# An Topic-level Random Walk Framework For Scene Image Co-Segmentation

**Zehuan Yuan[1], Tong Lu[1]*, and Palaiahnakote Shivakumara[2]**

*[1]National Key Lab. for Novel Software Technology, Nanjing University, Nanjing, 210046, China*

*[2]Faculty of Computer Science and Information Technology, University of Malaya*

# Introduction

## Task and motivation

In this paper, we simultaneously analyze multiple images from the same scene and decompose each complex scene image into disjoint but meaningful segments with each corresponding to an instance of a scene element (e.g., tree and car). Although promising results have been achieved, most of them still face difficulties when dealing with scene images due to large intra-class variability and complex scene structures.

## Contribution

Our main contributions include 1) the introduction of stable scene context into scene image co-segmentation, and 2) a new framework consisting of the VRN representation and the topic-level random walk on it to address the problem. Although topic-level random walk is familiar in mining social networks, it is novel for image co-segmentation to the best of our knowledge. According to the experiments on LabelMe and SUN , we have averagely 10% improvement over the state-of-the-art methods. Moreover, the proposed VRN is sparse with few hubs [9] and thus is efficient for large scene datasets compared to popular pixel-label methods.

## Framework

We propose a fully automatic co-segmentation method that exploits both the appearance consistency of the same class and the spatial scene context constraints of different classes. The core of our method is to derive a directed *flowing-graph* named Visual Relation Network (VRN) to characterize "soup of segments" [1] and their relations. The statement of the *flowing-graph* means that the weight of any edge varies over the state variables of its linked nodes. meaningful segments are actually the hubs of the graph. Thereby, co-segmentation from multiple scene images can be formulated as voting on the large-scale network. By considering "classes" as "topics", we address it by a topic-level random walk algorithm on VRN to search for the meaningful segments that have high ranking scores. Then for each image, we use a greedy strategy to search for the optimized segment combination from the selected meaningful segments. The overview of the entire framework is shown in Fig. 1.

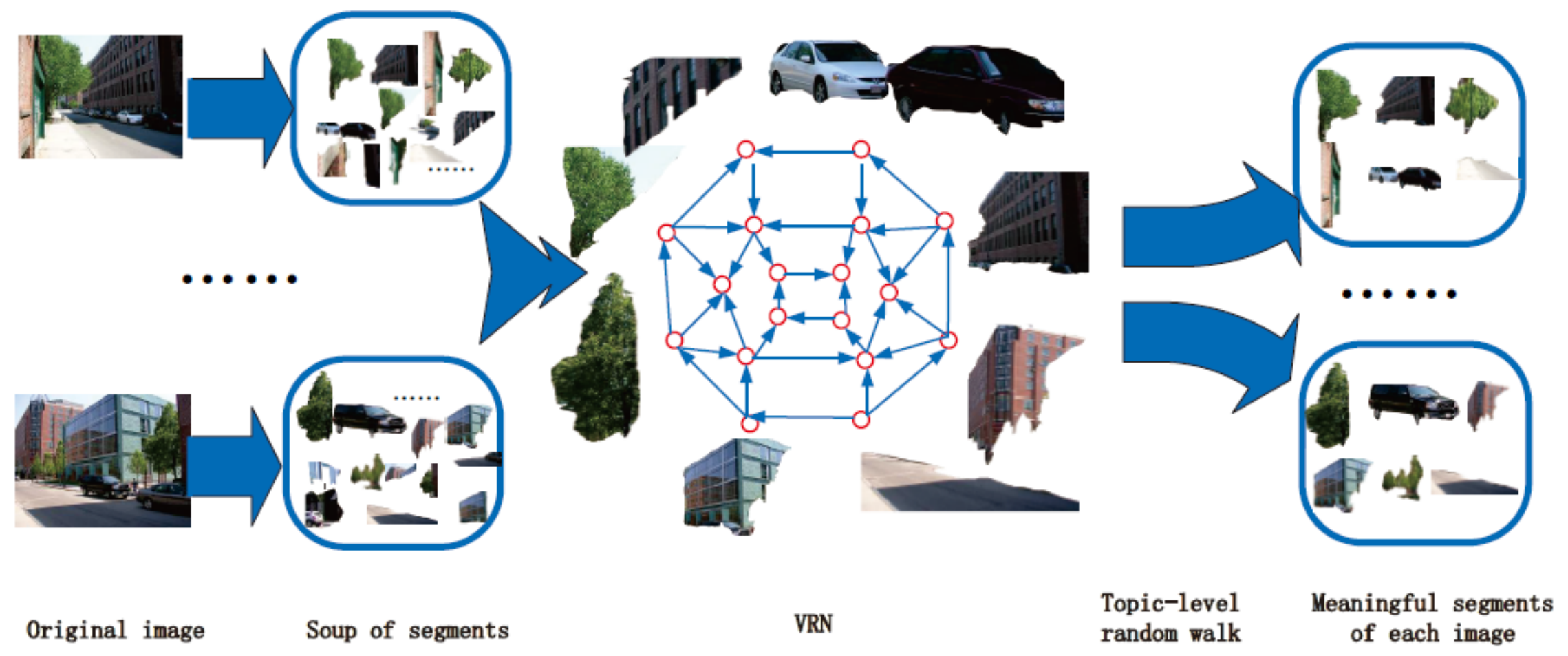

*Fig.1. The overview of the proposed method.*

# Methods

## Establishing VRN

VRN is a weighted directed graph $(V;E;W)$ with V as its vertex set, $E$ as its edge set, and $W$ as the edge weight set. Any node $a_i$ represents the *i*-th segment in the "soup of segments" of image $a$. For any image, we adopt the same strategy as [1] to obtain its "soup of segments". A segment 1) is described by the appearance $A$ that is characterized by its pHOG, color distribution and texton distribution, and 2) has a class variable $t$ belonging to $\{1,2,\cdots,T\}$ and a distribution P that describes the probability of the segment belonging to one class in $\{1,2,\cdots,T\}$.

**Edge construction across images** A similarity edge connecting $a_i$ and $b_j$ will be created, namely, the edge $a_i \to b_j$ as shown in Fig. 2 with its weight initialized as $S(a_i;\, b_j) = \frac{1}{|c|}\sum_c K_{\chi^2}(A_c(a_i),\ A_c(b_j))$, where $A_c$ denotes the *c*-th type of appearance features

**Establishing *part-of* edges** Given two segments $a_i$ and $a_j$ with an high overlapped scale, we add a directed edge $(a_i, a_j)$ if the segment $a_i$ is smaller, otherwise $(a_j, a_i)$ is added.

**Establishing class-level context edges** Given two segments $a_i$ and $a_j$ with no overlap, we add two directed edges $(a_i, a_j)$ and $(a_j, a_i)$ into $E$. With each element corresponding to a class pair, the weight of $(a_i, a_j)$ is a class-level vector $c(t_i, t_j)$ indicating the strength of spatial relation if $a_i$ and $a_j$ are equal to $t_i$ and $t_j$, respectively

$$c(t_i, t_j) = \frac{p(t_i|a_i)}{|a_i||a_j|} \sum_{\substack{(x,y)\in a_i \\ (x^*,y^*)\in a_j}} M_{t_j|t_i}(x^* - x, y^* - y)$$

where $|a_i|$ and $|a_j|$ are the number of pixels in $a_i$ and $a_j$, respectively. $p(t_i|a_i)$ is the probability of $a_i$ under class $t_i$, and $M_{t_j|t_i}(\cdot,\cdot)$ is the relative location map [2] of $t_j$ given $t_i$.

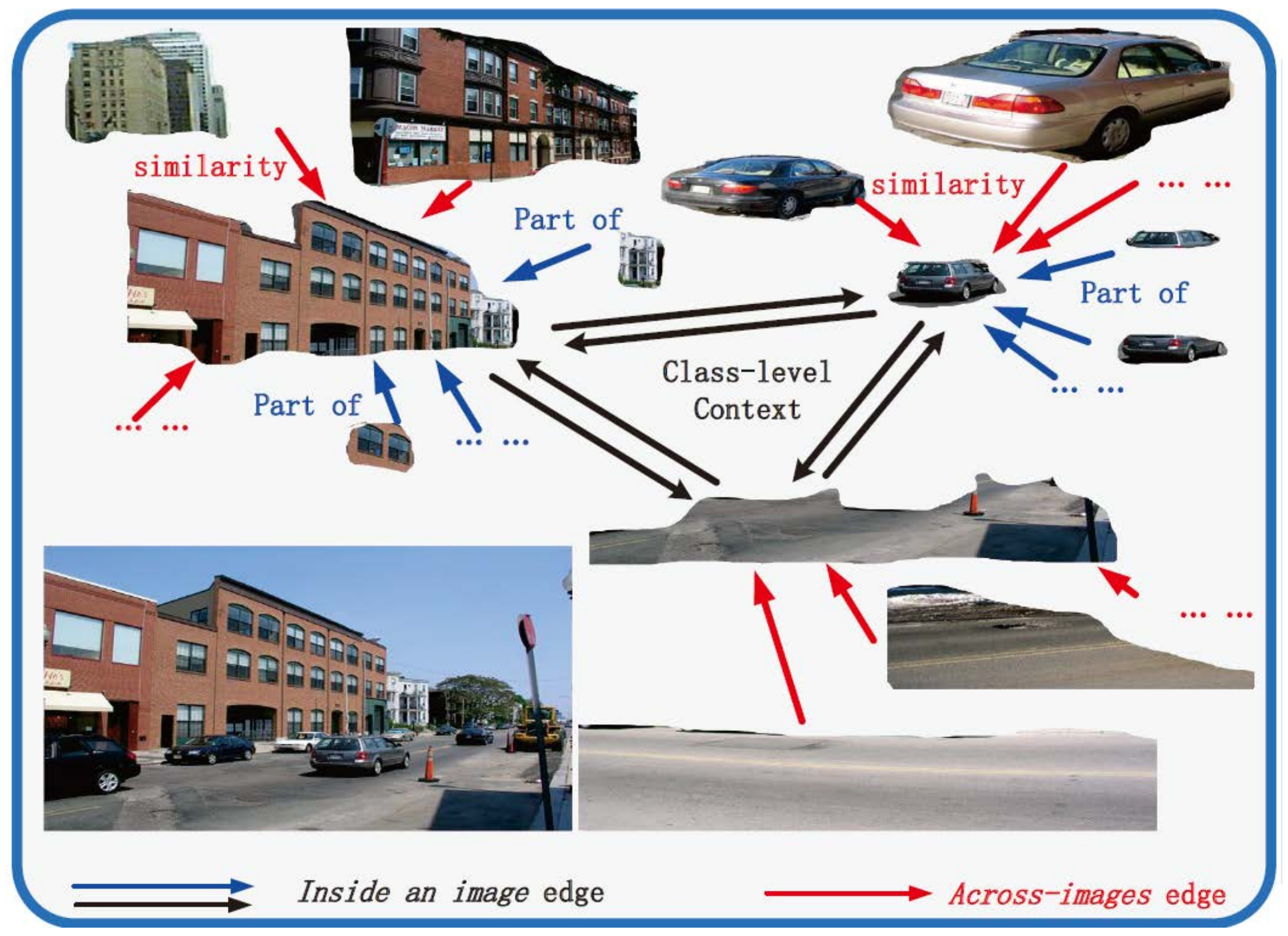

*Fig.2. An example VRN.*

## Topic-level random walk on VRN

Meaningful segments perform the role of hubs to which many other nodes are directed. Given a node $a_i$, we introduce a ranking score vector $\{r[a_i,t]_{t=1,\cdots,T}\}$ to represent the importance of $a_i$ under class $t$

$$r(a_i,t) = \varepsilon\frac{p(t|a_i)}{|V|} + (1-\varepsilon)(\kappa \sum_{(b_j,a_i)\in E} r(b_j,t) w_{b_j a_i} +$$

$$(1-\kappa) \sum_{(a_j,a_i)\in E} V(a_j,a_i,t))$$

$$V(a_j,a_i,t) = \begin{cases} \sum_{t_j} \tau r(a_j,t_j) & (a_j,a_i) \text{ is } \textit{part of} \\ \sum_{t_j} c_{a_j a_i}(t_j,t) r(a_j,t_j) & \text{Otherwise} \end{cases}$$

## Segments selection and class inference

To search for meaningful segments for any scene image $a$, we adopt a greedy algorithm and the details are shown in Tab. 1.

*Tab.1. The greedy algorithm to choose meaningful segments*

```
Input: Image set $D$ and all the candidate segments $S$
Output: The selected segments $segC$ for every image in $D$
For any  image $a$ in $D$
  (1) Calculate the overall score $r_{overall}$ of a segment $a_i$ and assign a
      class label $\bar{t}$ to it by $\bar{t} = argmax\, r(a_i,t); m_a = \frac{1}{|T|}\sum_t r(a_i,t)$;
      $v_a = \frac{1}{|T|}\sum_t (r(a_i,t) - m_a)^2$; $r_{overall} = v_a * max\, r(a_i,t)$
  (2) Sort all $a_i$ by $r_{overall}$ and initialize the selected segments set $segC$=[];
  (3) Select $a_i$ in the descending order of $r_{overall}$;
  (4) For $a_j \in segC$  calculate its overlap, if overlap > 0.1 return (3);
  (5) Add $a_i$ into $segC$, if $\cup\, segC$ < 0.9 * image size of $a$, return (3).
End
```

# Results

We evaluate our method on three datasets: MSRC-v2, LabelMe and SUN. Here we only give few results. Readers can refer more results and evaluations in the paper.

## Evaluation criterion

Firstly, we adopt the segmentation accuracy to quantitatively evaluate our results. For each class, we denote the ground-truth segments and the obtained segments with $G$ and $C$ , respectively. Then the segmentation accuracy can be defined as the ratio of the intersection of $G$ and $C$ to the union of them. Besides, purity score is also adopted to measure the coherency of class labeling of our method over the entire dataset. For each selected segment, its ground-truth class label is the one that the majority of pixels in it belong to. Note that different class labels may be potentially assigned to the selected segments with the same ground-truth class label in different images.

## Example Segmentation result on LabelMe

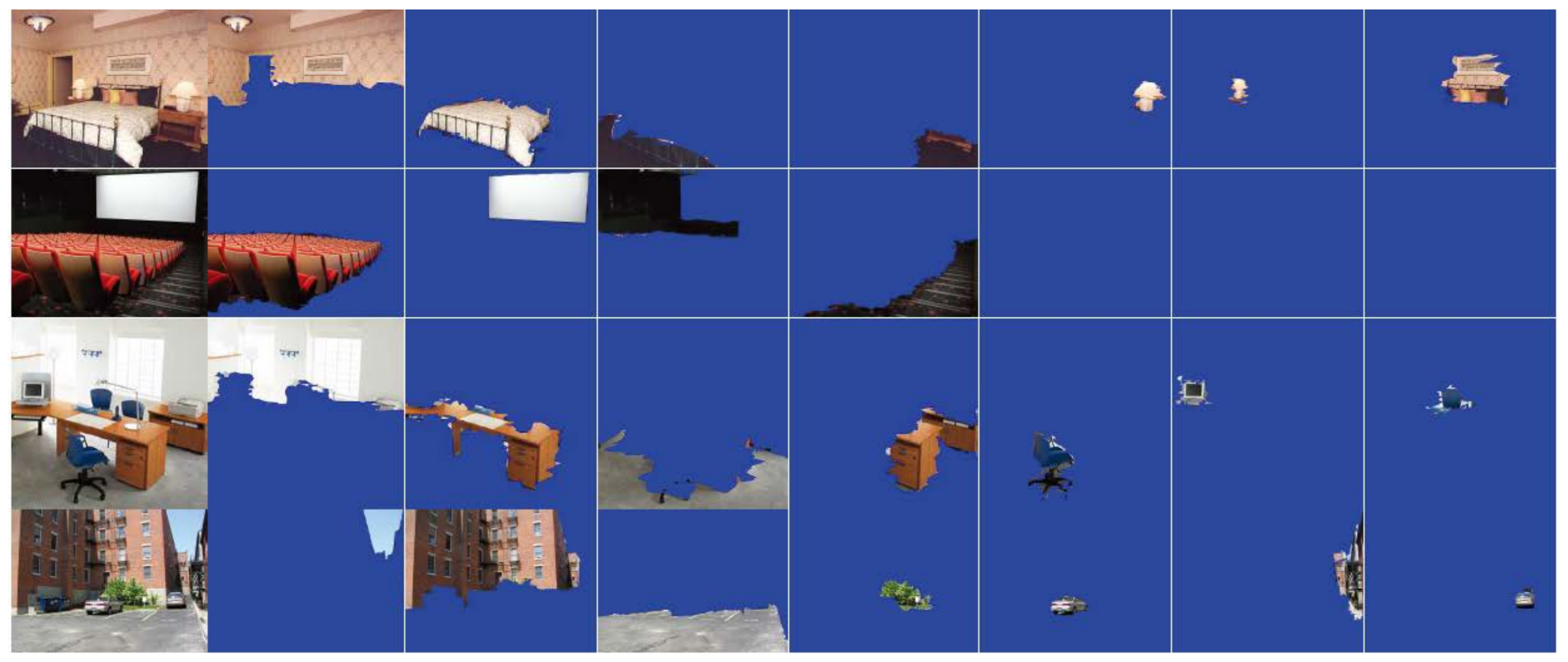

*Fig.3. The overview of the proposed method. The first column represents an example image, while the rest columns are the results after co-segmentation. The selected segments are ranked in a decreasing order by their overall importance scores from left to right*

*Tab.2. Segmentation Accuracy comparisons on LabelMe*

| Scene | Propose | (b)Cont | (c)Part | (d)Appr |
|---|---|---|---|---|
| Office | **0.35** | 0.29 | 0.22 | 0.20 |
| Theater | **0.44** | 0.40 | 0.34 | 0.29 |
| Bathroom | **0.45** | 0.39 | 0.38 | 0.39 |
| Bedroom | **0.39** | 0.32 | 0.30 | 0.25 |
| Street | **0.52** | 0.42 | 0.40 | 0.39 |
| Indoor | **0.44** | 0.35 | 0.32 | 0.30 |

# Conclusions

In this paper, we present a novel visual relation network to model the relationship between scene segment candidates and perform topic-level random walk on the network to exploit scene co-segmentation. The experiments on different datasets show the effectiveness of our method. However, if unfortunately most of the candidate segments are "garbage" ones, the accuracy will be according decreased during image co-segmentation. Potentially it can be avoided by enriching "soup of segments". Our further work is to improve the accuracy of our unsupervised scene image co-segmentation by including more class-level context cues.

# Reference

1. **Russell, B.C., Freeman, W.T., Efros, A.A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR (2). pp. 1605–1614 (2006)**
2. **Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-class segmentation with relative location prior. International Journal of Computer Vision 80(3), 300–316 (2008)**