

Supplementary Materials for: Unsupervised Object Class Discovery via Saliency-Guided Multiple Class Learning

Jun-Yan Zhu^{1,2}, Jiajun Wu^{3,1}, Yichen Wei¹, Eric Chang¹ and Zhuowen Tu^{1,4}

¹Microsoft Research Asia

²Dept. of Computer Science and Technology, Tsinghua University

³Institute for Interdisciplinary Information Sciences, Tsinghua University

⁴Lab of Neuro Imaging and Dept. of Computer Science, UCLA

{junyanzhu89, jiajunwu.cs}@gmail.com, {yichenw, echang, zhuowent}@microsoft.com

1. Proof for Theorems

Now we will do discriminative learning with the presence of hidden variables. Our step is similar to standard EM[3] while the primary difference is that we are given labels $Y = \{y_1, \dots, y_n\}$ in addition to observations $X = \{x_1, \dots, x_n\}$, and we want to estimate the model θ that minimizes the negative log-likelihood function $\mathcal{L}(\theta; Y, X) = -\log \Pr(Y|X; \theta)$. We proceed by integrating H out:

Theorem 1. *The discriminative expectation maximization (DiscEM) algorithm optimizes the training set log likelihood $\mathcal{L}(\theta; Y, X)$ w.r.t. model parameters θ in the presence of hidden variable H , via:*

$$\frac{d}{d\theta} \mathcal{L}(\theta; Y, X) = E_{H \sim \Pr(H|Y, X; \theta)} \frac{d}{d\theta} \mathcal{L}(\theta; Y, X, H) \quad (1)$$

where $\mathcal{L}(\theta; Y, X, H) = -\log \Pr(Y, H|X; \theta)$. Notice that $\Pr(H|Y, X; \theta) = \frac{\Pr(Y, H|X; \theta)}{\Pr(Y|X; \theta)}$ and X, Y are given.

Proof.

$$\begin{aligned} \frac{d}{d\theta} \mathcal{L}(\theta; Y, X) &= -\frac{d}{d\theta} \log \Pr(Y|X; \theta) = -\frac{1}{\Pr(Y|X; \theta)} \frac{d}{d\theta} \Pr(Y|X; \theta) \\ &= -\frac{1}{\Pr(Y|X; \theta)} \frac{d}{d\theta} \sum_H \Pr(Y, H|X; \theta) = -\frac{1}{\Pr(Y|X; \theta)} \sum_H \frac{d}{d\theta} \Pr(Y, H|X; \theta) \\ &= -\frac{1}{\Pr(Y|X; \theta)} \sum_H \Pr(Y, H|X; \theta) \frac{d}{d\theta} \log \Pr(Y, H|X; \theta) \\ &= E_{H \sim \Pr(H|Y, X; \theta)} \frac{d}{d\theta} (-\log \Pr(Y, H|X; \theta)) = E_{H \sim \Pr(H|Y, X; \theta)} \frac{d}{d\theta} \mathcal{L}(\theta; Y, X, H). \end{aligned} \quad (2)$$

□

The general form of DiscEM is similar to the standard EM. We iteratively improve an initial estimate θ_0 with successively better estimates $\theta_1, \theta_2, \dots$, and so on until convergence. Each phase r consists of two steps:

- **E** step: Compute $\Pr(H|Y, X; \theta)$ via previous estimate θ_r .
- **M** step: Update θ_{r+1} by minimizing $\mathcal{L}(\theta; Y, X)$ using eqn.(1).

Note that in the above formulation, parameter θ can be purely discriminative, *i.e.* they are parameters of classifiers. In this way, DiscEM can take the advantages of discriminative learning algorithms. This contrasts DiscEM to other conditional-EM frameworks[8, 13], where the task is to learn generative parameters through a discriminative objective. Compared with standard supervised algorithms, DiscEM can better handle hidden variables and embrace the weakly supervised learning setting.

Assuming all the data are conditionally independent, we could further derive as:

$$\frac{d}{d\theta} \mathcal{L}(\theta; Y, X) = -\frac{d}{d\theta} \log \Pr(Y|X; \theta) = -\frac{d}{d\theta} \sum_{i=1}^n \log \Pr(y_i|x_i; \theta) = \sum_{i=1}^n E_{h_i \sim \Pr(h_i|y_i, x_i; \theta)} \left[-\frac{d}{d\theta} \log \Pr(y_i, h_i|x_i; \theta) \right]. \quad (3)$$

Then we give the main insight connecting MIL-Boost[16] and DiscEM:

Theorem 2. *When the instance-level model (5) and the bag-level model (7) are used, MIL-Boost's update rule (8) is equivalent to DiscEM, which reads:*

$$\frac{d}{d\theta} \log \Pr(y_i|x_i; \theta) = \begin{cases} \sum_{j=1}^m \frac{-1}{1-p_{ij}} \frac{d}{d\theta} p_{ij} & \text{if } y_i = -1 \\ \sum_{j=1}^m \frac{1-p_i}{p_i(1-p_{ij})} \frac{d}{d\theta} p_{ij} & \text{if } y_i = 1 \end{cases} \quad (4)$$

Before the proof, we first recall MIL-Boost[16]. Standard boosting [7, 10] assumes an additive model on instance-level decisions: $h_{ij} = h(x_{ij})$ where $h(x_{ij}) = \sum_t \lambda_t h_t(x_{ij})$ is a weighted vote of weak classifiers $h_t : \mathcal{X} \rightarrow \mathcal{Y}$. Assuming that $y_{ij} \in \mathcal{Y}$ is the hidden instance label, its probability as positive is given by:

$$p_{ij} = \Pr(y_{ij} = 1|x_{ij}; h) = \frac{1}{1 + \exp(-h_{ij})}. \quad (5)$$

The bag-level probability is computed via a Noisy-OR (NOR) model:

$$p_i = \Pr(y_i = 1|x_i; h) = 1 - \prod_{j=1}^m (1 - p_{ij}). \quad (6)$$

Since the bag label is given in the training set, we can optimize the negative log-likelihood function:

$$\mathcal{L}_{MIL} = -\sum_{i=1}^n (\mathbf{1}(y_i = 1) \log p_i + \mathbf{1}(y_i = -1) \log (1 - p_i)) \quad (7)$$

by greedy search for h^t over a weak classifier candidate pool, followed by a line search for λ_t . $\mathbf{1}(\cdot)$ is an indicator function. According to the AnyBoost[10] framework, the weight w_{ij} on each instance x_{ij} is updated as:

$$w_{ij} = -\frac{\partial \mathcal{L}_{MIL}}{\partial h_{ij}} = \begin{cases} -\frac{1}{1-p_{ij}} \frac{\partial p_{ij}}{\partial h_{ij}} & \text{if } y_i = -1 \\ \frac{1-p_i}{p_i(1-p_{ij})} \frac{\partial p_{ij}}{\partial h_{ij}} & \text{if } y_i = 1 \end{cases} \quad (8)$$

After we review the formulation of MIL-Boost[16], we show the proof of **Theorem 2**.

Proof. Recall that the data is a set of bags $X = \{x_1, \dots, x_n\}$, where each bag X_i contains a set of instances $\{x_{i1}, \dots, x_{im}\}$. Label y_i is given for bag x_i while y_{ij} is hidden variable for the instance x_{ij} . We denote the $H_i = \{y_{i1}, \dots, y_{im}\}$ as the hidden variables for bag x_i and $H_I = \{H_1, \dots, H_n\}$ as all the instance-level hidden variables. For the negative bags, each instance x_{ij} is known to be negative; for the positive bags, at least one instance is positive. In other words, given $y_i = -1$, we know $y_{ij} = -1$ for every j . We assume instances in a bag are independent. For shorthand we write $p(y_{ij}) = \Pr(y_{ij}|x_{ij}; \theta)$ and $p_{ij} = p(y_{ij} = 1)$.

Thus, for negative bags we know $y_{ij} = -1$. After some rearrangement, it becomes:

$$\frac{d}{d\theta} \log \Pr(y_i = -1|x_i; \theta) = \sum_{j=1}^m \frac{d}{d\theta} \log \Pr(y_{ij} = -1|x_{ij}; \theta) = \sum_{j=1}^m \frac{d}{d\theta} \log(1 - p_{ij}) = \sum_{j=1}^m \frac{-\frac{d}{d\theta} p_{ij}}{1 - p_{ij}}. \quad (9)$$

Next we derive the expression for positive bags. The hidden variables H_i are conditionally dependent given y_i , but otherwise we assume they are independent, *i.e.* $\Pr(H_i|x_i; \theta) = \prod_{j=1}^m \Pr(y_{ij}|x_{ij}; \theta)$. We observe that $\Pr(H_i = -1, y_i = 1|x_i; \theta) = 0$ (the event is impossible) and $\Pr(H_i, y_i = 1|x_i; \theta) = \Pr(H_i|x_i; \theta)$ for all $H_i \neq -1$ (If $H_i \neq -1$ we then know $y_i = 1$). This leads to:

$$\Pr(H_i|y_i = 1, x_i; \theta) = \begin{cases} 0 & \text{if } H_i = -1 \\ \prod_{j=1}^m p(h_{ij})/p_i & \text{otherwise} \end{cases} \quad (10)$$

In the above we use the NOR model (eqn.(7)) in MIL-Boost[16]: $p_i = \Pr(y_i = 1|x_i; \theta) = 1 - \prod_{j=1}^m (1 - p_{ij})$. We now expand eqn.(3) for positive bags:

$$\begin{aligned} & \frac{d}{d\theta} \log \Pr(y_i = 1|x_i; \theta) \\ &= \sum_{H_i} \Pr(H_i|y_i = 1, x_i; \theta) \frac{d}{d\theta} \log \Pr(y_i = 1, H_i|x_i; \theta) = \sum_{H_i \neq -1} \prod_{k=1}^m \frac{p(h_{ik})}{p_i} \frac{d}{d\theta} \log \prod_{j=1}^m p(h_{ij}) \\ &= \frac{1}{p_i} \sum_{j=1}^m \left[\sum_{H_i} \prod_{k=1}^m p(h_{ik}) \frac{d}{d\theta} \log p(h_{ij}) - \sum_{H_i = -1} \prod_{k=1}^m p(h_{ik}) \frac{d}{d\theta} \log p(h_{ij}) \right] \\ &= \frac{1}{p_i} \sum_{j=1}^m \left[\sum_{h_{ij}} p(h_{ij}) \frac{d}{d\theta} \log p(h_{ij}) - \prod_{k=1}^m (1 - p_{ik}) \frac{d}{d\theta} \log (1 - p_{ij}) \right] \\ &= \frac{1}{p_i} \sum_{j=1}^m \left[p_{ij} \frac{d}{d\theta} \log p_{ij} + (1 - p_{ij}) \frac{d}{d\theta} \log (1 - p_{ij}) - (1 - p_i) \frac{d}{d\theta} \log (1 - p_{ij}) \right] \\ &= \frac{1}{p_i} \sum_{j=1}^m \left[\frac{d}{d\theta} p_{ij} - \frac{d}{d\theta} p_{ij} - (1 - p_i) \frac{d}{d\theta} \log (1 - p_{ij}) \right] = \sum_{j=1}^m \frac{1 - p_i}{p_i(1 - p_{ij})} \frac{d}{d\theta} p_{ij}. \end{aligned} \quad (11)$$

Based on eqn.(9) and eqn.(11), we prove the **Theorem 2** for both negative bags and positive bags. \square

The above DiscEM formulation of MIL-Boost partly explains its success. However, since MIL-Boost combines weak classifiers, which can not easily attain the optimum in the **M** step, it has to incorporate a gradient descent strategy in the function space [10]. When strong classifiers (such as SVM or Boosting itself) are available, we can directly employ the DiscEM formulation, *i.e.* alternating between **E** step (applying a trained classifier to obtain instance-level probability estimation) and **M** step (train a new classifier based on the estimation), without retaining history information.

2. More Experimental Results

We show more experimental results, in the similar organization of experiment section in the paper.

2.1. Simultaneous categorization and localization

In addition to the purity measurement used in the paper (Table 1), we also compare categorization performance results on two additional metrics. Results show that our approach bMCL consistently outperforms other methods by a large margin.

Clustering accuracy is widely used in previous clustering algorithms [17, 21] and in multiple instance clustering methods [20, 18, 19] to evaluate the clustering performance. Comparison results are reported in Table 1.

Normalized Mutual Information(NMI) is a symmetric measure to quantify the statistical information shared between two distributions[15]. It is also used to evaluate the clustering performance in previous multiple instance clustering methods [20, 18, 19]. Comparison results are reported in Table 2.

	bMCL	SD	M ³ IC	BAMIC	UnSL
SIVAL1	95.3	78.7	39.3	37.7	25.3
SIVAL2	84.0	65.7	38.7	33.3	34.0
SIVAL3	74.7	62.7	37.0	38.7	26.0
SIVAL4	94.0	86.0	33.0	37.7	26.3
SIVAL5	75.3	70.3	35.3	36.7	23.3
CC	73.9	63.5	38.2	46.1	53.3
3D1	81.1	64.0	46.0	43.2	34.7
3D2	78.4	76.6	52.3	51.4	35.0

Table 1: Categorization results measured by the mean clustering accuracy. We compare bMCL with recent MIC approaches M³IC[18], BAMIC[20], one state-of-the-art unsupervised discovery method, UnSL[9] and SD (saliency detection baseline), more reasonable than [12].

	bMCL	SD	M ³ IC	BAMIC	UnSL
SIVAL1	89.9	72.7	11.4	12.4	10.8
SIVAL2	73.2	57.3	10.1	5.8	19.1
SIVAL3	64.9	42.4	8.7	11.3	6.1
SIVAL4	87.2	75.4	7.4	13.3	10.6
SIVAL5	61.4	52.3	8.3	9.1	11.1
CC	77.3	59.7	15.8	23.0	59.7
3D1	69.7	52.3	20.3	15.4	23.6
3D2	87.9	75.8	22.4	25.9	29.4

Table 2: Categorization results measured by the mean Normalized Mutual Information (NMI). We compare bMCL with recent MIC approaches M³IC[18], BAMIC[20], one state-of-the-art unsupervised discovery method, UnSL[9] and SD (saliency detection baseline), more reasonable than [12].

We show illustrative results from a few object classes in Figure 1, 2, and 3. See the paper (Section 6.1 and Figure 3) for more discussions regarding such results. Please notice that MIC methods (M^3IC [18] and BAGIC[20]) cannot perform the object localization.

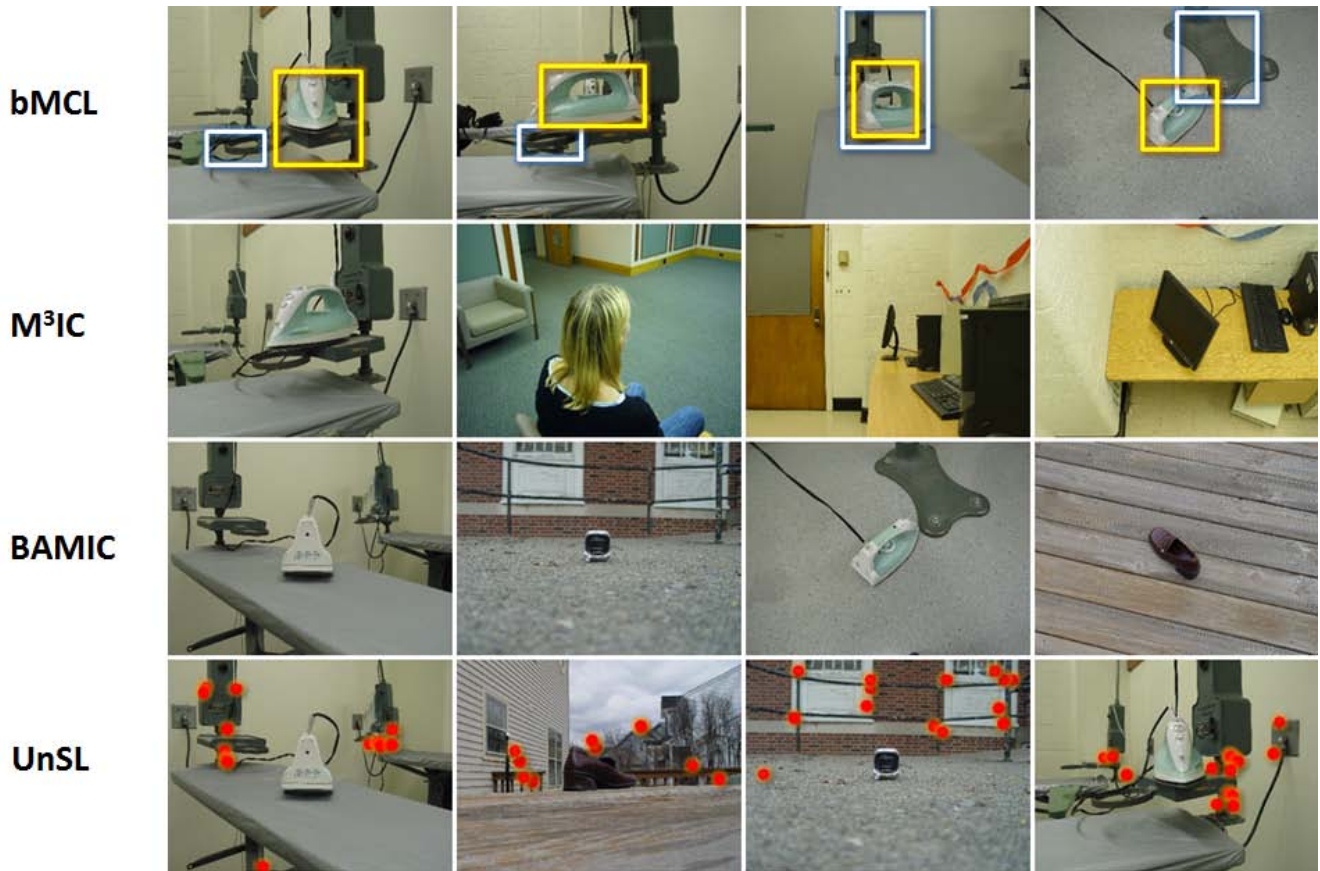


Figure 1: Illustrative categorization results of four methods in an object class from 3D object category dataset [14]. From top to down: bMCL, M^3IC [18], BAGIC [20] and UnSL [9]. In bMCL, the yellow rectangle is the localized object and the white rectangle is the most salient window computed by [6]. In UnSL, the learned object keypoints are overlaid (red points).

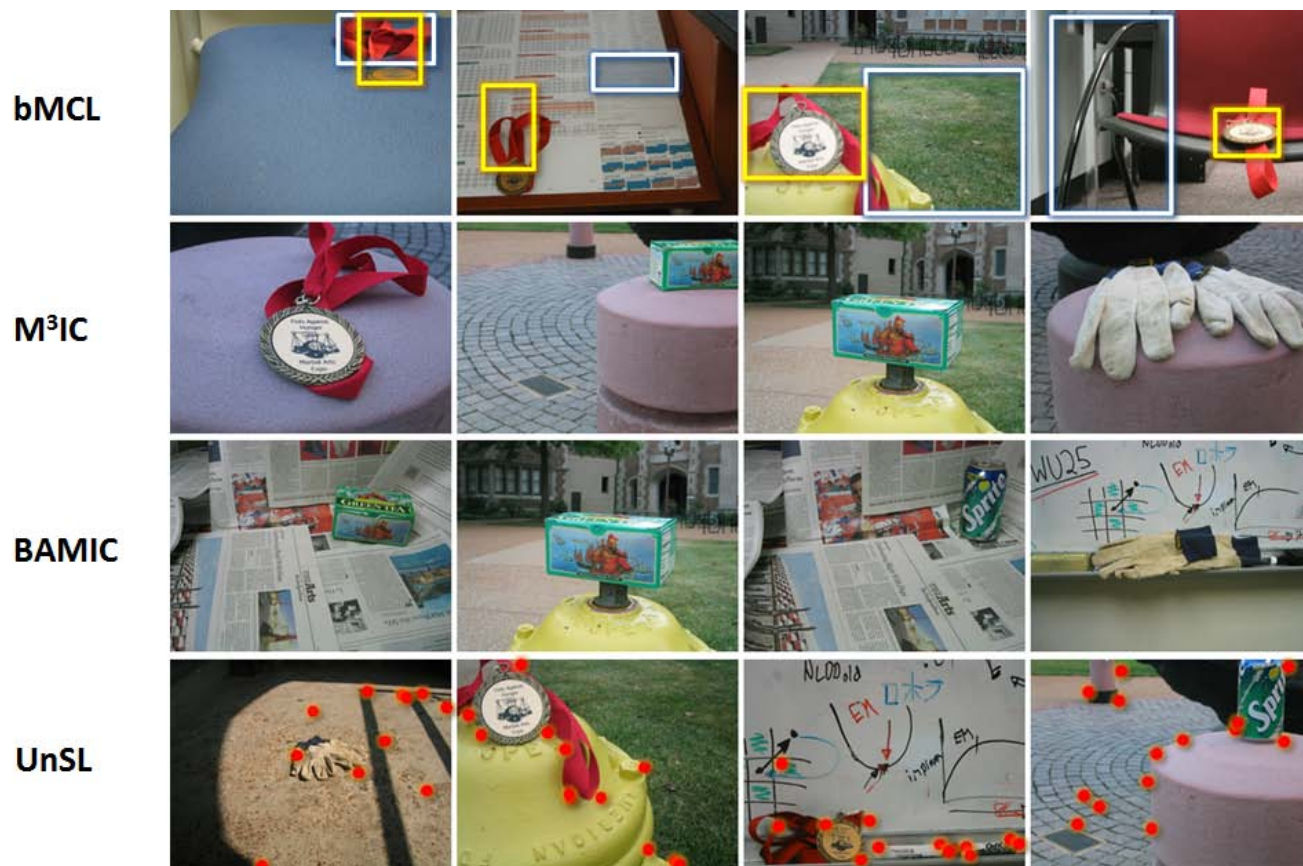


Figure 2: Illustrative categorization results of four methods in an object class from SIVAL dataset [11]. From top to down: bMCL, M³IC [18], BAMIC [20] and UnSL [9]. In bMCL, the yellow rectangle is the localized object and the white rectangle is the most salient window computed by [6]. In UnSL, the learned object keypoints are overlaid (red points).

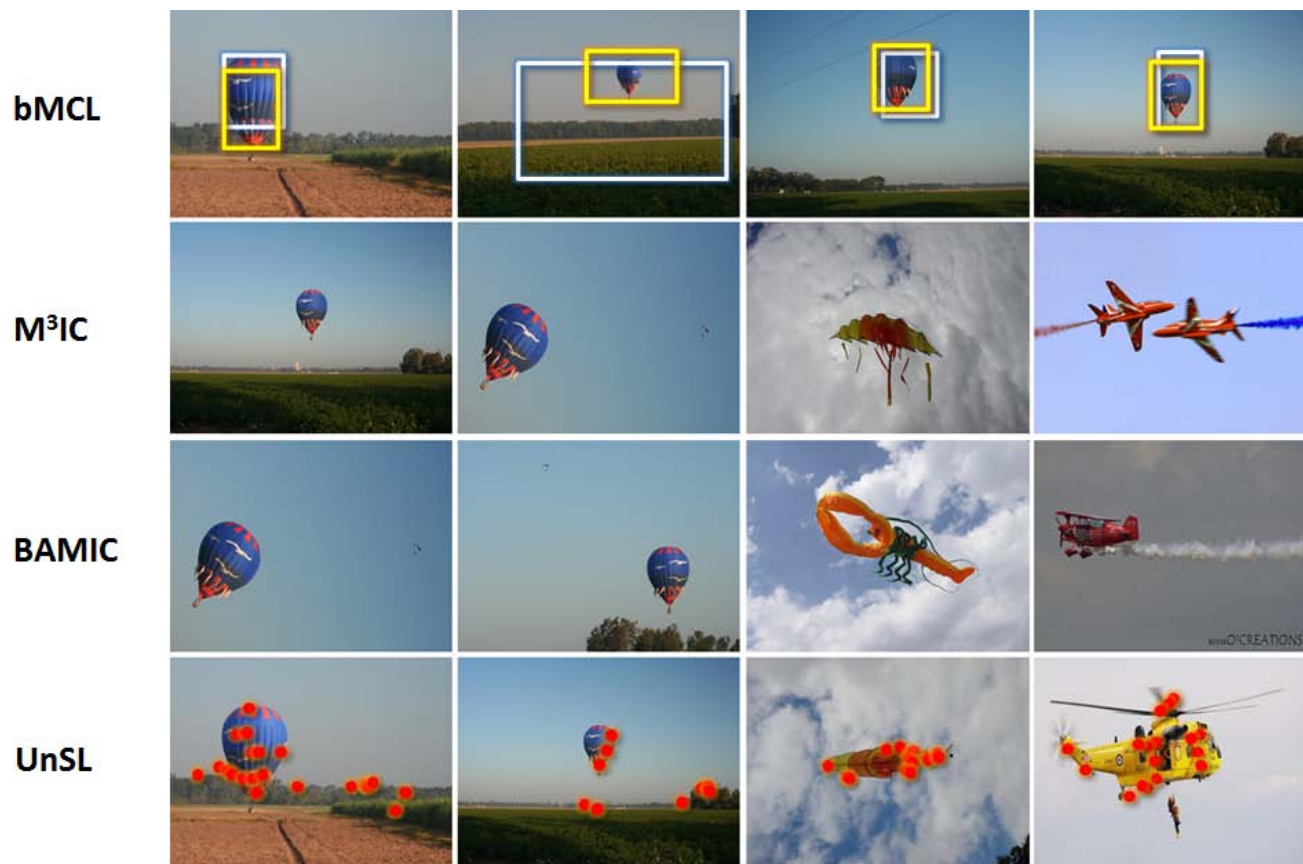
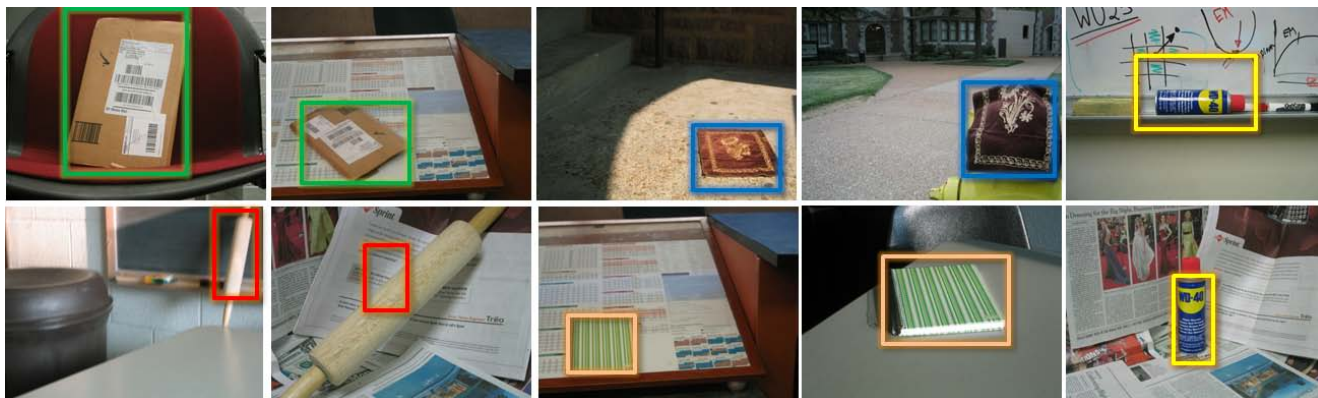


Figure 3: Illustrative categorization results of four methods in an object class from CMU-Cornell iCoseg dataset [1]. From top to down: bMCL, M³IC [18], BAMIC [20] and UnSL [9]. In bMCL, the yellow rectangle is the localized object and the white rectangle is the most salient window computed by [6]. In UnSL, the learned object keypoints are overlaid (red points).

2.2. Detecting novel objects using learned detectors

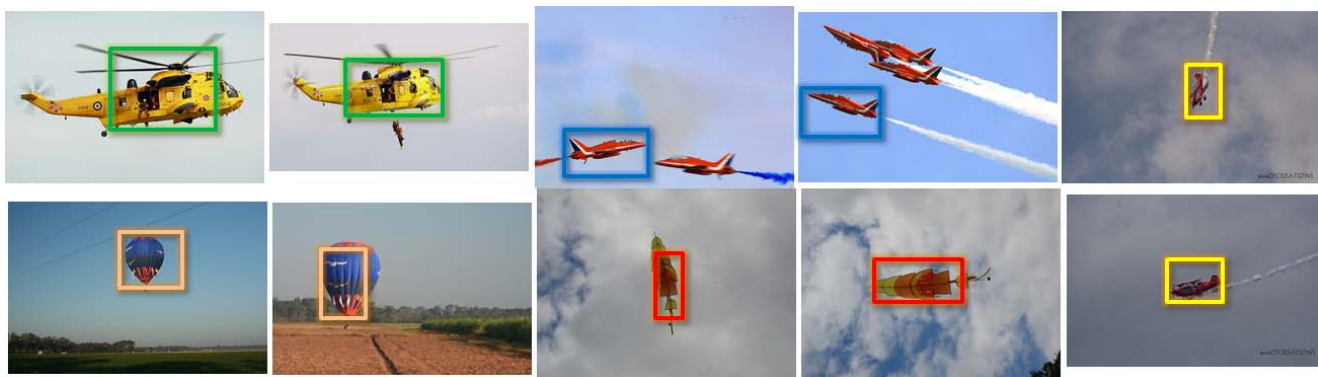
Figure 4 shows object detection results using learned object detectors in bMCL.



(a) SIVAL dataset



(b) 3D object category dataset



(c) CMU-Cornell iCoseg dataset

Figure 4: Object detection results using learned object detectors. Each color represents an object class.

2.3. Co-saliency

Figure 5 illustrates the co-saliency results of bMCL and the results of two state-of-the-art saliency methods [2, 6] on SIVAL dataset[11].

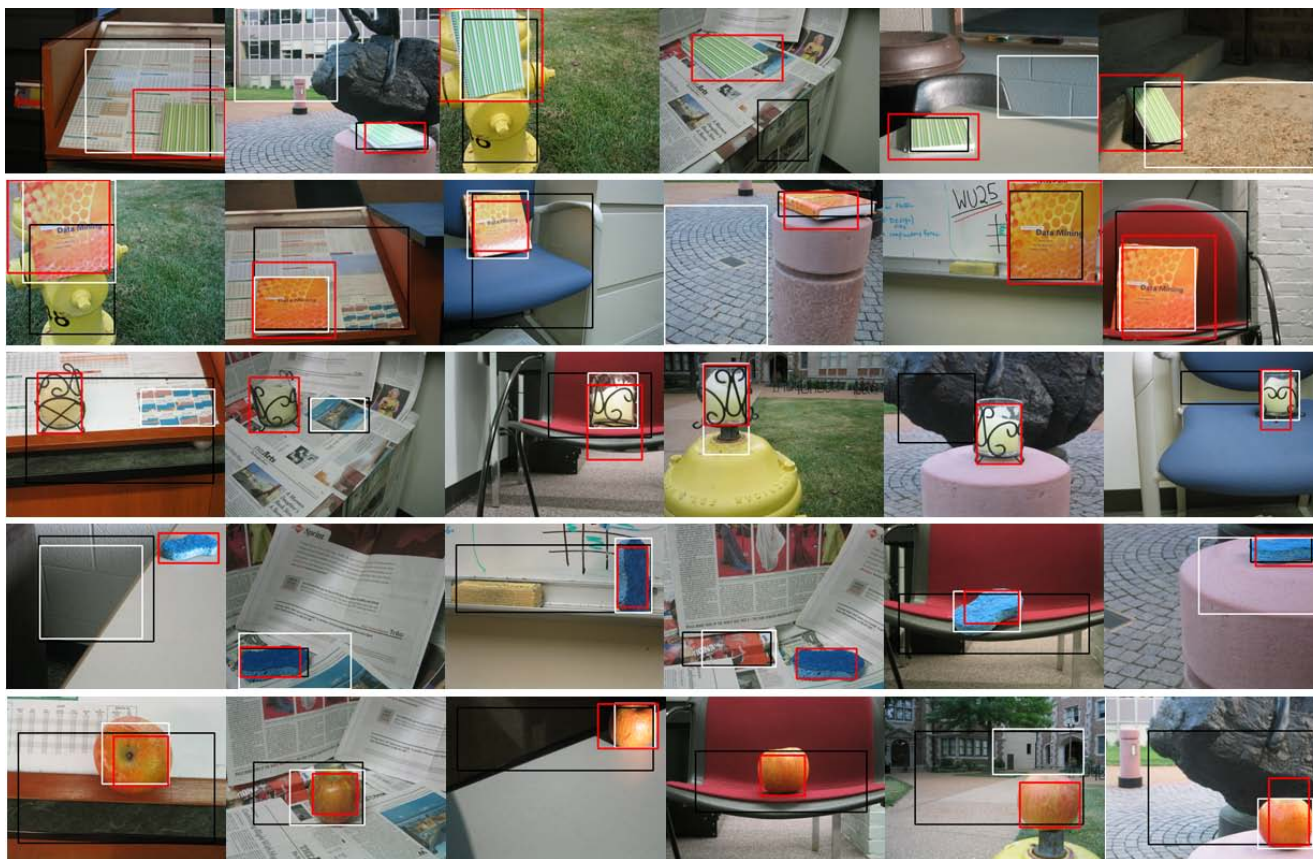


Figure 5: bMCL's co-saliency results and results of two state-of-the-art saliency methods. Red rectangles: bMCL co-saliency results. Black rectangles: results obtained by [2]. White rectangles: results obtained by [6]. SIVAL[11] categories from top to down: stripednotebook, dataminingbook, candlewithholder, bluescrunge, apple.

2.4. Weakly supervised learning with a single object class

Figure 6 and 7 show the localization results on PASCAL VOC 07[4] and PASCAL VOC 06[5] classes:



(a) aeroplane



(b) horse



(c) sofa



(d) train



(e) motorbike

Figure 6: Red rectangles: object localization results of bMCL with a single object class on challenging PASCAL VOC 07[4].



(a) cow



(b) car

Figure 7: Red rectangles: object localization results of bMCL with a single object class on challenging PASCAL VOC 06[5].

3. Datasets

We use the SIVAL dataset [11], CMU-Cornell iCoseg dataset [1], and 3D object category dataset [14] in the multi-class object discovery experiment. Table 3 shows the details of each dataset.

Table 3: Experiment names, dataset names, used categories, and the numbers of images.

Exp	Dataset	Classes	Size
SIVAL1	SIVAL	ajaxorange	60
		checkeredscarf	60
		bluescrunge	60
		glazedwoodpot	60
		juliespot	60
SIVAL2	SIVAL	dirtyworkgloves	60
		greenteabox	60
		goldmedal	60
		smileyfacedoll	60
		spritecan	60
SIVAL3	SIVAL	cardboardbox	60
		feltflowerrug	60
		stripednotebook	60
		wd40can	60
		woodrollingpin	60
SIVAL4	SIVAL	apple	60
		candlewithholder	60
		fabricsoftenerbox	60
		rapbook	60
		translucentbowl	60
SIVAL5	SIVAL	banana	60
		cokecan	60
		dataminingbook	60
		dirtyrunningshoe	60
		largespoon	60
CC	CMU-Cornell iCoseg	025_1	12
		025_2	39
		026	22
		032	19
		041	25
3D1	3D Object Category	cellphone_91	24
		head_9	24
		iron_7	24
		monitor_4	15
		shoe_1	24
3D2	3D Object Category	bicycle_9	24
		car_8	16
		mouse_8	23
		stapler_5	24
		toaster_10	24

References

- [1] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010. 7, 12
- [2] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, 2011. 9
- [3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statist. Soc. Series B*, 39(1):1–38, 1977. 1
- [4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>. 10
- [5] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. <http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf>. 10, 11
- [6] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *ICCV*, 2011. 5, 6, 7, 9
- [7] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. of Comp. and Sys. Sci.*, 55(1):119–139, 1997. 2
- [8] T. Jebara and A. Pentland. Maximum conditional likelihood via bound maximization and the cem algorithm. In *NIPS*, 1998. 2
- [9] G. Kim, C. Faloutsos, and M. Hebert. Unsupervised modeling of object categories using link analysis techniques. In *CVPR*, 2008. 4, 5, 6, 7
- [10] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent. In *NIPS*, 2000. 2, 3
- [11] R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J. E. Fritts. Localized content based image retrieval. *IEEE Trans. PAMI*, 30(11), 2008. 6, 9, 12
- [12] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition. In *CVPR*, 2004. 4
- [13] J. Salojarvi, K. Puolamaki, and S. Kaski. Expectation maximization algorithms for conditional likelihoods. In *NIPS*, 2005. 2
- [14] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007. 5, 12
- [15] A. Strehl, J. Ghosh, and C. Cardie. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002. 4
- [16] P. A. Viola, J. C. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, 2006. 2, 3
- [17] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *NIPS*, 2005. 4
- [18] D. Zhang, F. Wang, L. Si, and T. Li. Maximum margin multiple instance clustering. In *IJCAI*, 2009. 4, 5, 6, 7
- [19] D. Zhang, F. Wang, L. Si, and T. Li. Maximum margin multiple instance clustering with its applications to image and text clustering. *IEEE Transaction on Neural Networks*, 22(5):739–751, 2011. 4
- [20] M.-L. Zhang and Z.-H. Zhou. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence*, 31:47–68, August 2009. 4, 5, 6, 7
- [21] B. Zhao, F. Wang, and C. Zhang. Efficient multiclass maximum margin clustering. In *ICML*, 2008. 4