Object Extraction from Bounding Box Prior with Double Sparse Reconstruction

Lingzheng Dai, Jundi Ding, Jian Yang, Fanlong Zhang and Junxia Li
Dept. of CSE, Nanjing University of Science and Technology, Nanjing 210094, PR China
lingzhengdai@163.com,{dingjundi2010,csjyang}@njust.edu.cn,csfzhang@126.com,junxiali99@163.com

Abstract

Extracting objects from natural images has long been an active problem in image processing. Despite various attempts, it has not been completely solved up to date. Current state-of-the-art object proposal methods tend to extract a set of object segments from an image, and often these are consequential differences among these results for each image. Another type of methods strive to detect one object into a bounding box where some background parts are often covered. For these two methodologies, we observe: 1) there are generally some regions overlapped among different proposals, which are usually from one object; they could be as object 'segment hypotheses'; 2) pixels outside the detected bounding box could be as 'background hypotheses' as they are with high probability from the background. With them, we formulate the object extraction as a "double" sparse reconstruction problem in terms of the bounding box results. The idea is that object regions should be with small reconstruction errors to segment hypotheses bases, simultaneously, they should have large reconstruction errors to background hypotheses bases. Comprehensive experiments and evaluations on PASCAL VOC object segmentation dataset and GrabCut-50 database demonstrate the superiority of our built method. In particular, we achieve the state-of-the-art performance for the object segmentation with bounding box prior on these two benchmark datasets.

1. Introduction

Extracting objects from natural images has long been one of the most fundamental and critical problems in computer vision and image processing [3][18][28]. It plays a key role in vision applications, including object recognition, classification [11][27] [29] [30] , etc. However, experiments on PASCAL [32] or GrabCut-50 [12] database show that it is still an unsolved and challenging problem. This is mainly due to that photographs of natural scenes reflect real-world variations and are characterized by large ranges of color, texture and shapes.

Two paradigms have shaped this field of object extraction. In the former, a plethora of low-level object proposal methods [2] [3] [18] have aroused the interest of researchers. Given an input image, its output is a set of visually consistent object-level segments, often called object candidates. Some subsequent operations (*e.g.,* support vector machines) are used to rank and classify these object candidates for picking out a best one. However, when the semantic regions appear with large appearance diversity, to obtain a single proposal to cover the whole object accurately is a non-trivial task, due to bottom-up object proposals tend to yield the appearance consistency instead of semantic ones.

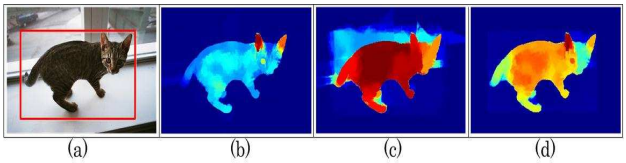

Figure 1. (a) Original image; (b) Background Reconstruction Map (BRM); (c) Segment Reconstruction Map (SRM); (d) Final object confidence map.

The latter paradigm is the current detection-based techniques such as Deformable Part Models [19] that have attracted wide interest in computer vision. Based on semi-local orientation histograms (*e.g.*, SIFT [4], HOG [5]), these methods are capable of bounding an object as a whole in one box by using the scanning-window architecture. The question is that inside the bounding box, many background pixels are also covered.

Intuitively, it is beneficial to jointly use the bounding box prior and object proposals for extracting the object entirely from natural images. On the one hand, benefitting from the 'object entirety' performance, as shown in Fig. 1(a), it is sensible for us to directly extract objects from the bounding box prior; on the other hand, the output of object proposals methods is a pool of possibly-overlapping region proposals that can be used as the search space for objects in the image. Given the bounding box prior, although many background components are covered in it, we can easily observe that the pixels outside the bounding box are with high probability from the background. Considering this, they could be as the *background hypotheses*. Then, we are easy to formu-

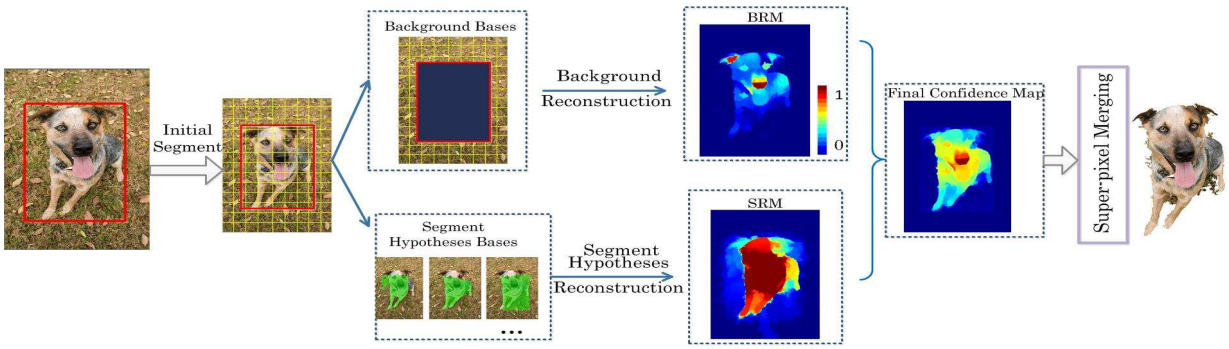

Figure 2. Framework of our double sparse reconstruction based segmentation model. Given an image and bounding box prior of the object, we first segment the image into super-pixels. Super-pixels outside the bounding box and many object proposals generated by CPMC method, respectively, are used to construct the background and segment hypotheses. Background hypotheses are used to reconstruct the background reconstruction map (BRM), and meanwhile the segment reconstruction map (SRM) is reconstructed through segment hypotheses. The output of BRM and SRM are obtained by the proposed double sparse reconstruction framework, including the reconstruction coefficients. They are then integrated to yield the final object confidence map. Region merging procedure is finally applied on the object confidence map to make the extracted object with a well-preserved boundary.

late the object extraction as a reconstruction problem. With respect to the background hypotheses, the object pixels or regions should have a large reconstruction error, shown as brighter regions in Fig .1(b). In addition, object pixels may be similar with background parts in some low-level cues in many situations. The background reconstruction errors of them are close [14]. To solve this, we also consider the object candidates that are generated by object proposals methods as some object *segment hypotheses*. The reason is that these object candidates generally exhibit a high overlapping which is usually from one object. Then, with respect to the segment hypotheses, the object pixels or regions should have small reconstruction errors, shown as brighter regions in Fig .1(c).

The goal of this paper is to greatly enhance the performance for extracting object with the bounding box prior. We propose a *double sparse reconstruction* method to object extraction in terms of this prior, which inherits the merits of both above two terms – background and segment hypotheses. As shown in Fig. 2, we first apply the SLIC [7] method to a testing image and obtain the super-pixels. Super-pixels outside the bounding box prior are used to construct the background hypotheses. In our model, the first reconstruction map is designed to be reconstructed by the background hypotheses, where super-pixels that exhibit large sparse reconstruction errors are predicted as object regions and that derive small errors to be background components. The reconstruction map measures how likely each super-pixel to be a object region. We refer to it as Background Reconstruction Map (BRM). And meanwhile a set of object proposals are generated by applying CPMC [2] method to build the segment hypotheses. These hypotheses are fed into our model to calculate the second reconstruction map, called Segment Reconstruction Map (SRM), in which the regions that derive a small reconstruction error with respect to them will be with high probability to be object. The proposed framework is able to take full advantage of the object bounding box prior information and object proposals. The output of BRM and SRM are unified obtained by our double sparse reconstruction framework, including the reconstruction coefficients. These two maps are integrated to generate the final object confidence map. Region merging procedure is finally used on the confidence map to make the extracted object with a well-preserved boundary. Figure 2 shows a pipeline of our framework. In summarize, our contributions are as follows:

• Current popular tendency is to detect the object in one bounding box as well as many background pixels. We here build a reconstruction framework that can refine the bounding box detected results, and can precisely put the object out from the bounding box. It is perhaps more useful for further vision tasks, such as object recognition and image understanding.

• Our reconstruction framework not only considers the background hypotheses, but also uses the object segment hypotheses. In this way, some challenging cases – where pixels of the object and its background are very similar or pixels of within-objects are not similar – can also be well tackled (see the extracted dog in Figure 2).

• Comprehensive experiments and evaluations on two challenging object segmentation datasets PASCAL VOC object segmentation dataset [32] and GrabCut-50 image segmentation database [12] demonstrate that our built framework is superior to the state-of-the-art methods.

2. Related Work

Our method could be viewed as a semi-supervised work because it aims to utilize some object-and-background prior to guide the object reconstruction. Existing methods that

use some brush strokes, or bounding box prior to predict starting seeds or locations of objects are related to our work.

Seed-based approaches: The seed-based methods include: GraphCut (GC) [15], constrained parametric min-cut (CPMC) [2] and Laplacian Coordinates [24]. Considering the image as a graph, GC seeks to find the minimum cut between seeded regions, where the similarity between neighboring pixels is encoded as edges of this weighted graph. GC uses a max-flow/min-cut algorithm to find this cut in order to segment images. Using a graph-cut based model, Carreira et al. [2] seek to generate a pool of object hypotheses by hypothesizing a set of placements of fore- and background seeds. For each configuration, segmentation results are obtained by solving a constrained parametric min-cut. Particularly, in their model the smoothness term borrows the definition from gPb [10] of similarity between adjacent pixels. Recently, Laplacian Coordinates (LC)[24] is proposed to minimize a novel quadratic energy function defined from an affinity graph of pixels. In their model, the average distance of pairwise pixels is minimized and anisotropic propagation of seeds labels is controlled well. Generally, careful assignment of these seeds is a non-trivial job, which influences the segmentation performance critically.

Bounding box based approaches: Some work, on the other hand, use the bounding box prior to guide object segmentation, including [16] [12] [13] [36]. Compared with seed strokes, bounding box prior is intuitive to users due to its availability of taking only tow mouse clicks and the emerging of object-detection techniques. In GrabCut [16], this bounding box prior is integrated into the energy function and the model is iteratively optimized by Expectation Maximization (EM) method. Further, the object boundary is refined by border matting in order to get the final segmentation results. The authors [12] further presented a new graph-cut framework. They investigate the effectiveness of the sufficiently tight bounding box and integrate this information as a constraint into their energy function. To optimize their model, a new rounding algorithm - pinpointing is handed as the optimization strategy. In [13], segmentation task is tackled as an adaptive figure-ground classification algorithm using a user provided bounding box. It compiles various foreground priors and one common background prior seamlessly. With the different foreground priors, many hypotheses are generated with evaluation score functions. At last, the one with the maximum segmentation quality score is selected as the best segmentation. Recently, Tang *et al.* [36] propose an alternative approach to color clustering using kernel K-means energy. Compared with histogram or GMM fitting used by [16], they argue that the fore/background regions can be clustered better using this energy. Probably the most similar work to us is Xia [6], which proposes to generate the object shape by directly selecting the best overlapping segments that align well to the object boundary and thereafter integrate it into the subsequent graph-cut based inference algorithm to obtain the segmentation results. Segmentation performance of this method heavily relies on the shape based graph-cut process. We utilize the generated object candidates as segment hypotheses bases. However, unlike [6], we operationalize this idea by exploring the usefulness of each segment towards object extraction based on an object reconstruction model. Furthermore, the pairwise correlation information of segment hypotheses can be preserved in our method, which is crucial to produce accurate and reliable results.

3. Object Extraction via Double Sparse Reconstruction

In this section, we present the proposed method in detail. Given a test image, we first segment it into super-pixels. Then for each bounding box input, the super-pixels outside the bounding box are used to construct background hypotheses bases. Meanwhile, we compute a large pool of object candidates to construct segment hypotheses bases for each image, using the publicly available Constrained Parametric Min-Cuts algorithm (CPMC) [2]. These two bases are integrated into our model as reconstruction bases for predicting the object confidence map. The obtained confidence map is further refined through some techniques including multi-scale strategy and region merging for extracting object entirely.

3.1. Background Reconstruction

When the bounding box prior is provided, although some background pixels are covered in it, it can be observed that the pixels outside it are with high probability from the background, as shown in Fig. 3. This means that some background regions can be easily identified. We apply SLIC [7] method to the test image and segment it into many super-pixels. Then the super-pixels from outside bounding box are used to construct the background hypotheses. Intuitively, using background hypotheses as bases to reconstruct the foreground and background regions, the reconstruction errors between them shall be different. For this reason, we seek to distinguish the foreground from the background based on a sparse reconstruction model.

The first reconstruction map is designed to be reconstructed by using background hypotheses as bases. Formally, let an image $\mathbf{X}$ formed by initial super-pixels, i.e., $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M] \in \mathbb{R}^{N \times M}$, where M is the number of super-pixels and N is the feature dimension. Let us represent each super-pixel with mean color features and coordinates, i.e., $\mathbf{x}_i = \{L, a, b, R, G, B, x, y\}$, where both Lab and RGB color spaces are used to describe its features, and x, y denote its coordinates.

With the bounding box, the background hypotheses bases are formally formed as: $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M] \in$

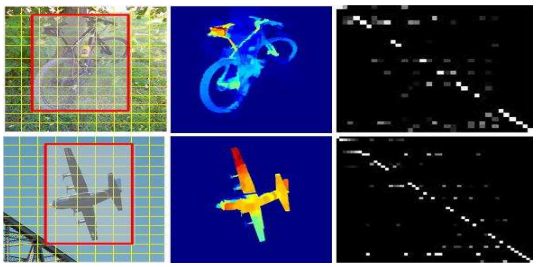

Figure 3. Visualization of our learnt sparse representation $\mathbf{Y}$ of Eq 1 and background reconstruction map (BRM) constructed using background hypotheses. From left to right: Original images with bounding box prior, BRM, the learnt sparse representation $\mathbf{Y}$. For $\mathbf{Y}$, brighter pixels represent the element of its column $\mathbf{y}_i$ having larger magnitude.

$\mathbb{R}^{N\times M}$, where $\mathbf{a}_i = \mathbf{x}_i$ if $\mathbf{a}_i$ belongs to the bounding box outside regions, otherwise $\mathbf{a}_i = \mathbf{0}$. For an image, we seek to represent each $\mathbf{x}$ by using a over-complete dictionary whose vectors are background bases themselves, i.e., $\mathbf{x} = \mathbf{A}\mathbf{y}$. It can be sought by solving the following optimization problem:

$$\widehat{\mathbf{y}}_i = \arg\min_{\mathbf{y}_i} \frac{1}{2}||\mathbf{x}_i - \mathbf{A}\mathbf{y}_i||_2^2 + \lambda_1||\mathbf{y}_i||_1 \qquad (1)$$

where λ_1 is the regularized parameter and ℓ_1 penalty can yield a sparse solution for $\mathbf{y}_i$. The ℓ_2 norm is used to minimizing the distance between the prediction reconstruction and each super-pixel. After obtaining the solution $\widehat{\mathbf{y}}_i$, it is easily to design a sparse representation based classifier [1] in terms of its reconstruction residual. The corresponding reconstruction residual is defined by

$$r(\widehat{\mathbf{y}}_i) = ||\mathbf{x}_i - \mathbf{A}\widehat{\mathbf{y}}_i||_2^2 \qquad (2)$$

An example of $\mathbf{Y} = [\widehat{\mathbf{y}}_1, \widehat{\mathbf{y}}_2, \ldots, \widehat{\mathbf{y}}_M]$ is presented in Fig. 3, from which we can see that all the representation of $\mathbf{y}_i$ of super-pixels outside the bounding box are located in the diagonal of the affinity matrix, while the regions within the bounding box are not in the diagonal. This means that the super-pixels outside the bounding box are represented by themselves, and super-pixel within the bounding box is represented by linear combination of the background hypotheses bases $\mathbf{A}$. The combination coefficients are the elements of $\mathbf{y}_i$, with larger magnitude showing brighter.

After obtaining the representation $\mathbf{Y}$, we can yield the $r(\widehat{\mathbf{y}}_i)$ for each super-pixel by (2). Within the bounding box, the super-pixels that belong to background regions can be well reconstructed by background hypotheses bases $\mathbf{A}$ through (1) and thus they may have small sparse reconstruction errors. On the contrary, the super-pixels that belong to object regions derive large sparse reconstruction errors, which gives us a straightforward way to express each super-pixel in the image with its reconstruction residual. We normalize the reconstruction residual value of each super-pixel into $[0, 1]$. Super-pixels that yield larger values are shown in brighter, meanwhile that have smaller values will illustrate darker, as shown in Fig. 3. We refer to it as background reconstruction map (BRM).

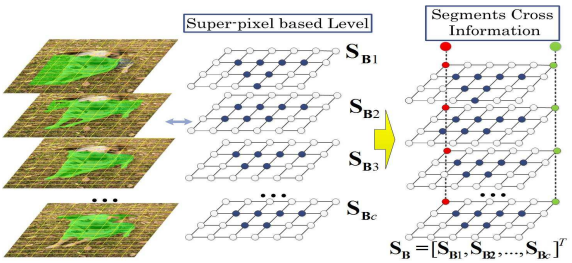

Figure 4. Segment hypotheses bases generation. We first use CPMC method to generate the proposals for the image. Given an input image, its output is a pool of visually consistent object-level segments, as shown in green regions. Then the super-pixels of each proposal that are covered by green region are characterized with 1, represented by solid nodes; otherwise they are with 0, represented by hollow nodes. They are all aligned to form the segment hypotheses bases $\mathbf{S}_\mathbf{B}$, each column of which is viewed as vector basis. Green and red columns are such examples.

However, the reconstruction residual measure that drives the BRM may not be robust to images that contain multiple instances of similar objects: super-pixels outside the bounding box would look similar to super-pixels inside it and yield low residuals for foreground. In this case, the reconstruction map can benefit from object candidates.

3.2. Segment Reconstruction

Give an input image, object proposal methods can generate a pool of visually consistent object-level segments, as shown in Fig. 4. It can be observed that there are often some regions overlapped among different segments, which are usually from one object. We are interested in using these candidates to construct segment hypotheses bases. The CPMC method [2] is used to generate the set of object proposals. Note that the procedure of ranking or classification of generated segments of [2] is not applied in our work. The second reconstruction map is designed to be reconstructed by using segment hypotheses as bases.

Different from previous work, such as [17] [6], the way we pursue is not only to make such segments more powerful by summing them but also to exploit the cross information between them, adapted to the sparse reconstruction model. To achieve this, extracting objects from images is formulated as a segment reconstruction problem, where the reconstruction map of each super-pixel is expressed as a linear combination of generated segment hypotheses bases, referred as Segment Reconstruction Map (SRM).

Suppose $S_1, S_2, \ldots, S_c \subset \mathbb{R}^2$ be the regions of the remaining segments cropped by the bounding box, let $T_i : \mathbb{R}^2 \rightarrow \{0, 1\}$ be the characteristic function of

each super-pixel $\mathbf{r}_j$ for all $j = 1, \ldots, M$ of S_i. Then we use vector $\mathbf{S_B}_i = [T_i(\mathbf{r}_1), T_i(\mathbf{r}_2), \ldots, T_i(\mathbf{r}_M)]^T \in \mathbb{R}^M$ to represent each segment hypothesis S_i, and $\mathbf{S_B} = [\mathbf{S_B}_1, \mathbf{S_B}_2, \ldots, \mathbf{S_B}_c]^T \in \mathbb{R}^{c \times M}$ be all segment hypotheses bases. An example of generating $\mathbf{S_B}$ is illustrated in Fig. 4.

After obtaining the $\mathbf{S_B}$, we can yield the averaging map of it. With them, $\mathbf{d}_i$ that represents each super-pixel of S-RM is predicted by finding the best linear combination of segment hypotheses bases. The SRM of all super-pixels are denoted as $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_M] \in \mathbb{R}^{c \times M}$, and we can obtain it by solving the following optimization problem:

$$\begin{aligned}\{\widehat{\mathbf{y}}_i, \widehat{\mathbf{d}}_i\} = \arg\min_{\mathbf{d}_i, \mathbf{y}_i} ||\mathbf{d}_i - \mathbf{S_B}\mathbf{y}_i||_1 + \lambda_1 ||\mathbf{y}_i||_1 \\ + \lambda_3 ||\mathbf{d}_i - \overline{\mathbf{S}}_\mathbf{B}(i)||_2^2\end{aligned} \tag{3}$$

where λ_3 is the regularized parameter and $\overline{\mathbf{S}}_\mathbf{B}(i) \in \mathbb{R}^c$ represents the average map of $\mathbf{S_B}$, its each element is the average magnitude of i-th column vector of $\mathbf{S_B}$.

Within the bounding box, the super-pixels that belong to the object regions can be well reconstructed by segment hypotheses bases $\mathbf{S_B}$ and thus have small sparse reconstruction errors. The elements of $\mathbf{y}_i$ corresponding to the object can be identified by (3). We predict the SRM not only by minimizing ℓ_2 distance between the prediction reconstruction map and the average map, but also finding the best linear combination of segment hypotheses bases to build it. The SRM of all super-pixels are together predicted by using the designed construction framework which is capable of capturing the correlations among all segment hypotheses. Fig. 5 depicts some SRM results, from which we can see that super-pixels from object regions exhibit brighter color than that of background components, indicating that they can be reconstructed well by segment hypotheses bases.

3.3. Double Sparse Reconstruction

To make full use of all the information produced by the set of generated segment hypotheses and extracted background regions seamlessly, the derived $\mathbf{S_B}$ and $\mathbf{A}$ should be integrated into an unified framework of object reconstruction. Here, our consideration for formulating the inference process is two-side: to inherit the advantages of sparse representation, the representation of the background and segment hypotheses bases is encouraged to be sparse; to make use of the cross-information of segment hypotheses, the segment reconstruction error of each super-pixel should be enforced to be sparsity-consistent simultaneously. By considering both sides, the joint reconstruction is achieved via the following problem:

$$\begin{aligned}\{\widehat{\mathbf{y}}_i, \widehat{\mathbf{d}}_i\} = \arg\min_{\mathbf{y}_i, \mathbf{d}_i} \frac{1}{2}||\mathbf{x}_i - \mathbf{A}\mathbf{y}_i||_2^2 + \lambda_1 ||\mathbf{y}_i||_1 + \\ \lambda_2 ||\mathbf{d}_i - \mathbf{S_B}\mathbf{y}_i||_1 + \lambda_3 ||\mathbf{d}_i - \overline{\mathbf{S}}_\mathbf{B}(i)||_2^2\end{aligned} \tag{4}$$

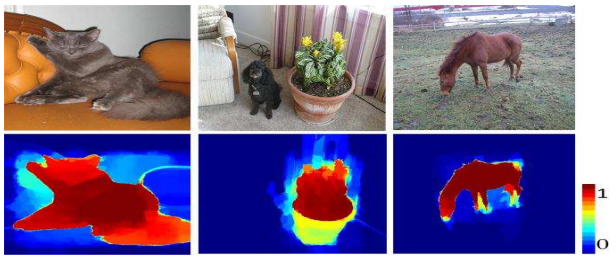

Figure 5. Some segment reconstruction map (SRM) results: (top) original images; (bottom) corresponding SRMs.

After feeding $\mathbf{S_B}$ and $\mathbf{A}$ into above model, we can obtain the output of BRM and SRM of each super-pixel by the unified sparse reconstruction framework, including the reconstruction coefficients. We refer to it as Double Sparse Reconstruction model. The proposed framework is able to take full advantages of background hypotheses information and object proposals.

These two output maps of (5), BRM and SRM, are integrated to generate the final object confidence map. Specifically, after obtaining the optimal solution of Eq. (4) for each super-pixel $\mathbf{x}$, we can directly use $\widehat{\mathbf{y}}_i$ and $\widehat{\mathbf{d}}_i = [d_{i1}, d_{i2}, \ldots, d_{ic}]^T$ to yield the object confidence map of the super-pixel as

$$\mathbf{o}_i = r(\widehat{\mathbf{y}}_i) + \alpha \Sigma_{j=1}^{c} d_{ij} \tag{5}$$

where, α is a parameter to balance the effect of two reconstruction map. Since the confidence map learnt from the reconstruction framework is defined at super-pixel level, objects are tend to be fragmented with heterogeneous parts and strong internal contours. In order to cope with this issue, we compute the super-pixels of the image at a multi-scale strategy. Then, the final object confidence map is calculated by averaging them in order to tackle the large ranges of object color, texture, shape, or other attributes.

After obtaining the object confidence map for an image, the result is projected back to the image. To obtain the object extraction result, the most easily choice is to directly use the map with above a threshold as object and the resident as the background. However, it is hard to set a fixed threshold to find the object for each image due to natural images are very complex.

As proposed in [17], self-similarly can be used to refine the segmentation. In this part, the top high scores serve as object seeds and the segmentation result can be then obtained by a region merging strategy. To achieve this, we use the method of [26] to merge the initial regions for more precise object extraction result.

3.4. Optimization Process

We aim to optimize Eq. (4). Obviously, there are only two parameters to optimize. We propose to optimize it w.r.t

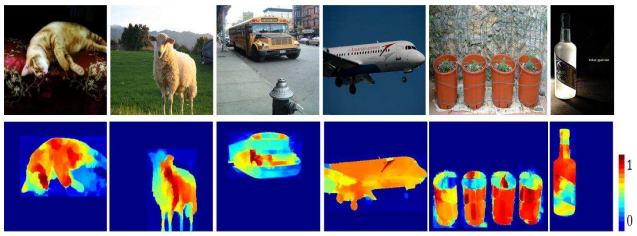

Figure 6. Some exemplar original images (top) from VOC dataset and object confidence map generated by Eq (5) (Best viewed in color).

representation coefficient $\mathbf{y}_i$ and $\mathbf{d}_i$ alternatively. The optimization procedure is keeping $\mathbf{d}_i$ fixed to optimize $\mathbf{y}_i$, and keeping $\mathbf{y}_i$ fixed to optimize $\mathbf{d}_i$ iteratively, repeating until convergence.

Specifically: it consists of the following iterations:

(i) Given $\mathbf{d}_i = \mathbf{d}_i^k$, optimize Eq. (4), update $\mathbf{y}_i^{k+1} \leftarrow \mathbf{y}_i^k$

(ii) Given $\mathbf{y}_i = \mathbf{y}_i^k$, optimize Eq. (4), update $\mathbf{d}_i^{k+1} \leftarrow \mathbf{d}_i^k$

The detailed Optimization Process of Eq. (4) is presented in supplementary material.

4. Experiments

In this section, we study the quality of the proposed double sparse reconstruction method for object extraction. We conduct comprehensive experiments on two publicly available datasts: PASCAL VOC object segmentation dataset [32] and GrabCut-50 dataset [12].

Confidence Map Details: This part presents the implementation details of final confidence map generation. As the confidence map learnt from Eq. (4) is processed at super-pixel level, objects are tend to be fragmented with heterogeneous parts and strong internal contours. In order to cope with this, we compute the super-pixels of an image at a multi-scale strategy. For each image, we first perform over-segmentation by SLIC [7] at eight different scales, with super-pixel number set respectively from 50 to 400. Then we run the double sparse reconstruction method eight times, and obtain the object confidence map by averaging the eight results. We set $\lambda_1 = 0.01$, $\lambda_2 = 0.01$, $\lambda_3 = 0.01$ and $\mu = 1$ in all experiments.

Some final object confidence maps are displayed in Fig. 6. It can be observed that pixels from object regions exhibit brighter color, further advocating the value of our method compared with directing merging segment hypotheses.

4.1. GrabCut-50

Comparison with Bounding Box based Methods: In this part, we provide comparisons against various bounding box based segmentation methods. We first conduct experiments on the popular interaction image segmentation dataset GrabCut-50, which is provided by [12] and includes 50 images with ground truth bounding boxes.

Figure 7. Some object extraction results of our approach on the GrabCut-50 dataset [12].

We use error-rate to evaluate the segmentation performance for different approaches, which is defined as the percentage of mislabeled pixels inside the bounding box. We compare our double sparse reconstruction method with following state-of-the-art: [16] [12] [13] [36] [6]. Results on this benchmark are reported on Table. 1. Our proposed method achieves the mean score 3.2% on this database, significantly outperforming the two bounding box based methods (GrabCut [16] and Tang *et al.* [36]). When more interactive priors are used, GrabCut-Pinpoint [12] and F-G Classification [13] on this dataset can achieve better performance. Our method also obtain lower error score (3.2% vs 3.7% for [12] and 5.4% for [13]). This verifies the effectiveness of our method for extracting object from image with bounding box prior. For fair comparison, we also report the results of method [6] that use graph-cut based procedure to increase the final segmentation performance (3.3%) by integrating the shape prior generated by CPMC proposals. Our method does not use this procedure and achieve competitive results with [6]. This demonstrates that our double sparse reconstruction model has the capability of directly predicting the reasonable object confidence map for object extracting without graph-cut based procedure which is commonly used by many interactive segmentation methods.

Some qualitative segmentation results are displayed in Fig. 7. The extracted objects by our method are highlighted by color mask. Our method successfully predicts the masks of objects that are in complex background. For example, for the banana in the clustered flowers, we can detect it entirely with well-preserved boundaries. It demonstrates that our framework performs well for extracting objects from natural images.

Comparison with Seed based Methods: As the procedure of region merging servers as a seed-based region merging strategy, in this part we additionally compare the proposed method to the state-of-the-art seed based segmentation approaches, including Laplacian Coordinates (LC) approach [24], Graph Cuts (GC), Power Watershed (PWS) [21], Maximun Spanning Forest with Kruskal's (MSFK) and Prim's (MSFP) algorithm [22] and Random Walker (RW) [23]. As these does not use the bounding box prior and for fair comparison, we directly use the published results

Method	Error-rate
GrabCut [16]	8.1 %
Kernel Segmentation [36]	9.7 %
Adaptive Kernel Segmentation [36]	7.1 %
GrabCut-Pinpoint [12]	3.7 %
F-G Classification [13]	5.4 %
Xia [6]	3.3 %
Ours	**3.2** %

Table 1. Error-rate of bounding box prior based algorithms GrabCut, GrabCut-Pinpoint, Adaptive Kernel Segmentation, F-G Classification, Xia, and our method on the Grabcut-50 dataset.

Method	RI	GCE	VoI
GC	0.9714	0.0268	0.1877
MSFK [22]	0.9690	0.0292	0.2013
MSFP	0.9689	0.0293	0.2018
PWS [21]	0.9704	0.0278	0.1931
RW [23]	0.9700	0.0280	0.1934
LC [24]	0.9715	**0.0262**	0.1836
Ours	**0.9765**	**0.0262**	**0.1654**

Table 2. PRI, GCE and VoI of seed based prior based algorithms Graph-cut, MSFK, MSFP, PWS, RW, LC and our method on the Grabcut-50 dataset

on [24] with three distinct region quality metrics to evaluate the segmentation quality of the proposed method: Probabilistic Rand Index (PRI) [9] (higher probability is better); Variation of Information (VoI) [8] (lower distance is better); and Global Consistency Error (GCE) [20] (lower distance is better).

We present the three quality metrics for each method in Table 2. Generally, our method outperforms previous seed-based approaches in this dataset, *e.g.,* PWS [21], MSFK, MSFP [22] and RW [23]. We also compare our approach with the state-of-the-art LC method [24], which is one of the best performing seed based approach in the GrabCut-50 image segmentation task. In terms of three quality metrics, our method achieves better results for two metrics, *e.g.,* 0.9765 vs 0.9715 [24] for PRI, 0.1654 vs 0.1836 [24] for VoI, and achieves the same GCE score 0.0262, indicating that our method achieves the best scores in three evaluation qualities. This further advocates the value of our double sparse reconstruction method for object extraction compared to seed based methods. Our region merging procedure only depends on the confidence map derived from the sparse reconstruction method, which is understandable: compared to the user input strokes location, the confidence map carries on more information in describing objects of natural images.

4.2. PASCAL VOC

In order to more thoroughly evaluate segmentation performance, we have experimented with our method on the PASCAL VOC object segmentation dataset [32]. We use images from the validation dataset to evaluate the method performance, where the bounding box for each object is

Method	IoU
GraphCut	63.1 %
SegmentsSum	56.7 %
Xia [6]	72.6 %
Our	**73.2 %**

Table 3. IoU of bounding box prior based algorithms GraphCut, SegmentsSum, Xia [6], and our method on the PASCAL VOC 2011 validation Set.

Method	[3]	[11]	[25]	[27]ad.	[27]	[18]	[3]	[2]	ours
N_c	1100	1100	1100	1100	1100	100	100	100	–
avg	71.6	71.4	67.4	63.1	58.9	63.7	61.7	59.0	73.3

Table 4. Jaccard similarity (%) of our method vs oracle scores of object proposals methods on VOC 2012 validation dataset [32].

provided. We use Intersection over Union (IoU) [32] measure to evaluate the performance of comparing methods. The weight α in Eq. (5) is set over the interval $[0.8, 2]$ for all experiments on VOC validation dataset. We vary the parameter α with a step size of 0.1. Each specific category shares a fixed α.

Compared with State-of-the-art Methods: A series of experiments have been conducted on the VOC 2011 validation database which contains 1,112 images. For these images, the bounding boxes are provided by ground truth. To demonstrate the effectiveness of the proposed method for utilizing the segment hypotheses to extract objects, we first compare our method with the result that directly merges segments generated by CPMC [2], named as SegmentsSum. The results are reported in Table 3. SegmentsSum achieves the Jaccard score of 56.7% on this database. Our method gives a huge boost in segmentation accuracy. It obtains 73.2% by leveraging the segment hypotheses based on a sparse reconstruction framework. For fair comparison, we further compare our method against the state-of-the-art methods GraphCut and [6]. With the bounding box prior, the method of GraphCut and [6] achieve 63.1% and 72.6% of average Jaccard score on this database, respectively. Our method also outperforms these two baselines GraphCut and [6] which integrates object shape guidance generated by CPMC method into their graph-cut-based optimization. This further demonstrates that the image objects can be effectively reconstructed by our proposed double sparse reconstruction method. And our methodology is more powerful than the segment hypothesis integrated approaches for extracting objects.

Some qualitative extraction results of our method are visualized in Fig. 8. Many objects from different categories are included which are often with intrinsic inhomogeneity. The extracted objects of our method are highlighted by color mask. It is visually clear that our method can produce satisfactory results for extracting the objects of large appearance or pose variations in natural images. However, in some cases, our method will fail, as shown in Fig. 9

Compared with Object Candidate Generation Meth-

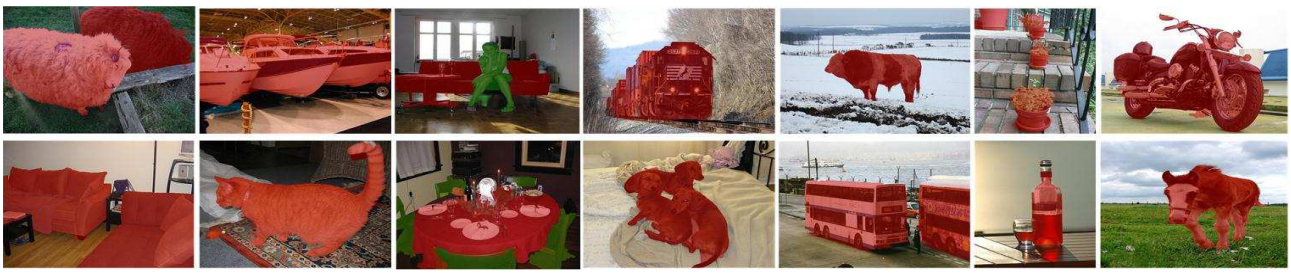

Figure 8. Object extraction results of our approach on VOC validation dataset: the extracted objects are highlighted by color mask overlaid on images, objects from different bounding boxes in an image are illustrated by different colors. (Best view in color)

ods: As the segment hypotheses are generated by the object proposals method [2] in our method, in this part, we will compare our proposed method to the state-of-the-art object proposals methods to demonstrate the effectiveness of our strategy for object extraction. Note that, to measure the quality of proposals, candidates generation methods report an "oracle" score that selected best candidate for each object (also Best Spatial Support score (BSS) [35]) among the pool with respect to the number of candidates. For example, when the number of object candidates is 100, each one is evaluated with the ground truth and then reports the best score among them. Obviously, the Jaccard score of our method is not an oracle score.

Table 4 reports the detailed comparison of Jaccard similarity of our method with the oracle scores of object proposal generation methods on VOC 2012 validation dataset. These methods include [2][3] [18] [11] [25] [27]. On this database, when the total number of candidates of these methods is 100, CPMC [2] achieves 59% of oracle score and the state-of-the-art object proposal method MCG [18] achieves 63.7% of oracle score. The proposed methodology achieves 73.3% of the mean Jaccard score (which is not an oracle score) on this dataset, clearly outperforming the state-of-the-arts. This verifies the effectiveness of our algorithm for obtaining accurate extraction results: the performance is even comparable with the oracle scores of the state-of-the-art object proposals methods.

Note that in this work, we do not mean to claim that our method is always superior over MCG method. It is predicated that in [18], MCG can achieve better results with generating more object candidates. For example, when they generate about 1000 object candidates for each image then report the best overlap ones, the oracle score increases to 76.0%. It is reasonable that they can achieve better results when the number of object candidates among the pool is getting larger. However it makes more difficult to pick out the best proposal.

5. Conclusion

This paper presents a double sparse reconstruction method to extract objects from images with the bounding box prior. Object regions can be well reconstructed by our model since they are with small reconstruction errors to segment hypotheses bases, simultaneously, large reconstruction errors to background hypotheses bases. Region merging procedure is finally used to make the reconstructed object with a well-preserved boundary. The proposed object extraction method is examined on two popular segmentation databases: PASCAL VOC object segmentation dataset and GrabCut-50 database, and experimental results indicate that (i) the proposed method is more robust than state-of-the-art semi-supervised methods for object extraction, and (ii) the proposed double sparse reconstruction scheme is more powerful than the segments integrated approaches for characterizing the correlation information between regions. Our future work will focus on how to obtain bounding box for natural image with more robustness.

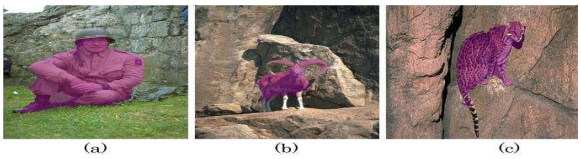

Figure 9. Some failure cases of our approach. The results are overlaid on the images with same color. (a) is due to the "hat" region, since it has very similar attributes with background regions, meanwhile little segment hypothesis has segment the hat into figure-ground. (b) is due to inaccurate super-pixels acquisition, since the legs of the object are so small. (c) is due to the large texture variation, region merging procedure cannot extract the tail accurately.

6. Acknowledgments

This work was partially supported by the National Science Fund for Distinguished Young Scholars under Grant Nos. 61125305, 91420201, 61472187, 61233011 and 61373063, the National Natural Science Fund of China(Grant Nos. 61103058), the Key Project of Chinese Ministry of Education under Grant No. 313030, the 973 Program No. 2014CB349303, Fundamental Research Funds for the Central Universities No. 30920140121005, and Program for Changjiang Scholars and Innovative Research Team in University No. IRT13072.

References

[1] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma. Robust Face Recognition via Sparse Representation. *TPAMI*, 2009.

[2] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 34(7): 1312-1328, July 2012.

[3] I. Endres and D. Hoiem. Category-independent object proposals with diverse ranking. *TPAMI*, 36(2): 222-234, 2014

[4] D. Lowe. Distinctive image feature from scale-invariant keypoints. *IJCV*, 2004.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005.

[6] W. Xia, C. Domokos, J. Dong, L. Cheong, and S. Yan. Semantic Segmentation without Annotation Segments. In *ICCV*, 2013.

[7] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels. *Technical Report 149300, EPEL,* 2010.

[8] M. Meila. Comparing clustering: An axiomatic view. In *ICML*, 2005.

[9] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *TPAMI*, 2007.

[10] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 33(5):898-916, May 2011.

[11] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Sematic segmentation using regions and parts. In *CVPR*, 2012.

[12] V. Lempitsky, P. Kohli, C. Rother and T. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009.

[13] Y. Chen, A. Chan and G. Wang. Adaptive figure-ground classification. In *CVPR*, 2012.

[14] X. Li, H. Lu, L. Zhang, X. Ruan and M. Yang Saliency Detection via Dense and Sparse Reconstruction. In *ICCV*, 2013.

[15] Y. Boykov and G. Funka-lea. Graph cuts and efficient n-d image segmentation. *IJCV*, 70(7): 109-131, 2006.

[16] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. *SIGGRAPH 04*, 23(3): 309-314, Mar. 2004.

[17] T. Brox, L. Bourdev, S.Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, 2011.

[18] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale Combinatorial Grouping. In *CVPR*, 2014

[19] P. Flzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, Vol. 32, pp. 1627-1645, 2010.

[20] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV* , volume 2, pages 416-423, July 2001.

[21] C. Couprie, L. Grady, L. Najman, and H. Talbot. Power Watershed: A unifying graph-based optimization framework. *TPAMI*, 33(7): 1384-1399, 2011.

[22] J. Cousty, G. Bertrand, L. Najman, and M. Couprie. Watershed cuts: Minimum spanning forest and the drop of water principle. *TPAMI*, 31(8): 1362-1374, 2009.

[23] L. Grady. Random walks for image segmentation. *TPAMI*, 28(11): 1768-1783, 2006.

[24] W. Casaca, L. Nonato, and G. Taubin. Laplacian Coordinates for Seeded Image Segmentation. In *CVPR*, 2014.

[25] J. Kim and K. Grauman. Shape sharing for object segmentation. In *ECCV*, 2012

[26] J. Ning, L. Zhang, D, Zhang and C. Wub. Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition*, 2010

[27] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

[28] M. Maire and S. X. Yu. Progressive multigrid eigensolvers for multiscale spectral segmentation. In *ICCV*, 2013.

[29] J. Carreira, F. Li, and C. Sminchisescu. Object recognition by sequential figure-ground ranking. *IJCV*, 98(7): 243-262,July 2012.

[30] C. Gu, J.Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, 2009.

[31] A. Y. S. Didler, R. Mottaghi and R. Urtasun. Bottom-up segmentation for top-down detection. In *CVPR*, 2013

[32] M. Everingham, L. Van Gool, C. Williams, J. Winn and A. Zisserman. The pascal object classes challenge (VOC). http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[33] Emmanuel Candes, J. Romberg L1-magic: Recovery of sparse signals via convex programming. 2005

[34] Z. C. Lin, M. M. Chen, and Y. Ma, The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrix. UIUC,Champaign, IL, USA, Tech. Rep. UILU-ENG-09-2215, Oct. 2009.

[35] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentation. In *BMVC*, 2007

[36] M. Tang, I. Ben Ayed, D. Marin, Y. Boykov. Secrets of GrabCut and Kernel K-means In *arXiv*, June 24, 2015.