Neurocomputing 219 (2017) 323–332

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Boundary-aware box refinement for object proposal generation

Xiaozhi Chen, Huimin Ma*, Chenzhuo Zhu, Xiang Wang, Zhichen Zhao

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Communicated by Jungong Han

Keywords:
Object proposals
Object detection
R-CNN

ABSTRACT

Object proposals have been widely used in object detection to speed up object searching. However, many of existing object proposal generators have pool localization quality, which weakens the performance of object detectors. In this paper, we present an effective approach to improve the localization quality of object proposals. We leverage the boundary-preserving property of superpixels and design an efficient algorithm for object proposal refinement. Our approach first performs bounding box alignment to adapt proposals to potential object boundaries, and then diversifies the proposals via multi-thresholding superpixel merging. The algorithm only takes 0.15 s and can be applied to any existing proposal methods to improve their localization quality. Extensive experiments on PASCAL VOC 2007 and ILSVRC 2013 datasets show our approach significantly and consistently improves the recall, localization accuracy, and detection performance of existing proposal methods. When combining with Region Proposal Network, our approach outperforms the state-of-the-art object detectors by a large margin.

1. Introduction

Object proposal generation has become a crucial technique for many vision recognition tasks, such as class-specific object detection and instance segmentation. The goad of object proposal generation is to select a small set of object candidates that cover most of the objects in an image. The advantages of object proposals over traditional sliding windows [1] lie in two aspects: reducing computations with fewer regions of interest, and improving accuracy by using more sophisticated features and classifiers. Recent object detectors [2–4] utilizing object proposals have achieved state-of-the-art performance on challenging PASCAL [5], ImageNet [6] and MS COCO [7] datasets.

Two distinct pipelines emerge for object proposal generation: bottom-up approaches based on low-level cues, and data-driven approaches based on Convolutional Neural Networks (CNN). Most bottom-up approaches generate proposals by window scoring or region grouping. Existing models belonging to this pipeline struggle to achieve a good balance between localization accuracy and computational efficiency. In particular, window scoring based methods, such as BING [8], are computationally efficient, but they suffer from pool localization quality, i.e., low recall under strict intersection over union (IoU) overlap criteria (e.g., IoU > 0.7). Region grouping based methods such as Selective Search [9] and MCG [10] are computationally expensive, while they usually achieve higher localization quality. Data-driven approaches [4,11] leverage rich convolutional neural network features to directly predict the regions of interest. Due to the powerful discrimination ability of CNN features, these models typically achieve higher recall than bottom-up proposals under loose overlap criteria (e.g., IoU < 0.7). However, similar to window scoring based methods, they usually have much lower recall at high overlap threshold.

Motivation. Most window scoring based and CNN-based methods fail to achieve high recall under strict overlap criteria. While region grouping based methods have better localization accuracy, they are usually computationally expensive. The goal of this paper is to improve localization accuracy of existing object proposals while preserving computational efficiency. Our work is inspired by the observation that superpixels have the good property of preserving object boundaries, which could benefit object localization. Prior superpixel-based methods typically exploit low-level features like color, texture, to compute region similarity for image partition, which is usually time-consuming. In contrast to those methods, we develop a very efficient algorithm by only using superpixel straddling to guide the superpixel merging process. Furthermore, instead of exploiting multiple segmentations to diversify proposals, we propose multi-thresholding superpixel merging for efficient diversification.

Overview. Our main idea is to utilize superpixel boundaries to refine candidate bounding boxes. Our approach consists of three stages: 1) Initialize a set of candidate bounding boxes using any existing proposal method. Note that our approach is agnostic to the proposal method. 2) Perform bounding box alignment to adapt proposals to boundaries of superpixels. 3) Diversify the proposals

* Corresponding author.
E-mail address: mhmpub@tsinghua.edu.cn (H. Ma).

http://dx.doi.org/10.1016/j.neucom.2016.09.045
Received 29 March 2016; Received in revised form 18 July 2016; Accepted 18 September 2016
Available online 23 September 2016

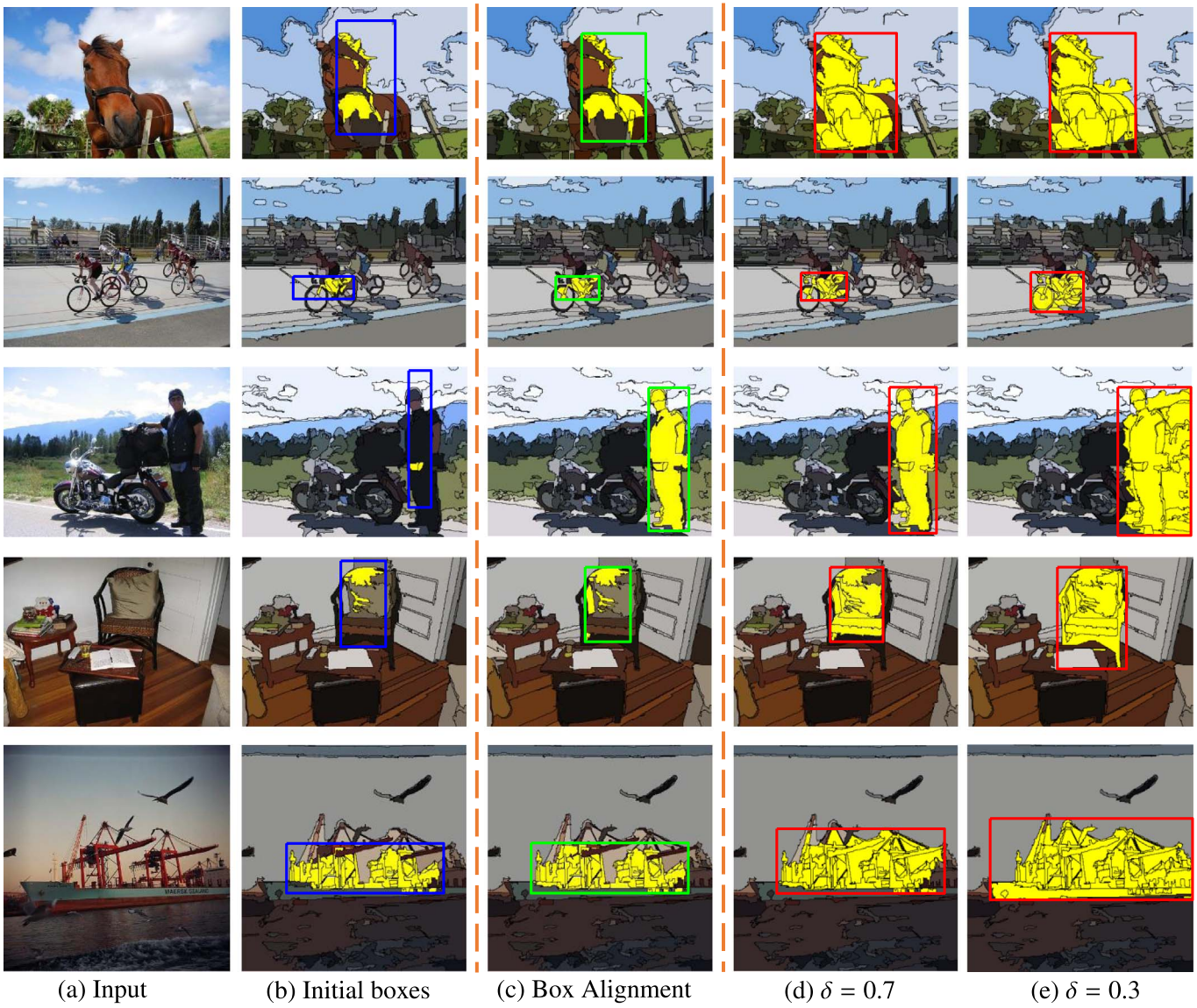


Fig. 1. Illustration of our method using several examples. (a) Input images. (b) Initial bounding boxes. (c) Boxes after alignment. (d-e) Proposals after straddling expansion by setting the threshold δ to 0.7 and 0.3, respectively. Superpixels wholly enclosed by a bounding box are indicated in yellow, (a) Input (b) Initial boxes (c) Box Alignment (d) $\delta = 0.7$ (e) $\delta = 0.3$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

using multi-thresholding superpixel merging. The overall process is illustrated in Fig. 1. The superpixel merging process only uses superpixel straddling feature, which can be computed very efficiently, thus our approach preserves computational efficiency when integrated into existing models. We also introduce superpixel tightness (ST) as an indicator to measure the localization accuracy of object proposals without access to ground truth boxes. The statistical behavior of the object proposals on PASCAL VOC 2007 dataset verifies the effectiveness of our approach.

Contributions. The main contributions of our work are as follows:

- A boundary-aware method is proposed to improve localization accuracy of object proposals. An efficient algorithm which takes only 0.15 s per image is presented for bounding box alignment and superpixel merging.
- An analysis of proposal localization is provided by introducing a superpixel tightness measure, which demonstrates the statistical behavior of our approach.
- Extensive experiments are conducted on PASCAL VOC 2007 and ImageNet ILSVRC 2013 datasets, which show that our approach significantly boosts the recall and localization quality of existing proposal methods. We further apply the improved object proposals to Fast R-CNN for object detection. **In particular, our approach outperforms the state-of-the-art Faster R-CNN [4] on VOC 2007 dataset, achieving 72.2% detection mAP**.

A preliminary version of this work was presented in [12]. Extensions are made in four aspects: 1) An extensive review of object proposal methods, including both bottom-up proposals and CNN-based proposals. 2) We further apply the approach to CNN-based proposals, in particular, the Region Proposal Network (RPN). 3) Extensive experiments on ImageNet ILSVRC 2013 datasets, which show the generalization ability of our approach. 4) The approach is extended by combining with Fast R-CNN for object detection. We show our approach outperforms the state-of-the-art Faster R-CNN [4] by 2.9% mAP on VOC 2007 dataset.

2. Related works

We first review two kinds of object proposal generation approaches,

namely *bottom-up* proposals and *data-driven* proposals. Then we briefly review related applications of object proposals.

Bottom-up proposals. This paradigm generates object proposals by exploiting low-level cues. In general, these methods can be divided into window scoring based and region grouping based methods. We refer the reader to [13,14] for an in-depth survey.

Window scoring based methods typically design a scoring function to rank a large set of candidate bounding boxes. The scoring functions are supposed to be able to distinguish objects from amorphous background stuff. To this end, various low-level cues are exploited in the scoring functions, such as color contrast, saliency [15], superpixels, location and size [16], binarized normed gradients [8], and edge maps [17]. While these approaches are computationally efficient, they can hardly achieve high recall under strict overlap criteria (e.g., IoU > 0.7). Their pool localization quality are mainly due to the fact that they usually use regular sampling to sample bounding boxes which can hardly locate object precisely. For instance, BING [8] has very high recall at IoU threshold of 0.5 and is also extremely fast in speed. However, its recall drops dramatically with the IoU threshold increasing due to its coarse quantization scheme [18,19]. To lessen loss caused by regular sampling, EdgeBoxes [17] refines top scoring bounding boxes via greedy iterative local search. But it gives no guarantee of alignment to object boundaries since the refinement is performed with a fixed searching step. Our approach aligns bounding boxes with potential object boundaries preserved by superpixels, thus obtains better localization. In fact, we will show that our approach can further improve EdgeBoxes.

Region grouping based methods perform superpixel merging or multiple segmentations to generate segment proposals. These approaches usually measure region similarity using diverse and complementary cues such as color, texture and location. Selective Search (SS) [9] performs superpixel segmentations using multiple scales and color spaces, and designs a hierarchical grouping algorithm to obtain region proposals. MCG [10] generates proposals by exploring combinatorial space in hierarchical segmentations and grouping multi-scale regions. Randomized Prim (RP) [20] produces candidate regions by computing random partial spanning trees in a superpixel connectivity graph. GOP [21] places promising object seeds and identifies candidate regions by employing signed geodesic distance transform. GLS [22] produces candidate regions by performing superpixels merging locally and graph cut globally. Furthermore, some other models [23–26] address it as multiple figure-ground segmentations and minimize a set of parametric energies. To improve proposal quality, most of these methods increase candidate diversity by employing segmentations in multiple scales and color spaces, which, however, requires much more computations (typically seconds). While our approach also leverages superpixels, we propose a very fast superpixel merging algorithm which does not require multiple segmentations to obtain diversity, thus saving computational cost.

Data-driven proposals. With the advances of deep neural networks on visual recognition [3,27–30], recent efforts on object proposals exploit data-driven approaches using powerful convolutional network features. MultiBox [11,31] learns to directly predict the coordinates of a fixed number of proposals and the corresponding confidences. Region Proposal Network (RPN) [4] regresses the coordinates of bounding box proposals relative to a set of pre-defined translation-invariant anchors. DeepBox [32] uses a small convolutional network to rerank bounding box proposals generated by EdgeBoxes [17]. DeepProposals [33] scores a large set of candidate bounding boxes using the last convolutional layers and refines their locations using the initial layers. DeepMask [34] designs a two-branch network to predict multiple figure-ground masks and confidences for dense image patches. Compared with bottom-up proposals, CNN-based proposals typically require fewer proposals (less than 1000) to achieve very high recall using at low overlap threshold (e.g., IoU < 0.6). However, the pool localization problem is also observed in CNN-based proposals. While CNN models have strong semantic extraction ability, they usually bear the loss of spatial information due to downsampling. As a result, most CNN-based proposals have relative lower recall under strict overlap criteria (e.g., IoU > 0.7). We show that when utilizing superpixels, our approach can effectively improve the localization quality of the state-of-the-art RPN [4] proposals, leading to significantly better detections on PASCAL VOC 2007.

Our approach is essentially a box refinement method that can be applied to any existing object proposals to improve their localization accuracy. With the boundary-preserving property of superpixels, our approach typically provides complementary cues to window scoring based and CNN-based proposals. Due to the diversification effect of our approach, it can further improve the recall of region grouping proposals.

Object proposals have been widely applied to vision recognition tasks, such as object detection [2–4], instance segmentation [35,36], and object tracking [37–40]. The state-of-the-art R-CNN [2,3] object detectors heavily rely on the quality of object proposals. Hosang et al. [13] investigates the effect of various object proposal methods on detection results and suggests that high recall and accurate localization are critical to achieving high-performance object detection. Hariharan et al. [35] combines object proposals with CNN to perform object detection and segmentation simultaneously. Yang et al. [37] formulates the object tracking problem as proposals selection and shows superior performance over previous works. In addition, it's interesting to see how object proposals are applied to 3D point cloud [41] and medical images [42,43]. While our object proposals can be applied to many vision recognition tasks, in this paper we focus on its application on object detection.

3. Proposed approach

We first present our box alignment method in Section 3.1 and the multi-thresholding superpixel merging algorithm in Section 3.2. Then we introduce a measure in Section 3.3 which demonstrate the statistical behavior of our method.

3.1. Box alignment

Given a pool of candidate bounding boxes generated from certain proposal method, we first use superpixels to align them with object boundaries. This step is typically beneficial to proposals which are generated without exploitation of the superpixel cues. For instance, BING proposal [8] selects a subset of box candidates from sliding windows, which are sampled uniformly with fixed sizes and aspect ratios, thus they possibly align badly with object boundaries.

Since the exact object boundaries are not accessible, we use superpixels as surrogate and align bounding boxes with the boundaries of superpixels. Given an image, we obtain a set of superpixels $\mathcal{S}_\theta$ via an over-segmentation with parameters θ. If a bounding box b is the minimum box enclosing a subset of the superpixels $\mathcal{S}_\theta$, then we say it aligns with superpixels $\mathcal{S}_\theta$.

Given an initial bounding box b, the goad of box alignment is to output a new box so that it not only aligns with superpixel boundaries, but also has the highest overlap with the initial box. To this end, we first compute the inner set $\mathcal{S}_{in}$ and straddling set $\mathcal{S}_{st}$ of box b, which are defined as

$$\mathcal{S}_{in} = \{s \in \mathcal{S}_\theta | SD(s, b) = 1\}, \mathcal{S}_{st} = \{s \in \mathcal{S}_\theta | 0 < SD(s, b) < 1\}, \quad (1)$$

where $SD(s, b) = |s \cap b|/|s|$ is the *straddling degree* of superpixel s with regard to box b. Intuitively, $\mathcal{S}_{in}$ represents the superpixels wholly enclosed by box b, and $\mathcal{S}_{st}$ represents the superpixels containing pixels both inside and outside b.

Let $b(\mathcal{S})$ denote the minimum box enclosing superpixels $\mathcal{S}$, and $O(b_i, b_j)$ the IoU overlap between b_i and b_j. We then sort the elements in the straddling set $\mathcal{S}_{st}$ according to the IoU overlaps, so that its

elements $\{s_1, \ldots s_K\}$ satisfy

$$O(b(\mathcal{S}_{in} \cup \{s_i\}), b) \geq O(b(\mathcal{S}_{in} \cup \{s_j\}), b), \quad \forall\, i < j. \tag{2}$$

Let $b(S_{in})$ denote the minimum box enclosing the inner set. The box alignment process is to merge superpixels greedily from the sorted straddling set, in order to expand the bounding box from $b(S_{in})$ to the one which is closest to the given box b. By this means, we obtain a new bounding box $b^\star$ which aligns with superpixel boundaries and also has the highest overlap with the initial box b. We summarize the specific procedure in Algorithm 1.

Algorithm 1. Box Alignment

Input: initial box b, superpixels $\mathcal{S}_\theta$
Output: aligned box $b^\star$
1: compute inner set: $\mathcal{S} \leftarrow \mathcal{S}_{in}$
2: obtain sorted straddling set: $\{s_1, \ldots s_K\}$
3: $k \leftarrow 1$
4: $o \leftarrow O(b(\mathcal{S}), b)$
5: $\hat{o} \leftarrow O(b(\mathcal{S} \cup \{s_k\}), b)$
6: **while** $\hat{o} \geq o$ **do**
7: $o \leftarrow \hat{o}$
8: $\mathcal{S} \leftarrow \mathcal{S} \cup \{s_k\}$
9: $k \leftarrow k + 1$
10: $\hat{o} \leftarrow O(b(\mathcal{S} \cup \{s_k\}), b)$
11: **end while**
12: $b^\star \leftarrow b(\mathcal{S})$

We illustrate some examples in the third column of Fig. 1. Intuitively, box alignment is capable of "dragging" the coarse bounding boxes back to the main part of an object, as shown in Row 1–3 in Fig. 1. In some cases (e.g., Row 3 in Fig. 1), the box alignment procedure is already sufficient to re-localize the object precisely.

3.2. Multi-thresholding superpixel merging

As box alignment only refines proposals to align it with superpixels boundaries, it can not improve the localization quality of some initial proposals which have small overlap with objects. Therefore, we propose the multi-thresholding superpixel merging algorithm as the second stage to further diversify proposals. Given an aligned bounding box $b^\star$, we perform superpixel merging based on its straddling with superpixels. Formally, given a threshold δ, we define *straddling expansion* as the following refinement:

$$\mathcal{S}_\delta(b^\star) = \mathcal{S}_{in}(b^\star) \cup \{s \in \mathcal{S}_\theta | SD(s, b^\star) \geq \delta\}. \tag{3}$$

By computing the minimum box enclosing $\mathcal{S}_\delta(b^\star)$ we obtain a new box $\hat{b}$. Some examples of straddling expansion with different δ values are visualized in Fig. 1. Intuitively, large value of δ produces a minor variant of b, which is desired for proposals that already have moderate overlap with an object. On the other hand, a distinct box can be obtained with small value of δ, which can increase the possibility of jumping out of a "local minima" for inaccurate box.

As a fixed threshold is not always optimal for all bounding boxes, we use multiple δ's to perform straddling expansion. By this means, multiple bounding boxes are generated for each initial bounding box. In practice, we perform straddling expansion five times by setting the threshold δ to $0.1 \times i$, $i = 1, 2, \ldots, 5$, which are determined via cross-validation. As a consequence, this generates five sets of bounding boxes. To reduce redundancy, we sort each set by adding some randomness. Specifically, let $\hat{b}_i$ be the bounding box refined from the initial box b_i using certain value of δ. We score $\hat{b}_i$ with value $i \times R$, where R is a random number in range [0,1]. We obtain a ranked list of candidate boxes by sorting all the boxes in ascending order. Non-maximal suppression (NMS) is performed after ranking to obtain a final set of proposals.

The unique benefit of our box refinement method is that it can naturally generate bounding boxes aligning with object boundaries preserved by superpixels. This property differentiate our method from EdgeBoxes [17] which performs fixed-step local search. Moreover, unlike existing superpixel merging methods (e.g., SS [44]), straddling expansion doesn't require extracting low-level features such as color and texture to measure regions similarity. Only straddling degrees are computed for superpixels, thus the algorithm is very efficient.

3.3. Analysis of localization

The advantage of region grouping proposals over window scoring proposals mainly lies in its higher localization quality. Owing to the leverage of superpixels, our approach is also able to obtain high localization accuracy. To understand the effect of our method, we introduce an indicator, *superpixel tightness (ST)*, which measures how tight a bounding box encloses an object. Formally, the *ST* of a bounding box b is defined as the proportion of the area of superpixels wholly enclosed by box b to the area of b:

$$ST(b) = \sum_{s \in \mathcal{S}_\theta} \frac{|s| \cdot \delta(|s| - |s \cap b|)}{|b|}, \tag{4}$$

where $\delta(x)$ is the Dirac delta function which takes value of 1 if x=0 and 0 otherwise. For superpixels s wholly enclosed by b, we sum up the number of pixels contained in the superpixels and divide by the area of b. $ST(b)$ is 0 when none of the superpixels is wholly enclosed by b. The *ST* measure is similar to the superpixels straddling cue introduced in [16]. Our *ST* measure differs from it by disregarding superpixels straddling the box.

The superpixel tightness measure can serve as an indicator of the object proposal localization quality. To show this, we compute the *ST* statistics on PASCAL VOC 2007 dataset. We first plot the distributions of superpixel tightness for ground truth bounding boxes and background regions respectively in Fig. 2(a). We sample background regions randomly from sliding windows and the overlap of the regions with ground truth boxes should be less than 0.5. It can be seen from Fig. 2(a) that objects have diverse degrees of superpixel tightness while the majority of background regions incline to low value of *ST*.

Based on this observation, a high-quality object proposal generator should produce box candidates with superpixel tightness distribution similar to that of ground truth objects. However, most window scoring methods fail to make it. For demonstration, we plot the *ST* distributions for proposals generated by several methods in Fig. 2(b). In particular, we test BING [8], OBJ [16], EB [17], RP [20], GOP [21], SS [9] and MCG [10]. Fig. 2(b) clearly shows that all window scoring proposals (i.e., OBJ, BING, and EB) have strong bias to low tightness while the region grouping based proposals have distributions more similar to the ground truth distribution. This accords with their difference in localization quality. Therefore, we can use *ST* distribution as an indicator of the localization quality.

Similar to region grouping methods, our superpixel merging method is able to generate object proposals with *ST* distribution close to the ground truth. We plot the *ST* distributions after applying our approach to the bounding box proposals generated by BING [8], OBJ [16], SS [44] and MCG [10] in Fig. 3. We can observe a shift of *ST* distribution from low *ST* to high *ST* after applying our box refinement to BING and OBJ. Note that even for SS and MCG, which are region grouping methods, we also obtain *ST* distributions closer to the ground truth distribution.

4. Experiments

Implementation. We compute superpixel segmentation using [45] in Lab color space at a single scale. Specifically, the segmentation

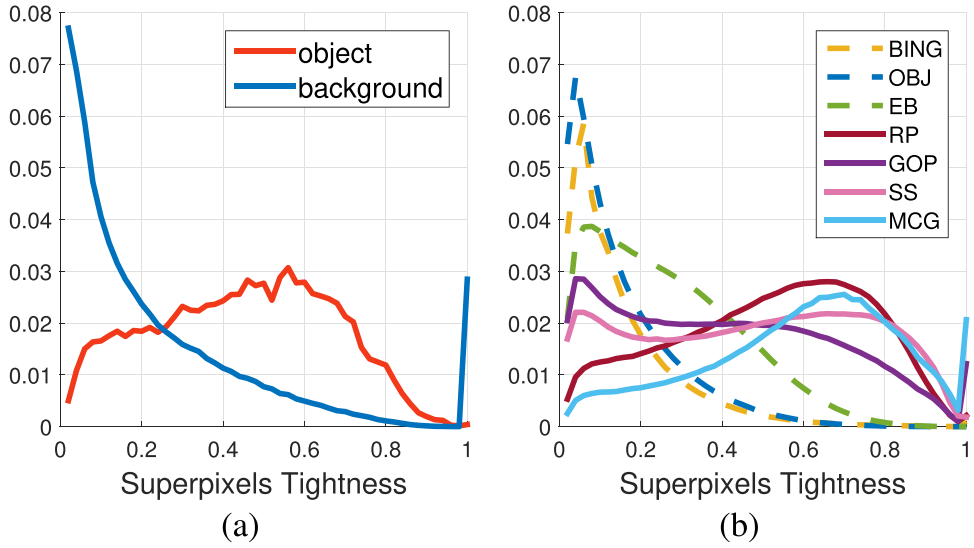


Fig. 2. Distributions of superpixel tightness for (a) ground truth objects and background regions on PASCAL VOC 2007 test set, and (b) 1 K object proposals generated by several window scoring based models (in dashed lines) and region grouping based models (in solid lines). The values at *ST*=0, which imply the proportion of bounding boxes that contain no superpixels entirely, are ignored in the figures for clarity. Best viewed in color.

parameters $\theta = (\sigma, k, min)$ are set to $\sigma = 0.8$, $k = 100$, min=100. For non-maximal suppression, we set the IoU threshold to 0.8 for window scoring proposals, and 0.9 for region grouping and CNN-based proposals, respectively. We found this setting can generate a moderate budget of object proposals with high accuracy.

Datasets. We evaluate our method on PASCAL VOC 2007 [46] and ImageNet ILSVRC 2013 [47] datasets. The VOC 2007 test set contains 4952 images and 14,976 object instances from 20 categories. Note that we follow Hosang et al. [13,14] to evaluate all objects including those annotated to be "difficult", which are removed in some other settings [8,17]. The ImageNet 2013 validation set has 200 categories with bounding box annotations in ~20,000 images.

Metrics. We evaluate recall and localization accuracy of object proposals, as well as the final detection performance by training Fast R-CNN [3] with proposals. Recall is computed as the fraction of ground truth bounding boxes covered by proposals above certain IoU overlap threshold. We use α-recall to denote recall at IoU overlap threshold of α. We use the recall vs proposal curve to depict recall for different number of proposals and recall vs overlap curve to illustrate the variation of recall under different IoU overlap criteria. We also evaluate Average Recall (AR) [13], which is computed as the area under "recall vs overlap" curve in overlap range 0.5–1.0. To evaluate localization accuracy, we also compute the Average Best Overlap (ABO) [44] for object proposals. Best overlap for each ground truth object is computed as its highest IoU overlap with object proposals.

To evaluate the impact of the proposals's localization quality on the object detection performance, we finally conduct detection experiments on PASCAL VOC 2007 by combining the state-of-the-art Fast R-CNN [3] model with object proposals.

4.1. Model analysis

We start by verifying the effectiveness of the two components of our method, namely box alignment and multi-thresholding superpixel merging. For this experiment, we use BING [8] for bounding box initialization. We compare two variants: BING with box alignment, and BING with both box alignment and superpixel merging. Recall curves are shown in Fig. 4. In particular, box alignment improves BING by 4.6% in average recall (AR) with 1000 proposals. After adding superpixel merging, we obtain another ~15% improvement. Note that our

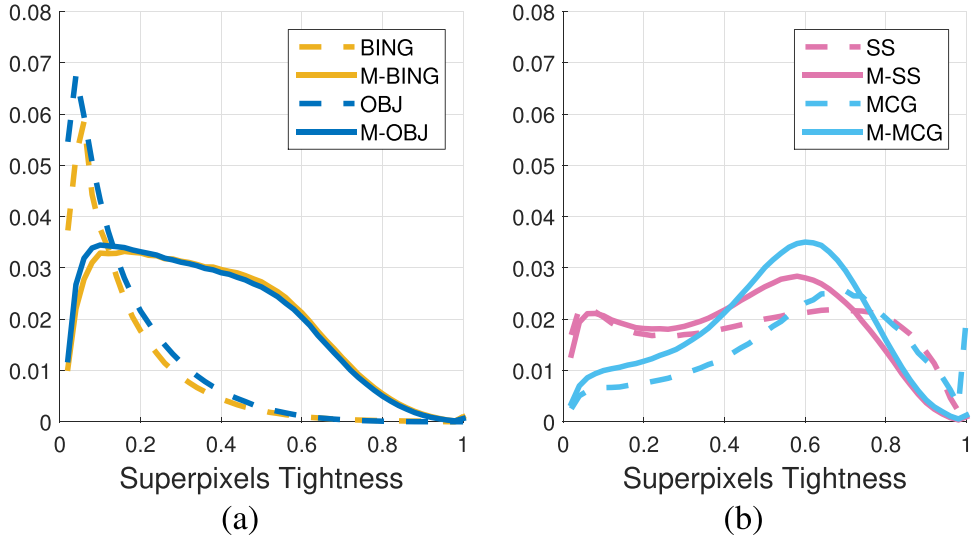


Fig. 3. Distributions of superpixel tightness before (in dashed lines) and after (in solid lines) applying straddling expansion for four baseline models: (a) BING [8] and OBJ [16], (b) SS [9] and MCG [10]. Best viewed in color.

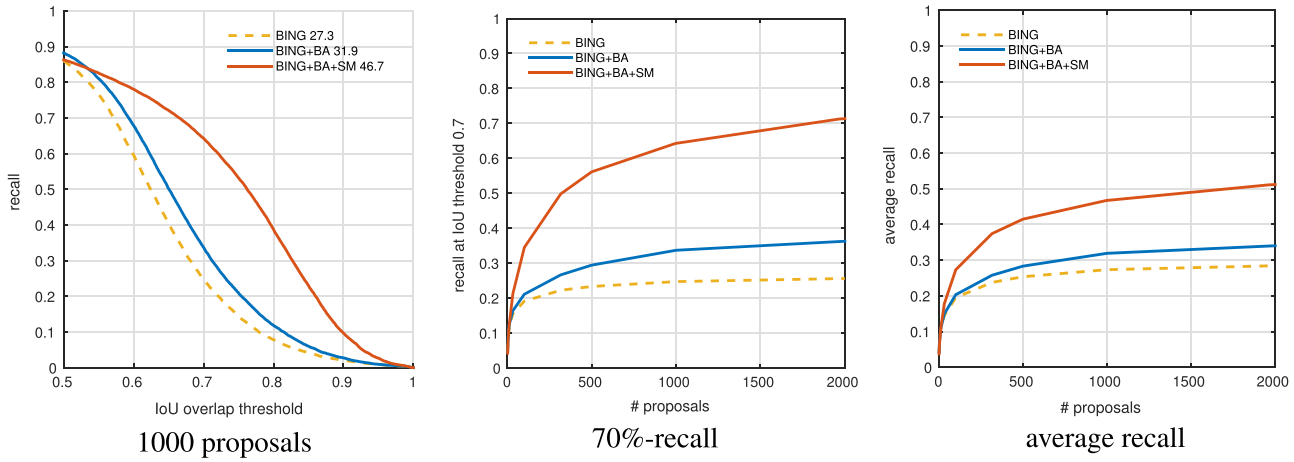


Fig. 4. Recall for BING and its improved versions using our approach. 'BING+BA': BING with box alignment; 'BING+BA+SM': BING with box alignment and superpixel merging. For the recall vs overlap curves, numbers next to labels indicate average recall (AR). Best viewed in color.

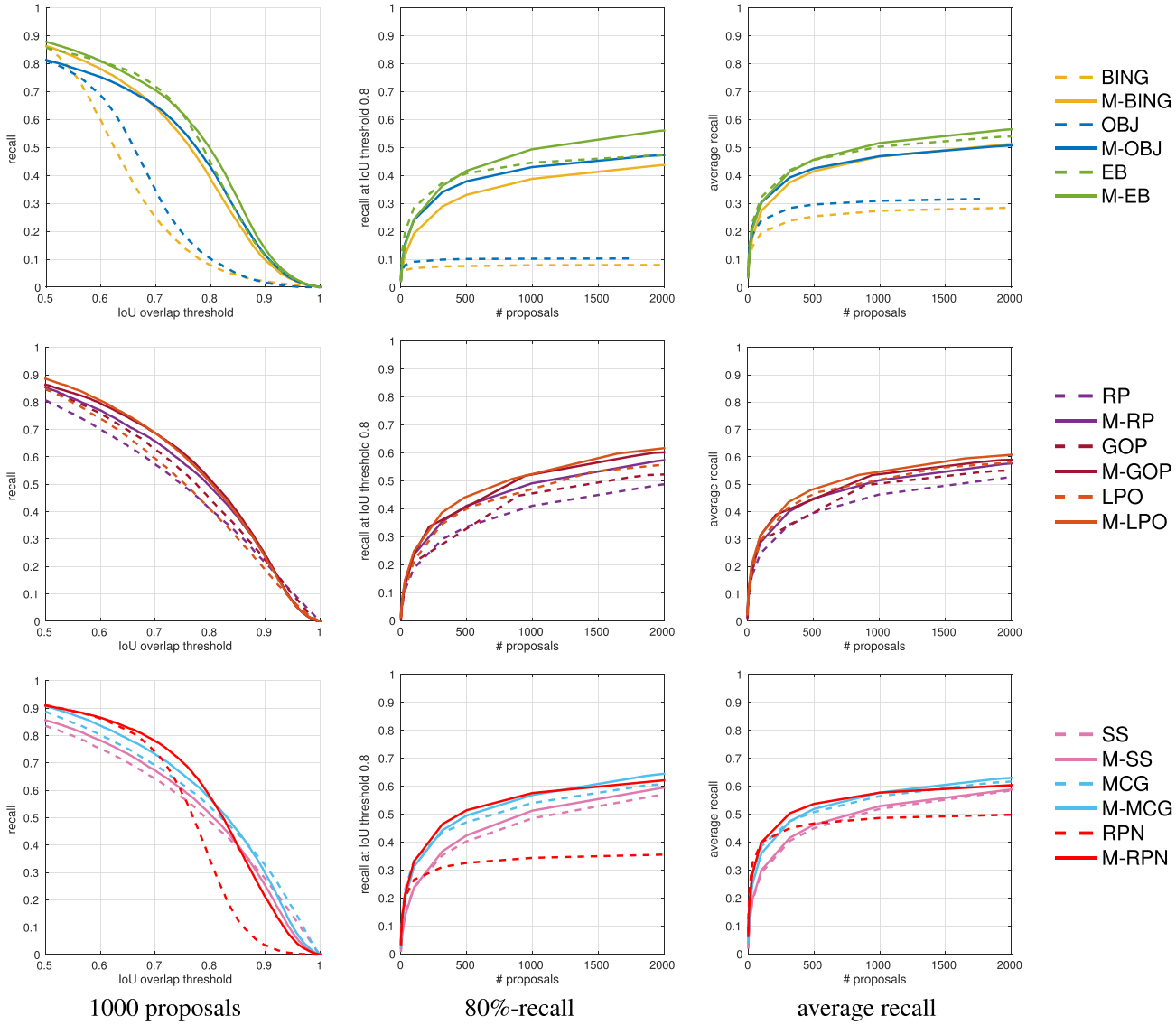


Fig. 5. Recall on PASCAL VOC 2007 *test* set for numerous models (in dashed lines) and their improved versions (in solid lines) using our method. From left to right: recall using 1000 proposals, recall at IoU of 0.8, average recall vs number of proposals. Best viewed in color.

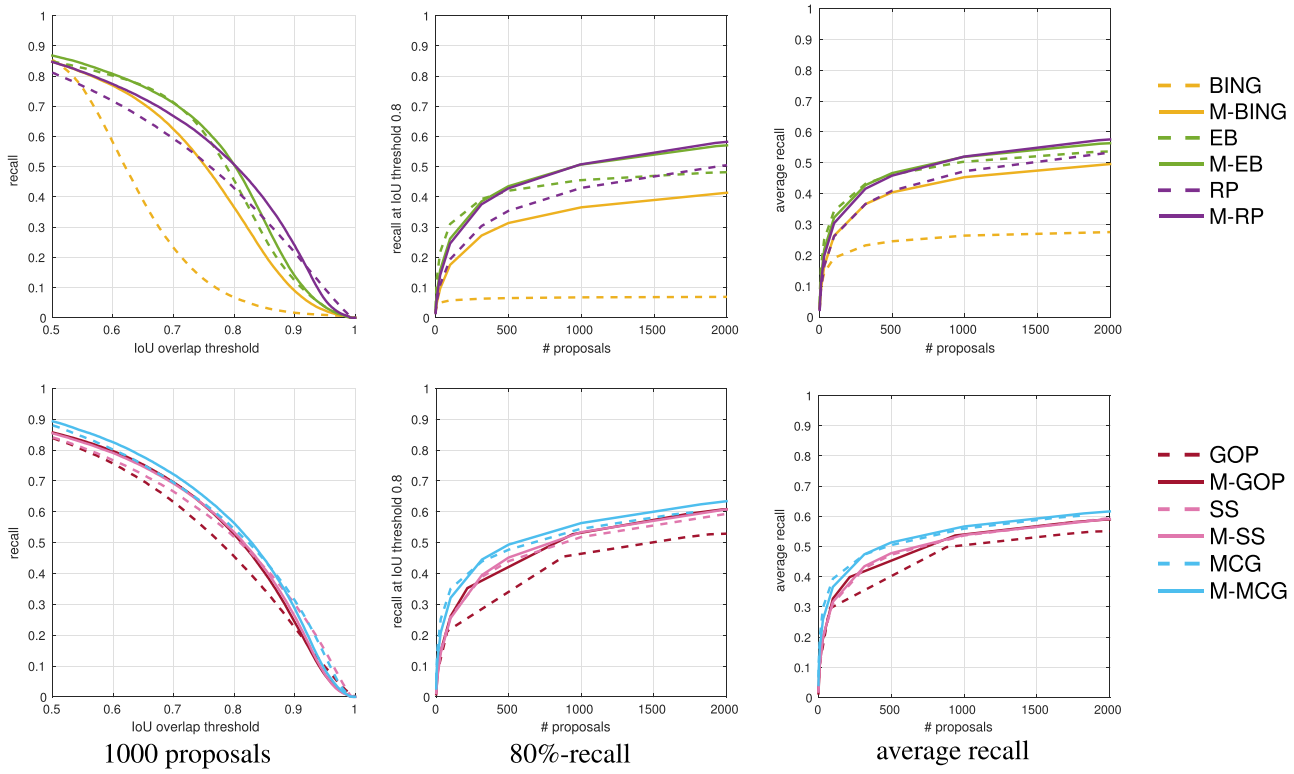


Fig. 6. Recall on ILSVRC 2013 *validation* set for numerous models (in dashed lines) and their improved versions (in solid lines) using our method. From left to right: recall using 1000 proposals, recall at IoU of 0.8, average recall vs number of proposals. Best viewed in color.

Table 1
Proposal results on PASCAL VOC 2007 *test* set. We report Average Recall (AR), Average Best Overlap (ABO) and recal at IoU of 0.8 for three budgets of proposals: 500, 1000, 2000. Numbers are shown for methods before/after applying our approach.

Method	# prop=500			# prop=1000			# prop=2000		
	AR	ABO	80%-recall	AR	ABO	80%-recall	AR	ABO	80%-recall
BING [8]	25.4/ **41.5**	58.9/ **67.0**	7.6/ **33.0**	27.3/ **46.7**	61.7/ **71.2**	7.9/ **38.7**	28.4/ **51.1**	63.1/ **74.3**	7.9/ **43.6**
OBJ [48]	29.6/ **42.5**	59.6/ **65.9**	10.1/ **37.9**	30.9/ **46.9**	61.8/ **69.6**	10.2/ **42.9**	31.6/ **50.4**	63.3/ **72.6**	10.3/ **47.0**
EB [17]	45.5/ **45.7**	67.7/ **69.5**	40.7/ **41.6**	50.2/ **51.5**	72.2/ **73.7**	44.5/ **49.3**	53.8/ **56.4**	75.4/ **76.9**	47.1/ **55.9**
RP [20]	39.6/ **44.8**	64.8/ **68.3**	33.6/ **41.0**	46.3/ **51.4**	70.0/ **73.3**	41.0/ **49.1**	52.6/ **57.4**	74.5/ **77.4**	48.6/ **57.1**
GOP [21]	28.8/ **38.9**	55.4/ **62.4**	20.2/ **33.6**	49.7/ **53.4**	72.3/ **74.3**	44.7/ **51.8**	54.3/ **58.1**	75.1/ **77.3**	51.2/ **58.7**
LPO [26]	45.2/ **47.9**	70.1/ **71.2**	38.6/ **44.0**	47.1/ **53.4**	71.5/ **75.1**	40.8/ **50.7**	56.1/ **59.4**	76.7/ **78.6**	53.2/ **59.7**
SS [44]	45.0/ **46.2**	68.0/ **69.2**	40.3/ **42.4**	51.9/ **52.9**	73.1/ **74.0**	48.5/ **51.2**	58.4/ **58.8**	77.5/ **77.9**	57.1/ **59.4**
MCG [10]	50.7/ **51.9**	73.1/ **73.6**	47.1/ **49.5**	56.4/ **57.7**	76.9/ **77.7**	54.0/ **56.8**	61.2/ **62.5**	79.8/ **80.6**	60.2/ **63.7**
RPN [4]	46.7/ **53.7**	71.1/ **74.6**	32.6/ **51.4**	48.6/ **57.7**	72.8/ **77.3**	34.4/ **57.6**	49.8/ **60.4**	73.8/ **79.0**	35.5/ **62.1**

full approach significantly boosts the recall at IoU of 0.7 from ~25% to ~65%, which implies the effectiveness of our method in improving localization accuracy.

Speed. We evaluate the runtime of our method on a 3.5 GHz CPU. The total time for our approach is 0.15 s using single thread, including 0.04 s for colorspace conversion, 0.1 s for superpixel segmentation, and 0.01 s for box alignment and superpixel merging. Therefore our method brings little computational overhead to existing models.

4.2. Recall evaluation

We integrate our method into numerous existing models. We evaluate proposal recall using recall vs overlap curve for 1000 proposals, recall vs proposal curve at IoU of 0.8, and AR vs proposal curve. Results are presented in Fig. 5 for PASCAL VOC 2007 and Fig. 6 for ImageNet 2013 datasets.

Baselines. As our approach can be applied to any object proposal methods, we conduct experiments by integrating it into numerous baseline models. In particular, we test on OBJ [16], BING [8], EB [17], RP [20], GOP [21], LPO [26], SS [9], MCG [10] and RPN [4]. Correspondingly, the improved versions are named M-OBJ, M-BING, M-EB, M-RP, M-GOP, M-LPO, M-SS, M-MCG and M-RPN, respectively. Note that the baselines cover a wide range of existing object proposal methods, including bottom-up proposals and CNN-based proposals.

PASCAL VOC 2007. Proposal recall plots and statistics are reported in Fig. 5 and Table 1. From Fig. 5 (Row 1) we observe that our approach significantly improves window scoring based models to a similar performance level with high recall consistently. In particular, for BING [8] and OBJ [16], which are typically tuned for low overlap, our method improves their recall at high overlap (i.e., IoU=0.8) significantly while preserving high recall at low overlap. For EB [17] which is tuned for IoU of 0.7, our method also obtains higher recall across a wider range of IoU threshold without losing edge at IoU of 0.7.

Despite most region grouping proposals already have quite good localization, our approach can further improves their performances (see Row 2–3 in Fig. 5). In particular, we achieve 6~13% improvement of recall at IoU of 0.8, which is very strict overlap criteria, in a wide

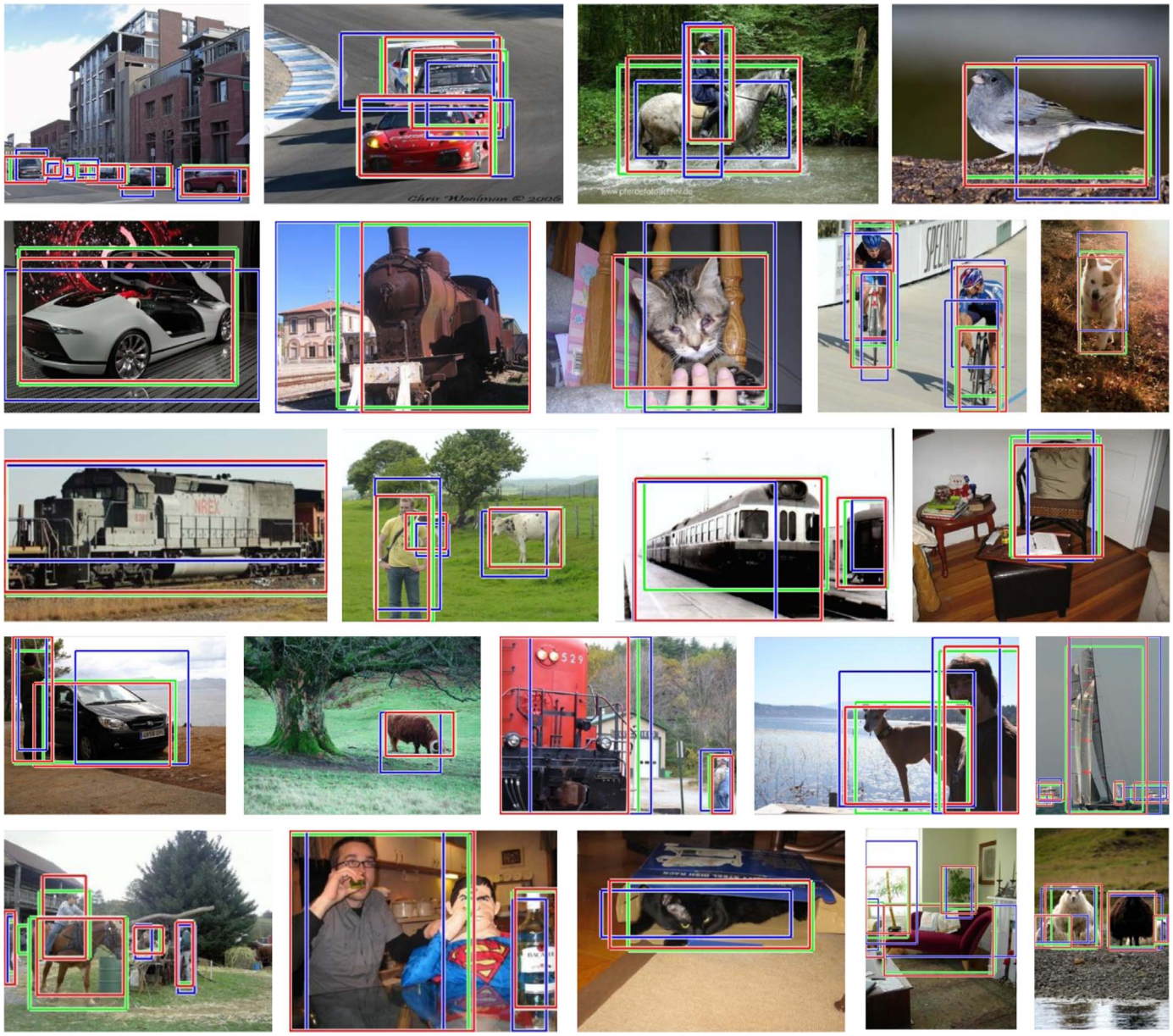


Fig. 7. Qualitative results of BING (blue) and its improved version M-BING (red) with 1000 proposals on PASCAL VOC 2007 *test* set. Ground truth objects are indicated in green. All presented proposals are the ones closest to each ground truth object. Note the improved localization after using our method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Proposal results on ILSVRC 2013 *validation* set. We report Average Recall (AR), Average Best Overlap (ABO) and recal at IoU of 0.8 for three budgets of proposals: 500, 1000, 2000. Numbers are shown for methods before/after applying our approach.

Method	#prop=500			#prop=1000			#prop=2000		
	AR	ABO	80%-recall	AR	ABO	80%-recall	AR	ABO	80%-recall
BING [8]	24.6/ **40.4**	58.2/ **66.0**	6.5/ **31.4**	26.4/ **45.4**	60.7/ **69.9**	6.8/ **36.6**	27.5/ **49.6**	62.3/ **73.1**	6.9/ **41.4**
EB [17]	46.2/ **46.7**	67.7/ **69.5**	42.0/ **43.5**	50.3/ **51.9**	71.8/ **73.4**	45.5/ **50.7**	53.4/ **56.2**	74.7/ **76.3**	48.0/ **56.8**
RP [20]	40.9/ **45.8**	65.3/ **68.4**	35.4/ **42.8**	47.3/ **52.0**	70.3/ **73.2**	43.0/ **50.8**	53.2/ **57.3**	74.5/ **76.9**	50.4/ **57.9**
GOP [21]	29.5/ **39.9**	55.3/ **62.5**	21.3/ **35.3**	49.8/ **53.7**	71.8/ **73.8**	45.5/ **52.7**	54.2/ **58.1**	74.6/ **76.7**	51.7/ **59.4**
SS [44]	47.3/ **47.8**	69.0/ **69.6**	44.0/ **45.2**	53.7/ **53.8**	73.7/ **74.0**	51.8/ **53.4**	**59.4**/59.1	**77.6**/**77.6**	59.3/ **60.7**
MCG [10]	50.5/ **51.4**	72.6/ **72.7**	47.7/ **49.4**	55.9/ **56.7**	76.3/ **76.6**	54.5/ **56.4**	59.6/ **61.0**	78.6/ **79.5**	59.2/ **62.4**

range of proposal budgets for RP [20], GOP [21] and LPO [26]. For SS [9] and MCG [10], which are state-of-the-art bottom-up proposals as reported in [13], we also their recall at IoU of 0.8 by 3%.

We also apply our method to Region Proposal Network (RPN) [4], which is proposed in Faster R-CNN and achieves state-of-the-art detection performance. As shown in Fig. 5, our method significantly boosts the recall of RPN under strict overlap criteria (e.g., IoU > 0.7). **Note that when using 1000 proposals, recall at IoU of 0.8 is improved from 34.4% to 57.6% and AR is improved from 48.6% to 57.5%**. This demonstrates the effectiveness of combining

Table 3
Object detection results on PASCAL VOC 2007 *test* set with Fast R-CNN and VGG-16 network. 1000 proposals are used for all methods. RPN† denotes the original Faster R-CNN [4], which is trained end to end using 2000 proposals and tested with 300 proposals. RPN‡ is fine-tuned using Fast R-CNN with 1000 fixed RPN proposals generated by RPN† model.

Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Person	Plant	Sheep	Sofa	Train	Tv	mAP
BING [8]	64.4	72.3	61.4	46.5	46.1	67.9	74.6	75.8	35.7	69.0	50.5	72.5	72.1	64.1	64.2	33.0	58.7	57.4	70.7	64.6	61.1
M-BING	69.7	76.9	66.5	57.6	40.0	80.8	76.7	82.5	42.6	74.3	66.1	78.8	83.4	74.9	67.2	36.6	64.1	**70.9**	74.4	71.5	67.8
EB [17]	69.4	**79.5**	65.7	57.7	44.3	79.6	77.3	83.1	45.8	76.8	64.2	79.6	79.5	75.1	68.7	41.7	70.3	69.2	76.4	71.4	68.8
M-EB	72.4	78.8	68.3	57.8	41.7	**84.5**	78.9	78.1	45.6	78.9	70.3	80.9	83.4	76.1	71.3	40.8	70.0	67.1	77.4	72.1	69.7
RP [20]	74.2	76.4	64.1	52.1	32.3	80.3	74.3	84.4	45.4	74.8	68.6	77.3	78.7	74.9	63.7	32.9	63.5	68.9	77.7	69.8	66.7
M-RP	69.3	79.3	68.8	56.4	42.1	81.3	76.9	83.5	47.4	75.7	70.2	80.9	79.2	76.0	67.3	38.4	67.7	68.3	77.7	70.2	68.8
GOP [21]	64.5	77.1	64.5	52.9	36.7	77.4	76.6	83.6	45.9	73.2	66.2	81.2	**84.6**	75.6	66.5	37.0	63.1	70.7	76.6	71.4	67.3
M-GOP	69.3	78.9	67.5	55.2	39.6	82.0	78.2	83.1	46.7	74.3	67.7	81.5	79.6	76.9	67.7	38.4	61.4	67.3	77.4	70.4	68.2
LPO [26]	69.0	76.3	65.9	54.4	34.6	78.4	77.9	84.0	43.5	73.6	69.4	76.6	79.4	75.6	66.9	39.3	64.6	69.0	76.6	70.8	67.3
M-LPO	73.4	78.4	71.2	58.8	41.2	81.4	78.4	84.2	47.8	73.8	68.0	81.3	80.4	75.9	71.4	40.0	70.4	69.4	77.5	70.9	69.7
SS [44]	73.3	77.4	66.8	56.5	39.7	80.5	77.2	83.0	42.4	73.5	69.1	81.5	80.1	76.1	67.0	34.9	69.7	68.3	75.9	64.0	67.8
M-SS	69.3	77.9	67.4	56.0	42.1	82.4	78.1	82.6	46.9	74.3	68.4	80.0	79.8	73.4	67.6	39.9	65.8	67.3	76.7	72.2	68.4
MCG [10]	68.7	77.0	66.1	50.0	42.9	81.4	76.6	79.5	44.7	74.6	69.3	75.3	78.7	73.4	68.3	37.5	63.8	69.6	76.8	**72.3**	67.3
M-MCG	73.6	78.6	70.8	57.9	46.6	80.5	78.3	78.6	46.5	75.7	**70.5**	77.0	79.9	75.7	71.6	40.8	68.2	69.6	76.8	71.7	69.4
RPN† [4]	70.4	78.0	67.9	58.5	49.0	80.0	79.6	78.0	**50.1**	74.9	66.3	78.7	82.9	**78.0**	**76.6**	40.3	70.1	64.0	76.0	66.0	69.3
RPN‡ [4]	68.6	78.2	72.0	57.9	**54.6**	80.6	82.5	83.8	49.1	77.9	63.8	80.0	84.0	73.5	75.8	41.9	**71.0**	62.2	77.1	71.6	70.3
M-RPN	**74.8**	79.2	**72.9**	**64.5**	50.5	83.3	**84.2**	**85.1**	49.7	**80.3**	68.1	**81.9**	84.0	76.7	75.7	**46.0**	**71.0**	67.3	**77.9**	71.6	**72.2**

powerful CNN features and low-level superpixel cues.

For illustration, some qualitative examples for BING and its improved version M-BING are shown in Fig. 7.

ImageNet 2013. We extensively conduct experiments on ImageNet ILSVRC 2013 validation set. As the dataset is much larger, we only apply our method to baselines that have pre-computed proposals available online [13]. Results are presented in Fig. 6 and Table 2. Similar to VOC 2007, we observe significant improvement in recall and ABO for all methods. In particular, when using 1000 proposals, the maximum gains are obtained when applying our method to BING [8], achieving 19%, 9.2%, and 29.8% improvement in AR, ABO and 80%-recall, respectively. For EB [17], RP [20], and GOP [21], which are quite efficient in computation but have moderate localization quality, our method improve their AR from 47%~50% to 52%~54%, achieving similar performance with SS [44]. The consistent improvement on ImageNet 2013 dataset implies the generalization ability of our method.

4.3. Object detection performance

We also evaluate object detection performance on PASCAL VOC 2007 by combining object proposals with Fast R-CNN [3]. We use VGG-16 [49] network and 1000 proposals per image for the experiments. Comparisons of average precision (AP) for numerous baselines and our improved versions are reported in Table 3. Our method achieves consistent improvements in terms of mean AP for all proposal methods. In particular, we obtain about 2% mAP improvement for most methods.

When applying to BING [8], our approach significantly improves the mAP from 61.1% to 67.8%. Note that the state-of-the-art SS [44] method also has mAP of 67.8%. However, a combination of BING and our box refinement method is much more efficient than SS, as it only takes about 0.15 s while SS requires 2 s. In fact, BING++ [18] shows that when utilizing GPU implementation, the approach can be achieved to be real-time.

For RPN [4], we compare our results with two variants. One is the original Faster R-CNN, where RPN is trained end to end together with Fast R-CNN. Following [4], this version is trained with 2000 proposals and tested with 300 proposals. Another is the unsharing feature version, i.e., the object detector is fine-tuned with 1000 fixed RPN proposals generated by the first variant. Note that the second variant is different from the unsharing feature version implemented in [4], as in our implementation RPN is first trained end to end in Faster R-CNN and the object detector is trained once more by fixing RPN proposals. As shown in Table 3, RPN integrated with our method outperforms the end-to-end Faster R-CNN model and the unsharing feature version by about 3% and 2%, respectively, achieving **72.2%** mAP. This gain solely comes from the improved localization quality of proposals, which suggests that the localization quality of proposals affects the detection performance significantly and our method is effective in improving object detection with complementary superpixel cues.

5. Conclusions

We have proposed an effective approach to improve the localization quality of object proposals. Our approach leverages the property of superpixels to perform boundary-aware proposal refinement. We design a box alignment algorithm to align proposals with superpixel boundaries, and a fast superpixel merging method to diversify proposals. Our method is very efficient and agnostic to any proposal methods.

Experiments on PASCAL VOC 2007 and ILSVRC 2013 datasets show the effectiveness of the proposed method. It significantly boosts the recall and localization accuracy of most existing proposal methods. We also conduct object detection experiments by combining the improved proposals with Fast R-CNN. Our approach integrated with RPN obtains the highest detection mAP on VOC 2007 test set. The experiments demonstrate the effectiveness of combining powerful CNN features and low-level superpixel cues in improving object proposal localization quality and object detection. In future, we will integrate superpixel cues into an end-to-end network to further boost the object detection performance.

Acknowledgment

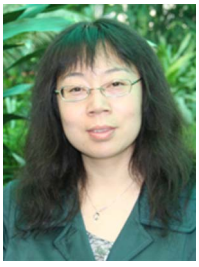
This work was supported by National Natural Science Foundation of China (No. 61171113).

References

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE TPAMI.
[2] R.B. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE CVPR, 2014.
[3] R. Girshick, Fast r-cnn, in: ICCV, 2015.
[4] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems (NIPS), 2015.
[5] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, IJCV.
[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale

Visual Recognition Challenge, IJCV.
[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: ECCV, 2014.
[8] M.-M. Cheng, Z. Zhang, W.-Y. Lin, P.H.S. Torr, BING: Binarized normed gradients for objectness estimation at 300fps, in: IEEE CVPR, 2014.
[9] J. Uijlings, K. van de Sande, T. Gevers, A. Smeulders, Selective search for object recognition, IJCV.
[10] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marqus, J. Malik, Multiscale combinatorial grouping, in: IEEE CVPR, 2014.
[11] A.T.D.A. Dumitru Erhan, Christian Szegedy, Scalable object detection using deep neural networks, in: CVPR, 2014.
[12] X. Chen, H. Ma, X. Wang, Z. Zhao, Improving object proposals with multi-thresholding straddling expansion, in: CVPR, 2015.
[13] J. Hosang, R. Benenson, P. Dollár, B. Schiele, What makes for effective detection proposals?, PAMI.
[14] J. Hosang, R. Benenson, B. Schiele, How good are detection proposals, really?, in: BMVC, 2014.
[15] W. Li, P. Dong, B. Xiao, L. Zhou, Object recognition based on the region of interest and optimal bag of words model, Neurocomputing 172 (2016) 271–280.
[16] B. Alexe, T. Deselaers, V. Ferrari, Measuring the objectness of image windows, IEEE TPAMI.
[17] C.L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, in: ECCV, 2014.
[18] Z. Zhang, Y. Liu, T. Bolukbasi, M.-M. Cheng, V. Saligrama, Bing++: A fast high quality object proposal generator at 100fps, arXiv:1511.04511.
[19] Q. Zhao, Z. Liu, B. Yin, Cracking bing and beyond, in: BMVC, 2014.
[20] S. Manen, M. Guillaumin, L. Van Gool, Prime object proposals with randomized prim's algorithm, in: IEEE ICCV, 2013.
[21] P. Krähenbühl, V. Koltun, Geodesic object proposals, in: ECCV, 2014, pp. 725–739.
[22] P. Rantalankila, J. Kannala, E. Rahtu, Generating object segmentation proposals using global and local search, in: IEEE CVPR, 2014.
[23] J. Carreira, C. Sminchisescu, Cpmc: Automatic object segmentation using constrained parametric min-cuts, IEEE TPAMI.
[24] I. Endres, D. Hoiem, Category-independent object proposals with diverse ranking, IEEE TPAMI 36 (2) (2014) 222–234.
[25] A. Humayun, F. Li, J.M. Rehg, Rigor: Recycling inference in graph cuts for generating object regions, in: IEEE CVPR, 2014.
[26] P. Krähenbühl, V. Koltun, Learning to propose objects, in: CVPR, 2015.
[27] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: a review, Neurocomputing.
[28] X. Jiang, Y. Pang, X. Li, J. Pan, Speed up deep neural network based pedestrian detection by sharing features across multi-scale models, Neurocomputing 185 (2016) 163–170.
[29] S.A.A. Shah, M. Bennamoun, F. Boussaid, Iterative deep learning for image set based face and object recognition, Neurocomputing 174 (2016) 866–874 (Part B).
[30] G. Wu, W. Lu, G. Gao, C. Zhao, J. Liu, Regional deep learning model for visual tracking, Neurocomputing 175 (2016) 310–323 (Part A).
[31] D.E.D.A. Christian Szegedy, Scott Reed, Scalable, high-quality object detection, in: arXiv:1412.1441, 2014.
[32] W. Kuo, B. Hariharan, J. Malik, Deepbox:learning objectness with convolutional networks, in: ICCV, 2015.
[33] M.P.T.T.L.V.G. Amir Ghodrati, Ali Diba, Deepproposal: Hunting objects by cascading deep convolutional layers, in: ICCV, 2015.
[34] P.O. Pinheiro, R. Collobert, P. Dollar, Learning to segment object candidates, in: NIPS, 2015.
[35] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in: IEEE ECCV, 2014.
[36] X. Liang, Y. Wei, X. Shen, Z. Jie, J. Feng, L. Lin, S. Yan, Reversible recursive instance-level object segmentation, in: IEEE CVPR, 2016.
[37] Y. Hua, K. Alahari, C. Schmid, Online object tracking with proposal selection, in: IEEE ICCV, 2015.
[38] B. Zhang, A. Perina, Z. Li, V. Murino, J. Liu, R. Ji, Bounding multiple gaussians uncertainty with application to object tracking, Int. J. Comput. Vis. 118 (3) (2016) 364–379.
[39] B. Zhang, Z. Li, A. Perina, A.D. Bue, V. Murino, J. Liu, Adaptive local movement modeling for robust object tracking, IEEE Trans. Circuits Syst. Video Technol. PP 99 (2016) 1–1.
[40] B. Zhang, A. Perina, V. Murino, A.D. Bue, Sparse representation classification with manifold constraints transfer, in: IEEE CVPR, 2015, pp. 4557–4565.
[41] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, R. Urtasun, 3d object proposals for accurate object class detection, in: NIPS, 2015.
[42] S. Wang, M. Chen, Y. Li, Y. Zhang, L. Han, J. Wu, S. Du, Detection of dendritic spines using wavelet-based conditional symmetric analysis and regularized morphological shared-weight neural networks, Comput. Math. Methods Med. 2015 (2) (2015) 1–12.
[43] X.X.Z. Non-Member, Y.Z. Non-Member, G.J. Non-Member, J.Y. Non-Member, Z.D. Non-Member, S.W. Non-Member, G.Z. Non-Member, P.P. Non-Member, Detection of abnormal mr brains based on wavelet entropy and feature selection, IEEJ Trans. Electr. Electron. Eng.
[44] K.E. Van de Sande, J.R. Uijlings, T. Gevers, A.W. Smeulders, Segmentation as selective search for object recognition, in: IEEE ICCV, 2011.
[45] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, IJCV 59 (2) (2004) 167–181.
[46] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, ⟨http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html⟩.
[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.
[48] B. Alexe, T. Deselaers, V. Ferrari, What is an object?, in: IEEE CVPR, 2010.
[49] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556.

Xiaozhi Chen received the B.S. degree in Electronic Engineering from Tsinghua University, Beijing, China in 2012, where he is currently pursuing the Ph.D. degree. His research interests include computer vision and machine learning.



Huimin Ma received the M.S. and Ph.D. degrees in Mechanical Electronic Engineering from Beijing Institute of Technology, Beijing, China in 1998 and 2001 respectively. She is an associate professor in the Department of Electronic Engineering of Tsinghua University, and the director of 3D Image Simulation Lab. She worked as an visiting scholar in University of Pittsburgh in 2011. She is also the executive director and the vice secretary general of China Society of Image and Graphics. Her research and teaching interests include 3D object recognition and tracking, system modeling and simulation, psychological base of image cognition.

Chenzhuo Zhu is an undergraduate student in Department of Electronic Engineering in Tsinghua University, Beijing, China. His research interests include computer vision and machine learning.

Xiang Wang received the B.S. degree in Electronic Engineering from Tsinghua University, Beijing, China in 2014, where he is currently pursuing the Ph.D. degree. His research interests include computer vision especially on salient object detection and machine learning.

Zhichen Zhao received the B.S. degree in Electronic Engineering from Tsinghua University, Beijing, China in 2014, where he is currently pursuing the Master degree. His research interests include computer vision especially on action recognition.