Learning Co-segmentation by Segment Swapping for Retrieval and Discovery

Xi Shen[1] Alexei A. Efros[2] Armand Joulin[3] Mathieu Aubry[1]

[1]LIGM (UMR 8049), École des Ponts ParisTech [2] UC Berkeley [3]Meta AI

1. Introduction

Motivation

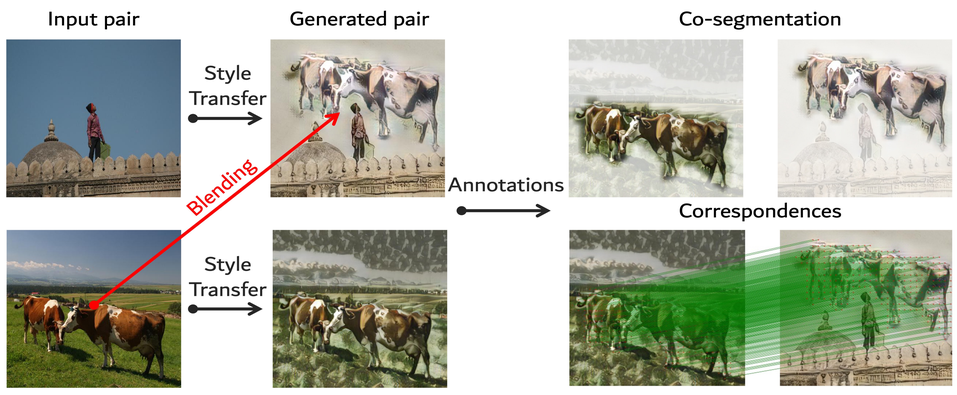
Discover influences via repetition of visual details.

Place recognition with repeated segments.

Challenges

- No training data available.

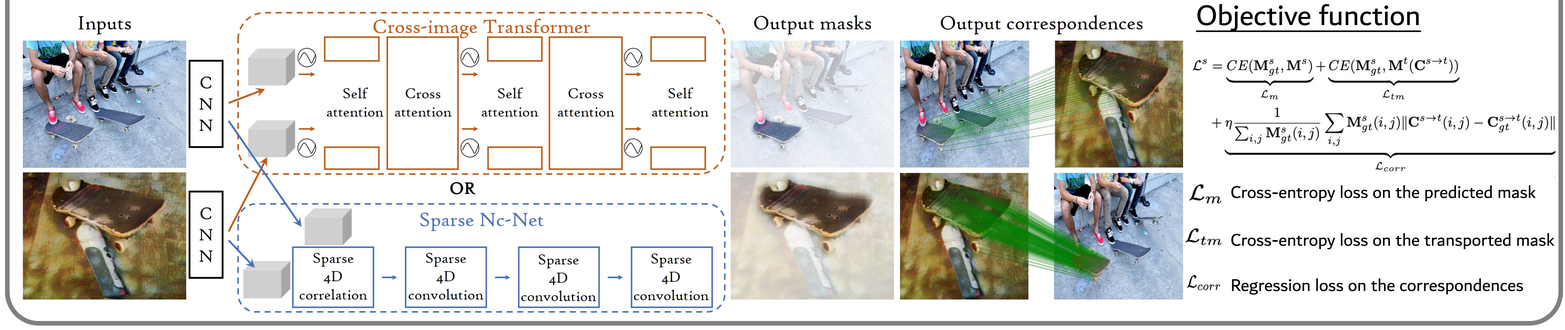
Key idea : training with synthetic pairs



Main contributions

- A method of generating synthetic pairs to learn co-segmentation
- A transformer-based architecture for co-segmentation producing competitive performances on art detail retrieval and place recognition
- Spectral clustering on a correspondence graph for discovery in image collections.

2. Learning co-segmentation



3. Experiments

Feat. + Methods	mAP Retrieval	mAP Det.(IoU > 0.3)
Shen et al. [54] + cos [54]	75.5	75.3
Shen et al. [54] + discovery [54]	76.6	76.4
MocoV2 [9] + cos [54]	79.0	78.7
MocoV2 [9] + discovery [54]	80.8	79.6
Ours + Unsupervised segments		
Transformer	81.8	79.4
Sparse-Ncnet	82.8	73.4
Ours + COCO segments [36]		
Transformer	**84.4**	**81.8**
Sparse-Ncnet	83.3	73.7

Art detail retrieval on Brueghel dataset, predicted masks are shown in transparency.

Score between a pair of images

$$\mathcal{S}(\mathbf{I}^s, \mathbf{I}^t) = \sum_{i,j} \underbrace{\mathbf{M}^s_{joint}(i,j)}_{\text{Mask}} \underbrace{cos(\mathbf{F}^s(i,j), \mathbf{F}^t(\mathbf{C}^{s\to t}(i,j)))}_{\text{Feat. similarity}}$$

Where $\mathbf{M}^s_{joint}$ is the product of the source and the transported target mask :

$$\mathbf{M}^s_{joint}(i,j) = \mathbf{M}^t(\mathbf{C}^{s\to t}(i,j))\mathbf{M}^s(i,j)$$

$\mathbf{F}^s(i,j)$ Feature at coordinate (i, j) in the source

$\mathbf{F}^t(\mathbf{C}^{s\to t}(i,j))$ Feature at warped coordinate (i, j) in the target

Place recognition on Tokyo24/7 (*Top*) and Pitts30K (*Bottom*) dataset, predicted masks are shown in transparency.

Method	Supervision	Tokyo 24/7 [61] R@1	R@5	R@10	Pitts30k-test [62] R@1	R@5	R@10
AP-GEM [20, 44]	Image location	40.3	55.6	65.4	75.3	89.3	92.5
DenseVLAD [20, 61]	Image location	59.4	67.3	72.1	77.7	88.3	91.6
NetVLAD [3, 20]	Image location	73.3	82.9	86.0	86.0	93.2	95.1
CRN [16, 29]	Image location	75.2	83.8	87.3	-	-	-
SARE [16, 38]	Image location	79.7	86.7	90.5	-	-	-
IBL [16]	Image location	**85.4**	**91.1**	**93.3**	-	-	-
Re-ranking Top-100 from NetVLAD [3, 20]							
Patch-NetVLAD [20]	Image location	81.9	85.7	87.9	88.6	94.5	95.8
Patch-NetVLAD [20] + RANSAC	Image location	86.0	88.6	**90.5**	**88.7**	94.5	95.9
SuperGlue [20, 50]*	Pose+Depth	**88.2**	**90.2**	90.2	**88.7**	**95.1**	**96.4**
Ours + Unsupervised segments							
Transformer	Segment swapping	74.0	82.9	86.0	85.2	93.5	95.4
Nc-Net	Segment swapping	84.1	87.0	88.9	86.4	94.3	95.6
Ours + COCO segments [36]							
Transformer	Segment swapping	80.0	86.0	87.9	84.7	93.5	95.6
Nc-Net	Segment swapping	85.4	88.3	89.2	86.8	94.4	95.8

* uses learnt keypoint detector Superpoint [14]

Method	ECSSD [57] max F_β	IoU	Acc.	DUTS [74] max F_β	IoU	Acc.	DUT-OMRON [77] max F_β	IoU	Acc.
HS [76]	0.673	0.508	0.847	0.504	0.369	0.826	0.561	0.433	0.843
wCtr [80]	0.684	0.517	0.862	0.522	0.392	0.835	0.541	0.416	0.838
WSC [36]	0.683	0.498	0.852	0.528	0.384	0.862	0.523	0.387	**0.865**
DeepUSPS [43]	0.584	0.440	0.795	0.425	0.305	0.773	0.414	0.305	0.779
BigBiGAN [73]	0.782	0.672	0.899	0.608	0.498	0.878	0.549	0.453	0.856
E-BigBiGAN [73]	0.797	0.684	0.906	0.624	0.511	0.882	0.563	0.464	0.860
LOST [58]	0.758	0.654	0.895	0.611	0.518	0.871	0.473	0.410	0.797
LOST [58] + Bilateral Solver [5] (Ours)	**0.837**	**0.723**	**0.916**	**0.697**	**0.572**	**0.887**	**0.578**	**0.489**	0.818

Validation of unsupervised segments on unsupervised saliency detection benchmarks : ECSSD, DUTS, DUT-OMRON.

4. Object discovery

Discovered clusters on Brueghel dataset.

Correspondences graph

Nodes : correspondences

Edges :

$$\epsilon_{i,j} = \frac{1}{2} m_i m_j exp(-\frac{\|x^s_i - x^s_j\|}{\sigma})(exp(-\frac{\|x^t_i - \mathbf{C}^{t_i\to t_j}(x^t_j)\|}{\sigma}) + exp(-\frac{\|x^t_j - \mathbf{C}^{t_i\to t_j}(x^t_i)\|}{\sigma}))$$

Method	Airplane $\mathcal{P}$	Airplane $\mathcal{J}$	Car $\mathcal{P}$	Car $\mathcal{J}$	Horse $\mathcal{P}$	Horse $\mathcal{J}$	Avg $\mathcal{P}$	Avg $\mathcal{J}$
DOCS [37]*	0.946	0.64	0.940	0.83	0.914	0.65	0.933	0.70
Sun et al. [59]	0.886	0.36	0.870	0.73	0.876	0.55	0.877	0.55
Joulin et al. [28]	0.493	0.15	0.587	0.37	0.638	0.30	0.572	0.27
Kim et al. [30]	0.802	0.08	0.689	0.00	0.751	0.06	0.754	0.05
Rubinstein et al. [51]	0.880	0.56	0.854	0.64	0.828	0.52	0.827	0.43
Chen et al. [9]	0.902	0.40	0.876	0.65	0.893	0.58	0.890	0.54
Quan et al. [45]	0.910	0.56	0.885	0.67	0.893	0.58	0.896	0.60
Hati et al. [19]	0.777	0.33	0.621	0.43	0.738	0.20	0.712	0.32
Chang et al. [7]	0.726	0.27	0.759	0.36	0.797	0.36	0.761	0.33
Lee et al. [33]	0.528	0.36	0.647	0.42	0.701	0.39	0.623	0.39
Jerripothula et al. [26]	0.905	0.61	0.880	0.71	0.883	0.61	0.889	0.64
Jerripothula et al. [25]	0.818	0.48	0.847	0.69	0.813	0.50	0.826	0.56
Hsu et al. [22]	0.936	0.66	0.914	0.79	0.876	0.59	0.909	0.68
Chen et al. [10]	**0.941**	0.65	**0.940**	**0.82**	**0.922**	**0.63**	**0.935**	**0.70**
Ours + Unsupervised segments								
transformer	**0.941**	0.66	0.919	0.79	0.887	0.57	0.916	0.67
Nc-Net	0.682	0.19	0.791	0.56	0.774	0.27	0.749	0.34
Ours + COCO segments [38]								
transformer	**0.941**	**0.67**	0.928	**0.82**	0.916	0.60	0.928	**0.70**
Nc-Net	0.655	0.23	0.857	0.61	0.873	0.43	0.795	0.42

* learned with strong supervision (i.e., manually annotated object masks)

Co-segmentation results in Internet dataset for the Horse, Airplane and Car categories.

5. Training pairs

(a) Source Image (b) Target Image (c) Stylised Source Image (d) Stylised Target Image

Project page (code, more experimental and visual results): http://imagine.enpc.fr/~shenx/SegSwap/