], builds on the discriminative framework of [1], [17] which utilizes a more powerful top down maximum margin machinery in an unsupervised fashion.

Contrary to the conventional approach of multi task learning, [41], [42] where two (or more) similar tasks are jointly learned using a shared representation, we instead leverage two representations at different scales and enable the transfer of information implicit in these representations during a one shot optimization scheme. More precisely, the proposed formulation exploits the semantic, localization cues of bounding boxes to guide cosegmentation and leverages low level segmentation appearance cues cues at superpixel level to improve colocalization.

Our contributions are as follows: 1) We propose a novel framework that simultaneously learns to localize and segment common objects in images. The unified framework obtains competitive results compared to the state of the art for cosegmentation on three benchmark datasets and second best result for colocalization on Pascal VOC 2007. 2) We show a novel mechanism to constrain the two problems via linear constraints in an unsupervised way that lifts the output performance of cosegmentation and colocalization by more than 10 points. 3) We provide an extensive evaluation of our approach that shows contribution of novel terms in the objective function and the effect of different cues on three benchmark datasets.
2 RELATED WORK

Supervised Setting. Numerous works have used off-the-shelf object detectors to guide segmentation process. Ladicky et al. [4] used object detections as higher order potentials in a CRF-based segmentation system by encouraging all pixels in the foreground of a detected object to share the same category label as that of the detection. Xu et al. [23] also used bounding box as a weak supervision for semantic segmentation. Vicente et al. [1] introduced the idea of using bounding box for cosegmentation in a supervised setting. Alternatively, segmentation cues have been used before to help detection [6], [31]. Parkhi et al. [6] uses color models from predefined rectangles on cat and dog faces to do GrabCut [44] and improve the predicted bounding box. Hariharan et al. [22] used CNN to simultaneously detect and segment by classifying image regions. All these approaches require ground truth annotation either in the form of bounding boxes or segmented objects during training phase which is challenging to obtain on large scale.

Weakly Supervised Setting. Rother et al. [7] first introduced the idea of cosegmentation in a relatively simple setting where the same object lies in front of different backgrounds in a pair of images. Since then, many work [2], [12], [14], [28], [33], [34] have been proposed to improve cosegmentation performance which can be broadly classified into discriminative and similarity based approaches. Similarity based approaches [7], [10], [11], [12] exploit the information of having common foreground across images and seek to segment it out by learning the foreground distribution or matching it across images [10], [29]. For example, Faktor & Irani [29] propose to discover the co-occurring regions first, and then perform cosegmentation by mapping between the co-occurring regions across the different images. In contrast, discriminative techniques [2], [3] mainly rely on separating a set of images into most separable clusters while taking care of local spatial consistency. For example, Joulin et al. [2] leverages discriminative clustering [1] to segment out the most discriminative parts in a set of images. However, most of these approaches are tailored only for cosegmentation task and do not use localization cues.

Colocalization is a similar problem [9] where the aim is to localize the common object, given a set of images. It was proposed under different names before. For example, object co-detection [5] is similar, but is given additional bounding boxes and correspondence annotations. Deselaers et al. [21] generated candidate bounding boxes and tried to select the correct box within each image using a conditional random field. Cho et al. [30], in contrast, localizes the common object by matching common object parts. However, all these approaches are designed for colocalization alone.

Our work is mainly inspired by the discriminative framework, proposed first for cosegmentation in Joulin et al. [2] and later extended for colocalization by Tang et al. [9] & Joulin et al. [45]. We first briefly explain the two main components of the discriminative framework of [2].

Discriminative clustering. Xu et al. [17] first proposed the idea of using supervised classifier such as SVM to perform unsupervised clustering. It formulates the clustering problem into the following optimization problem:

$$\min_{y \in \{0,1\}^n, \alpha \in \mathbb{R}^d} l(y, \mathcal{X} \alpha + b) + \beta \|\alpha\|^2,$$

where $\mathcal{X}$ is an $n \times d$ feature matrix (also known as design matrix), $l : \mathbb{R}^d \rightarrow \mathbb{R}$ is some loss function, and $\alpha$ a weight vector in $\mathbb{R}^d$ and scalar $b$ are the parameters of a linear classifier. When $l$ is the square loss, [1] shows that the problem is equivalent to

$$\min_{y \in \{0,1\}^n} y^T D y,$$

where

$$D = \Pi[\mathcal{I}_d - \mathcal{X} \mathcal{X}^T \Pi \mathcal{X} + \beta \mathcal{I}_d]^{-1} \mathcal{X}^T \Pi,$$

Note that $\mathcal{I}_d$ is an identity matrix of dimension $d$, $\Pi = \mathcal{I}_d - \frac{1}{n} \mathbf{1}\mathbf{1}^T$ is the usual centering projection matrix and $D$ is positive semi-definite. We refer to [1] for more details.

Local Spatial Similarity. To enforce spatial consistency, a similarity term is combined with the discriminative term $y^T D y$. The
The similarity term $y^T \mathcal{L} y$ is based on the idea of normalised cut [8] that encourages nearby superpixels with similar appearance to have the same label. Thus, a similarity matrix $\mathcal{W}^i$ is defined to represent local interactions between superpixels of same image. For any pair of $(a, b)$ of superpixels in image $i$ and for positions $p_a$ and color vectors $c_a$,

$$\mathcal{W}^i_{ab} = \exp(-\lambda_p \|p_a - p_b\|^2_2 - \lambda_c |c_a - c_b|^2)$$

The $\lambda_p$ is set empirically to .001 & $\lambda_c$ to .05. Normalised laplacian matrix is given by:

$$\mathcal{L} = \mathcal{I}_N - Q^{-1/2} \mathcal{W} Q^{-1/2}$$

(4)

where $\mathcal{I}_N$ is an identity matrix of dimension $d$, $Q$ is the corresponding diagonal degree matrix $I$.

Rest of the paper is organized as follows: Section 3 describes our novel joint framework. Section 4 gives the implementation details while section 5 evaluates it for the task of cosegmentation on three benchmark datasets. We then move on to colocalization experiments. Lastly, we conclude with discussions of empirical and qualitative results.

3 Joint Colocalization & Cosegmentation

**Notation.** We use italic Roman or Greek letters (e.g., $x$ or $\gamma$) for scalars, bold italic fonts (e.g., $\mathbf{y} = (y_1, \ldots, y_n)^T$) for vectors, and calligraphic ones (e.g., $\mathcal{C}$) for matrices. We assume we have $m$ bounding boxes per image.

3.1 Formulation for One Image

For the sake of simplicity and clarity, let us first consider a single image, and a set of $m$ bounding boxes per image, with a binary vector $z$ in $\{0, 1\}^m$ such that $z_i = 1$ when bounding box $i$ in $\{1, \ldots, m\}$ is in the foreground and $z_i = 0$ otherwise. We oversegment the image into $n$ superpixels and define the global superpixel binary vector $\mathbf{y}$ in $\{0, 1\}^n$ such that $y_j = 1$ when superpixel number $j$ in $\{1, \ldots, n\}$ is in the foreground and $y_j = 0$ otherwise. We also compute a normalized saliency map $M$ (with values in $[0, 1]$), and define: $s = -\log(M)$.

An Image as a collection of bounding boxes. We define our optimization problem (in particular, linear constraints) over bounding box and superpixel level. This requires an additional indexing of superpixels on bounding box level and thus, we maintain the following encoding of superpixels: for each bounding box, we maintain the set $S_i$ of its superpixels and define the corresponding indicator vector $x_i$ in $\{0, 1\}^{|S_i|}$ such that $x_{ij} = 1$ when superpixel $j$ of bounding box $i$ is in the foreground, and $x_{ij} = 0$ otherwise. Note that $x$ (indexing at bounding box level) and $y$ (indexing at image level) are related by a linear constraint. We define an indicator projection matrix $\mathcal{P}$ that encodes the occurrence of a superpixel in all bounding box by 1 and 0 as follows: for every box $i$, we define a matrix $\mathcal{P}_i$ of dimensions $|S_i| \times n$ such that $\mathcal{P}_{ij} = 1$ if superpixel $j$ is present in bounding box $i$ and 0 otherwise.

Optimization Problem. We propose to combine the objective function defined for cosegmentation and colocalization and thus, define:

$$E(\mathbf{y}, z) = \mathbf{y}^T (D_s + \alpha \mathcal{L}_s) \mathbf{y} + z^T D_b z + \nu y^T s_a + \mu z^T s_b,$$

(5)

The quadratic term $z^T D_b z$ penalizes the selection of bounding boxes whose features are not easily linearly separable from the other boxes. Similarly, minimizing $y^T D_s y$ encourages the most discriminative superpixels to be in the foreground. Minimizing the similarity term $y^T \mathcal{L}_s y$ encourages nearby similar superpixels to have same label whereas the linear terms $y^T s_a$ and $z^T s_b$ encourage selection of salient superpixels and bounding box respectively. Given the feature matrix for superpixels and bounding box, the matrix $D_s$ and $D_b$ are computed by Equation 3 whereas $\mathcal{L}_s$ is computed by Eq.4. We define the features and value of scalars later in the implementation detail.

We now impose appropriate constraints and define the optimization problem as follows:

$$\min_{\mathbf{y}, z} E(\mathbf{y}, z)$$

under the constraints:

$$\gamma |S_i| z_i \leq \sum_{j \in S_i} x_{ij} \leq (1 - \gamma) |S_i| z_i$$

for $i = 1, \ldots, m$, (6)

$$\sum_{i,j \in S_i} x_{ij} \leq \sum_{i,j \in S_i} z_{ij},$$

for $j = 1, \ldots, n$, (7)

$$\sum_{i} \mathcal{P}_{ij} \mathbf{y} = \mathbf{x}_i,$$

for $i = 1, \ldots, m$. (8)

$$\sum_{i=1}^{m} z_i = 1$$

(9)

The constraint (6) guarantees that when a bounding box is in the background, so are all its superpixels, and when it is in the foreground, a proportion of at least $\gamma$ and at most $(1-\gamma)$ of its superpixels are in the foreground as well, with $0 \leq \gamma \leq 1$. We set $\gamma$ to 0.1. The constraint (7) guarantees that a superpixel is in the foreground for only one box, the foreground box that contains it (only one of the variables $z_i$ in the summation can be equal to 1). For each bounding box $i$, the constraint (8) relates the two indexing of superpixels, $x$ and $y$, by a projection matrix $\mathcal{P}_i$ defined earlier. The constraint (9) guarantees that there is exactly one foreground box per image. We illustrate the above optimization problem by a toy example of 1 image and 2 bounding boxes in appendix at the end.

In equations (5)-(9), we obtain an integer quadratic program. Thus, we relax the boolean constraints, allowing $\mathbf{y}$ and $z$ to take any value between 0 and 1. The optimization problem becomes convex since all the matrix defined in equation(5)are positive semi-definite [2] and the constraints are linear. Given the solution to the quadratic program, we obtain the bounding box by choosing $z_i$ with highest value. For superpixels, since the value of $x$ (and thus $y$) are upper bounded by $z$, we first normalize $y$ and then, round the values to 0 (background) and 1 (foreground).

Why Joint Optimization. We briefly visit the intuition behind joint optimization. Note that the superpixel variables $x$ and $y$ are bounded by bounding box variable $z$ in Eq. 6 and 7. If the discriminative colocalization part considers some bounding box $z_i$ to be background and sets it to close to 0, this, in principle, enforces the cosegmentation part that superpixels in this bounding box are more likely to be background (= 0) as defined by the right hand side of Equation 6: $\sum_{j \in S_i} x_{ij} \leq \delta |S_i| z_i$. Similarly, the segmentation cues influence the final score of $z_i$ variable if the superpixels inside this bounding box are highly discriminative and more likely to be foreground.
4 Implementation Details

We will release source code of our implementation at the time of publication. We use superpixels obtained from publicly available implementation of [16]. This reduces the size of the matrix $D_s,C_s$ and allows us to optimize at superpixel level. Using the publicly available implementation of [20], we generate 20 bounding boxes for each image. We use unsupervised method of [13] to compute off the shelf saliency maps in our experiments.

Features. Following [2], we densely extract SIFT features at every 4 pixels and kernelize them using Chi-square distance. For each bounding box, we extract 4096 dimensional feature vector using AlexNet [24].

Hyperparameters Following [9], we set $\mu$, the balancing scalar for box saliency, to .001. To set $\alpha$, we follow [2] and set it $\alpha = .1$ for foreground objects with fairly uniform colors, and $\alpha = .001$ corresponding to objects with sharp color variations. We empirically set scalar $\nu = .005$ by optimizing over a small set.

5 Evaluation of Joint Framework

The goal of this section is two fold: First, we propose several baselines that help understand the individual contribution of various cues in the optimization problem defined in section 3.1. Second, we empirically validate and show that learning the two problems jointly significantly improve the performance over learning them individually.

5.1 Cosegmentation Experiments

5.1.1 Baseline Methods

In the section 3.1, we make the following two changes to the cosegmentation framework of [2]: First, we add the saliency cues as a linear term $s_p$ to the framework of [2]. Second, we propose to optimize the objective function of [2] with a quadratic program (QP) solver whereas in [2], it is optimized with a semi-definite programming (SDP) solver [18]. To illustrate the importance of saliency cues and better understand the different optimization techniques, we propose the following baseline methods:

B1. Discriminative clustering [1] objective is usually optimized with a SDP solver as the semi-definite relaxations are strong and do not suffer from trivial solutions. To validate this, we optimize the objective function of [2] with a quadratic program (QP) solver and compare the results with the SDP solver of [18].

B2. To quantify the impact of saliency cues, we propose to solve a linear program that obtains an image segmentation by finding the most salient pixels. Thus, we minimize a linear saliency term $s_A$ under a linear constraint that minimum number of foreground pixels should be greater than a fraction of total image pixels. This basically means choosing a fraction of the most salient pixels in an image. We set the fraction to $A$ as a rough measure of total foreground pixels in MSRC [15] and Object Discovery dataset [10].

B3. To illustrate the benefits of combining discriminative framework and saliency cues, we solve a QP that optimizes the new objective function of [2] that includes the saliency cues.

We denote the results obtained from our joint framework by Ours. In addition to the baselines proposed above and JBP10 [2], we compare our method with four state of the art approaches RJKL13 [10], WHG13 [14], FI13 [29] and QHZN16 [28]. Unless stated otherwise, we measure the segmentation accuracy as the percentage of pixels labeled accurately i.e. average precision (AP).

5.1.2 Benchmark Datasets.

We evaluate the cosegmentation performance of our framework on three benchmark datasets: MSRC [15], Object Discovery dataset [10] and PASCAL-VOC 2010. MSRC contains a subset of 10 object classes, each containing 24 to 30 images. The Object Discovery dataset [10] was collected by downloading images from Internet for airplane, car and horse. It is significantly larger and thus, diverse in terms of viewpoints, texture, color etc. Faktor & Irani [29] collected a subset of PASCAL-VOC 2010 dataset to evaluate the cosegmentation performance. This subset is obtained by choosing images in which the total size of a co-object is at least 1% of the image size. Overall, it contains 1037 images from the 20 PASCAL classes.

In Table 1, 2 and 3, we show our results and comparison with other approaches on these three datasets. Note that the results mentioned for JBP10 [2] are obtained by running their open source code and verified with the authors while for others, we simply cite their numbers from their paper. WHG13 [14] shows results on MSRC in two modes: supervised and unsupervised. We compare with unsupervised performance on six classes from their paper. For fair comparison with the state of the art [28] on Object Discovery dataset and PASCAL-VOC 2010, we report performance obtained by applying grab cut [44] based post processing on our output.

MSRC Dataset In Table 2, we observe that the B1 is consistently outperformed by the SDP solver of JBP10 [2] on both datasets by an average margin of 5 % AP. However, B3 consistently improves the performance of JBP10 [2] by an average of 5 % AP. This shows that the objective function of [2], combined with saliency

### Table 1: Comparison on Object Discovery dataset.

| Class   | JBP10 | B1  | B2  | B3  | Ours | RJKL13 | WHG13 |
|---------|-------|-----|-----|-----|------|--------|-------|
| Horse   | 69    | 64  | 70  | 72  | 87   | 82.8   | 89.3  |
| Plane   | 65    | 71  | 64  | 71  | 86   | 85.3   | 91.0  |
| Car     | 75    | 63  | 79  | 79  | 87   | 82.0   | 88.5  |
| Avg.    | 69.7  | 66  | 71.3| 74.0| 86.6 | 83.4   | 89.6  |

### Table 2: Comparison on MSRC dataset.

| Class   | JBP10 | B1  | B2  | B3  | Ours | RJKL13 | WHG13 |
|---------|-------|-----|-----|-----|------|--------|-------|
| Dog     | 73    | 62  | 72  | 75  | 87   | 92     | -     |
| Chair   | 74    | 69  | 73  | 78  | 84   | 84     | -     |
| Sheep   | 83    | 73  | 74  | 80  | 92   | 90     | -     |
| Bike    | 64    | 65  | 64  | 66  | 76   | 78     | 74.8  |
| Plane   | 53    | 73  | 65  | 74  | 84   | 82     | 87.3  |
| Cow     | 80    | 69  | 76  | 80  | 89   | 92     | 89.7  |
| Car     | 68    | 51  | 75  | 78  | 82   | 82     | 90.0  |
| Face    | 75    | 62  | 73  | 76  | 82   | 82     | 89.3  |
| Cat     | 70    | 65  | 72  | 76  | 84   | 90     | 88.3  |
| Bird    | 78    | 65  | 73  | 77  | 90   | 90     | -     |
| Avg.    | 71.1  | 65.4| 71.5| 75.9| 85.0 | 86.2   | 86.6  |
cues, can be optimized efficiently and accurately with a QP solver. Also, only saliency based segmentation, B2, gives a reasonable accuracy of 71% AP. Compared to JBP10 [2], our framework improves the average precision on MSRC dataset by almost 14%. Our results compete well with RJKL13 [10], on 6 out of 10 classes on MSRC dataset.

Experiments on Object Discovery Dataset In Table 1, we observe the same trend. We improve upon the result of JBP10 [2] by 14% and consistent gains over the baselines demonstrate the robustness of the model using localization cues. We outperform RJKL13 on all three classes. We compare well with QHZN16 [28] on classes Car and Horse but worse on airplane class.

Experiments on Pascal VOC 2010 As argued by Faktor & Irani [29], average precision metric is not reliable to evaluate cosegmentation algorithm on this subset as 90% of overall image content lies in background. Therefore, in addition to average precision(AP), we also evaluate our algorithm using Intersection over union (IoU) metric, also known as Jaccard similarity, in Table 3. We compare with two state of the art approaches that have shown results on this dataset yet: FI13 [29] and QHZN16. We outperform FI13 [29] in both metrics but perform slightly worse than QHZN16 [28].

TABLE 3: Comparison on Pascal VOC2010

| Metric   | Ours | FI13 | QHZN16 |
|----------|------|------|--------|
| Mean IoU | 47   | 46   | 52     |
| AP       | 86   | 84   | 89     |

Qualitative Results In Figure 2 and 3, we provide some examples of our end result. Figure 2 illustrates that our framework considers saliency as one of the various helpful cues and is robust to not salient objects or incorrect saliency maps. For example, in the very first example in Figure 2, the smaller cow is not at all salient and yet, our method rightly segments it as a foreground. In Figure 3, we show some examples from MSRC dataset. on top, we have an original image shown with the selected bounding box and underneath, we show the segmentation of the whole image.

5.2 Colocalization Experiments

Evaluation Metrics. We conduct colocalization experiments on PASCAL VOC 2007 [35]. We use two evaluation metrics to compare with state of the art colocalization techniques:

1) The standard Intersection over union (IoU) metric for object detection(intersection of predicted bounding box area and ground-truth bounding box area divided by the area of their union)

2) Correct Localization (CorLoc) metric, an evaluation metric used in related work [9], [30], and defined as the percentage of images correctly localized according to the criterion: \(IoU > .5\).

5.2.1 Baseline Methods

We analyze individual components of our colocalization model by removing various terms in the objective function and consider the following baselines:

Sal. This baseline only minimizes the saliency term for bounding boxes, without any segmentation cue, and picks the most salient one in each image. It is important as it gives an approximate idea about which object classes are more salient in the dataset.

Sal-Disc. This baseline includes the saliency and discriminative term for boxes, without any segmentation cues.

TJLF14 Tang et al.,TJLF14 [9] tackles colocalization alone without any segmentation spatial support. It quantifies how much we gain in colocalization performance by leveraging segmentation cues.
5.2.2 Colocalization evaluation on Pascal VOC 2007

Following the experimental setup defined in [9], [21], [30], we evaluate our method on the PASCAL07-6x2 subset to compare to previous methods for colocalization. This subset consists of all images from 6 classes (aeroplane, bicycle, boat, bus, horse, and motorbike) of the PASCAL VOC 2007 [35]. Each of the 12 class/viewpoint combinations contains between 21 and 50 images for a total of 463 images. Compared to the Object Discovery dataset, it is significantly more challenging due to considerable clutter, occlusion, and diverse viewpoints.

**CorLoc Metric** In Table 4, we report our experiments based on CorLoc Metric and compare them with our baselines proposed before. Note that we use plane instead of Aeroplane and bike instead of Motorbike. We see that results using stripped down versions of our model are not consistent and less reliable. In particular, we observe that on salient classes, saliency term alone, \( Sal \), is a very strong baseline. This can also be partially predicted by looking at the individual images of Bike class. However, it fails badly on classes such as Boat and Bus which are more cluttered, occluded and exhibit huge change in scale space. \( Sal + Disc \) improves upon the saliency baseline only in some classes such as Plane but overall fails to improve upon TJLF14 without segmentation cues. Our results improve upon both \( Sal \) and \( Sal + Disc \) in all classes. This again validates our hypothesis of leveraging segmentation cues to lift the colocalization performance. Our results outperforms TJLF14 [9] on most classes.

**IoU Metric.** In Figure 4, we show failure cases of our colocalization results based on CorLoc Metric. In the first row, we show several instances where the localization is near perfect and yet, the bounding box only achieves the CorLoc score of approximately .45, to .49 and thus, counts as a failure case according to CorLoc metric. This is mainly because it does not include the tail or a wing of aeroplane inside bounding box. We observe similar cases in class Horse too. To further support our argument quantitatively, we compare our results based on IoU metric with [30] on Pascal VOC 2007 in Table 5. We could not compare with TJLF14 on IoU metric as their source code and hyper-parameters are not publicly available. IoU metric gives a value of .45 to instances such as shown in Figure 3 whereas CorLoc gives it a score of 0.

**Comparison to state of the art.** Cho et al., CSP15 [30], outperforms all approaches by a huge margin on CorLoc metric where it obtains an absolute score of 64. This is partially because it leverages part based matching by Hough Transform where the predicted bounding box is selected by a heuristic standout score. In contrast, the discriminative framework of ours does not incorporate any constraints on including parts of objects in the predicted bounding box. This is also partially evident in Table 5 where the margin between our performance and CSP15 on IoU metric is almost half that of CorLoc. Moreover, CSP15, by design, is tailored for colocalization only whereas our framework tackles both colocalization and cosegmentation.

### TABLE 4: Comparison on Pascal VOC-2007

| Class       | \( Sal \) | \( Sal + Disc \) | TJLF14 | Ours |
|-------------|-----------|------------------|--------|------|
| Plane-Left  | 32        | 41               | 42     | 42   |
| Plane-Right | 20        | 43               | 51     | 59   |
| Boat-Left   | 02        | 06               | 11     | 16   |
| Boat-Right  | 09        | 09               | 12     | 21   |
| Bike-Left   | 48        | 30               | 51     | 41   |
| Bike-Right  | 47        | 41               | 65     | 56   |
| Horse-Left  | 31        | 25               | 44     | 38   |
| Horse-Right | 36        | 34               | 52     | 52   |
| Bus-Left    | 14        | 14               | 38     | 53   |
| Bus-Right   | 39        | 39               | 57     | 65   |
| Bicycle-Left| 25        | 27               | 25     | 35   |
| Bicycle-Right| 28       | 32               | 24     | 45   |
| Mean CorLoc | 28        | 29               | 39     | 44   |

### TABLE 5: IoU score on Pascal VOC2007

| Metric | Ours | CSP15 |
|--------|------|-------|
| Mean IoU | 45  | 56    |
projection matrix and \( x \) contains superpixel \( 1 \) respectively. And segment common objects or actions in images and videos providing a single framework that simultaneously classify, localize and segment objects. The proposed formulation is based on two

\[ \text{CONCLUSION & Future Work} \]

We proposed a novel framework that jointly learns to localize and segment objects. The proposed formulation is based on two different levels of visual representations and uses linear constraints as a means to transfer information implicit in these representations in an unsupervised manner. Although we demonstrate the effectiveness of our approach with a variant of maximum margin clustering, the key idea of transferring knowledge between tasks at different granularity is general and can be incorporated in the framework of constrained CNN \([43]\). Future work could also extend our model by including an image-video classifier, thereby providing a single framework that simultaneously classify, localize and segment common objects or actions in images and videos respectively.

\[ \text{Acknowledgement} \]

This work started as a Master’s thesis project at INRIA Willow team and was partially supported by ERC Advanced grants VideoWorld and Allegro. Many thanks to Armand Joulin for numerous helpful discussions and comments on this paper. The author also thanks anonymous reviewers for their comments.

\[ \text{Appendix} \]

We illustrate our joint colocalization and cosegmentation framework by a simple toy example. Suppose the image contains 5 superpixels. Thus, the global image level superpixel indexing is defined by \( y = (y_1, y_2, y_3, y_4, y_5)^T \). Also, assume that there are two bounding boxes per image and that bounding box \( 1, z_1 \), contains superpixel 1,2,4 while bounding box 2, \( z_2 \), contains superpixel 1,2,4. Thus, bounding box indexing for first proposal \( z_1 \) is defined by \( x_1 = (y_1, y_3, y_4)^T \) and for \( z_2 \) is defined by \( x_2 = (y_1, y_2, y_4)^T \). Vector \( x \) is obtained by concatenating \( x_1 \) and \( x_2 \). Then, vector \( x_1 \) and vector \( y \) are related by an indicator projection matrix \( P \) as follows:

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
y_4 \\
y_5
\end{bmatrix}
\]

Note that matrix \( P \) basically tells how many times a superpixel occurs in all bounding boxes or equivalently, how many times \( y_i \) is duplicated in the vector \( x \). We now translate the other three constraints from the paper one by one. Note that \( |S_i| = 3 \) since each bounding box contains 3 superpixels, \( m = 2 \) and \( n = 5 \).

To keep it short, we only demonstrate the constraints for the superpixels of the first bounding box \( i = 1 \).

\[
\gamma |S_i| z_i \leq \sum_{j \in S_i} x_{ij} \leq (1 - \gamma) |S_i| z_i \quad \text{for} \quad i = 1
\]

\[
\Rightarrow \gamma * 3z_1 \leq (x_{11} + x_{12} + x_{13}) \leq (1 - \gamma) * 3z_1
\]

\[
\Rightarrow \gamma * 3z_1 \leq (y_1 + y_3 + y_4) \leq (1 - \gamma) * 3z_1 \quad \text{(By} \ P_1 y = x_1)\]

Similarly, the second constraint for superpixels is equivalent to:

\[
\sum_{i,j \in S_i} x_{ij} \leq \sum_{i,j \in S_i} z_i, \text{for} \quad j = 1, 2, 3, 4, 5
\]

\[
(x_{11} + x_{21}) \leq (z_1 + z_2) \Rightarrow 2y_1 \leq (z_1 + z_2)
\]

\[
x_{22} \leq z_2 \Rightarrow y_2 \leq z_2
\]

\[
x_{12} \leq z_1 \Rightarrow y_3 \leq z_1
\]

\[
(x_{13} + x_{23}) \leq (z_1 + z_2) \Rightarrow 2y_4 \leq (z_1 + z_2)
\]

Finally, for the bounding boxes, we have:

\[
\sum_{i=1}^m z_i = 1 \Rightarrow z_1 + z_2 = 1
\]

\[ \text{References} \]

[1] F. Bach and Z. Harchaoui. DIFFRAC: a discriminative and flexible framework for clustering. In Proc. Neural Info. Proc. Systems, 2007.

[2] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image cosegmentation. In CVPR, 2010. 1, 2, 3, 4, 5

[3] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In CVPR, 2012. 2

[4] L. Ladicky, P. Sturgess, K. Alahari, C. Russel and P.H. Torr. What where and how many? combining object detectors and crfs. In ECCV, 2010. 1, 2

[5] S. Y. Bao, Y. Xiang and S. Savarese. Object co-detection. In ECCV, 2012. 2

[6] O.M. Parkhi, A. Vedaldi, C. Jawahar and A. Zisserman. The truth about cats and dogs. In ICCV, 2011. 2

[7] C. Rother, V. Kolmogorov, T. Minka, and A. Blake. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In CVPR, 2006. 2

[8] J. Shi and J. Malik. Normalized cuts and image segmentation. PAMI, 22(8):888–905, 2000. 3

[9] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In CVPR, 2014. 1, 2, 4, 5, 6

Fig. 4: Some Failure Cases of colocalization according to CorLoc Metric
[10] M. Rubinstein, A. Joulin J. Kopf C. Liu Unsupervised Joint Object Discovery and Segmentation in Internet Images In CVPR, 2013. 1, 2, 4, 5
[11] S. Vicente and C. Rother and V. Kolmogorov Object Cosegmentation In CVPR, 2011. 2
[12] J. Rubio and J. Serrat and A. Lopez and N. Paragios, Unsupervised co-segmentation through region matching In CVPR, 2012. 2
[13] Q. Yan, L. Xu, J. Shi and J. Jia Hierarchical Saliency Detection In CVPR, 2013. 4
[14] F. Wang, Q. Huang, and L. Guibas Image Co-Segmentation via Consistent Functional Maps In ICCV, 2013. 2, 4
[15] J. Shotton, J. Winn, C. Rother, and A. Criminisi, TextureBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation In ECCV, 2006 4
[16] A. Vedaldi and S. Soatto, Quick shift and kernel methods for mode seeking InECCV, 2008 4
[17] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, Maximum margin clustering InNIPS, 2005 1, 2
[18] M. Journee, F. Bach, P.-A. Abisal and R. Sepulchre, Low rank optimization for semidefinite convex problems In Technical Report, 2008 4
[19] R. Girshick, J. Donahue, T. Darrell and J. Malik Rich feature hierarchies for accurate object detection and semantic segmentation In CVPR, 2014 4
[20] B. Alexe, T. Deselaers and V. Ferrari Measuring the objectness of image windows PAMI, 34(11):2189–2202, 2012 4
[21] T. Deselaers, B. Alexe and V. Ferrari Weakly supervised localization and learning with generic knowledge IJCV, 2012 2, 6
[22] B. Hariharan, P. Arbelaez, R. Girshick and J. Malik, Simultaneous Detection and Segmentation In ECCV, 2014 2
[23] J. Xu, A. Schwing and R. Urtasun, Learning to Segment Under Various Forms of Weak Supervision In CVPR, 2015 2
[24] A. Krizhevsky, I. Sutskever, and G. Hinton ImageNet Classification with Deep Convolutional Neural Networks In NIPS, 2012 4
[25] A. Joulin, L. Maaten, A. Jabri and N. Vasilache Learning Visual Features from Large Weakly Supervised Data In ECCV, 2016 1
[26] S. X. Yu, R. Gross, and J. Shi Concurrent Object Recognition and Segmentation by Graph Partitioning In NIPS, 2003 1
[27] M. Maire, S. X. Yu and P. Perona Object Detection and Segmentation from Joint Embedding of Parts and Pixels In ICCV, 2011 1
[28] R. Quan, J. Han, D. Zhang and F. Nie Object Co-Segmentation via Graph Optimized-Flexible Manifold Ranking In CVPR, 2016 1, 2, 4, 5
[29] A. Faktor and M. Irani Co-segmentation by Composition In ICCV, 2013 1, 2, 4, 5
[30] M. Cho, S. Kwak, C. Schmid and J. Ponce Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals In CVPR, 2015 1, 2, 5, 6
[31] Y. Li, L. Liu, C. Shen and A. van den Hengel Image Co-localization by Mimicking a Good Detector’s Confidence Score Distribution In ECCV, 2016 1, 2
[32] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman Return of the Devil in the Details: Delving Deep into Convolutional Nets In BMVC, 2014 1
[33] L. Mukherjee, V. Singh and C. R. Dyer Half-integrality based algorithms for cosegmentation of images In CVPR, 2009 2
[34] L. Mukherjee, V. Singh and J. Peng Scale invariant cosegmentation for image groups In CVPR, 2011 2
[35] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results In Tech-Report, 2007 5, 6
[36] C. Wang, W. Ren, K. Huang and T. Tan Weakly Supervised Object Localization with Latent Category Learning In ECCV, 2014 1
[37] M. Pandey and S. Lazebnik Scene recognition and weakly supervised object localization with deformable part-based models In ICCV, 2011 1
[38] P. Siva, C. Russell, T. Xiang and L. Agapito Looking Beyond the Image: Unsupervised Learning for Object Saliency and Detection In CVPR, 2013 1
[39] W. Yang, Y. Wang, and Mori, G. Recognizing human actions from still images with latent poses In CVPR, 2010 1
[40] A. Karpathy, A. Joulin and L. Fei-Fei Deep Fragment Embeddings for Bidirectional Image Sentence Mapping In NIPS, 2014 1
[41] R. Caruana Algorithms and Applications for Multitask Learning In ICML, 1996 1
[42] M. Lapin and B. Schiele and M. Hein Scalable Multitask Representation Learning for Scene Classification In CVPR, 2014 1
[43] D. Pathak, F. Krähenbühl and Darrell, T. Constrained Convolutional Neural Networks for Weakly Supervised Segmentation In ICCV, 2015 7
[44] C. Rothe, V. Kolmogorov, T. Minka and A. Blake GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts In SIGGRAPH, 2004 2, 4
[45] A. Joulin, K. Tang and L. Fei-Fei Efficient Image and Video Co-localization with Frank-Wolfe Algorithm In ECCV, 2014 2
[46] J. Long, E. Shelhamer, and T. Darrell Fully Convolutional Networks for Semantic Segmentation In CVPR, 2015 1

Fig. 5: More Qualitative results.