The Visual Computer
https://doi.org/10.1007/s00371-017-1467-5

ORIGINAL ARTICLE

Co-segmentation for space-time co-located collections

Hadar Averbuch-Elor[1] · **Johannes Kopf**[2] · **Tamir Hazan**[3] · **Daniel Cohen-Or**[4]

Abstract

We present a co-segmentation technique for space-time co-located image collections. These prevalent collections capture various dynamic events, usually by multiple photographers, and may contain multiple co-occurring objects which are not necessarily part of the intended foreground object, resulting in ambiguities for traditional co-segmentation techniques. Thus, to disambiguate what the common foreground object is, we introduce a weakly supervised technique, where we assume only a small seed, given in the form of a single segmented image. We take a distributed approach, where local belief models are propagated and reinforced with similar images. Our technique progressively expands the foreground and background belief models across the entire collection. The technique exploits the power of the entire set of image without building a global model, and thus successfully overcomes large variability in appearance of the common foreground object. We demonstrate that our method outperforms previous co-segmentation techniques on challenging space-time co-located collections, including dense benchmark datasets which were adapted for our novel problem setting.

Keywords Image co-segmentation · Foreground extraction · Non-rigid and deformable motion analysis · Belief propagation

1 Introduction

Nowadays, *Crowdcam* photography is both abundant and prevalent [1,2]. A crowd of people capturing various events form collections with great variety in content. However, they normally share a common theme. We refer to a collection of images that was captured about the same time and space as "Space-time Co-located" images, and we assume that such a co-located collection contains a significant subset of images that share a common foreground object, but other objects may also co-occur throughout the collection. See Fig. 1 for such an example, where the Duchess of Cambridge is photographed in her wedding, and some of the images contain, for instance, her husband, Duke of Cambridge.

Foreground extraction is one of the most fundamental problems in computer vision, receiving ongoing attention for several decades now. Technically, the problem of cutting out the common foreground object from a collection of images is known and has been referred to as co-segmentation [6,9,26]. A traditional co-segmentation problem assumes that objects which are both co-occurring and salient necessarily belong to the foreground regions. However, the space-time co-location of the images leads to a more challenging setting, where the premise of common co-segmentation techniques is no longer valid, as the foreground object is not well-defined. Therefore, we ask the user to provide a segmented template image to specify what the intended foreground object is.

The foreground object varies considerably in appearance across the entire space-time co-located collection. Thus, we do not use a single *global* model to represent it, but instead take a distributed *local* approach. We decompose each image

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00371-017-1467-5) contains supplementary material, which is available to authorized users.

✉ Hadar Averbuch-Elor
hadar.a.elor@gmail.com; averbuch1@mail.tau.ac.il

Johannes Kopf
jkopf@fb.com

Tamir Hazan
tamir.hazan@technion.ac.il

Daniel Cohen-Or
dcor@tau.ac.il

[1] Electrical Engineering School, Tel Aviv University, Tel Aviv, Israel

[2] Facebook, Seattle, WA, USA

[3] Faculty of Industrial Engineering and Management, Technion, Haifa, Israel

[4] Computer Science School, Tel Aviv University, Tel Aviv, Israel

Published online: 03 January 2018

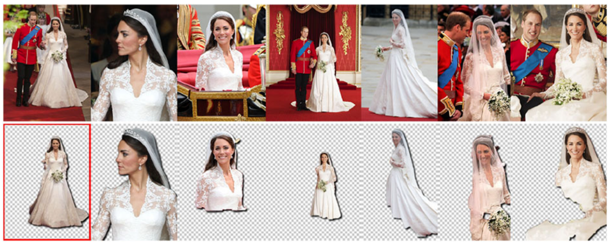


Fig. 1 The appearance of the Duchess of Cambridge varies throughout the images that capture her wedding ceremony. Starting from a single image template (marked with a red border), our method progressively expands the foreground belief model across the entire collection

into parts at multiple scales. Parts store local beliefs about the foreground and background models. These beliefs are iteratively propagated to similar parts within and among images. In each iteration, one image is selected as the current seed. See Fig. 2 which illustrates the progression of beliefs in the network of images.

The propagation of beliefs from a given seed is formulated as a convex belief propagation (CBP) optimization. Foreground and background likelihood maps of neighboring images are first inferred independently (see Sect. 4.2). These beliefs are then reinforced across images to consolidate local models and thus allow for more refined likelihood estimates (see Sect. 4.3). To allow for a joint image inference, we extend the CBP algorithm to include quadratic terms.

We show that when starting from a reliable seed model, we can progressively expand the foreground belief model across the entire collection. This gradual progression succeeds to co-segment the collection, outperforming state-of-the-art co-segmentation techniques on rich benchmark datasets which were adapted for our problem setting. We also provide an extensive evaluation on various space-time co-located collections which contain repeated elements that do not necessarily belong to the semantic foreground region. Our analysis demonstrates the advantages of our technique over previous methods, and in particular illustrates its robustness against significant cluttered backgrounds.

Explicitly stated, the main contributions of our work are the introduction of the novel co-segmentation problem for space-time co-located image collections, and technically, a distributed approach that can handle the great variability in appearance of the foreground object.

2 Related work

Segmenting and extracting the foreground object from an image is a fundamental and challenging problem, which has received significant ongoing attention. Extracting the foreground object requires some guidance or supervision since in most cases it is unclear what the semantic intent is. When several images that share a common foreground are given, the problem is referred to as co-segmentation [25]. Many solutions have been proposed to the co-segmentation problem, which can be applied to image collections of varying sizes and characteristics [5,6,9,17,26]. Co-segmentation techniques learn the appearance commonalities in the collection to infer the common foreground object or objects. To initialize the learning process, unsupervised techniques are usually based on objectness [28] or visual saliency [6,27] cues to estimate the target object.

State-of-the-art co-segmentation methods are based on recent advancements in feature matching and correspondence techniques [9,26,27]. Additional cues may also be considered, such as depth in the co-segmentation work of Fu et al. [11]. Rubinstein et al. [26] proposed to combine saliency and dense correspondences to co-segment large and noisy internet collections. Faktor and Irani [9] also use dense correspondences; however, they compute statistical significance of the shared regions, rather than computing saliency separately per image. These techniques are unsupervised, and they assume that recurrent and unique segments necessarily belong to the object of interest. However, in many collections this is not the case, and some minimal semantic supervision is then required. Batra et al. [3], for example, aimed at *topi-*

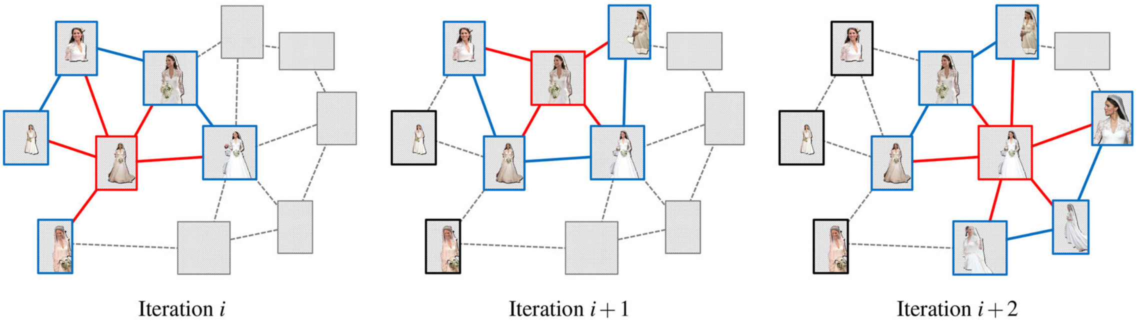


Fig. 2 Our technique iteratively propagates beliefs to images (framed in blue) which are adjacent to the current seed (framed in red). In each iteration, object likelihood maps are first inferred from the seed image to each one of its adjacent images (illustrated with red edges) and then these maps are propagated across similar images (illustrated with blue edges) to reinforce the inference

cally related images, and their supervision was given in the form of multiple user scribbles. In our work, we deal with images that belong to the same instance, and not to a general class, which exhibit great variability in appearance. We use the segmentation of a single image in the collection to guide the process and target the intended object.

The work of Kim and Xing [15] is most closely related to ours. In their work they address the problem of multiple foreground co-segmentation, where K objects of interest repeatedly occur over an entire image collection. They show promising results when roughly 20% percent of their images are manually annotated. In our work, we target a single object of interest, while space-time co-located collections often contain several repeated elements that clutter and distract common means to distinct the foreground. Unlike their global optimization framework, that solves for all the segmentations at once, our technique gradually progresses, and each image in turn guides the segmentation of its adjacent images. In this sense of progression, the method of Kuettel et. al [18] is similar to ours. However, their method is strongly based on the semantic hierarchy of ImageNet, while we aim at segmenting an unstructured space-time co-located image collection.

There are other space-time co-located settings where images share a common foreground. One setting is a video sequence [7,10,24,30], where the coherence among the frames is very high. It is worth noting that even then, the extraction of the foreground object is surprisingly difficult. Few recent techniques (e.g., [8,20]) address a multi-view setting, where multiple synchronized videos capture a certain scene. Kim and Xing [16] presented an approach to co-segment multiple photo streams, which are captured by various users at varying places and times. Similarly to our work, they also iteratively use a belief propagation model over the image graph. Another setting is multi-view object segmentation, e.g., [4,12], where the common object is captured from a calibrated set of cameras. These techniques commonly employ 3D reconstruction of the scene to co-segment the object of interest. In our setting, the images are rather sparse in scene-space and not necessarily captured all at once, which makes any attempt to reconstruct the target object highly improbable.

3 Co-segmentation using iterative propagation

We describe the foreground and background models, denoted by F and B, using local beliefs that are propagated within and across images. To define a topology over which we propagate beliefs, we construct a parts-level graph G^{p}, where nodes are image parts from all images, and edges connect corresponding parts in different images or spatially neighboring parts within an image. Furthermore, we define an associated image-level graph G^{i}, where the nodes correspond to the images, and two images are connected by an edge if there exists at least one edge in G^{p} that connects the pair of images. In Sect. 4.3, we describe in more detail how these inter-image connections are built. In short, they connect corresponding parts, as illustrated in Fig. 3. The F/B likelihoods are iteratively propagated throughout the part-level graph G^{p}, while the propagation flow is determined according to the image-level graph G^{i}. The graph topology is illustrated in Fig. 3.

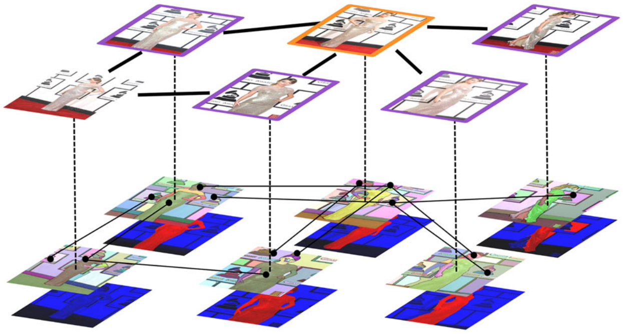


Fig. 3 The image-level graph G^{i} (on top) defines a topology of images over which local belief models are iteratively propagated. In each iteration, a seed image (marked with an orange border) propagates the F/B likelihood maps to its adjacent images (marked with a purple border). From these likelihood estimates, we extract the common foreground object (in red) and choose the next seed image

In what follows, we first explicitly define the graph topology. We then describe how these beliefs are gradually spread across the entire image collection, starting from the user-segmented template image.

3.1 Propagation graph topology

The basis for the propagation is image parts. To obtain the parts, we use the hierarchical image segmentation method of Arbeláez et al. [23]. We threshold the ultrametric contour map, which defines the hierarchy of image regions, at a relatively fine level ($\lambda_i = 0.15$). See Fig. 6 (on the left) for an illustration of the parts obtained at a number of different levels. The level we use for the image parts is illustrated in the left-most image. Although a fine level yields a large number and perhaps less meaningful parts, we would like to avoid issues concerning parts which are not accurate. A coarser level often merges between foreground and background parts.

We construct a parts-level graph G^{p}, where edges connect corresponding parts or spatially neighboring parts within an image. To compute reliable correspondences between image parts, we use the non-rigid dense correspondence technique (NRDC) [13], which outputs a confidence measure (with values between 0 and 1) along with each displacement value. We consider corresponding pixels to be those with a confidence which exceeds a certain threshold, which we set empirically to 0.5.

Two images are connected by an edge in the associated image-level graph G^{i} if there exists at least one edge in G^{p} that connects the pair of images.

3.2 Iterative likelihood propagation

We assign each part in G^{p} a foreground likelihood. Initially all parts are equally likely to be foreground or background (except the parts in the user-segmented template image, whose F-likelihood is either exactly 0 or 1).

The likelihoods are iteratively propagated throughout the graphs. In each iteration, a seed image is selected and its likelihoods are propagated to the adjacent neighbors in G^{i}. In the first iteration, the seed image is always the user-segmented template image. In subsequent iterations the seed image randomly picked from the neighbors of the current seed. Within an iteration, the seed image likelihoods are considered fixed. Note that the template image likelihoods remain fixed throughout the whole algorithm.

The details of this propagation are described in the next section. The new estimates are first derived separately, according to Sect. 4.2, and are then jointly refined, according to Sect. 4.3. These new likelihood estimates are combined with previous estimates, where estimates are amortized along their propagation, and get exponentially lower weights over time, as we have more confidence in beliefs that are closer to our template image.

After propagating the likelihoods, we update the foreground–background segmentation for the next seed using a modified implementation of graph-cuts [18], where the unary terms are initialized according to the obtained likelihoods.

The algorithm above is terminated once all images have been propagated to at least once. To avoid error accumulation, we execute the full pipeline multiple times (five in our implementation). The final results are obtained by averaging all the likelihood estimates follows by a graph-cut segmentation.

4 Likelihood inference propagation

Our algorithm uses convex belief propagation and further extends the variational approximate inference program to include quadratic terms. Therefore, in Sect. 4.1, we briefly introduce notations used in later sections. In Sect. 4.2, we present an approach to infer an object likelihood map of a single target image from a seed image. Finally, in Sect. 4.3, we introduce a technique to propagate the likelihood maps across similar images to improve the accuracy and reliability of these inferred maps.

4.1 Convex belief propagation

Markov random fields (MRFs) consider joint distributions over discrete product spaces $Y = Y_1 \times \cdots \times Y_n$. The joint probability is defined by combining potential functions over subsets of variables. Throughout this work we consider two types of potential functions: single variable functions, $\theta_i(y_i)$, which correspond to the n vertices in a graph, $i \in \{1, ..., n\}$, and functions over pairs of variables $\theta_{i,j}(y_i, y_j)$ that correspond to the graph edges, $(i, j) \in E$. The joint distribution is then given by Gibbs probability model:

$$p(y_1, ..., y_n) \propto \exp\Big(\sum_{i\in V}\theta_i(y_i) + \sum_{i,j\in E}\theta_{i,j}(y_i, y_j)\Big). \quad (1)$$

Many computer vision tasks require to infer various quantities from the Gibbs distribution, e.g., the marginal probability $p(y_i) = \sum_{y\setminus y_i} p(y_1, ..., y_n)$.

Convex belief propagation [14,29] is a message-passing algorithm that computes the optimal beliefs $b_i(y_i)$ which approximate the Gibbs marginal probabilities. Furthermore, under certain conditions, these beliefs are precisely the Gibbs marginal probabilities $p(y_i)$. For completeness, in the Supplementary Material, we define these conditions and explicitly describe the optimization program.

4.2 Single target image inference

In the following we present the basic component of our method, which infers an object likelihood map of a target image from an image seed. We construct a Markov random field (MRF) on the parts of the target image and use a convex belief propagation to infer the likelihood of these parts to be labeled as foreground.

Each part can be labeled as either foreground or background, i.e., $y_i \in \{-1, +1\}$. First, we describe the local potentials of each part $\theta_i(y_i)$, which describe the likelihood of the part to belong to the foreground or the background. Then, we describe the pairwise potentials $\theta_{i,j}(y_i, y_j)$, which account for the spatial relations between adjacent parts. We infer the foreground–background beliefs of the parts in the target image $b_i(y_i)$ by executing the standard convex belief propagation algorithm [14,29].

4.2.1 Local potentials

The local potentials $\theta_i(y_i)$ express the extent of agreement of a part with the foreground or background models. To define parts in the seed image, we use the technique of Arbeláez et. al [23] at multiple levels to obtain a large bag of candidate parts of different scales. Let i be a part in the target image, and s be a part in the source image seed. Then for each source part s, we compute its compatibility with a target part i, and denote it by $p_{\text{comp}}(i, s)$.

To construct the foreground/background likelihood of each part in the target image i, we consider the F/B parts of the source seed, and set

$$\theta_i(f) = \max_{s \in F} p_{\text{comp}}(i, s) \quad \text{and} \quad \theta_i(b) = \max_{s \in B} p_{\text{comp}}(i, s),$$

where f and b are the two labels that can be assigned to y_i.

We define our compatibility measure as follows:

$$p_{\text{comp}}(i, s) = p_{\text{corr}}(i, s) + \delta \cdot p_{\text{sim}}(i, s), \tag{2}$$

where δ is a balancing coefficient that controls the amount of *enrichment* of the available set of correspondences. The term $p_{\text{corr}}(i, s)$ measures the fraction of pixels that are matched between parts i and s. This is measured according to

$$p_{\text{corr}}(i, s) = N(i, s)\, |s|^{-1}, \tag{3}$$

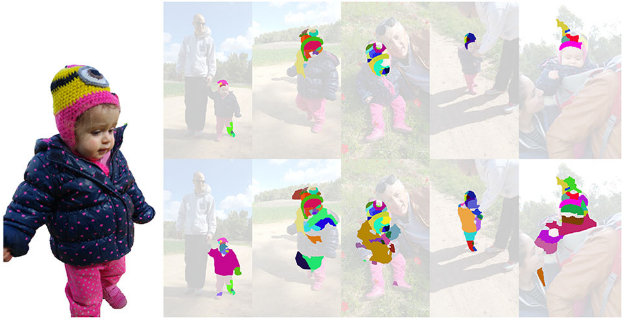


Fig. 4 Corresponding foreground parts of adjacent images in G^{i} according to p_{corr} only (top row) vs our enriched compatibility measure (bottom row) that contains significantly more compatible parts. The foreground source is displayed on the left

where $N(i, s)$ is the number of corresponding pixels, and $|s|$ is the number of pixels in part s. As mentioned before, the matching is based on NRDC.

We identified that highly compatible parts are rather sparse, and thereby $p_{\text{corr}}(i, s)$ is almost always zero in many source-target pairs. Nonetheless, we can exploit these sparse correspondences to discover new compatible parts with the term $p_{\text{sim}}(i, s)$. See Fig. 4 for an illustration of the compatible target parts in the foreground regions without (top row) and with (bottom row) our enrichment term $p_{\text{sim}}(i, s)$. In practice, since the background does not necessarily appear in both source and target image, $\delta > 0$ only for regions $s \in F$.

In these foreground regions, the term $p_{\text{sim}}(i, s)$ aims at revealing a similarity between parts whose appearance and spatial displacement highly agree. Similarity in appearance, in our context, is measured according to the Bhattacharyya coefficient of the RGB histograms, following the method of [21]. In order for parts i and s to highly agree in appearance, we further require that $i \in$ top-k(s), where the number of nearest neighbors is set to three. To recognize parts whose spatial displacement agree, we utilize the set of corresponding pixels in the foreground regions. We approximate the pixel values of the part corresponding to s according to the known correspondences. Formally, for each $s \in F$, let $i(s)$ be the estimated corresponding region in the target. Thus, for a similarity between parts i and s, we require that $i \cap i(s) \neq \emptyset$.

To simplify computations, we assume $i(s)$ to be a circle within the target image, which we compute according to the closest and farthest foreground correspondences. These two corresponding points define a relative scale between the two images. To compute the circle center, we compute the relative offset from the closest corresponding point (using the relative scale). The radius is determined according to the distance to the nearest corresponding point in the target. See Fig. 5a for an illustration of the estimated corresponding region $i(s)$.

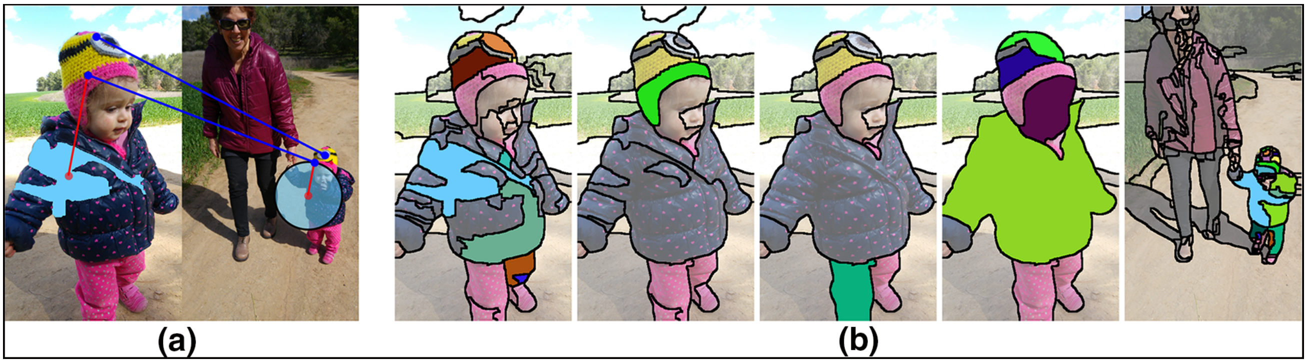
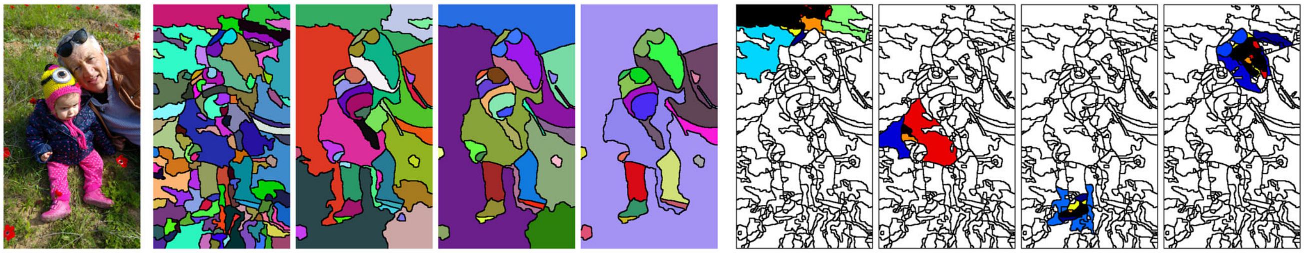


Fig. 5 **a** Based on two reliable correspondences (in blue), the relative offset (in red) to a part (light blue) defines the region where the correspondent part is expected (marked with a light blue circle). **b** The multi-scale parts of the source seed image are matched to the parts of the target image (on the right). The parts that yield maximum compatibility are highlighted in unique colors (corresponding to the highlighted parts of the target image on the right)



Target image Target parts in different scales Visualizing the spatial consistencies

Fig. 6 Given the image on the left, parts are obtained using hierarchical segmentation. For each target image in the collection, we obtain a large number of parts, as in the left-most image. On the right are visualizations of various parts (colored in black) and the potentials which are induced to their neighboring parts. The neighboring parts are colored according to their proximities which are expressed in Eq. (4) (warm colors correspond to strong proximities, while cool colors correspond to weaker proximities)

Putting it all together, $p_{\text{sim}}(i,s)$ is measured according to

$$p_{\text{sim}}(i,s) = \begin{cases} \sum_{u=1}^{16^3} \sqrt{\text{hist}(i)_u \cdot \text{hist}(s)_u} & i \in \text{top-k}(s),\, i \cap i(s) \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

In our experiments, we set $\delta = 0.1$ for all the foreground regions. See Fig. 5b for an illustration of the multi-scale source parts that obtained maximum compatibility with parts in the target image.

4.2.2 Pairwise potentials

The pairwise potential function $\theta_{i,j}(y_i, y_j)$ induces spatial consistency from the part generation process within the image. As previously mentioned, we obtain parts at multiple scales by thresholding at varying levels λ_i in the ultrametric contour map (see Fig. 6 for an illustration).

To measure spatial consistencies between adjacent parts in the target, we can compute how *quickly* these two parts merge into one by examining the level λ_{merge} where the two parts become one. Hence, we define a pairwise relation between adjacent parts in each target image according to:

$$\theta_{i,j}^{\text{intra}}(y_i, y_j) = \exp\left(-\tau\left(\lambda_{\text{merge}} - \lambda_{\text{min}}\right)\right) \cdot y_i y_j \qquad (4)$$

where the parameter $\tau = 4$ was set empirically, and the finest level we examine to measure the spatial consistencies is $\lambda_{\text{min}} = 0.2$ (a merge there would induce the strongest proximity between the parts). See the heat-maps in Fig. 6 for an illustration of $\theta_{i,j}^{\text{intra}}(y_i, y_j)$ on a few randomly chosen target parts. We empirically multiply the pairwise terms by a balancing coefficient of 0.05 to better balance between the local and pairwise terms involved in the single target image inference optimization.

4.3 Joint multi-target inference

In Sect. 4.2, we presented our approach to infer an object likelihood map from a seed image. In our setting, similar regions may co-occur across multiple images. Therefore, to improve the accuracy and reliability of the likelihood maps obtained by a single inference step, we propagate the inferred maps onto adjacent images in the image graph G^{i}. The output beliefs of each inferred target image are sent to its neighbors as a heat-map (i.e., per part foreground–background probability). Thus, our likelihood maps are complemented with joint inference across neighboring images. As stated earlier, neighboring images are images which are connected by an edge on the image graph G^{i}.

Algorithm 1 One propagation iteration

```
Input: $G^{\text{i}}$, $G^{\text{p}}$, and seed image $I^{seed}$
for each adjacent image $I$ do
    $b_i(y_i) \leftarrow BP(\theta_i(y_i), \theta_{i,j}^{intra}(y_i, y_j))$ {compute beliefs for $y_i \in I$}
    $\theta_i^*(y_i) \leftarrow b_i(y_i)$
end for
$b_i^*(y_i) \leftarrow BP(\theta_i^*(y_i), \theta_{i,j}^{b}(y_i, y_j))$ {compute beliefs jointly}
```

To differentiate the different types of edges on G^{p}, we denote the edges that connect parts across images by E^b. A joint inference is encouraged by a pairwise potential function between matched parts in E^b. Since the labels satisfy $y_i, y_j \in \{-1, +1\}$, this can be done with the potential function

$$\theta_{i,j}^{b}(y_i, y_j) = \Big(p_{\text{corr}}(i,j) + p_{\text{corr}}(j,i)\Big) y_i y_j \qquad (5)$$

Simply stated, the local potentials propagate the output beliefs of one target as input potentials of its neighboring images. Algorithm 1 describes how these output beliefs are refined through a joint inference. The initial beliefs, denoted in Algorithm 1 by $b_i(y_i)$, are computed for each adjacent image in the image graph G^{i} independently. These initial beliefs become the local potentials, denoted in Algorithm 1 by $\theta_i^*(y_i)$. A joint inference thus yields the output beliefs $b_i^*(y_i)$. Our intuition is that the output beliefs $b_i(y_i)$, which are concluded by running a convex belief propagation within its image, serve as a source seed signal to the neighboring image. We multiply the weights of the pairwise term by two to better balance between the two terms involved in the joint optimization. Our two-step approach introduces novel nonlinearities to the propagation algorithm. To control these nonlinearities, we extend the variational approximate inference program to include quadratic terms.

We also determine the conditions for which these quadratic terms still define a concave inference program and prove that repeated convex belief propagation iterations across images achieve its global optimum. We refer the interested reader to the Supplementary Material for more details on our extended variational approximate inference program.

5 Results and evaluation

We empirically tested our algorithm on various datasets and compared our segmentation results to two state-of-the-art unsupervised co-segmentation techniques [9,26] and to another semi-supervised co-segmentation technique [15]. The merit of comparing our weakly supervised technique with unsupervised ones is twofold; first, it serves as a qualitative calibration of the results on the new co-located setting, and second, it clearly demonstrates the necessity of some minimal supervision to define the semantic foreground region.

It should be noted that although [15] discuss an unsupervised approach as well, they only provide implementations for the semi-supervised approach. To be compatible to our input, we provide their method with one input template mask. We measure the performance on different datasets, including benchmark datasets that were adapted for our novel problem setting. Once our work is published, we will make the full implementation of our method publicly available, along with the datasets that were used in the experiments.

Space-time images We evaluated our technique on various challenging space-time co-located image collections depicting various dynamic events. Some of them (Bride, Singer and Broadway) were downloaded from the internet, while others (Toddler, Baby, Singer with Guitarist and Peru) were casually captured by multiple photographers. These images contain repeated elements that do not necessarily belong to the semantic foreground region, and the appearance of the foreground varies greatly throughout the collections. We provide thumbnails for the *full* seven sets, together with results and comparisons, in the Supplementary Material. We demonstrate results starting from various template images. Please refer to these results for assessing the high quality of our results.

For a quantitative analysis, we manually annotated the foreground regions of three of our collections (Bride, Toddler, and Singer), and report the precision P (percentage of correctly labeled pixels) and Jaccard similarity J (intersection over union of result and ground-truth segmentations) as in previous works (see Table 1). It should be noted that, to strengthen the evaluation, we perform three independent runs for the semi-supervised techniques, starting from different random seeds, and report the average scores. Figure 7 shows a sample of results, where the left-most image is provided as template for the semi-supervised techniques. As can be observed from our results, the unsupervised co-segmentation techniques fail almost completely on our co-located collections. Regarding the semi-supervised technique, as Fig. 7 demonstrates, when both the foreground and background regions highly resemble those of their counterparts in the given template, then the results of [15] are somewhat comparable to ours. As soon as the backgrounds differ or there are additional models that were not in the template, their method includes many outliers, as can be seen in Fig. 7. Unlike their method, we avoid defining strict global models that hold for all the images in the collection, and thus allow flexibility that is required to deal with the variability across the collection.

Table 1 Comparison against co-segmentation techniques on an annotated subset of our space-time co-located collections

	Bride		Singer		Toddler	
	P	J	P	J	P	J
[26]	47.3	16.6	28.9	16.6	49.6	25.7
[9]	71.1	42.9	68.4	30.4	81.8	44.1
[15]	63.9	27.4	88.8	63.4	66.7	38.9
Ours	88.9	76.3	94.2	83.0	94.5	74.2

Multiple foreground objects We also compared our performance to [15] using their data. We use their main example, which also corresponds to our problem setting. The results are displayed in Fig. 8 where we mark the multiple foreground objects in different colors. We execute our method multiple times with different seeds to meet their input. As we can see here and in general, our method has less false-positives and is more resistant to cluttered backgrounds. If we are able to spread our beliefs toward the target image, then we succeed in capturing the object rather well. Quantitatively, our technique cuts the precision error by more than half (from 6.89% down to 2.68%). However, if there is not enough confidence that reaches the target image, then the object remains undetected, as can be observed in the uncolored basket of apples in the rightmost image. This is the main weakness of our propagation technique. If meaningful connections do not exist, the beliefs fail to spread within the collection.

Sampled video collections The DAVIS dataset [22] is a recent benchmark for video segmentation techniques, containing 50 sequences that exhibit various challenges including occlu-

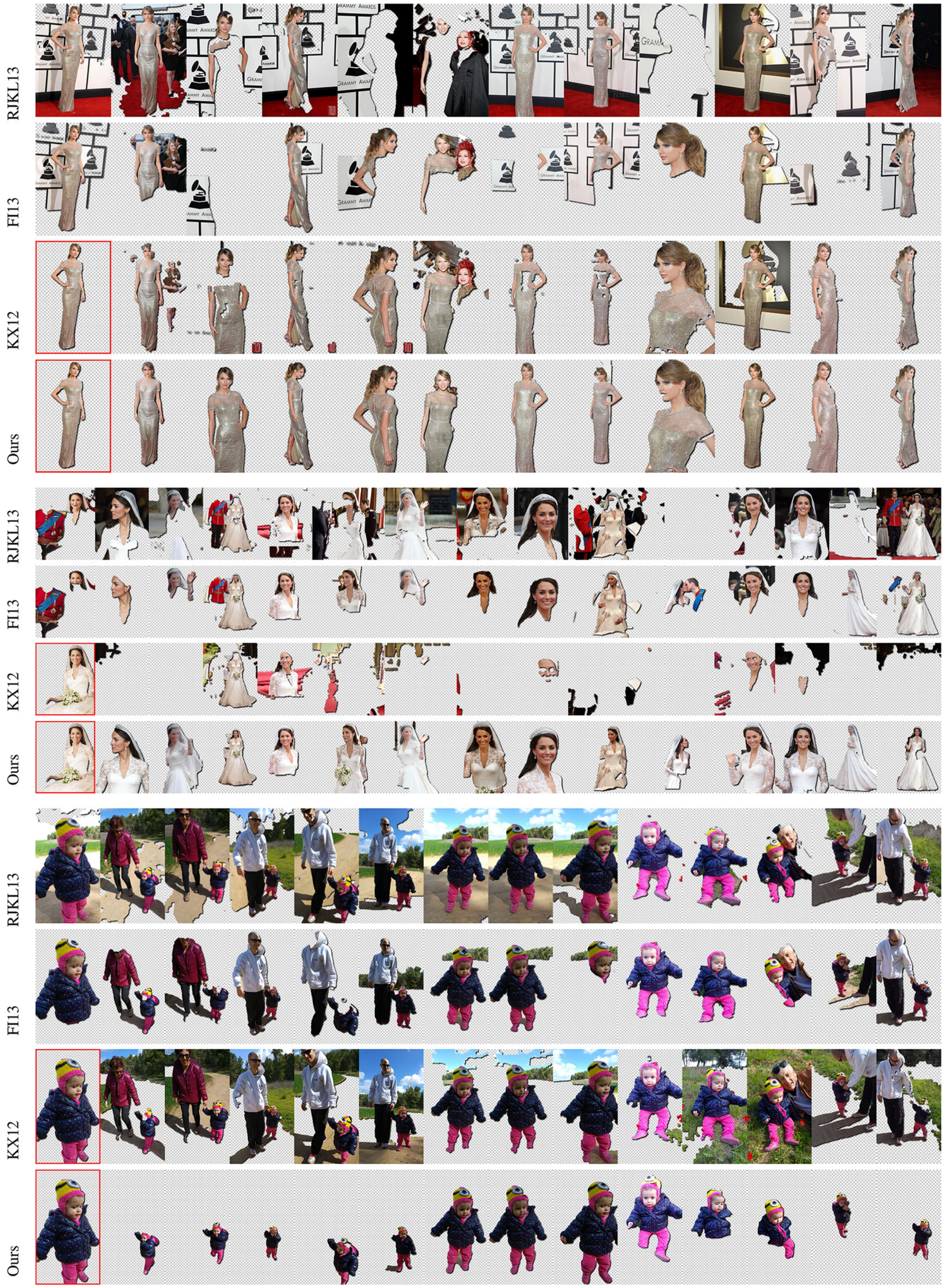


Fig. 7 Comparison to state-of-the-art co-segmentation techniques. The top three rows illustrate the results obtained by the RJKL13 [26], FI13 [9], and KX12 [15], respectively. The bottom row illustrate our results. For both our technique and [15], the left-most image is provided as template. Please refer to the Supplementary Material for an extensive and interactive comparison

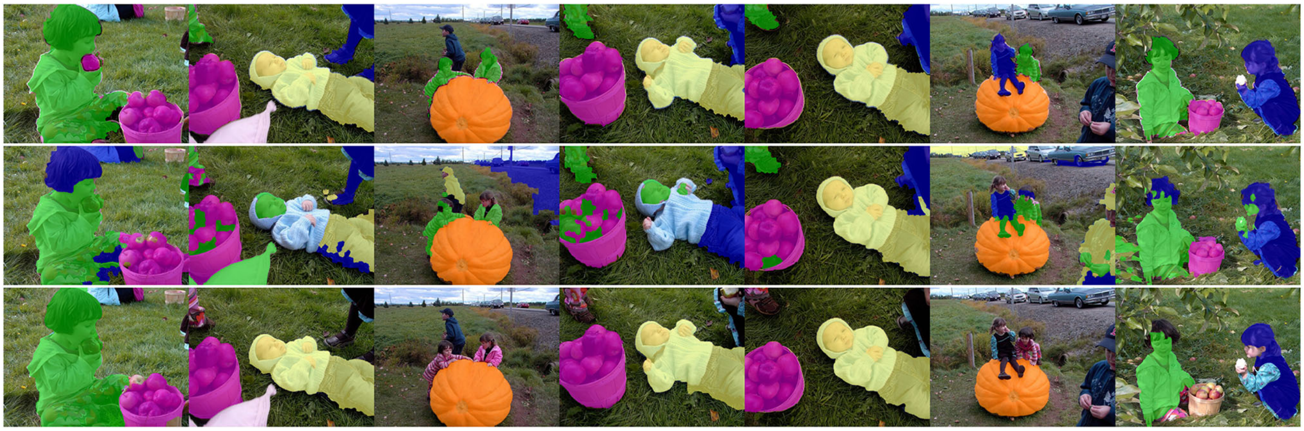


Fig. 8 Comparison to [15] on their dataset. The top row illustrates the ground-truth labellings of multiple foreground objects: apple basket (in pink), pumpkin (in orange), baby (in yellow) and two girls (in green and blue). The bottom rows illustrate their results (middle row) and ours (bottom row) On average, our method yields higher P scores ($97.32 \gg 93.11\%$) and comparable J scores (49.02–49.61%)

sions and appearance changes. The dataset comes with per-frame, per-pixel ground-truth annotations. We *sparsified* these sequences (taking every 10th frame) to construct a large number of datasets that are somewhat related to our problem setting. Table 2 shows the intersection-over-union (IoU) scores on a representative subset and the average over all 50 collections. Note that this score is also referred to as "Jaccard(J) per sequence." Similar to the input provided to video segmentation techniques in the mask propagation task, we also provide the semi-supervised techniques with a manual segmentation of the first frame. However, on our sparsified collections, subsequent frames are quite different, as illustrated in Fig. 9 and in Supplementary Material.

Our extensive evaluation on the adapted DAVIS benchmark clearly illustrates, first of all, the difficulty of the problem setting, as the image structure is not temporally-coherent, and unlike dense video techniques, we cannot benefit from any temporal priors. Furthermore, it demonstrates the robustness of our technique, as it achieves the highest scores on most of the datasets, as well as the highest average score on all 50 collections. It is important to note that in many cases our scores are comparable to video segmentation techniques. For example, the recent video segmentation technique by Marki et al. [19] obtains an average intersection-over-union (IoU) score of 0.67 on the full dense collections, while our average score (on the sparsified collections) is 0.53.

Table 2 Comparison against RJKL13 [26], FI13 [9], and KX12 [15] on the sparsified DAVIS benchmarks

	RJKL13	FI13	KX12	Ours
Bear	0.19	0.05	0.73	**0.92**
Blackswan	0.30	0.07	**0.74**	0.64
bmx-trees	0.04	0.06	0.08	**0.27**
bmx-bumps	0.03	0.17	0.24	**0.28**
Breakdance-flare	0.10	0.08	**0.25**	0.08
Breakdance	0.09	0.09	**0.50**	0.29
Bus	0.50	**0.84**	0.73	**0.79**
Dance-twirl	0.13	0.04	0.15	**0.40**
Libby	0.38	0.14	0.29	**0.41**
Dog	0.36	0.61	**0.71**	0.57
Drift-chicane	0.02	0.00	**0.02**	0.00
Drift-straight	0.13	**0.31**	0.11	0.26
Mallard-water	0.06	0.37	0.46	**0.69**
Mallard-fly	0.01	0.05	**0.47**	0.13
Elephant	0.13	0.00	0.28	**0.45**
Flamingo	0.18	0.23	0.43	**0.64**
Goat	0.07	0.04	0.42	**0.64**
Hike	0.17	0.00	0.36	**0.89**
Paragliding	0.30	0.13	0.80	**0.82**
Soccerball	0.02	0.00	0.37	**0.67**
Surf	0.12	0.94	0.63	**0.96**
Average	0.16	0.22	0.36	**0.53**

Following previous work, we report the IoU scores on a representative subset and the average is computed over all 50 sequences

Bold signifies best score

6 Conclusions and future work

In this work, we have presented a co-segmentation method that takes a distributed approach. Common co-segmentation methods gather information from all the image in the collection, analyze it globally, building a common model, and then infer the common foreground objects in all, or part of, the images. Here, there is no global model. The beliefs are propagated across the collection without forming a global model of the foreground object. Each image independently collects the beliefs from its neighbors and consequentially infers its own model for the foreground object. Although our

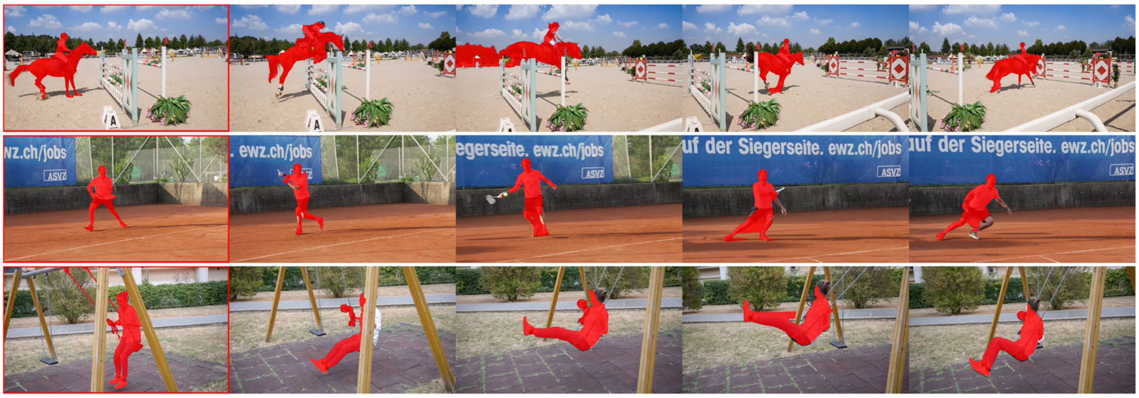


Fig. 9 Qualitative results of our technique on a sequence of sparse frames sampled from the Davis dataset [22], where the first frame is provided as template

method is distributed, currently there is a seed model, which clearly does not concur to the claim of having a distributed method. However, some supervision is necessarily required to define the semantic target model. Currently, it is provided as a single segmented image, but the seed model can possibly be provided in other forms.

We have shown that our approach outperforms state-of-the-art co-segmentation methods. However, as our results demonstrate, there are limitations as the object cut-outs are imperfect. The entire object is not always inferred and also portions of the background may contaminate the extracted object. To alleviate these limitations, there are two possible avenues for future research: (i) one in high level, to better learn the semantics of the object, perhaps using data-driven approaches, e.g., convolutional networks, and (ii) in low level, seeking for better alternatives to graph-cuts and its inherent limitations.

In the future, we hope to explore our approach on massive collections, which may include thousands of photographs capturing interesting dynamic events, for example, a collection of images of a parade, where a 3D reconstruction is not applicable. The larger number of images is not just a quantitative difference, but qualitative as well, as the collection can become dense with stronger local connections. For such massive collections, the foreground object does not have to be only a single object. We can propagate multi-target beliefs over the image network, like we demonstrated in our comparison to Kim and Xing [15]. Finally, the distributed nature of our method leads itself to parallel computation, which can be effective for large-scale collections.

References

1. Arpa, A., Ballan, L., Sukthankar, R., Taubin, G., Pollefeys, M., Raskar, R.: Crowdcam: instantaneous navigation of crowd images using angled graph. In: International Conference on 3D Vision-3DV 2013, pp. 422–429. IEEE (2013)
2. Basha, T., Moses, Y., Avidan, S.: Photo sequencing. In: Computer Vision–ECCV 2012, pp. 654–667. Springer (2012)
3. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: icoseg: Interactive co-segmentation with intelligent scribble guidance. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176. IEEE (2010)
4. Campbell, N.D., Vogiatzis, G., Hernández, C., Cipolla, R.: Automatic 3d object segmentation in multiple views using volumetric graph-cuts. Image Vis. Comput. **28**(1), 14–25 (2010)
5. Chang, K.Y., Liu, T.L., Lai, S.H.: From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2129–2136. IEEE (2011)
6. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2015)
7. Chiu, W.C., Fritz, M.: Multi-class video co-segmentation with a generative multi-video model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 321–328 (2013)
8. Djelouah, A., Franco, J.S., Boyer, E., Pérez, P., Drettakis, G.: Cotemporal multi-view video segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 360–369. IEEE (2016)
9. Faktor, A., Irani, M.: Co-segmentation by composition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1297–1304 (2013)
10. Fan, Q., Zhong, F., Lischinski, D., Cohen-Or, D., Chen, B.: Jump-cut: non-successive mask transfer and interpolation for video cutout. ACM Trans. Gr. (TOG) **34**(6), 195 (2015)
11. Fu, H., Xu, D., Lin, S., Liu, J.: Object-based rgbd image co-segmentation with mutex constraint. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4428–4436 (2015)
12. Gang, Z., Long, Q.: Silhouette extraction from multiple images of an unknown background. In: Proceedings of the Asian Conference of Computer Vision, Citeseer (2004)
13. HaCohen, Y., Shechtman, E., Goldman, D.B., Lischinski, D.: Non-rigid dense correspondence with applications for image enhancement. ACM Trans. Gr. (TOG) **30**(4), 70 (2011)
14. Heskes, T.: Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. J. Artif. Intell. Res. **26**(1), 153–190 (2006)
15. Kim, G., Xing, E.P.: On multiple foreground cosegmentation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 837–844. IEEE (2012)

16. Kim, G., Xing, E.P.: Jointly aligning and segmenting multiple web photo streams for the inference of collective photo storylines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 620–627 (2013)
17. Kim, G., Xing, E.P., Fei-Fei, L., Kanade, T.: Distributed cosegmentation via submodular optimization on anisotropic diffusion. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 169–176. IEEE (2011)
18. Kuettel, D., Guillaumin, M., Ferrari, V.: Segmentation propagation in imagenet. In: Computer Vision–ECCV 2012, pp. 459–473. Springer (2012)
19. Maerki, N., Perazzi, F., Wang, O., Sorkine-Hornung, A.: Bilateral space video segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
20. Mustafa, A., Hilton, A.: Semantically coherent co-segmentation and reconstruction of dynamic scenes. In: CVPR 2017 Proceedings (2017)
21. Ning, J., Zhang, L., Zhang, D., Wu, C.: Interactive image segmentation by maximal similarity based region merging. Pattern Recogn. **43**(2), 445–456 (2010)
22. Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L.V., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Computer Vision and Pattern Recognition (2016)
23. Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(1), 128–140 (2017)
24. Ramakanth, S.A., Babu, R.V.: Seamseg: Video object segmentation using patch seams. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 376–383. IEEE (2014)
25. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 993–1000. IEEE (2006)
26. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1939–1946. IEEE (2013)
27. Rubio, J.C., Serrat, J., López, A., Paragios, N.: Unsupervised cosegmentation through region matching. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 749–756. IEEE (2012)
28. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2217–2224. IEEE (2011)
29. Wainwright, M.J., Jaakkola, T.S., Willsky, A.S.: A new class of upper bounds on the log partition function. Trans. Inf. Theory **51**(7), 2313–2335 (2005)
30. Zhang, D., Javed, O., Shah, M.: Video object co-segmentation by regulated maximum weight cliques. In: European Conference on Computer Vision, pp. 551–566. Springer (2014)

Hadar Averbuch-Elor is a Ph.D. student at the School of Electrical Engineering, Tel Aviv University. She received the B.Sc. (cum laude) degree in electrical engineering from the Technion in 2012. She worked as an computer vision algorithms developer in the defense industry from 2011 to 2015. She was a research intern at Facebook, Seattle, during the summer of 2016. Her research interests include computer vision and computer graphics, focusing on unstructured image collections and unsupervised techniques.

Johannes Kopf is a research scientist in the Computational Photography research group. Prior to joining Facebook, Johannes received a Ph.D. from the University of Konstanz in Germany and worked at Microsoft Research for 8 years. He received the Eurographics Young Researcher Award in 2013 and the ACM SIGGRAPH Significant New Researcher Award in 2015 for contributions to the fields of digital imaging and video. His work covers a broad range within computer graphics and vision. His interests are in a variety of areas including computational photography, digital imaging and video, image-based rendering, and image and texture synthesis.

Tamir Hazan received his Ph.D. from the Hebrew University of Jerusalem (2010) and he is currently an assistant professor at the Technion Institute of Technology, Israel. His research describes efficient methods for reasoning about structured models. More specifically, his research focuses on mathematically founded solutions to modern real-life problems that demonstrate non-traditional statistical behavior. Recent examples are perturbation models that allow efficient learning of high-dimensional statistics, deep learning of infinite networks and primal-dual optimization for high-dimensional inference problem.

Daniel Cohen-Or is a professor at the School of Computer Science, Tel Aviv University. He received the B.Sc. (cum laude) degree in mathematics and computer Science and the M.Sc. (cum laude) degree in computer science, both from Ben-Gurion University, in 1985 and 1986, respectively. He received the Ph.D. from the Department of Computer Science at State University of New York at Stony Brook in 1991. He received the 2005 Eurographics Outstanding Technical Contributions Award. In 2015, he was named a Thomson Reuters Highly Cited Researcher. Currently, his main interests are in few areas: image synthesis, analysis and reconstruction, motion and transformations, shapes and surfaces.