Deep Descriptor Transforming for Image Co-Localization

Xiu-Shen Wei[1*], Chen-Lin Zhang[1*], Yao Li[2], Chen-Wei Xie[1], Jianxin Wu[1], Chunhua Shen[2], Zhi-Hua Zhou[1]

1 National Key Laboratory for Novel Software Technology, Nanjing University

2 The University of Adelaide, Adelaide, Australia

IJCAI-17 MELBOURNE

LAMDA Learning And Mining from DatA

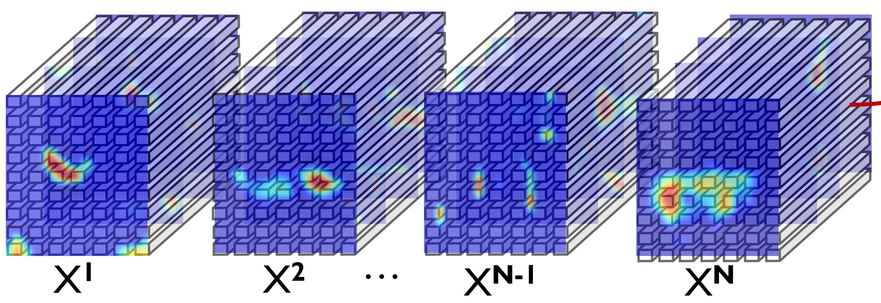
1 – What is image co-localization?

Image co-localization (a.k.a. unsupervised object discovery) is a fundamental computer vision problem, which simultaneously localizes objects of the same category across a set of distinct images.

2 – The proposed method: DDT

Input images

CNN pre-trained models $X^n = \left\{ \boldsymbol{x}^n_{(i,j)} \in \mathcal{R}^d \right\}$



Deep descriptors

$$\bar{\boldsymbol{x}} = \frac{1}{K} \sum_n \sum_{i,j} \boldsymbol{x}^n_{(i,j)}$$

Calculating descriptors' mean vector

$$\mathrm{Cov}(\boldsymbol{x}) = \frac{1}{K} \sum_n \sum_{i,j} (\boldsymbol{x}^n_{(i,j)} - \bar{\boldsymbol{x}})(\boldsymbol{x}^n_{(i,j)} - \bar{\boldsymbol{x}})^\top$$

$$\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_d \text{ of } \mathrm{Cov}(\boldsymbol{x})$$

Obtaining the covariance matrix and then getting the eigenvectors

$$p^1_{(i,j)} = \boldsymbol{\xi}_1^\top \left(\boldsymbol{x}_{(i,j)} - \bar{\boldsymbol{x}} \right)$$

Transforming

$$P^1 = \begin{bmatrix} p^1_{(1,1)} & p^1_{(1,2)} & \cdots & p^1_{(1,w)} \\ p^1_{(2,1)} & p^1_{(2,2)} & \cdots & p^1_{(2,w)} \\ \vdots & \vdots & \ddots & \vdots \\ p^1_{(h,1)} & p^1_{(h,2)} & \cdots & p^1_{(h,w)} \end{bmatrix}$$

Indicator matrix for co-localization

Deep Descriptor Transforming

DDT results

3 – Contributions

- ✓ We propose a simple yet effective method, i.e., Deep Descriptor Transforming, for image co-localization. DDT does **not** require image labels, negative images or redundant object proposals.
- ✓ To our knowledge, this is the first work to demonstrate the possibility of convolutional activations/descriptors in pre-trained models being able to act as a detector for the common object, which also reveals another probability of deep pre-trained network reusing.
- ✓ For the co-localization performance, DDT consistently outperforms state-of-the-arts of image co-localization methods by a large margin and also weakly supervised object localization methods.
- ✓ DDT has a good generalization ability for unseen categories and robustness for dealing with noisy data.

4 – Experiments

Four benchmark co-localization datasets:

Object Discovery, PASCAL VOC 07, PASCAL VOC 12, ImageNet-Subset

Col-localization results:

Table 1: Comparisons of CorLoc on *Object Discovery*.

Methods	*Airplane*	*Car*	*Horse*	**Mean**
[Joulin *et al.*, 2010]	32.93	66.29	54.84	51.35
[Joulin *et al.*, 2012]	57.32	64.04	52.69	58.02
[Rubinstein *et al.*, 2013]	74.39	87.64	63.44	75.16
[Tang *et al.*, 2014]	71.95	93.26	64.52	76.58
SCDA	87.80	86.52	75.37	83.20
[Cho *et al.*, 2015]	82.93	94.38	75.27	84.19
Our DDT	**91.46**	**95.51**	**77.42**	**88.13**

Table 5: Comparisons of on image sets disjoint with ImageNet.

Methods	*Chipmunk*	*Rhino*	*Stoat*	*Racoon*	*Rake*	*Wheelchair*	**Mean**
[Cho *et al.*, 2015]	26.6	81.8	44.2	30.1	8.3	35.3	37.7
SCDA	32.3	71.6	52.9	34.0	7.6	28.3	37.8
[Li *et al.*, 2016]	44.9	81.8	67.3	41.8	14.5	39.3	48.3
Our DDT	**70.3**	**93.2**	**80.8**	**71.8**	**30.3**	**68.2**	**69.1**

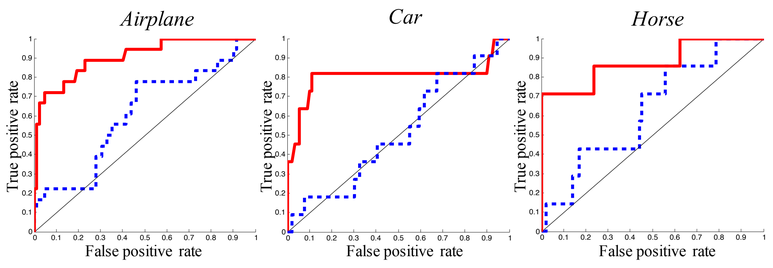
Table 2: Comparisons of the CorLoc metric with state-of-the-art co-localization methods on *VOC 2007*.

Methods	*aero*	*bike*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*	*table*	*dog*	*horse*	*mbike*	*person*	*plant*	*sheep*	*sofa*	*train*	*tv*	**Mean**
[Joulin *et al.*, 2014]	32.8	17.3	20.9	18.2	4.5	26.9	32.7	41.0	5.8	29.1	**34.5**	31.6	26.1	40.4	17.9	11.8	25.0	27.5	35.6	12.1	24.6
SCDA	54.4	27.2	43.4	13.5	2.8	39.3	44.5	48.0	6.2	32.0	16.3	49.8	51.5	49.7	7.7	6.1	22.1	22.6	46.4	6.1	29.5
[Cho *et al.*, 2015]	50.3	42.8	30.0	18.5	4.0	62.3	**64.5**	42.5	8.6	**49.0**	12.2	44.0	64.1	57.2	15.3	9.4	30.9	34.0	61.6	**31.5**	36.6
[Li *et al.*, 2016]	**73.1**	45.0	43.4	**27.7**	6.8	53.3	58.3	45.0	6.2	48.0	14.3	47.3	69.4	66.8	**24.3**	12.8	**51.5**	25.5	65.2	16.8	40.0
Our DDT	67.3	**63.3**	**61.3**	22.7	**8.5**	**64.8**	57.0	**80.5**	**9.4**	**49.0**	22.5	**72.6**	**73.8**	**69.0**	7.2	**15.0**	35.3	**54.7**	**75.0**	29.4	**46.9**

Table 3: Comparisons of the CorLoc metric with state-of-the-art co-localization methods on *VOC 2012*.

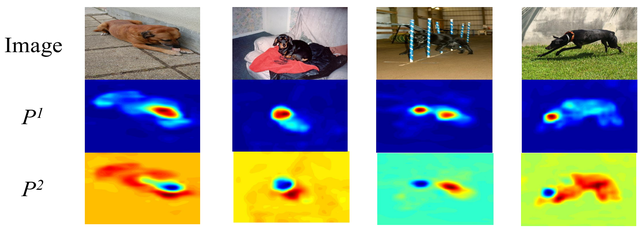
Methods	*aero*	*bike*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*	*table*	*dog*	*horse*	*mbike*	*person*	*plant*	*sheep*	*sofa*	*train*	*tv*	**Mean**
SCDA	60.8	41.7	38.6	21.8	7.4	67.6	38.8	57.4	16.0	34.0	23.9	53.8	47.3	54.8	7.9	9.9	25.3	23.2	50.2	10.1	34.5
[Cho *et al.*, 2015]	57.0	41.2	36.0	26.9	5.0	81.1	**54.6**	50.9	18.2	54.0	**31.2**	44.9	61.8	48.0	13.0	11.7	51.4	45.3	64.6	**39.2**	41.8
[Li *et al.*, 2016]	65.7	57.8	47.9	28.9	6.0	74.9	48.4	48.4	14.6	**54.4**	23.9	50.2	**69.9**	68.4	**24.0**	14.2	**52.7**	30.9	72.4	21.6	43.8
Our DDT	**76.7**	**67.1**	**57.9**	**30.5**	**13.0**	**81.9**	48.3	**75.7**	**18.4**	48.8	27.5	**71.8**	66.8	**73.7**	6.1	**18.5**	38.0	**54.7**	**78.6**	34.6	**49.4**

Detecting noisy images:



ROC curves of our DDT (the red line) at identifying noisy images on *Object Discovery*.

5 – Further study



From P^1 to P^2

-- Part based co-localization

6 – Conclusions

- ✓ DDT revealed another reusability of deep pre-trained networks.
- ✓ It offered further understanding and insights about CNNs.
- ✓ Our proposed DDT method is both efficient and effective.
- ✓ The generalization ability and robustness of DDT ensure its effectiveness and powerful reusability in real-world applications.