Convolutional Oriented Boundaries: From Image Segmentation to High-Level Tasks

Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool, *Life Fellow, IEEE*

Abstract—We present Convolutional Oriented Boundaries (COB), which produces multiscale oriented contours and region hierarchies starting from generic image classification Convolutional Neural Networks (CNNs). COB is computationally efficient, because it requires a single CNN forward pass for multi-scale contour detection and it uses a novel sparse boundary representation for hierarchical segmentation; it gives a significant leap in performance over the state-of-the-art, and it generalizes very well to unseen categories and datasets. Particularly, we show that learning to estimate not only contour strength but also orientation provides more accurate results. We perform extensive experiments for low-level applications on BSDS, PASCAL Context, PASCAL Segmentation, and NYUD to evaluate boundary detection performance, showing that COB provides state-of-the-art contours and region hierarchies in all datasets. We also evaluate COB on high-level tasks when coupled with multiple pipelines for object proposals, semantic contours, semantic segmentation, and object detection on various databases (MS-COCO, SBD, PASCAL VOC'07), showing that COB also improves the results for all tasks.

Index Terms—Contour detection, contour orientation estimation, hierarchical image segmentation, object proposals, semantic contours

arXiv:1701.04658v1 [cs.CV] 17 Jan 2017

1 Introduction

The adoption of Convolutional Neural Networks (CNNs) has caused a profound change and a large leap forward in performance throughout the majority of fields in computer vision. In the case of a traditionally category-agnostic field such as contour detection, it has recently fostered the appearance of systems [1], [2], [3], [4], [5], [6] that rely on large-scale category-specific information in the form of deep architectures pre-trained on Imagenet for image classification [7], [8], [9], [10].

This paper proposes Convolutional Oriented Boundaries (COB), a generic CNN architecture that allows end-to-end learning of multiscale oriented contours, and we show how it translates top performing base CNN networks into high-quality contours; allowing to bring future improvements in base CNN architectures into semantic grouping. We then propose a sparse boundary representation for efficient construction of hierarchical regions from the contour signal. Our overall approach is both efficient (it runs in 0.8 seconds per image) and highly accurate (it produces state-of-the-art contours and regions on PASCAL and on the BSDS). Figure 1 shows an overview of our system.

For the last fifteen years, the Berkeley Segmentation Dataset and Benchmark (BSDS) [11] has been the experimental testbed of choice for the study of boundary detection and image segmentation. However, the current large-capacity and very accurate models have underlined the limitations of the BSDS as the primary benchmark for grouping. Its 300 train images are inadequate for training systems with tens of millions of parameters and, critically, current state-of-the-art techniques are reaching human performance for boundary detection on its 200 test images.

In terms of scale and difficulty, the next natural frontier for perceptual grouping is the PASCAL VOC dataset [12], an influential benchmark for image classification, object detection, and semantic segmentation which has a *trainval* set with more than 10 000 challenging and varied images. A first step in that direction was taken by Hariharan et al. [13], who annotated the VOC dataset for category-specific boundary detection on the foreground objects. More recently, the PASCAL Context dataset [14] extended this annotation effort to all the background categories, providing thus fully-parsed images which are a direct VOC counterpart to the human ground-truth of the BSDS. In this direction, this paper investigates the transition from the BSDS to PASCAL Context in the evaluation of image segmentation.

We derive valuable insights from studying perceptual grouping in a larger and more challenging empirical framework. Among them, we observe that COB leverages increasingly deeper state-of-the-art architectures, such as the recent Residual Networks [10], to produce improved results. This indicates that our approach is generic and can directly benefit from future advances in CNNs. We also observe that, in PASCAL, the globalization strategy of contour strength by spectral graph partitioning proposed in [15] and used in state-of-the-art methods [1], [16] is unnecessary in the presence of the high-level knowledge conveyed by pre-trained CNNs and oriented contours, thus removing a significant computational bottleneck for high-quality contours.

We conduct two types of experiments, the first of which regards low-level vision applications, such as contour detection and generic segmentation on PASCAL Context and the BSDS500. We extend the evaluation to the NYUD RGB-D dataset, showing that that the pipeline of COB can benefit

- *K.-K. Maninis, J. Pont-Tuset, and L. Van Gool are with the Computer Vision Laboratory, ETHZ, Switzerland.*
- *P. Arbeláez is with the Department of Biomedical Engineering, Universidad de los Andes, Colombia.*

Contacts: see http://www.vision.ee.ethz.ch/~cvlsegmentation/

ALIGN – EXHIBIT 1021

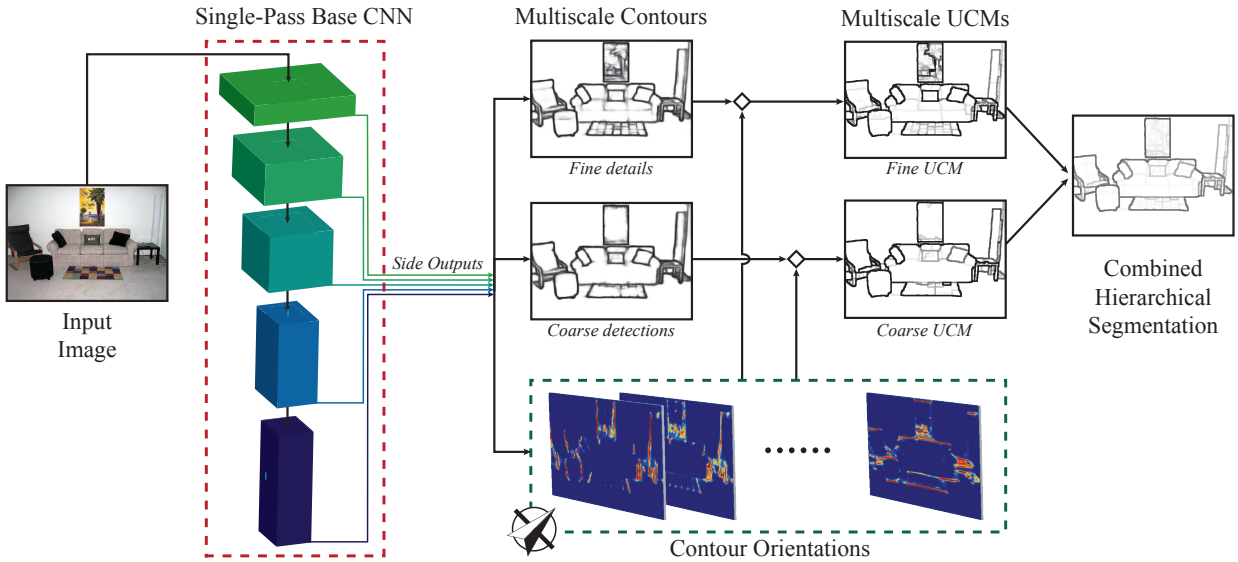


Fig. 1. **Overview of COB**: From a single pass of a base CNN, we obtain multiscale oriented contours. We combine them to build Ultrametric Contour Maps (UCMs) at different scales and fuse them into a single hierarchical segmentation structure.

from depth embeddings. We also include evaluation of Object Contour detection on the VOC'12 database. In all cases, COB demonstrates state-of-the-art performance on contours and regions while being computationally efficient.

In a second set of experiments, we study the interplay of COB with various downstream recognition applications. We use our hierarchical regions as input to the combinatorial grouping algorithm of [16] and obtain state-of-the-art segmented object proposals on PASCAL VOC'12 Segmentation by a significant margin. Furthermore, we provide empirical evidence for the generalization power of COB by evaluating our object proposals without any retraining in the even larger and more challenging MS-COCO dataset, where we also report competitive results compared to the state of the art. We have also studied the effects of COB when coupled with well known pipelines, showing that injecting COB detections to them lead to improvements on Semantic Segmentation, and Object Detection. Finally, we report a new state of the art on Semantic Boundary detection.

Our efforts on segmentation through CNNs have also found application in retinal image segmentation [17], obtaining state-of-the-art and super-human performance in vessel and optic disc segmentation, which further highlights their generality.

The COB code, pre-computed results, pre-trained models, and benchmarks are publicly available at www.vision.ee.ethz.ch/~cvlsegmentation/.

2 Related Work

The latest wave of contour detectors takes advantage of deep learning to obtain state-of-the-art results [1], [2], [3], [4], [5], [6], [18]. Ganin and Lempitsky [6] use a deep architecture to extract features of image patches. They approach contour detection as a multiclass classification task, by matching the extracted features to predefined ground-truth features. The authors of [3], [4] make use of features generated by pre-trained CNNs to regress contours. They prove that object-level information provides powerful cues for the prediction of contours. Shen et al. [5] learn deep features using shape information. Xie and Tu [2] provide an end-to-end deep framework to boost the efficiency and accuracy of contour detection, using convolutional feature maps and a novel loss function. Kokkinos [1] builds upon [2] and improves the results by tuning the loss function, running the detector at multiple scales, and adding globalization. COB is different from this previous work in that we obtain multiscale information in a single pass of the network on the whole image, it combines the per-pixel classification with contour orientation estimation, and its output is richer than a linear combination of cues at different scales.

At the core of all these deep learning approaches, lies a *base CNN*, starting from the seminal AlexNet [7] (8 layers), through the more complex VGGNet [9] (16 layers) and inception architecture of GoogLeNet [8] (22 layers), to the very recent and very deep ResNets [10] (up to 1001 layers). Image classification results, which originally motivated these architectures, have been continuously improved by exploring deeper and more complex networks. In this work, we present results both using VGGNet and ResNet, showing that COB is modular and can incorporate and benefit from future improvements in the base CNN.

Recent work has also explored weakly supervised or unsupervised learning of contours: Khoreva et al. [19] learn from the results of generic contour detectors coupled with object detectors; and Li et al. [20] train contour detectors from motion boundaries acquired from video sequences. Yang et al. [21] use conditional random fields to refine the inaccurately localized boundary annotations of PASCAL. Our approach uses full supervision from BSDS and PASCAL Context for contour localization and orientation.

COB exploits the duality between contour detection and segmentation hierarchies, initially studied by Najman and Schmitt [22]. Arbeláez et al. [15] showed its usefulness for jointly optimizing contours and regions. Pont-Tuset et al. [16] leveraged multi-resolution contour detection and proved its interest also for generating object proposals. We differentiate from these approaches in two aspects. First, our sparse boundary representation translates into a clean

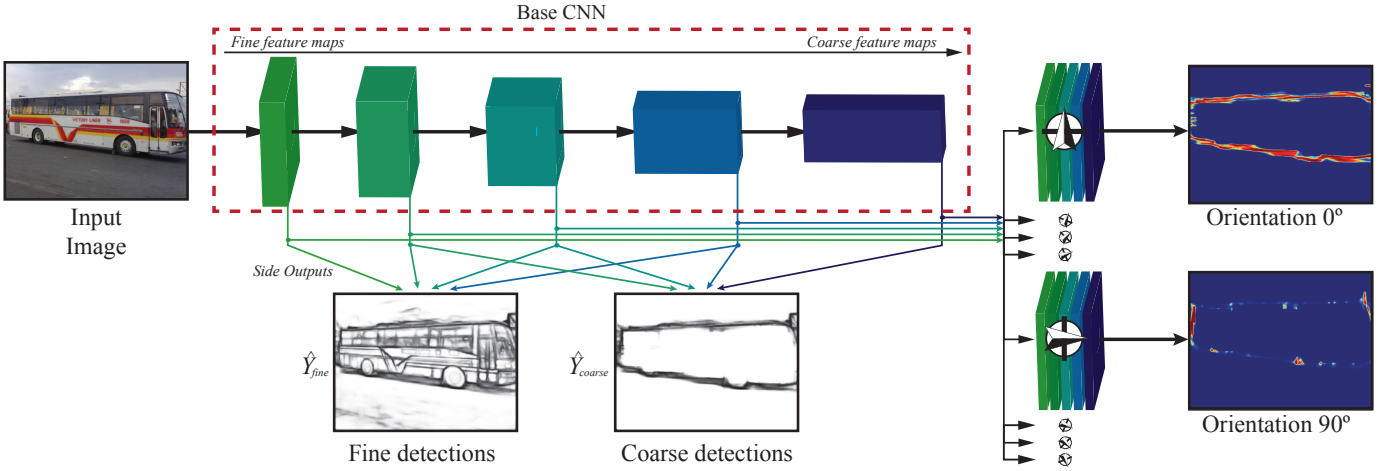


Fig. 2. **Our deep learning architecture** (best viewed in color). The connections show the different stages that are used to generate the multiscale contours. Orientations further require additional convolutional layers in multiple stages of the network.

and highly efficient implementation of hierarchical segmentation. Second, by leveraging high-level knowledge from the CNNs in the estimation of contour strength and orientation, our method benefits naturally from global information, which allows bypassing the globalization step (output of normalized cuts), a bottleneck in terms of computational cost, but a cornerstone of previous aproaches.

3 Deep Multiscale Oriented Contours

CNNs are by construction multi-scale feature extractors. If one examines the standard architecture of a CNN consisting of convolutional and spatial pooling layers, it becomes clear that as we move deeper, feature maps capture more global information due to the decrease in resolution. For contour detection, this architecture implies local and fine-scale contours at shallow levels, coarser spatial resolution and larger receptive fields for the units when going deeper and, consequently, more global information for predicting boundary strength and orientation. CNNs have therefore a built-in globalization strategy for contour detection, analogous to the hand-engineered globalization of contour strength through spectral graph partitioning in [15], [16].

Figure 2 depicts how we make use of information provided by the intermediate layers of a CNN to detect contours and their orientations at multiple scales. Different groups of feature maps contain different, scale-specific information, which we combine to build a multiscale oriented contour detector. The remainder of this section is devoted to introducing the recent approaches to contour detection using deep learning, to presenting our CNN architecture to produce contour detection at different scales, and to explaining how we estimate the orientation of the edges; all in a single CNN forward pass at the image level.

Training deep contour detectors: The recent success of [2] is based on a CNN to accurately regress the contours of an image. Within this framework, the idea of employing a CNN in an image-to-image fashion without any post-processing has proven successful, and lead to a big leap in performance for the task of contour detection. Their network, HED, produces scale-specific contour images (side outputs) for different scales of a network, and combines their activations linearly to produce a contour probability map. Using the notation of the authors, we denote the training dataset by $S = \{(X_n, Y_n), n = 1, \ldots, N\}$, with X_n being the input image and $Y_n = \{y_j^{(n)}, j = 1, \ldots, |X_n|\}, y_j^{(n)} \in \{0, 1\}$ the predicted pixelwise labels. For simplicity, we drop the subscript n. Each of the M side outputs minimizes the objective function:

$$\ell_{side}^{(m)}\left(\mathbf{W}, \mathbf{w}^{(m)}\right) = -\beta \sum_{j \in Y_+} \log P\left(y_j = 1 | X; \mathbf{W}, \mathbf{w}^{(m)}\right) - (1-\beta) \sum_{j \in Y_-} \log P\left(y_j = 0 | X; \mathbf{W}, \mathbf{w}^{(m)}\right) \quad (1)$$

where $\ell_{side}^{(m)}$ is the loss function for scale $m \in \{1, \ldots, M\}$, $\mathbf{W}$ denotes the standard set of parameters of the CNN, and $\{\mathbf{w}^{(m)}, m = 1, \ldots, M\}$ the corresponding weights of the the m-th side output. The multiplier β is used to handle the imbalance of the substantially greater number of background compared to contour pixels. Y_+ and Y_- denote the contour and background sets of the ground-truth Y, respectively. The probability $P(\cdot)$ is obtained by applying a sigmoid $\sigma(\cdot)$ to the activations of the side outputs $\hat{A}_{side}^{(m)} = \{a_j^{(m)}, j = 1, \ldots, |Y|\}$. In HED, the activations are finally fused linearly, as: $\hat{Y}_{fuse} = \sigma\left(\Sigma_{m=1}^{M} h_m \hat{A}_{side}^{(m)}\right)$ where $\mathbf{h} = \{h_m, m = 1, \ldots, M\}$ are the fusion weights. The fusion output is also trained to resemble the ground-truth applying the same loss function of Equation 1, by optimizing the complete set of parameters, including the fusion weights $\mathbf{h_m}$. We instead take advantage of the common CNN architectures to regress both the strength of the coarse and detailed (fine) contours, as well as the contour orientations. COB combines these output channels non-linearly to a single hierarchical segmentation. Inside this segmentation, the placement of each region in the hierarchy is determined by the strength of the boundaries to the neighbouring regions. All in all, COB combines efficiently contour strengths and orientations into a segmentation hierarchy, which unlike an open contour map, can further facilitate high-level vision tasks related to segmented object proposals. In the rest of the paper

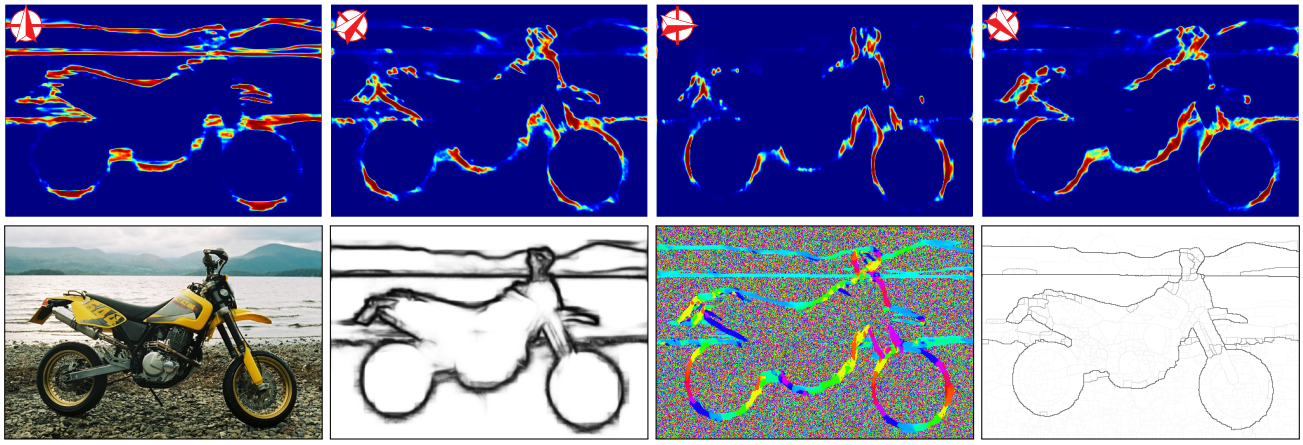


Fig. 3. Illustration of contour orientation learning. Row 1 shows the responses B_k for 4 out of the 8 orientation bins. Row 2, from left to right: original image, contour strength, learned orientation map into 8 orientations, and hierarchical boundaries.

we use the class-balancing cross-entropy loss function of Equation 1.

Multiscale contours: We start from a deep network pre-trained on ImageNet [23], such as VGG [9] or ResNet [10]. The fully connected layers used for classification are removed, and so are the batch normalization layers, since we operate on one image per iteration. Therefore, the network consists mainly of convolutional layers coupled with ReLU activations, divided into 5 stages. We will refer to this architecture as the "base CNN" of our implementation. Each stage is handled as a different scale, since it contains feature maps of a similar size. At the end of a stage, there is a max pooling layer, which reduces the spatial dimensions of the produced feature maps to a half. As discussed before, the CNN naturally contains multiscale information, which we exploit to build a multiscale contour regressor.

We separately supervise the output of the last layer of each stage (side activation), comparing it to the ground truth using the loss function of Equation 1. This way, we enforce each side activation to produce an intermediate contour map at different resolution. The idea of supervising intermediate parts of a CNN has successfully been used in previous approaches, for a variety of tasks [2], [8], [24]. In the 5-scale base CNN illustrated in Figure 2, we linearly combine the side activations of the 4 finest and 4 coarsest scales to a fine-scale and a coarse-scale output ($\hat{Y}_{fine}$ and $\hat{Y}_{coarse}$, respectively) with trainable weights. The finer scale contains better localized contours, whereas the coarse scale leads to less noisy detections. To train the two sets of weights of the linear combinations, we freeze the pre-trained weights of the base CNN.

Estimation of Contour Orientations: In order to predict accurate contour orientations, we propose an extension of the CNN that we use to predict contour strength. We define the task as pixel-wise image-to-image multiscale classification into K bins. We connect K different branches (sub-networks) to the base network, each of which is associated with one orientation bin, and has access to feature maps that are generated from the intermediate convolutional layers at M different scales. We assign the parts of the CNN associated with each orientation a different task from the base network: classify the pixels of the contours that match a specific orientation. In order to design these orientation-specific subtasks, we classify each pixel of the human contour annotations into K different orientations. The orientation of each contour pixel is obtained by approximating the ground-truth boundaries with polygons, and assigning each pixel the orientation of the closest polygonal segment, as shown in Figure 5. As in the case of multiscale contours, the weights of the base network remain frozen when training these sub-networks.

Each sub-network consists of M convolutional layers, each of them appended on different scales of the base network. Thus we need $M * K$ additional layers. In our setup, we use $K = 8$ and $M = 5$. All K orientations are regressed in parallel, and since they are associated with a certain angle, we post-process them to obtain the orientation map. Specifically, the orientation map is obtained as:

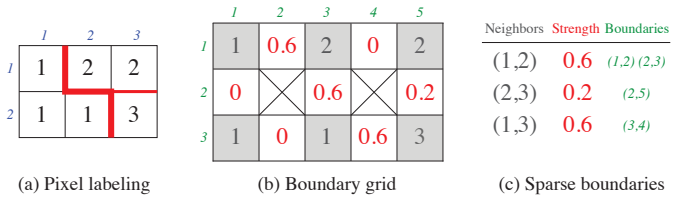
$$O(x, y) = \mathcal{T}\left(\arg\max_k B_k\left(x, y\right)\right), k = 1, \ldots, K \qquad (2)$$

where $B_k(x, y)$ denotes the response of the k-th orientation bin of the CNN at the pixels with coordinates (x, y) and $\mathcal{T}(\cdot)$ is the transformation function which associates each bin with its central angle. For the cases where two neighboring bins lead to strong responses, we compute the angle as their weighted average. At pixels where there is no response for any of the orientations, we assign random values between 0 and π, not to bias the orientations. The different orientations as well as the resulting orientation map (color-coded) are illustrated in Figure 3.

In [15], [16], [25] the orientations are computed by means of local gradient filters. In Section 5 we show that our learned orientations are significantly more accurate and lead to better region segmentations.

4 Fast Hierarchical Regions

This section is devoted to building an efficient hierarchical image segmentation algorithm from the multiscale contours and the orientations extracted in the previous section. We build on the concept of Ultrametric Contour Map (UCM) [15], which transforms a contour detection probability map into a hierarchical boundary map, which



(a) Pixel labeling (b) Boundary grid (c) Sparse boundaries

Fig. 4. **Image Partition Representation:**
(a) Pixel labeling, each pixel gets assigned a region label. (b) Boundary grid, markers of the boundary positions. (c) Sparse boundaries, lists of boundary coordinates between neighboring regions.

gets partitions at different granularities when thresholding at various contour strength values. Despite the success of UCMs, their low speed limits their applicability.

In the remainder of this section we first describe an alternative representation of an image partition that allows us to reduce the computation time of multiscale UCMs by an order of magnitude, to less than one second. Then, we present the global algorithm to build a hierarchy of regions from the multiscale contours and the orientations presented in Section 3. As we will show in the experimental section, the resulting algorithm improves the state of the art significantly, at a fraction of the computational time of [16].

Sparse Boundary Representation of Hierarchies of Regions: An image partition is a clustering of the set of pixels into different sets, which we call regions. The most straightforward way of representing it in a computer is by a matrix of labels, as in the example in Figure 4(a), with three regions on an image of size 2×3. The boundaries of this partition are the edge elements, or *edgels*, between the pixels with different labels (highlighted in red). We can assign different *strengths* to these boundaries (thicknesses of the red lines), which indicate the *confidence* of that piece of being a boundary. By iteratively *erasing* these boundaries in order of increasing strength we obtain different partitions, which we call *hierarchy of regions*, or Ultrametric Contour Maps.

These boundaries are usually stored in the *boundary grid* (Figure 4(b)), a matrix of double the size of the image (minus one), in which the odd coordinates represent pixels (gray areas), and the positions in between represent boundaries (red numbers) and junctions (crossed positions).

UCMs use this representation to store their boundary *strength* values, that is, each boundary position stores the threshold value beyond which that edgel *disappears* and the two neighboring regions merge. This way, simply *binarizing* a UCM we have a partition represented as a boundary grid.

This representation, while useful during prototyping, becomes very inefficient at run time, where the percentage of *activated* boundaries is very sparse. Not only are we wasting memory by storing those *empty* boundaries, but it also makes operating on them very inefficient by having to *sweep* over the entire matrix to perform a modification on a single boundary piece.

Inspired by how sparse matrices are handled, we designed the *sparse boundaries* representation (Figure 4(c)). It stores a look-up table for pairs of neigboring regions, their boundary strength, and the list of coordinates the boundary occupies. Apart from being more compact in terms of memory, this representation enables efficient operations on specific pieces of a boundary, since one only needs to perform a search in the look-up table and scan the activated coordinates; instead of sweeping the whole boundary grid.

Fast Hierarchies from Multiscale Oriented Contours: The deep CNN presented in Section 3 provides different levels of detail for the image contours. A linear combination of the layers is the straightforward way of providing a single contour signal [2]. The approach in this work is to combine the region hierarchies extracted from the contour signals at each layer instead of the contours directly. We were inspired by the framework proposed in [16], in which a UCM is obtained from contours computed at different image scales and then combined into a single hierarchy. However,instead of processing the image at multiple resolutions, we use the different outputs that are computed in a single pass of the proposed CNN architecture.

A drawback of the original framework [16] is that the manipulation of the hierarchies is very slow (in the order of seconds), so the operations on the UCMs had to be discretized and performed at a low number of contour strengths. By using the fast sparse boundary representation, we can operate on all contour strengths, yielding better results at a fraction of the original cost. Moreover, we use the learned contour orientations for the computation of the Oriented Watershed Transform (OWT), further boosting performance.

5 Experiments on Low-Level Applications

This section presents the empirical evidence that supports our approach for low-level applications (image segmentation and contour detection). First, Section 5.1 explores ablated and baseline techniques in order to isolate and quantify the improvements due to different components of our system. Section 5.2 further analyzes and evaluates the proposed contour orientations. In Section 5.3, Section 5.4, and Section 5.5 we compare our results against the state of the art in generic RGB image segmentation, RGB object boundary detection, and RGB-D image segmentation, respectively. In all three cases, we obtain the best results to date by a significant margin. Finally, Section 5.6 analyzes the effect of the various components in terms of speed on COB.

In terms of datasets, we extend the main BSDS benchmark [26] to the PASCAL Context dataset [14], which contains carefully localized pixel-wise semantic annotations for the entire image on the PASCAL VOC 2010 detection trainval set. This results in 459 semantic categories across 10 103 images, which is an order of magnitude ($20\times$) larger than the BSDS. In order to allow training and optimization of large capacity models, we split the data into train, validation, and test sets as follows: *VOC train* corresponds to the official PASCAL Context train with 4 998 images, *VOC val* corresponds to half the official PASCAL Context validation set with 2 607 images and *VOC test* corresponds to the second half with 2 498 images. In the remainder of the paper, we refer to this dataset division. Note that, in contrast to the BSDS, in this setting the notion of boundary is defined as separation between different semantic categories and not between their parts.

In all our experiments for boundary detection and image segmentation, we used the standard evaluation benchmark evaluating boundaries (F_b [26]) and regions (F_{op} [27]).

Through the literature, the tolerance in the boundary localization metric F_b is altered (the `maxDist` parameter), depending on the database and the quality of the annotations. To avoid confusion, we list the value of this parameter for all our experiments in Table 1.

We used the publicly available *Caffe* [28] framework for training and testing CNNs, and all the state-of-the-art results are computed using the publicly-available code provided by the respective authors.

5.1 Control Experiments/Ablation Analysis

This section presents the control experiments and ablation analysis to assess the performance of all subsystems of our method. We train on *VOC train*, and evaluate on *VOC val* set. We report the standard F measure at Optimal Dataset Scale (ODS) and Optimal Image Scale (OIS), as well as the Average Precision (AP), both evaluating boundaries (F_b [26]) and regions (F_{op} [27]).

Table 2 shows the evaluation results of the different variants, highlighting whether we include globalization and/or trained orientations. As a first baseline, we test the performance of MCG [16], which uses Structured Edges [25] as input contour signal. We then substitute SE by the newer HED [2], trained on *VOC train* as input contours and denote it MCG-HED. Note that the aforementioned baselines require multiple passes of the contour detector (3 scales).

In the direction of using the side outputs of the base CNN architecture as multiscale contour detections in one pass, we tested the baseline of naively taking the 5 side outputs directly as the contour detections. We trained both VGGNet [9] and ResNet50 [10] on *VOC train* and combined the 5 side outputs with our fast hierarchical regions of Section 4 (VGGNet-Side and ResNet50-Side).

We finally evaluate different variants of our system, as presented in Section 3. We first compare our system with two different base architectures: Ours(VGGNet) and Ours(ResNet50). We train the base networks for 30000 iterations, with stochastic gradient descent and a momentum of 0.9. We observe that the deeper architecture of ResNet translates into better boundaries and regions. Using the even deeper counterparts of ResNet lead to negligible gain in accuracy while significantly sacrificing speed, thus we did not explore this direction extensively.

Database	Task	train	test	`maxDist`
BSDS500	Generic Segmentation	300	200	0.0075
VOC Context	Generic Segmentation	7 605	2 498	0.0075
VOC'12 Segm.	Object Contours	1 464	1 449	0.01
NYUD-v2	RGB-D Segmentation	795	654	0.011
SBD	Semantic Contours	8 498	2 857	0.02
COCO	Object Proposals	-	40 504	-
VOC'07	Object Detection	5 011	4 952	-

TABLE 1
Datasets and Parameters: The list of databases used to evaluate our approach on various low-level and high-level tasks. We report the number of images used for training and testing our algorithm, along with the tolerance for contour localization used in the literature, when applicable. In all our experiments, we keep those numbers unchanged.

We then evaluate the influence of our trained orientations and globalization, by testing the four possible combinations (the orientations are further evaluated in the next section). Our method using ResNet50 together with trained orientations leads to the best results both for boundaries and for regions. The experiments also show that, when coupled with trained orientations, globalization even decreases performance, so we can safely remove it and get a significant speed up. Our technique with trained orientations and without globalization is therefore selected as our final system and will be referred to in the sequel as Convolutional Oriented Boundaries (COB).

5.2 Contour Orientation

We evaluate contour orientation results by the classification accuracy into 8 different orientations, to isolate their performance from the global system. We compute the ground-truth orientations as depicted in Figure 5 by means of the sparse boundaries representation. We then sweep all ground-truth boundary pixels and compare the estimated orientation with the ground-truth one. Since the orientations are not well-balanced classes (much more horizontal and vertical contours), we compute the classification accuracy per each of the 8 classes and then compute the mean.

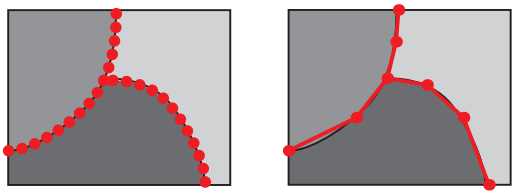
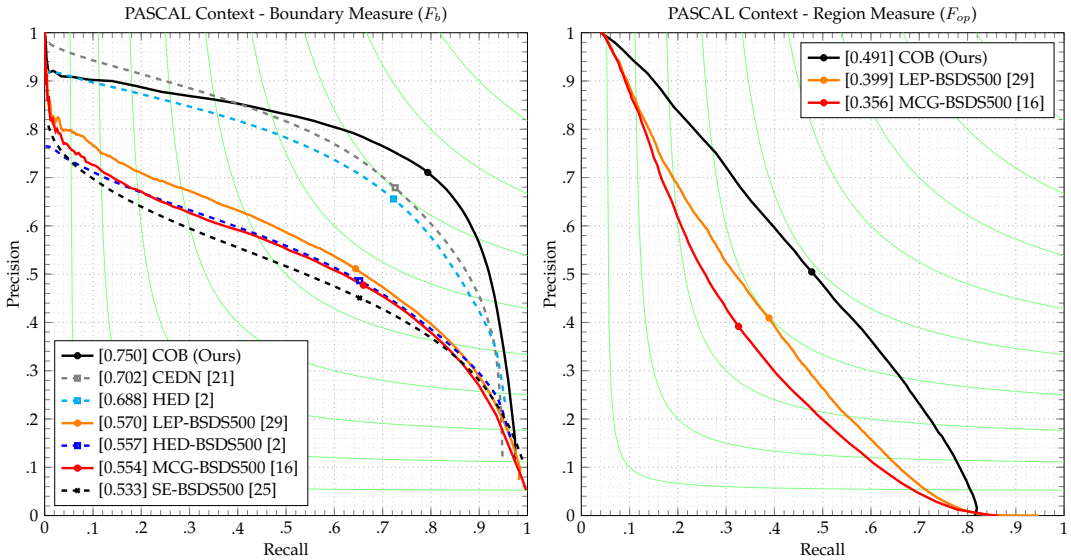


Fig. 5. **Polygon simplification:** From all boundary points (left) to simplified polygons (right), which are used to compute the ground-truth orientation robustly.

Figure 6 shows the classification accuracy with respect to the confidence of the estimation. We compare our proposed technique against the local gradient estimation used in previous literature [15], [16], [25]. As a baseline, we plot the result a random guess of the orientations would get. We observe that our estimation is significantly better than the previous approach. As a summary measure, we compute the area under the curve of the accuracy (ours 58.6%, local gradients 41.2%, random 12.5%), which corroborates the superior results from our technique.

			Boundaries - F_b			Regions - F_{op}		
Method	Global.	Orient.	ODS	OIS	AP	ODS	OIS	AP
MCG [16]	✓	✗	0.548	0.594	0.519	0.355	0.419	0.263
MCG-HED	✓	✗	0.691	0.727	0.693	0.459	0.520	0.374
VGGNet-Side	✓	✗	0.644	0.683	0.664	0.439	0.505	0.351
ResNet50-Side	✓	✗	0.676	0.711	0.681	0.456	0.521	0.374
Ours (VGGNet)	✗	✓	0.705	0.735	0.741	0.466	0.533	0.384
Ours (ResNet50)	✗	✗	0.734	0.767	0.757	0.475	0.545	0.405
Ours (ResNet50)	✓	✗	0.726	0.759	0.725	0.461	0.531	0.395
Ours (ResNet50)	✓	✓	0.732	0.763	0.731	0.481	**0.554**	**0.418**
Ours (ResNet50)	✗	✓	**0.737**	**0.768**	**0.758**	**0.483**	0.553	0.417

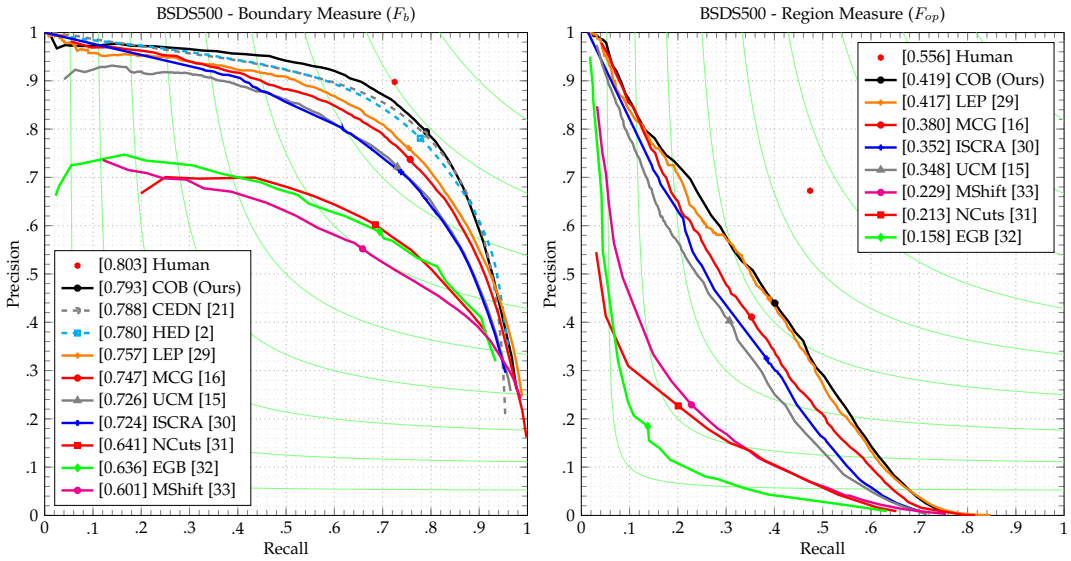
TABLE 2
Ablation analysis on *VOC val*: Comparison of different ablated and baseline versions of our system.



Method	Boundaries - F_b ODS	OIS	AP
COB	**0.750**	**0.781**	**0.773**
CEDN [21]	0.702	0.718	0.744
HED [2]	0.688	0.707	0.704
LEP-B [29]	0.570	0.636	0.547
HED-B [2]	0.557	0.588	0.518
MCG-B [16]	0.554	0.609	0.528
SE-B [25]	0.533	0.568	0.496

Method	Regions - F_{op} ODS	OIS	AP
COB	**0.491**	**0.565**	**0.439**
LEP-B [29]	0.399	0.461	0.304
MCG-B [16]	0.356	0.419	0.275

Fig. 7. **PASCAL Context *VOC test* Evaluation**: Precision-recall curves for evaluation of boundaries (F_b [26]), and regions (F_{op} [27]). Contours in dashed lines and boundaries (from segmentation) in solid lines. ODS, OIS, and AP summary measures.



Method	Boundaries - F_b ODS	OIS	AP
COB (Ours)	**0.793**	**0.820**	**0.859**
CEDN [21]	0.788	0.804	0.834
HED [2]	0.780	0.796	0.834
LEP [29]	0.757	0.793	0.828
MCG [16]	0.747	0.779	0.759
UCM [15]	0.726	0.760	0.727
ISCRA [30]	0.724	0.752	0.783
NCuts [31]	0.641	0.674	0.447
EGB [32]	0.636	0.674	0.581
MShift [33]	0.601	0.644	0.493

Method	Regions - F_{op} ODS	OIS	AP
COB (Ours)	**0.419**	**0.478**	**0.343**
LEP [29]	0.417	0.468	0.334
MCG [16]	0.380	0.433	0.271
ISCRA [30]	0.352	0.418	0.275
UCM [15]	0.348	0.385	0.235
MShift [33]	0.229	0.292	0.122
NCuts [31]	0.213	0.270	0.096
EGB [32]	0.158	0.240	0.080

Fig. 8. **BSDS500 Test Evaluation**: Precision-recall curves for evaluation of boundaries (F_b [26]), and regions (F_{op} [27]). ODS, OIS, and AP summary measures.

5.3 Generic Image Segmentation

We present our results for contour detection and generic image segmentation on PASCAL Context [14] as well as on the BSDS500 [11], which is the most established benchmark for perceptual grouping.

PASCAL Context: We train COB in the *VOC train*, and perform hyper-parameter selection on *VOC val*. We report the final results on the unseen *VOC test* when trained on *VOC trainval*, using the previously tuned hyper-parameters. We compare our approach to several methods trained on the BSDS [2], [16], [25], [29] and we also re-train the current state-of-the-art contour detection methods HED [2] and the recent CEDN [21] on *VOC trainval* using the code provided by the respective authors.

Figure 7 presents the evaluation results of COB compared to the state of the art, showing that it outperforms all others by a considerable margin both in terms of boundaries and in terms of regions. The lower performance of the methods trained on the BSDS quantifies the difficulty of the task when moving to a larger and more challenging dataset.

BSDS500: We retrain COB using only the 300 *trainval* images of the BSDS, after data augmentation as suggested in [2], keeping the architecture decided in Section 5.1. For comparison to HED [2], we used the model that the authors provide online. We also compare with CEDN [21], by evaluating the results provided by the authors.

Figure 8 presents the evaluation results, which show that we also obtain state-of-the-art results in this dataset. The smaller margins are in all likelihood due to the fact that we almost reach human performance for the task of contour detection on the BSDS, which motivates the shift to PASCAL Context to achieve further progress in the field.

Qualitative Results: Figure 9 shows some qualitative results of our hierarchical contours. Please note that COB

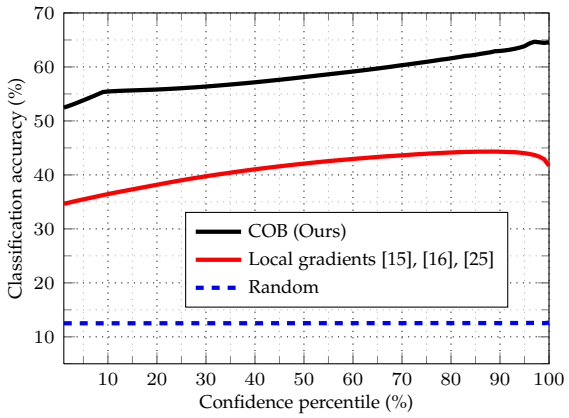


Fig. 6. **Contour orientation**: Classification accuracy into orientations quantized in 8 bins.

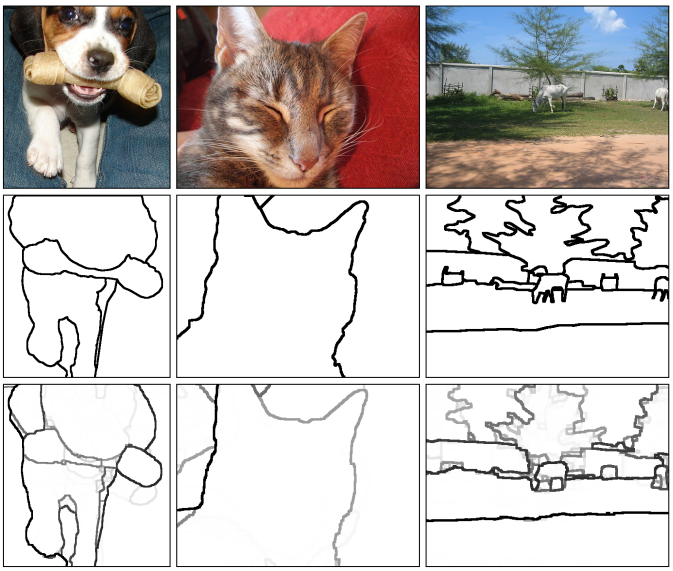


Fig. 9. **Qualitative results on PASCAL - Hierarchical Regions**. Row 1: original images, Row 2: ground-truth boundaries, Row 3: hierarchical regions with COB.

is capable of correctly distinguishing between internal contours and external, more semantical, object boundaries.

5.4 Object boundary detection

Concurrent works with the conference version of our paper [34] presented results on object boundary detection [19], [21] on the PASCAL VOC'12 Segmentation database. The database consists of 1464 training and 1449 validation images, including pixel-wise annotations of the instances and the semantic classes of the objects. The goal is to detect the boundaries of the objects that belong to the 20 classes of PASCAL, without distinguishing the semantics. Different from generic image segmentation, boundaries that do not belong to an object are treated as background.

We retrain COB on VOC'12 train set and report the results on the validation set. We use the instance level annotation of the database, and extract contours from the semantic segmentation annotations of the database. The uncertain areas (annotated with value of 255) are treated as background. We compare to several baselines, together with recent state-of-the-art results. Specifically, Khoreva et al. [19] retrained HED [2] on object contours, and Yang et al. [21] proposed a novel encoder-decoder architecture to tackle the same task. We evaluate the best pre-computed results provided by the authors in both cases. The results are quantified in Figure 11. We observe that COB obtains state-of-the-art results in all metrics. CEDN [21] performs better in the high precision regime. However, the authors used extra images from the SBD dataset [13] for training their detector. Also, CEDN is trained on an improved version of the ground truth, aligning the uncertain areas of VOC'12 with the the true image boundaries by using a CRF. We report results of COB trained only on VOC'12 train set, to be consistent with the results of Khoreva et al. [19]. In this experiment, we use `maxDist` of 0.01, as is adopted by the literature [19], [36].

Figure 10 illustrates some qualitative results, as well as the differences of generic segmentation and object boundary detection. We show our results on images of the VOC'12 val set using the model trained on PASCAL Context for generic image segmentation, and we compare qualitatively to the model retrained on the 20 classes of VOC'12 for object boundary detection. In the latter case, the detections are focused on the 20 object classes, disregarding strong contour cues of the background that are detected by the generic segmentation model.

5.5 RGB-D boundary detection on NYUD-v2 dataset

The NYUD (v2) dataset [37] consists of 1449 RGB-D indoor images, divided into splits of 795 training and 654 testing images, with the corresponding semantic and instance level segmentations. Gupta et al. [38] adopted this dataset for contour detection. In their experiments, they obtained the respective boundary annotations from the instance-level segmentations of the dataset. We evaluated the performance

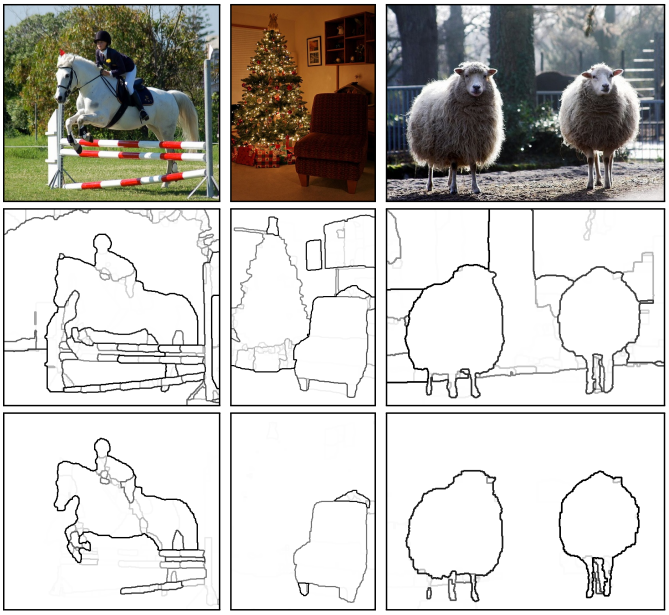
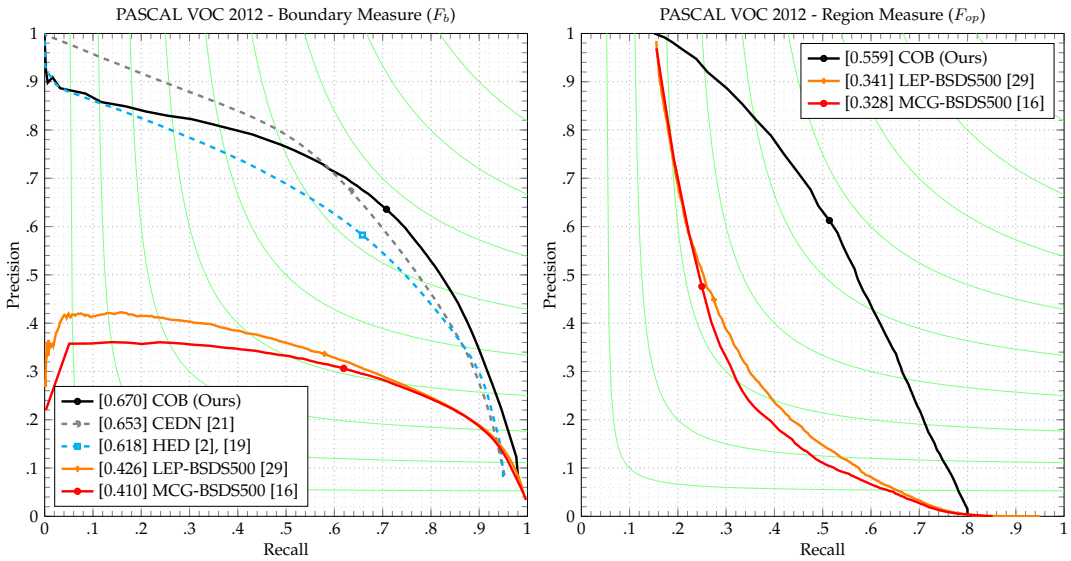


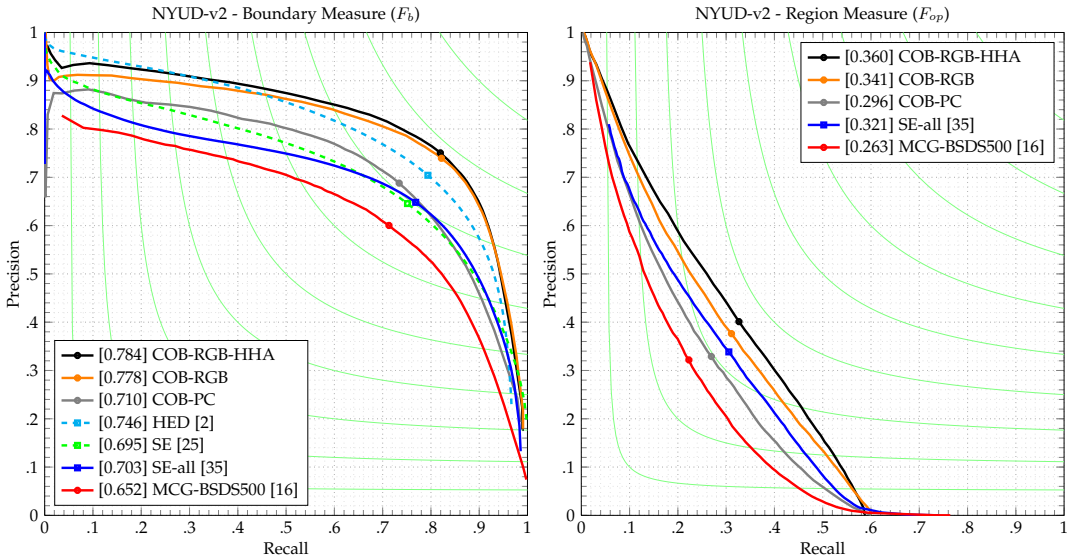
Fig. 10. **Qualitative results for Object Boundaries**. Row 1: original images, Row 2: Generic Image Segmentation results, Row 3: Object Boundary results.



Method	Boundaries - F_b ODS	OIS	AP
COB (Ours)	**0.670**	**0.728**	**0.687**
CEDN [21]	0.653	0.683	0.686
HED [2], [19]	0.618	0.650	0.628
LEP-B [29]	0.426	0.517	0.325
MCG-B [16]	0.410	0.486	0.294

Method	Regions - F_{op} ODS	OIS	AP
COB (Ours)	**0.559**	**0.644**	**0.391**
LEP-BSDS500 [29]	0.341	0.456	0.156
MCG-BSDS500 [16]	0.328	0.431	0.140

Fig. 11. **PASCAL VOC 2012 evaluation**: Precision-recall curves for evaluation of boundaries (F_b [26]), and regions (F_{op} [27]). ODS, OIS, and AP summary measures.



Method	Boundaries - F_b ODS	OIS	AP
COB-RGB-HHA	**0.784**	**0.805**	**0.825**
COB-RGB	0.778	0.799	0.814
COB-PC	0.710	0.735	0.734
HED [2]	0.746	0.761	0.786
SE [25]	0.695	0.708	0.719
SE-all [35]	0.703	0.716	0.707
MCG-B [16]	0.652	0.681	0.613

Method	Regions - F_{op} ODS	OIS	AP
COB-RGB-HHA	**0.360**	**0.398**	**0.270**
COB-RGB	0.341	0.376	0.252
COB-PC	0.296	0.332	0.203
SE-all [35]	0.321	0.354	0.179
MCG-BSDS500 [16]	0.263	0.296	0.159

Ablation: Method	Boundaries - F_b ODS	OIS	AP
ResNet50-RGB-HHA	0.745	**0.762**	**0.792**
ResNet50-RGB	**0.746**	0.761	0.789
ResNet50-RGBD	0.683	0.699	0.681

Fig. 12. **NYUD-v2 test**: Precision-recall curves for evaluation of boundaries (F_b [26]), and regions (F_{op} [27]). ODS, OIS, and AP summary measures.

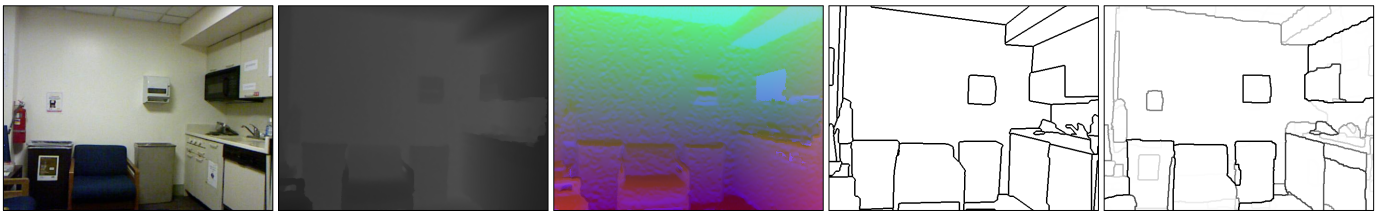


Fig. 13. **Data and results on NYUD-v2**. From left to right: RGB image, depth, HHA features [35], ground truth, and COB detections

by using the standard benchmarks of BSDS. Following [2], [25], [35], we increased the tolerance for incorrect localizations from 0.0075 of the image diagonal to 0.11, to compensate for inaccurate annotations of boundaries.

We use the extra information of depth to train different variants of COB on the NYUD dataset. Gupta et al. [35] used the camera parameters of the images to encode the depth information in three channels: horizontal disparity, height above ground, and the angle of the local surface normal with the inferred gravity direction at each pixel (HHA). We retrain three different variants of the CNN: (a) Only using RGB data (ResNet50-RGB), (b) Incorporating depth information into a fourth channel (ResNet50-RGBD), and (c) Concatenating RGB and HHA channels and operate on 6 channels directly (ResNet50-RGB-HHA). Figure 13 illustrates an overview of the data, along with depth and HHA features that we used, as well as the results obtained by COB. In Figure 12 we show the ablation analysis by directly evaluating the CNN output, without any post-processing. We observe that the CNNs retrained on RGB and RGB-HHA channels obtain significantly better results than the one trained on RGB-D data, showing that HHA features provide an appropriate encoding for depth information. We retrain the full pipeline of COB (including orientations) on NYUD and we report the precision-recall curves. We compare with various state-of-the-art methods, showing significant improvements. Specifically, we compare with the SE [25] detector retrained on the RGB-D data of NYUD, the detector proposed by [35] trained on RGB and depth normal gradients, and the best result reported on NYUD by HED [2], where the authors trained two different variants of the detector on RGB and HHA modalities respectively and averaged the obtained results. For completeness, we report results obtained by the original MCG [16] without any retraining on NYUD. The best result is obtained by the variant of COB trained on both RGB and HHA modalities. Compared to its RGB-only counterpart, the particular model achieves higher accuracy, suggesting that the depth embeddings are useful cues to discern contours when RGB modality alone is unable to do so. It is noteworthy that the post-processing step (orientations and UCMs) further boosts the performance of COB. For example, performance increases from 0.745 (ResNet50-RGB-HHA) to 0.784 (COB-RGB-HHA) by plugging in the orientations and the UCM pipeline to the trained ResNet50 architecture. We also report the results of the model trained on PASCAL Context (COB-PC) and operating only on RGB data, showing that it performs fairly well without any retraining on the NYUD dataset.

5.6 Efficiency Analysis

Contour detection and image segmentation, as a preprocessing step towards high-level applications, need to be computationally efficient. The previous state-of-the-art in hierarchical image segmentation [15], [16] was of limited use in practice due to its computational load.

As a core in our system, the forward pass of our network to compute the contour strength and 8 orientations takes 0.28 seconds on a NVidia Titan X GPU. Table 3 shows the timing comparison between the full system COB (Ours) and some related baselines on PASCAL Context. We divide the timing into different relevant parts, namely, the contour detection step, the Oriented Watershed Transform (OWT) and Ultrametric Contour Map (UCM) computation, and the globalization (normalized cuts) step.

Steps	(1) MCG [16]	(2) MCG-HED	(3) Fast UCMs	(4) COB (Ours)
Contours	3.08	0.39*	0.39*	0.28*
OWT, UCM	11.33	11.58	1.63	0.51
Globalize	9.96	9.97	9.92	0.00
Total Time	24.37	21.94	11.94	**0.79**

TABLE 3
Timing experiments: Comparing our approach to different baselines. Times computed using a GPU are marked with an asterisk.

Column (1) shows the timing for the original MCG [16], which uses Structured Edges (SE) [25]. As a first baseline, Column (2) displays the timing of MCG if we naively substitute SE by HED [2] at the three scales (running on a GPU). By applying the sparse boundaries representation we reduce the UCM and OWT time from 11.58 to 1.63 seconds (Column (3)). Our final technique COB, in which we remove the globalization step, computes the three scales in one pass and add contour orientations, takes 0.79 seconds in mean. Overall, comparing to previous state-of-the-art, we get a significant improvement at a fraction of the computation time (24.37 to 0.79 seconds).

6 Experiments on High-Level Vision Applications

This section is dedicated to present the interaction of COB boundaries and segments with higher vision tasks. In Section 6.1 we evaluate COB as object proposals by plugging in the detected UCMs into the combinatorial goruping pipeline of MCG [16]. In Section 6.2 we study the interplay of our boundary detector with semantic contours and semantic segmentation by combining COB with the Dilated Convolutions Network [53], and in Section 6.3 we couple the COB proposals with the Fast-RCNN [54] pipeline for object detection. In all cases, we show that COB co-operates well with existing approaches by improving their performance.

6.1 Object Proposals

Object proposals are an integral part of current object detection and semantic segmentation pipelines [48], [54], [55], as they provide a reduced search space on locations, scales, and shapes over the image. This section evaluates COB as a segmented and bounding box proposal technique, when using our high-quality region hierarchies in conjunction with the combinatorial grouping framework of MCG [16]. In terms of segmented object proposals, we compare against the most recent techniques SharpMask [39], DeepMask [40], POISE [41], MCG and SCG [16], LPO [42], GOP [44], SeSe [45], GLS [46], and RIGOR [47]. In terms of bounding box proposals, we compare also against SharpMask [39], DeepMask [40], EB [49], RPN [48] MCG and SCG [16], LPO [42], BING [50], SeSe [45], GLS [46], RIGOR [47], Obj [51], and RP [52]. Recent thorough comparisons of object proposal generation methods can be found in [56], [57].

We perform experiments on the PASCAL 2012 Segmentation dataset [12] and on the bigger and more challenging

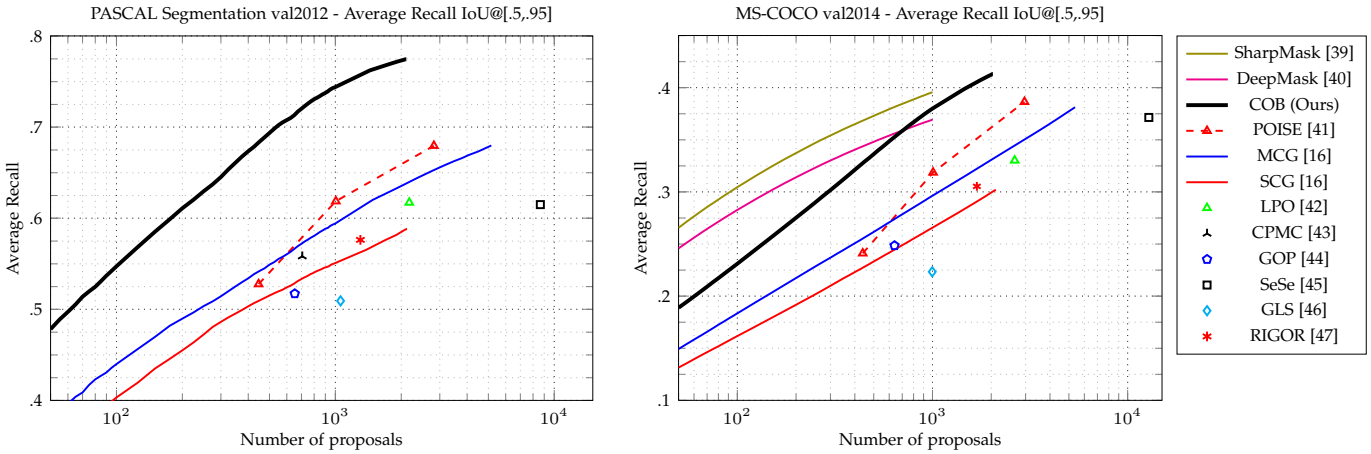


Fig. 14. **Segmented object proposals evaluation on PASCAL Segmentation val and MS-COCO val**: Dashed lines refer to methods that do not provide a ranked set of proposals, but they need to be re-parameterized.

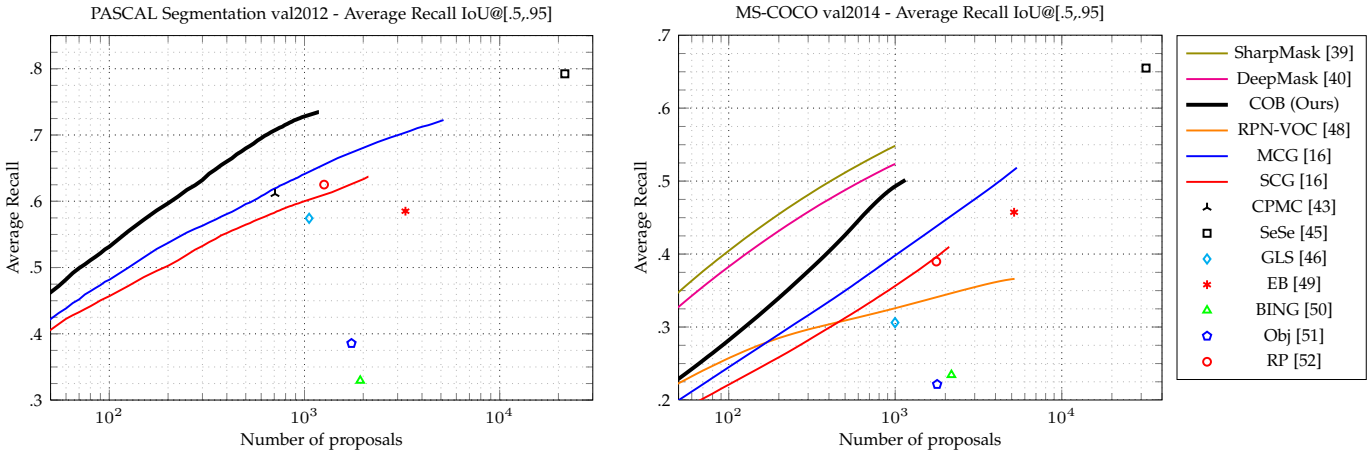


Fig. 15. **Bounding-box object proposals evaluation on PASCAL Segmentation val and MS-COCO val**: Dashed lines refer to methods that do not provide a ranked set of proposals, but they need to be re-parameterized. Note that COB is designed to detect segmented object proposals and not bounding-box proposals.

MS-COCO [58] (val2014 set). The hierarchies and combinatorial grouping are trained on PASCAL Context. To assess the generalization capability, we evaluate on MS-COCO, which contains a large number of previously unseen categories, without further retraining.

Figure 14 shows the average recall [57] with respect to the number of object proposals. In PASCAL VOC'12 Segmentation, the absolute gap of improvement of COB is at least of +13% with the second-best technique, and consistent in all the range of number of proposals. In MS-COCO, even though we did not train on any MS-COCO image, COB reaches competitive results for the task, with only very recent techniques [39], [40] reaching higher Average Recall when evaluating a low number of proposals. This shows that our contours, regions, and proposals are properly learning a generic concept of object rather than some specific categories.

Figure 15 shows the evaluation in terms of bounding box object proposals. COB is less competitive in terms of box proposals, however the algorithm was not specifically designed for detecting bounding boxes. We also show the comparison to RPN [48], which is trained on VOC'07, and thus does not generalize well in the classes of COCO.

6.2 Semantic Boundaries and Semantic Segmentation

The task of Semantic Boundaries, introduced by [13], requires not only detecting the boundaries, but also associating a semantic class to them. It can be thought as a combination of Boundary Detection and Semantic Segmentation, where except for the binary information of boundaries, one needs to label each of the detected pixels with the corresponding semantic class. The common approach to this task is to separately approach semantic segmentation and contour detection, and fuse the results of the two tasks [4], [13], [18]. Hariharan et al [13] tackled the task with generic object detectors and bottom up contours. Bertasius et al. [4], [18] show that results can be significantly improved when using deep-learning based semantic segmenters and contour detectors. Kokkinos [59] approaches the task with fully-convolutional networks trained end to end, although the results do not reach the current state of the art.

We also follow the most common approach of mixing the two tasks. We couple the COB boundaries with Semantic Segmentation results by dilated convolutions [53]. Specifically, we mask the boundaries with Semantic Segmentation results, with a tolerance of 0.02 of the image diagonal.

Technique	Plane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	MBike	Person	Plant	Sheep	Sofa	Train	TV	Mean maxF
COB-dil	**84.2**	**72.3**	**81.0**	64.2	**68.8**	81.7	71.5	**79.4**	**55.2**	**79.1**	40.8	**79.9**	**80.4**	**75.6**	**77.3**	**54.4**	**82.8**	51.7	72.1	62.4	**70.7**
DilatedConv [53]	83.7	71.8	78.8	**65.5**	66.3	**82.6**	**73.0**	77.3	47.3	76.8	37.2	78.4	79.4	75.2	73.8	46.2	79.5	46.6	**76.4**	**63.8**	69.0
BNF [18]	76.7	60.5	75.9	60.7	63.1	68.4	62.0	74.3	54.1	76.0	**42.9**	71.9	76.1	68.3	70.5	53.7	79.6	**51.9**	60.7	60.9	65.4
HFL [4]	73.6	61.1	74.2	57.0	58.7	70.2	60.8	71.8	46.3	72.1	36.0	70.9	72.9	67.5	69.9	44.1	73.1	42.2	62.2	60.4	62.2
[19]	65.9	54.1	63.6	47.9	47.0	60.4	50.9	56.5	40.4	56.0	30.0	57.5	58.0	57.4	59.5	39.0	64.2	35.4	51.0	42.4	51.9
[13]	41.5	46.7	15.6	17.1	36.5	42.7	40.3	22.6	18.8	27.0	12.5	18.2	35.4	29.4	48.1	13.8	26.9	11.0	22.0	31.3	27.9

TABLE 4
Semantic contours results: maximal F_b per class and mean maximal F_b is reported for all methods.

Technique	Plane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	MBike	Person	Plant	Sheep	Sofa	Train	TV	Mean AP
COB-dil	**85.7**	**69.3**	**77.6**	**59.7**	**64.1**	**82.9**	**69.7**	**80.5**	**41.8**	**79.4**	**26.0**	**78.9**	**81.5**	**74.7**	**77.3**	**43.8**	**82.8**	**39.3**	**73.3**	**56.4**	**67.2**
BNF [18]	75.9	46.0	70.5	48.9	48.6	65.3	53.5	65.2	38.2	69.7	20.9	62.3	72.2	56.6	63.3	38.5	75.7	31.4	45.6	48.1	54.8
HFL [4]	71.3	54.9	68.8	45.6	48.3	70.9	56.5	65.6	29.0	65.8	17.6	64.3	68.3	64.0	65.6	28.8	66.5	25.8	59.5	49.8	54.3
[19]	67.1	50.5	62.2	42.1	38.9	57.8	47.7	53.7	32.1	52.3	17.5	53.1	56.0	53.2	57.7	29.4	62.2	24.0	46.2	32.8	46.8
[13]	38.4	38.9	8.6	9.3	23.0	37.1	33.6	18.4	11.5	16.0	5.1	12.2	29.0	21.3	46.9	7.2	15.8	5.6	14.4	21.4	20.7

TABLE 5
Semantic contours results: Average Precision (AP) per class and mean AP (mAP) is reported for all methods.

Technique	BG	Plane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	MBike	Person	Plant	Sheep	Sofa	Train	TV	Mean
COB-dil	**92.6**	**84.0**	**75.0**	**82.3**	**70.2**	**74.8**	**92.5**	**84.6**	**89.3**	**48.5**	**88.3**	**56.8**	**87.4**	**85.0**	**84.5**	**83.5**	**63.0**	**81.7**	**53.8**	**85.6**	**73.8**	**78.0**
DilatedConv [53]	92.0	83.0	74.3	81.2	69.8	71.6	90.4	82.5	87.6	46.9	84.8	55.0	85.5	83.2	82.3	82.2	62.6	81.0	51.1	84.1	70.7	76.3

TABLE 6
Effect of COB on Semantic Segmentation. Per-class IoU and mean IoU are reported.

Technique	Plane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	MBike	Person	Plant	Sheep	Sofa	Train	TV	Mean
SeSe	**76.0**	76.8	65.3	**54.6**	38.0	76.5	78.2	81.6	40.1	74.1	66.5	78.9	81.8	74.5	66.2	32.9	65.6	67.7	73.4	**66.8**	66.8
COB	69.5	76.8	**69.7**	53.3	**44.6**	**80.5**	**81.3**	**83.1**	**45.3**	**74.2**	**69.4**	**80.1**	**84.2**	**76.7**	**72.8**	**35.9**	**67.1**	**68.4**	**75.1**	65.4	**68.7**

TABLE 7
VOC 2007 test evaluation: Object Detection performance (mAP) of Fast-RCNN [54], using object proposals from [45] (original) or COB.

We report results on the SBD [13] database, for semantic boundary detection and semantic segmentation, by using the standard benchmark. Tables 4 and 5 compare the results among various methods, in both metrics used in the benchmark (mean maximal F-measure and Average Precision) for all classes. The combination of COB with [53], denoted with COB-dil, achieves state-of-the-art results in both metrics. For fair comparison, we also include the results obtained by evaluating the semantic segmentation results obtained by [53] directly as contours. We show that COB fairly improves the result.

Having explored the performance of COB combined with the Dilated Convolution network on Semantic Boundaries, it is interesting to investigate the dual task: the effects of COB in semantic segmentation. This is done by treating the resulting COB UCMs as superpixels, and snapping the results of the semantic segmentation to those superpixels. The snapping is performed by majority voting of the regions, i.e superpixels that overlap more than 50% with the semantic class, are assigned the corresponding label. Table 6 reports the effects of such snapping on Semantic Segmentation, on the validation split of SBD dataset. Results improve consistently for all the classes, indicating that COB superpixels are further refining the semantic segmentation results on boundary locations.

In Figure 16 we present some qualitative results. Snapping to COB superpixels improves mainly on boundary locations, as well as on noisy semantic segmentation detections in places where COB superpixels are not present.

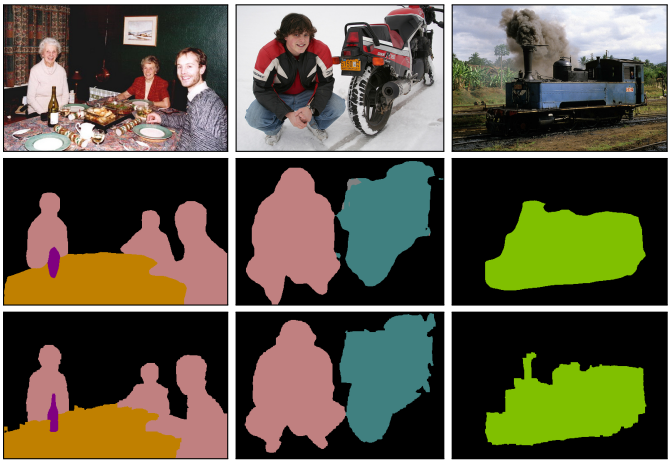


Fig. 16. **Qualitative results for Semantic Segmentation**. Row 1: original images, Row 2: Dilated Convolution Network, Row 3: Dilated Network with COB superpixels.

6.3 COB Object Proposals for Object Detection

Object Proposals have been extensively used to facilitate object detection [48], [54], [55]. Most common pipelines use object proposals in the form of a bounding box to regress

a class score and a refined prediction of the bounding box locations. Even though our approach provides segmented object proposals from a hierarchy of regions, it is possible to study their effect on common object detection pipelines by simply extracting the bounding box around them.

We evaluate the bounding box proposals generated by COB by feeding them into the Fast-RCNN [54] pipeline for Object Detection. The original approach uses the VGG network [9] together with the box proposals generated by the Selective Search [45] algorithm to predict class probability and refine the localization for each of them. The final detection performance is evaluated by performing non-maximum suppression on the detections.

Experiments are performed on the VOC'07 detection database. The database consists of 5011 training, and we report the performance on its 4952 testing images. In our experiments, we change the box proposals of Selective Search, to the ones generated by COB. We keep all the hyper-parameters of the original approach unchanged, both at training and test times. Table 7 quantitatively evaluates the effects of COB proposals in performance. We observe improvements in object detection performance (mean Average Precision - mAP), which further proves the high quality of the proposals generated by COB. We would like to emphasize that the latest developments on Object Detection use joint training of bounding box proposals and object class scores [48], [60], [61], [62], which together with training on external data achieves much higher results. Instead, we focus on proving the high quality of COB proposals compared to other object proposal techniques.

7 Conclusions

In this work, we have developed an approach to detect contours at multiple scales, together with their orientations, in a single forward pass of a convolutional neural network. We provide a fast framework for generating region hierarchies by efficiently combining multiscale oriented contour detections, thanks to a new sparse boundary representation. We shift from the BSDS to PASCAL in the evaluation to unwind all the potential of data-hungry methods such as CNNs and by observing that the performance on the BSDS is close to saturation.

Our technique achieves state-of-the-art performance by a significant margin for contour detection, the estimation of their orientation, and generic (RGB and RGB-D) image segmentation. We show that our architecture is modular by using two different CNN base architectures, which suggests that it will be able to transfer further improvements in CNN base architectures to perceptual grouping. We also show that our method does not require globalization, which was a speed bottleneck in previous approaches. The generalization of COB was further demonstrated when applied to high-level vision tasks (object proposals, object detection, and semantic contours and segmentation) in combination with recent pipelines, where the results are improved in all cases.

All our code, CNN models, pre-computed results, dataset splits, and benchmarks are publicly available at www.vision.ee.ethz.ch/~cvlsegmentation/.

Acknowledgments

Research funded by the EU Framework Programme for Research and Innovation - Horizon 2020 - Grant Agreement No. 645331 - EurEyeCase. The authors gratefully acknowledge support by armasuisse, and thank NVIDIA Corporation for donating the GPUs used in this project.

References

[1] I. Kokkinos, "Pushing the boundaries of boundary detection using deep learning," in *ICLR*, 2016.
[2] S. Xie and Z. Tu, "Holistically-nested edge detection," in *ICCV*, 2015.
[3] G. Bertasius, J. Shi, and L. Torresani, "Deepedge: A multi-scale bifurcated deep network for top-down contour detection," in *CVPR*, 2015.
[4] ——, "High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision," in *ICCV*, 2015.
[5] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *CVPR*, 2015.
[6] Y. Ganin and V. Lempitsky, "N^4-fields: Neural network nearest neighbor fields for image transforms," in *ACCV*, 2014.
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.
[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[11] 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*, 2001.
[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.
[13] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *ICCV*, 2011.
[14] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.
[15] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *TPAMI*, vol. 33, no. 5, pp. 898–916, 2011.
[16] J. Pont-Tuset, P. Arbeláez, J. Barron, F.Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *TPAMI*, vol. 39, no. 1, pp. 128 – 140, 2017.
[17] K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool, "Deep retinal image understanding," in *MICCAI*, 2016.
[18] G. Bertasius, J. Shi, and L. Torresani, "Semantic segmentation with boundary neural fields," in *CVPR*, 2016.
[19] A. Khoreva, R. Benenson, M. Omran, M. Hein, and B. Schiele, "Weakly supervised object boundaries," in *CVPR*, 2016.
[20] Y. Li, M. Paluri, J. M. Rehg, and P. Dollár, "Unsupervised learning of edges," in *CVPR*, 2016.
[21] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang, "Object contour detection with a fully convolutional encoder-decoder network," in *CVPR*, 2016.
[22] L. Najman and M. Schmitt, "Geodesic saliency of watershed contours and hierarchical segmentation," *TPAMI*, vol. 18, no. 12, pp. 1163–1173, 1996.
[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, 2015.
[24] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," *arXiv preprint arXiv:1409.5185*, 2014.
[25] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *TPAMI*, vol. 37, no. 8, pp. 1558–1570, 2015.

[26] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *TPAMI*, vol. 26, no. 5, pp. 530–549, 2004.

[27] J. Pont-Tuset and F. Marques, "Supervised evaluation of image segmentation and object proposal techniques," *TPAMI*, vol. 38, no. 7, pp. 1465–1478, 2016.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[29] Q. Zhao, "Segmenting natural images with the least effort as humans," in *BMVC*, 2015.

[30] Z. Ren and G. Shakhnarovich, "Image segmentation by cascaded region agglomeration," in *CVPR*, 2013.

[31] J. Shi and J. Malik, "Normalized cuts and image segmentation," *TPAMI*, vol. 22, no. 8, 2000.

[32] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, p. 2004, 2004.

[33] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *TPAMI*, vol. 24, no. 5, pp. 603 –619, 2002.

[34] K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool, "Convolutional oriented boundaries," in *ECCV*, 2016.

[35] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *ECCV*, 2014.

[36] J. Uijlings and V. Ferrari, "Situational object boundary detection," in *CVPR*, 2015.

[37] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[38] S. Gupta, P. Arbeláez, and J. Malik, "Perceptual organization and recognition of indoor scenes from RGB-D images," in *CVPR*, 2013.

[39] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *ECCV*, 2016.

[40] P. O. Pinheiro, R. Collobert, and P. Dollar, "Learning to segment object candidates," in *NIPS*, 2015.

[41] A. Humayun, F. Li, and J. M. Rehg, "The middle child problem: Revisiting parametric min-cut and seeds for object proposals," in *ICCV*, 2015.

[42] P. Krähenbühl and V. Koltun, "Learning to propose objects," in *CVPR*, 2015.

[43] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *TPAMI*, vol. 34, no. 7, pp. 1312–1328, 2012.

[44] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *ECCV*, 2014.

[45] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[46] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *CVPR*, 2014.

[47] A. Humayun, F. Li, and J. M. Rehg, "RIGOR: Recycling Inference in Graph Cuts for generating Object Regions," in *CVPR*, 2014.

[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[49] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.

[50] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014.

[51] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *TPAMI*, vol. 34, pp. 2189–2202, 2012.

[52] S. Manén, M. Guillaumin, and L. Van Gool, "Prime Object Proposals with Randomized Prim's Algorithm," in *ICCV*, 2013.

[53] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[54] R. Girshick, "Fast R-CNN," in *ICCV*, 2015.

[55] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[56] J. Pont-Tuset and L. Van Gool, "Boosting object proposals: From Pascal to COCO," in *ICCV*, 2015.

[57] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *TPAMI*, vol. 38, no. 4, pp. 814–830, 2016.

[58] T. Lin, M. Maire, S. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *arXiv:1405.0312*, 2014.

[59] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," *arXiv preprint arXiv:1609.02132*, 2016.

[60] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "Ssd: Single shot multibox detector," in *ECCV*, 2016.

[61] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016.

[62] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *ECCV*, 2016.

Kevis-Kokitsi Maninis is a PhD candidate at ETHZ, Switzerland, in Prof. Luc Van Gool's Computer Vision Lab (2015). He received the Diploma degree in Electrical and Computer Engineering from National Technical University of Athens (NTUA) in 2014. He worked as undergraduate research assistant in the Signal Processing and Computer Vision group of NTUA (2013-2014).

Jordi Pont-Tuset is a post-doctoral researcher at ETHZ, Switzerland, in Prof. Luc Van Gool's Computer Vision Lab (2015). He received the degree in Mathematics in 2008, the degree in Electrical Engineering in 2008, the M.Sc. in Research on Information and Communication Technologies in 2010, and the Ph.D with honors in 2014; all from the Universitat Politècnica de Catalunya, BarcelonaTech (UPC). He worked at Disney Research, Zürich (2014).

Pablo Arbeláez received a PhD with honors in Applied Mathematics from the Université Paris-Dauphine in 2005. He was a Research Scientist with the Computer Vision Group at UC Berkeley from 2007 to 2014. He currently holds a faculty position at Universidad de los Andes in Colombia. His research interests are in computer vision, where he has worked on a number of problems, including perceptual grouping, object recognition and the analysis of biomedical images.

Luc Van Gool got a degree in electromechanical engineering at the Katholieke Universiteit Leuven in 1981. Currently, he is professor at the Katholieke Universiteit Leuven, Belgium, and the ETHZ, Switzerland, Switzerland. He leads computer vision research at both places, where he also teaches computer vision. He has authored over 200 papers in this field. He has been a program committee member of several major computer vision conferences. His main interests include 3D reconstruction and modeling, object recognition, tracking, and gesture analysis. He received several Best Paper awards. He is a co-founder of 5 spin-off companies.