Local Shape Transfer for Image Co-segmentation

Wei Teng*[1]
tengw@buaa.edu.cn
Yu Zhang*[1]
zhangyulb@gmail.com
Xiaowu Chen†[1]
chen@buaa.edu.cn
Jia Li[12]
jiali@buaa.edu.cn
Zhiqiang He[3]
lirong2@lenovo.com

[1] State Key Laboratory of Virtual Reality Technology and Systems
Beihang University
Beijing, China

[2] International Research Institute for Multidisciplinary Science
Beihang University
Beijing, China

[3] Lenovo Research

Abstract

Image co-segmentation is a challenging computer vision task that aims to segment all pixels of the common objects in an image set. In real-world cases, however, the common objects often vary greatly in poses, locations and scales, making their global shapes highly inconsistent across images and difficult to be segmented. To address this problem, this paper proposes a novel co-segmentation approach that transfers patch-level local object shapes, which appear more consistently across different images. In our approach, we first employ dense correspondences to construct a patch neighbourhood system, which is refined using Locally Linear Embedding. Based on the patch relationships, an efficient algorithm is developed to jointly segment the objects in each image while transferring their local shapes across different images. Experiments show that our approach performs comparably with or better than the state-of-the-arts on iCoseg dataset [2], while achieving more than 31% relative improvements on a challenging benchmark Fashionista [31].

1 Introduction

Image co-segmentation is a young yet widely pursued topic in computer vision. In a word, it aims to segment all pixels of the common objects from a collection of images. With this tool, many high-level visual understanding tasks would be greatly facilitated, such as visual concept discovery [5] and fine-grained object recognition [14].

To extract the common objects, existing studies have explored various object cues. Among them, appearance cues such as color and texture are most popular due to their effectiveness for separating the foreground from the background. After extracting appearance descriptors in each image, the common objects can be discovered by either learning a shared descriptor distribution [10] or building correspondences among the descriptors in different images [28].

*The first two authors contribute equally. † correspondence should be addressed to Xiaowu Chen.

Pages 3.1-3.12
DOI: https://dx.doi.org/10.5244/C.30.3

Figure 1: The motivation of this paper. The common objects in these images have different poses, rendering their global shapes inconsistent. However, the local object shapes in different images are highly consistent and provide important cues for co-segmentation.

Despite the successes, appearance cues may fail to distinguish between the visually similar foreground and background regions, and often have difficulties in handling object categories with complex appearances.

To address these issues, some works explored shape cues for image co-segmentation, as they are reliable for removing foreground/background ambiguities [32]. In these works, the common objects in different images are assumed to be "generated" by a single shape model, which can take the form of shape priors [1, 16] or deformable shape templates [6]. However, due to the large variance of viewpoints, scales and object poses, global shapes of the common objects are often inconsistent and difficult to capture. As a result, template-based approaches may work less well in such scenario.

We observe that although the common objects may have inconsistent global shapes, their local shapes are often highly consistent and thus transferable (see Fig. 1). Based on this observation, this paper proposes a novel framework for image co-segmentation that transfers patch-level object shapes across different images. For each image patch, our approach seeks for transferable patch neighbours through image correspondences. To prune out unreliable matches, we further learn a sparser neighbourhood system for the image set using Locally Linear Embedding [22]. Given the patch correspondences, an efficient algorithm is proposed to incorporate patch-level consistencies into graph-cut based energy for jointly segmenting the common objects in each image.

The main contributions of this paper include 1) a novel framework that introduces local shape transfer for image co-segmentation, 2) a strategy for refining patch correspondences in an image set through Locally Linear Embedding, and 3) an algorithm that integrates patch consistencies into graph-cut based energy. Experiments show that our approach performs comparably or better than the state-of-the-arts on iCoseg [2] dataset. On challenging Fasionista [31] dataset with complex object appearance and pose, the proposed approach achieves more than 26% relative improvements over the leading co-segmentation approaches.

2 Related Work

Existing studies for image co-segmentation can be roughly categorized into matching-based and template-based groups. We briefly review them, while also discuss some tightly correlated shape transfer methods.

Template-based group assumes that there exists a single model that generalizes to represent all common objects in different images. Following this idea, some works proposed to learn shared distributions of appearance features. For example, Jolin *et al*. [10] learned linear models jointly for foreground and background based on color and texture features. Russell *et al*. [25] treated foreground objects as latent topics sharing similar visual words. Kim *et al*. [12] founds the latent foreground representations with a diffusion process. However, these models are often insufficient to capture object categories with complex appearances. To address this issue, several works [1, 6, 16] advocated using shape models to facilitate co-segmentation. A common practice is to learn shape prior maps, which indicates the likelihoods of the common objects appearing at different image locations. As object shapes are actually unknown in co-segmentation, the shape priors were iteratively refined using the current segmentations [1, 16]. In [6], a sophisticated model was designed to jointly segment the common objects and learn their deformable shape templates.

Template-based approaches are powerful since they output not only the segmented objects but also the learned foreground/background/shape models. However, the used models are often simple for tractability during learning or inference, thus may not adequately capture real-world object categories with various appearances and structures.

Matching-based group builds the region correspondences among different images. Some works enforced the matching constraint at object-level, assuming that the foreground feature histograms aggregated on different images are similar [21, 29]. However, this strategy may have difficulty applying to object categories with large variabilities. Another idea is to select the object proposals in each image that match consistently as the common objects [27], but with the cost of supervised training. More recent works adopted local region correspondences for image co-segmentation. For example, Wang *et al*. [28] proposed to match regions in functional space. Rubio *et al*. [24] proposed a MRF formulation to jointly address object co-segmentation and region matching. After obtaining the correspondences, they transfer the foreground/background labels among the matched pixels/superpixels. Faktor and Irani [7] adopted structured matching to detect the common object parts in different images, through which "co-saliency" maps were generated to guide segmentation in each image. However, this strategy may suppress object parts that are not "co-salient" in the whole image set.

Our approach is also based on local region matching. However, we differ from [24, 28] in transferring labels at patch-level rather than point-level. In this manner, structured consistency is imposed to preserve the local object shapes during transfer. Compared with [7], our approach does not assume the "co-saliency" of the common objects in the whole image set. In contrast, we only assume the co-occurrence of a common object part in a sparse set of neighbouring image patches. As a result, the proposed approach can effectively identify the whole foreground objects, as confirmed by the experiments.

Shape transfer is widely adopted for data-driven foreground/background segmentation. Most existing works proposed to transfer the masks of pre-segmented objects to the testing images, *e.g*., [26] and [15]. Beyond global object shapes, several works proposed to transfer local shape masks by sparse reconstruction [30] and non-parametric MRF [32] to handle local deformations. Our work is inspired by these successes, but operates in unsupervised manner without assuming pre-segmented images at hand.

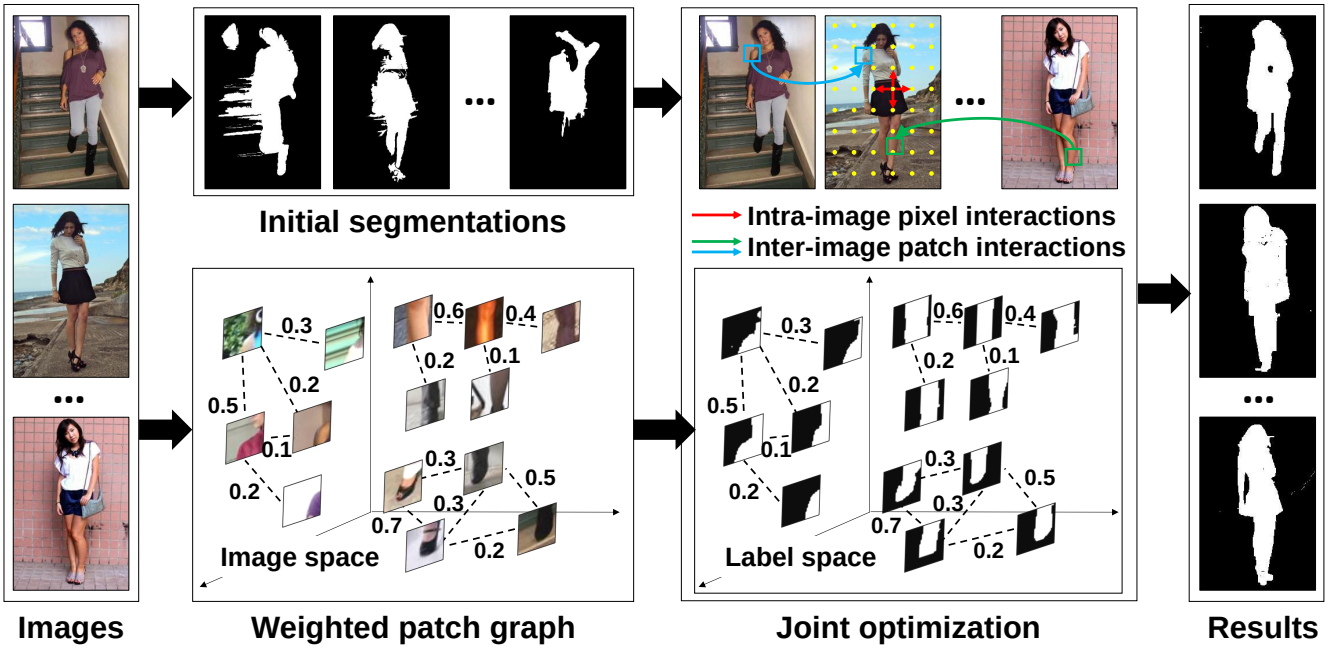

Figure 2: The framework of our approach. The initial foreground/background segmentation for each input image is estimated using [33]. Meanwhile, we construct a weighted graph among the patches sampled from different images using [13], where weights are learned by Locally Linear Embedding [22]. Finally, we optimize intra-image object segmentation and inter-image local shape transfer jointly while preserving the patch weights in label space.

3 Our Co-segmentation Framework

The pipeline of our approach is shown in Fig. 2. Given a set of M images, our framework first estimates coarse initial foreground segmentations by thresholding saliency maps [33]. Meanwhile, we build inter-image connections by constructing a weighted graph on patches, which implements local shape transfer. With the patch graph, the segmentations in all images are refined jointly. In the rest of this section, we first explain how local shape transfer helps image co-segmentation and the patch graph implementation, and then present the co-segmentation algorithm.

3.1 Local Shape Transfer for Image Co-segmentation

From the machine learning perspective, the global shapes of the common objects under different real-world conditions lie in high-dimensional space. In existing works, such shape spaces were often learned with sophisticated non-linear models (*e.g.*, random forests [18]). Our observation, however, is that the local object shapes can be well represented by their sparse neighbours in a linear and low-dimensional space. To implement this idea, we construct a neighbourhood system on image patches by finding dense pixel correspondences across images [13]. For the ith patch in a set of M images, the algorithm in [13] returns $M-1$ neighbouring patches, one in each different image. We denote the indices of these neighbours with $\mathcal{N}_i$.

Let $\vec{y}_i$ concatenate the binary segmentation labels in the ith patch, where 1 and 0 represents the foreground and background, respectively. Based on our assumption, the segmentation in the ith patch should be well reconstructed by its neighbours, *i.e.*, $\vec{y}_i \approx \frac{1}{M-1}\sum_{j\in\mathcal{N}_i}\vec{y}_j$.

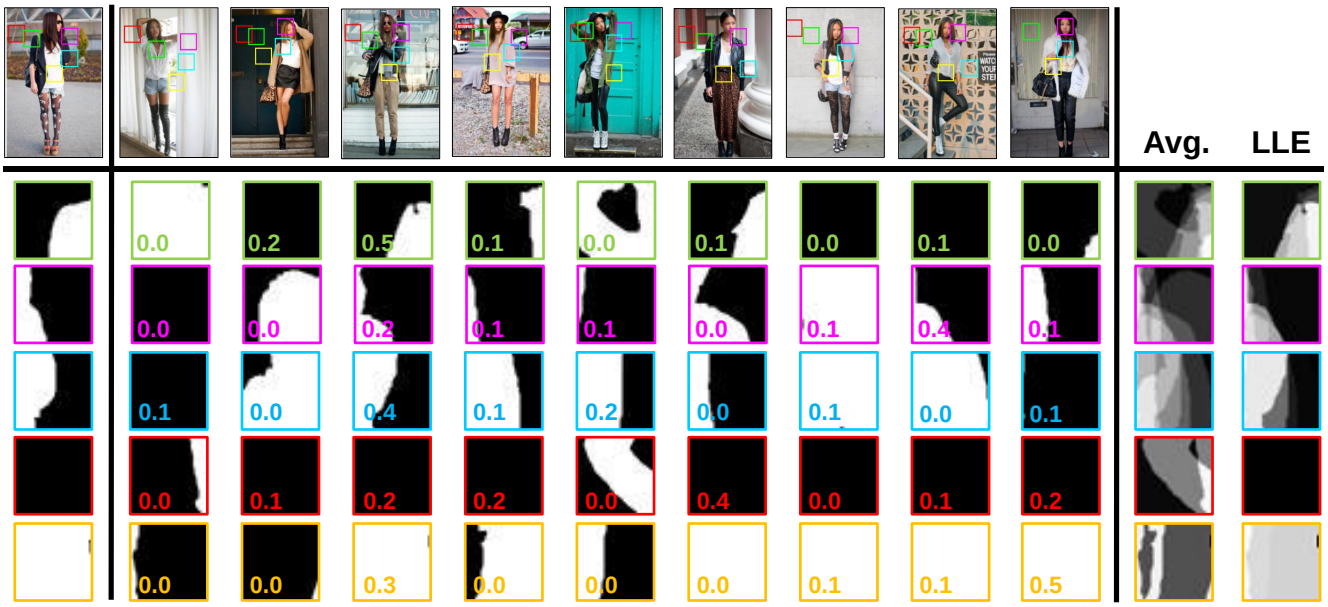

Figure 3: Illustration of local shape transfer. For the patches sampled from an image (left), we illustrate the neighbouring patches on different images (middle). Different colors represent different patches and their neighbours. The transferred segmentation mask using average pooling and Locally Linear Embedding are shown on the right. The learned weights at the bottom-left of each neighbouring patch show that Locally Linear Embedding is effective to suppress incorrect matches, which leads to more reliable shape transfer results.

This formulation assumes that the neighbouring patches are all good surrogates of the original image patch, which would be too ideal. As shown on the right of Fig. 3, aggregating the local shapes in many patches with inconsistent structures may confuse the shape transfer. To address this issue, we propose to learn a sparser but more reliable neighbouring relationships for each patch using Locally Linear Embedding [22]:

$$\min_{\mathbf{w}\succeq\mathbf{0}} \sum_{i=1}^{P} \left\| \vec{x}_i - \sum_{j\in\mathcal{N}_i} w_{ij}\vec{x}_j \right\|^2, \; s.t. \; \forall i, \; \sum_{j\in\mathcal{N}_i} w_{ij} = 1, \tag{1}$$

where P is the total number of patches sampled from the image set, $\vec{x}_i$ is the SIFT feature extracted from the ith patch. The simplex constraint imposes sparsity for neighbour selection. Given the learned neighbours, the local shapes are thus transferred by $\vec{y}_i \approx \sum_{j\in\mathcal{N}_i} w_{ij}\vec{y}_j$. Fig. 3 shows that this strategy leads to more consistent shape transfer results.

3.2 Co-segmentation with Local Shape Transfer

Given the patch graph, we refine the initial segmentation in each image by transferring the local shapes among different images. During transfer, the weights learned in the patch feature space are preserved when optimizing the label space. Formally, we minimize the objective

$$\min_{\mathbf{y}} \sum_{i=1}^{M} E_{\text{seg}}(\mathbf{y}^{[i]}) + \alpha \sum_{i=1}^{P} \left\| \vec{y}_i - \sum_{j\in N_i} w_{ij}\vec{y}_j \right\|^2, \; s.t. \; \mathbf{y} \in \{0,1\}^{|\mathbf{y}|}, \tag{2}$$

where $\mathbf{y}$ concatenates the foreground/background labels of all pixels in the image set, $\mathbf{y}^{[i]}$ is the part from the ith image. The energy E_{seg} implements intra-image foreground/background

segmentation, for which we use the popular Markov Random Field (MRF) energy, see [3] for details. The problem (2) is NP-hard and usually large scale as it operates on pixels. We propose an efficent algorithm to approximately solve it by half quadratic splitting [9].

Optimization algorithm. The core idea is to substitute the pixel labels $\mathbf{y}$ with auxiliary variable $\mathbf{z}$, while introducing additional constraints on patches as $\vec{z}_i = \vec{y}_i$. Relaxing this hard constraint, we have

$$\min_{\mathbf{y},\mathbf{z}} \sum_{i=1}^{M} E_{\text{seg}}(\mathbf{y}^{[i]}) + \alpha \sum_{i=1}^{P} \left\| \vec{z}_i - \sum_{j \in N_i} w_{ij} \vec{z}_j \right\|^2 + \lambda \sum_{i=1}^{P} \|\vec{z}_i - \vec{y}_i\|^2, \; s.t. \; \mathbf{y}, \mathbf{z} \in \{0,1\}^{|\mathbf{y}|}. \quad (3)$$

By iteratively solving $\mathbf{y}$ and $\mathbf{z}$ while keeping one of the them fixed, the original problem is decoupled into two simpler sub-problems. When $\mathbf{z}$ is fixed, the expanded sub-problem is

$$\min_{\mathbf{y}} \sum_{i=1}^{M} E_{\text{seg}}(\mathbf{y}^{[i]}) + \lambda \sum_{i=1}^{P} \left(\|\vec{y}_i\|^2 - 2(\vec{z}_i)^{\mathrm{T}} \vec{y}_i \right), \; s.t. \; \mathbf{y} \in \{0,1\}^{|\mathbf{y}|}. \quad (4)$$

Note that $\|\vec{y}_i\|^2 = \|\vec{y}_i\|$ as the patch labels are binary. Thus, the second term in (4) takes linear form *w.r.t.* $\mathbf{y}$, and can be directly merged into the unary potentials of the MRF energy. The remaining part is highly efficient by performing graph-cut [3] in each image in parallel.

The step of optimizing $\mathbf{z}$ requires solving a large-scale quadratic program. We discard the binary constraint, which results in closed-form solution of a quadratic program by solving a linear system. For efficiency, we approximate this solution by a sequence of label diffusions. In a diffusion step, the pixel labels $\vec{z}_i$ in the ith patch are optimized by fixing the labels of all other pixels. By setting the derivation *w.r.t.* $\vec{z}_i$ to zero, we obtain the following update rule

$$\vec{z}'_i = \frac{\alpha \left[\sum_{j \in \mathcal{N}_i} w_{ij} \vec{z}_j + \sum_{j:i \in \mathcal{N}_j} w_{ji} \left(\vec{z}_j - \sum_{k \in \mathcal{N}_j, k \neq i} w_{jk} \vec{z}_k \right) \right] + \lambda \vec{y}_i}{\alpha + \lambda + \sum_{j:i \in \mathcal{N}_j} w_{ji}^2}, \quad (5)$$

which can be written in compact form

$$\mathbf{Z}' = \{\lambda \mathbf{Y} + \alpha \left[\mathbf{W} + \mathbf{W}^{\mathrm{T}} - \mathbf{W}^{\mathrm{T}}\mathbf{W} + \mathrm{Diag}(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \right] \mathbf{Z}\} \left[(\alpha + \lambda)\mathbf{I} + \mathrm{Diag}(\mathbf{W}^{\mathrm{T}}\mathbf{W}) \right]^{-1}, \quad (6)$$

where the matrices $\mathbf{Z}$, $\mathbf{Z}'$ and $\mathbf{Y}$ concatenate in a row the column vectors $\vec{z}_i$, $\vec{z}'_i$ and $\vec{y}_i$, respectively, $\mathbf{W}$ is a $P \times P$ pairwise matrix of patch-wise neighbouring weights, and $\mathbf{I}$ is the identical matrix. The operator $\mathrm{Diag}(\cdot)$ creates a diagonal matrix by picking out the diagonal elements of the input matrix. We found 15 iterations of (6) to be adequate in practice. After diffusion, we normalize the soft labels $\mathbf{z}$ into $[0,1]$ separately for each image.

The two steps are repeated until near-convergence. Empirically, we terminate the optimization in 10 iterations and take the last discrete labels $\mathbf{y}$ as the final segmentations.

Implementation details. Input images are resized to have around 60000 pixels, on which we sample 17×17 patches uniformly with a stride of 5 pixels. The unary term of the MRF energy E_{seg} is the (log-negative) foreground/background color likelihoods generated by 12-components GMM models. Initially, the GMMs are learned on saliency-based segmentations. In each iteration, we update them using the latest segmentations. We follow [4] to define the pairwise term, which models color contrasts between adjacent pixels. Parameters α and λ are empirically set to 1 and 0.3, respectively.

iCoseg	Ours	[8]	[28]	[24]	[27]	[19]	[10]
Alaska Bear	0.861	**0.935**	0.904	0.864	0.900	-	0.748
Red Sox Players	**0.972**	0.965	0.942	0.905	0.909	0.957	0.730
Stonehenge1	**0.936**	0.930	0.925	0.873	0.633	0.927	0.566
Stonehenge2	0.844	0.835	0.872	0.884	**0.888**	0.849	0.860
Liverpool	0.905	**0.921**	0.894	0.826	0.875	-	0.764
Ferrari	0.892	0.917	**0.956**	0.843	0.899	0.900	0.850
Taj Mahal	0.878	0.887	0.926	0.887	0.911	**0.941**	0.737
Elephants	**0.961**	0.904	0.867	0.750	0.431	0.877	0.701
Pandas	0.835	0.812	0.886	0.600	**0.927**	0.928	0.840
Kite	**0.980**	0.966	0.939	0.898	0.903	0.946	0.870
Kite panda	0.905	0.838	0.931	0.783	0.902	**0.934**	0.732
Gymnastics	**0.984**	0.954	0.904	0.871	0.917	0.922	0.909
Skating	0.893	0.817	0.787	0.768	0.775	**0.966**	0.821
Hot Balloons	**0.969**	0.965	0.904	0.890	0.901	0.952	0.852
Liberty Statue	0.966	0.927	**0.968**	0.916	0.938	0.966	0.906
Brown bear	0.938	0.948	0.881	0.804	**0.953**	0.885	0.740
Average	**0.920**	0.907	0.905	0.839	0.853	-	0.789

Table 1: Comparison with leading co-segmentation approaches of correctly classified pixels on iCoseg dataset. All the numbers are taken from the original papers except [10], which is taken form [28].

4 Experimental Results

We evaluate the proposed approach on two public benchmarks:

The iCoseg dataset [2] contains 643 images of 38 object classes with pixel-level annotations. In each class, images have similar color but varying locations and scales. We test on a subset of 16 classes which are widely used by the leading co-segmentation approaches. For each class, all images are used for co-segmentation.

The Fashionista dataset [31] contains 685 street photographs of fashion models. In contrast to conventional co-segmentation datasets, it is extremely challenging with various human poses, background clutters and complex appearances. As existing co-segmentation approaches may have difficulty handling large amounts of images, we randomly partition the dataset into 23 groups with nearly 30 images per group. Evaluations are averaged over 10 random partitions.

We use two evaluation protocols: the ratio of correctly classified pixels for iCoseg, and the Intersection-over-Union overlap ratio for Fashionista. The former is chosen for throughout comparison with previous approaches, although the latter is more preferred as it was shown unbiased to the object size [17].

4.1 Comparison with State-of-the-Arts

The results are summarized in Table 1, 2 and 3, respectively. In Table 1, our approach obtains the best overall performance with leading accuracies on $6/16$ categories. We improve remarkably on challenging categories *elephants* and *gymnastics*, on which most previous

Ours	[7]	[6]	[23]
0.920	0.944	0.895	0.896

Table 2: Percentages of correctly classified pixels on iCoseg dataset.

Ours	[7]	[6]	[10]	[20]
0.756	0.501	0.576	0.358	0.642

Table 3: Intersection-over-Union overlap ratios on Fashionista dataset.

approaches work less well. As these object categories have large pose variance, the proposed local shape transfer strategy may handle them better.

Our approach outperforms other approaches based on local region matching [24, 28]. We believe that it is the patch-level structured consistency that makes difference. We also obtain better results than [8, 27, 29], although they used external training data. Note that [19] performs quite well on the reported 14 classes, achieving 92.49% average accuracy, while our approach obtains 92.52%. However, they also rely on training images to learn dictionaries while our approach is unsupervised. In Table 2, our approach performs better than [6, 23] and comparably with [7], which reports the best performance so far on iCoseg dataset. However, all of [6, 7, 23, 28] and our approach can locate the common objects quite well on this dataset, while the main differences are mainly due to finer localization of object boundaries. Some effective practices for co-segmentation that are missing in our current implementation (*e.g.*, multi-scale reasoning and joint GrabCut [7, 16]) may further improve our results on several categories, *e.g.* *Alaska bear* and *panda*, where the objects exhibit extremely inconsistent scales and viewpoints but similar colors.

To show the weaknesses of existing co-segmentation approaches and clarify our contributions, we apply the state-of-the-arts [6, 7, 10] on the Fashionista [31] dataset using the released codes. We also compare with a GrabCut [20] baseline using a bounding box with 8 pixels margin from the image borders. Evaluations are summarized in Table 3, where the numbers of [6, 7, 10] and our approach are averaged on all groups, while the number of the GrabCut baseline is directly taken from [32].

Table 3 shows that the leading co-segmentation approaches have difficulty generalizing well to this dataset. Our approach obtains promising performance on both iCoseg and Fashionista datasets. Notably, we obtain 51%, 31% and 111% relative improvements over [7], [6] and [10] on Fashionista, respectively. Due to the complexity and large variance of object appearance and pose, the template-based approaches [6, 10] may have difficulty learning a proper template to represent the object category, while [7] often detects incomplete object shapes and misses important object details. On the contrary, the proposed local shape transfer strategy can well handle the appearance and pose variances on this dataset. See Fig. 4 for visual comparisons.

Running time. Our approach takes around 50 minutes to process 30 images with resolution 300×200. Saliency estimation can be done in a few seconds. Building correspondences, learning graph weights and optimization take 40, 0.5 and 7 minutes, respectively. Thus, large speed improvement can be expected when integrating faster image matching algorithms. Empirical comparisons show that the current implementation runs faster than several state-of-the-arts [6, 7], which take more than one hour.

4.2 Sensitiveness Analysis

To study the sensitiveness of our approach, we conduct two additional experiments on Fashionista dataset. In the first experiment, we evaluate the effects of local shape transfer for image co-segmentation by sampling the weight α (see (2)) uniformly in log scale, and sum-

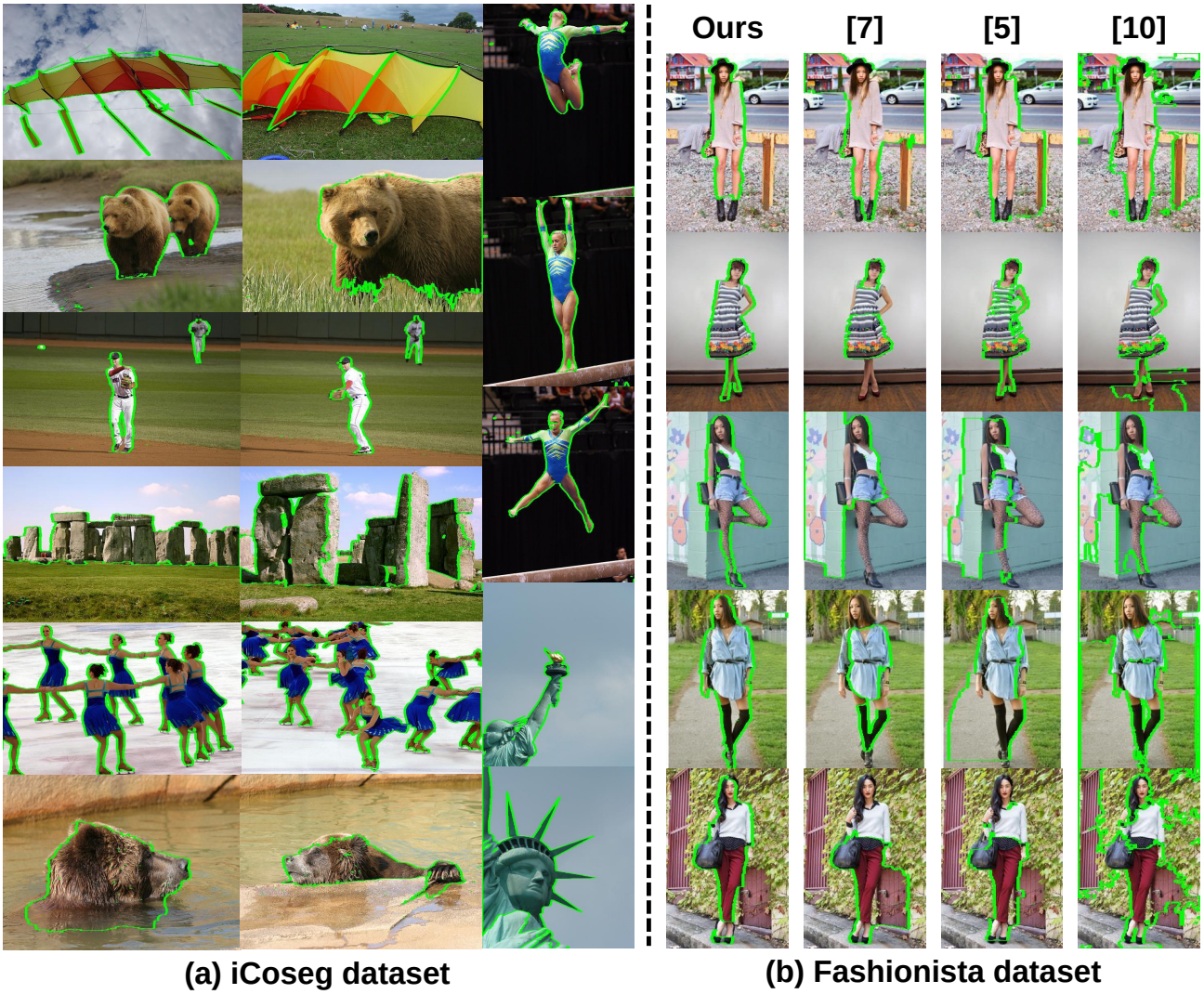

Figure 4: (a) Representative segmentations of our approach on iCoseg dataset. (b) Visual comparisons among different co-segmentaion approaches on Fashionista dataset.

marize the results in Fig. 5 (a). Note that when $\alpha = 0$, no shape transfer is employed and this variant can be seen as grab-cut with saliency cues. It is observed that non-zero assignment of α significantly improves the results, confirming the effectiveness of local shape transfer. It also shows that the results are relatively stable when $\alpha \geq 1$.

In the second experiment, we investigate the segmentation accuracy as a function of the number of images. To this end, we randomly select 10 images from a 50-image subset of Fashionista, and incrementally adds images at random to see the performance when the number of images increases. We repeat this step for 20 times and summarize the averaged accuracies at each number of images in Fig. 5 (b). As expected, the accuracy increases with more images, which provide more collective cues for co-segmentation. Note that the accuracy converges fast when $N \geq 20$, suggesting that a smaller number of images already enables our approach to effectively capture the inter-image relationships on this subset.

5 Conclusions

This paper proposes a novel approach for image co-segmentation. Compared with existing approaches, it does not assume a global model to represent the common objects but

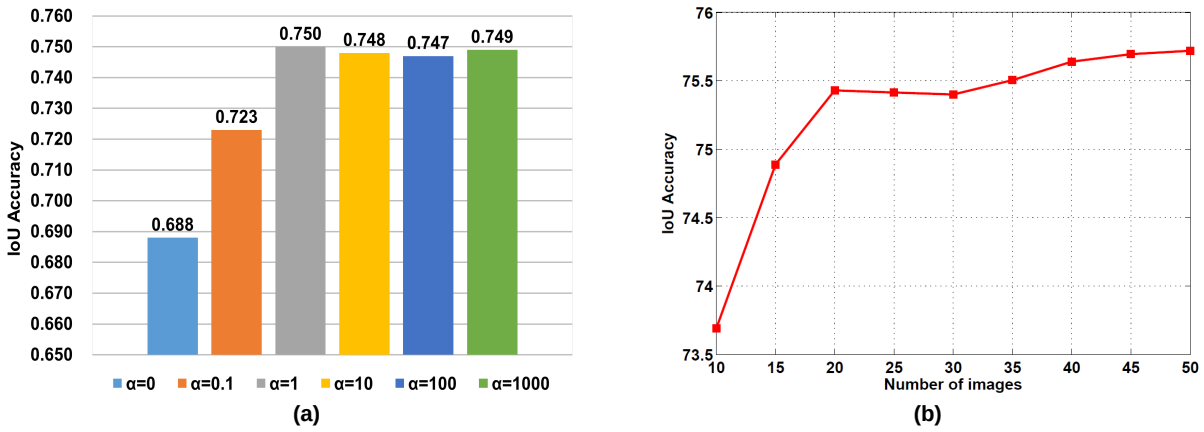

Figure 5: Sensitive analysis of the proposed approach. We show the segmentation accuracy as a function the weight for shape transfer α (left) and the number of images N (right).

transfer their local shapes among different images. To this end, our approach constructs a reliable patch neighbourhood system, and incorporates label consistencies among neighbouring patches in different images. Compared with the state-of-the-arts, our approach performs better or comparably on iCoseg dataset [2], while substantially better on the challenging Fashionista [31] dataset. For improvement, it is interesting to integrate multi-scale strategy into our approach, or extend it for multi-foreground object co-segmentation [11].

Acknowledgement

We thanks the reviewers for their valuable feedback. This work is supported in part by grants from NSFC (61325011) & (61421003), SRFDP (20131102130002) and Lenovo Outstanding Young Scientists Program (LOYS).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Classcut for unsupervised class segmentation. In *ECCV*, pages 380–393, 2010.

[2] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, pages 3169–3176, 2010.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[4] W. Casaca, L. G. Nonato, and G. Taubin. Laplacian coordinates for seeded image segmentation. In *CVPR*, pages 384–391, 2014.

[5] X. Chen, A. Shrivastava, and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *CVPR*, pages 2035–2042, 2014.

[6] J. Dai, Y. N. Wu, J. Zhou, and S. C. Zhu. Cosegmentation and cosketch by unsupervised learning. In *CVPR*, pages 1305–1312, 2013.

[7] A. Faktor and M. Irani. Co-segmentation by composition. In *ICCV*, pages 1297–1304, 2013.

[8] H. Fu, D. Xu, S. Lin, and J. Liu. Object-based RGBD image co-segmentation with mutex constraint. In *CVPR*, 2015.

[9] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(3):367–383, 1992.

[10] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, pages 1943–1950, 2010.

[11] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In *CVPR*, pages 837–844, 2012.

[12] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. pages 169–176, 2011.

[13] J. Kim, C. Liu, F. Sha, and K. Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *CVPR*, pages 2307–2314, 2013.

[14] J. Krause, H. Jin, J. Yang, and L. Fei-Fei. Fine-grained recognition without part annotations. In *CVPR*, pages 5546–5555, 2015.

[15] D. Kuettel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR*, pages 558–565, 2012.

[16] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in ImageNet. In *ECCV*, pages 459–473, 2012.

[17] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, pages 2192–2199, 2013.

[18] X. Liu, M. Song, D. Tao, J. Bu, and C. Chen. Random geometric prior forest for multiclass object segmentation. *IEEE Trans. Image Processing*, 24(10):3060–3070, 2015.

[19] L. Mukherjee, V. Singh, J. Xu, and M. D. Collins. Analyzing the subspace structure of related images: Concurrent segmentation of image sets. In *ECCV*, pages 128–142. Springer, 2012.

[20] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[21] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *CVPR*, pages 993–1000, 2006.

[22] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[23] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, pages 1939–1946, 2013.

[24] J. C. Rubio, J. Serrat, A. LÂt'opez, and N. Paragios. Unsupervised co-segmentation through region matching. In *CVPR*, pages 749–756, 2012.

[25] B. C. Russell., W. T. Freeman, A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, pages 1605–1614, 2006.

[26] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, pages 3001–3008, 2013.

[27] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *CVPR*, pages 2217–2224, 2011.

[28] F. Wang, Q. Huang, and L. J. Guibas. Image co-segmentation via consistent functional maps. In *CVPR*, pages 849–856, 2013.

[29] Z. Wang and R. Liu. Semi-supervised learning for large scale image cosegmentation. In *ICCV*, pages 393–400, 2013.

[30] W. Xia, C. Domokos, J. Xiong, L. F. Cheong, and S. Yan. Segmentation over detection via optimal sparse reconstructions. *IEEE Trans. Circuits Syst. Video Technol.*, 25(8): 1295–1308, 2015.

[31] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

[32] J. Yang, B. Price, S. Cohen, Z. Lin, and M. H. Yang. PatchCut: Data-driven object segmentation via local shape transfer. In *CVPR*, pages 1770–1778, 2015.

[33] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, and B. Price. Minimum barrier salient object detection at 80 fps. In *ICCV*, pages 1404–1412, 2015.