

Bi-Directional Cascade Network for Perceptual Edge Detection

Jianzhong He¹, Shiliang Zhang¹, Ming Yang², Yanhu Shan², Tiejun Huang^{1,3}

¹Peking University, ²Horizon Robotics, Inc, ³Peng Cheng Laboratory.

{jianzhonghe, slzhang.jdl, tjhuang}@pku.edu.cn, m-yang4@u.northwestern.edu, yanhu.shan@horizon.ai

Abstract

Exploiting multi-scale representations is critical to improve edge detection for objects at different scales. To extract edges at dramatically different scales, we propose a Bi-Directional Cascade Network (BDCN) structure, where an individual layer is supervised by labeled edges at its specific scale, rather than directly applying the same supervision to all CNN outputs. Furthermore, to enrich multi-scale representations learned by BDCN, we introduce a Scale Enhancement Module (SEM) which utilizes dilated convolution to generate multi-scale features, instead of using deeper CNNs or explicitly fusing multi-scale edge maps. These new approaches encourage the learning of multi-scale representations in different layers and detect edges that are well delineated by their scales. Learning scale dedicated layers also results in compact network with a fraction of parameters. We evaluate our method on three datasets, *i.e.*, BSDS500, NYUDv2, and Multicue, and achieve ODS F-measure of 0.828, 1.3% higher than current state-of-the-art on BSDS500. The code has been available¹.

1. Introduction

Edge detection targets on extracting object boundaries and perceptually salient edges from natural images, which preserve the gist of an image and ignore unintended details. Thus, it is important to a variety of mid- and high-level vision tasks, such as image segmentation [1, 41], object detection and recognition [13, 14], *etc.* Thanks to research efforts ranging from exploiting low-level visual cues with hand-crafted features [4, 22, 1, 28, 10] to recent deep models [3, 30, 23, 47], the accuracy of edge detection has been significantly boosted. For example, on the Berkeley Segmentation Data Set and Benchmarks 500 (BSDS500) [1], the detection performance has been boosted from 0.598 [7] to 0.815 [47] in ODS F-measure.

Nevertheless, there remain some open issues worthy of studying. As shown in Fig. 1, edges in one image stem



Figure 1. Some images and their ground truth edge maps in BSDS500 dataset. The scale of edges in one image varies considerably, like the boundaries of human body and hands.

from both object-level boundaries and meaningful local details, *e.g.*, the silhouette of human body and the shape of hand gestures. The variety of scale of edges makes it crucial to exploit multi-scale representations for edge detection. Recent neural net based methods [2, 42, 49] utilize hierarchical features learned by Convolutional Neural Networks (CNN) to obtain multi-scale representations. To generate more powerful multi-scale representation, some researchers adopt very deep networks, like ResNet50 [18], as the backbone model of the edge detector. Deeper models generally involve more parameters, making the network hard to train and expensive to infer. Another way is to build an image pyramid and fuse multi-level features, which may involve redundant computations. In another word, can we use a shallow or light network to achieve a comparable or even better performance?

Another issue is about the CNN training strategy for edge detection, *i.e.*, supervising predictions of different network layers by one general ground truth edge map [49, 30]. For instance, HED [49, 50] and RCF [30] compute edge prediction on each intermediate CNN output to spot edges at different scales, *i.e.*, the lower layers are expected to detect more local image patterns while higher layers capture object-level information with larger receptive fields. Since different network layers attend to depict patterns at different scales, it is not optimal to train those layers with the same supervision. In another word, existing works [49, 50, 30] enforce each layer of CNN to predict edges at all scales and ignore that one specific intermediate layer can only focus on edges at certain scales. Liu *et al.* [31] propose to relax the supervisions on intermediate layers using Canny [4] detectors with layer-specific scales. However, it is hard to decide layer-specific scales through human intervention.

¹<https://www.pkuvmc.com/dataset.html>.

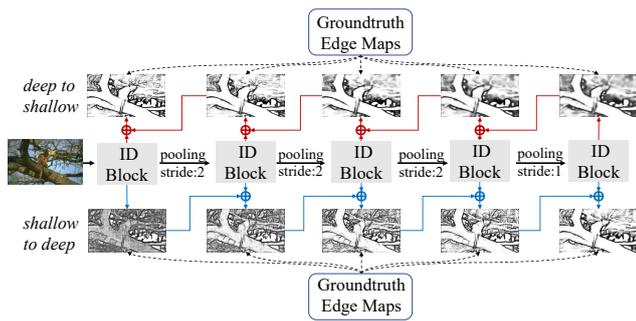


Figure 2. The overall architecture of BDCN. ID Block denotes the Incremental Detection Block, which is the basic component of BDCN. Each ID Block is trained by layer-specific supervisions inferred by a bi-directional cascade structure. This structure trains each ID Block to spot edges at a proper scale. The predictions of ID Blocks are fused as the final result.

Aiming to fully exploit the multiple scale cues with a shallow CNN, we introduce a Scale Enhancement Module (SEM) which consists of multiple parallel convolutions with different dilation rates. As shown in image segmentation [5], dilated convolution effectively increases the size of receptive fields of network neurons. By involving multiple dilated convolutions, SEM captures multi-scale spatial contexts. Compared with previous strategies, *i.e.*, introducing deeper networks and explicitly fusing multiple edge detections, SEM does not significantly increase network parameters and avoids the repetitive edge detection on image pyramids.

To address the second issue, each layer in CNN shall be trained by proper layer-specific supervision, *e.g.*, the shallow layers are trained to focus on meaningful details and deep layers should depict object-level boundaries. We propose a Bi-Directional Cascade Network (BDCN) architecture to achieve effective layer-specific edge learning. For each layer in BDCN, its layer-specific supervision is inferred by a bi-directional cascade structure, which propagates the outputs from its adjacent higher and lower layers, as shown in Fig. 2. In another word, each layer in BDCN predicts edges in an incremental way *w.r.t* scale. We hence call the basic block in BDCN, which is constructed by inserting several SEMs into a VGG-type block, as the Incremental Detection Block (ID Block). This bi-directional cascade structure enforces each layer to focus on a specific scale, allowing for a more rational training procedure.

By combining SEM and BDCN, our method achieves consistent performance on three widely used datasets, *i.e.*, *BSDS500*, *NYUDv2*, and *Multicue*. It achieves ODS F-measure of 0.828, 1.3% higher than current state-of-the-art CED [47] on *BSDS500*. It achieves 0.806 only using the trainval data of *BSDS500* for training, and outperforms the human perception (ODS F-measure 0.803). To our best knowledge, we are the first that outperforms human perception by training only on trainval data of *BSDS500*. More-

over, we achieve a better trade-off between model compactness and accuracy than existing methods relying on deeper models. With a shallow CNN structure, we obtain comparable performance with some well-known methods [3, 42, 2]. For example, we outperform HED [49] using only 1/6 of its parameters. This shows the validity of our proposed SEM, which enriches the multi-scale representations in CNN. This work is also an original effort studying a rational training strategy for edge detection, *i.e.*, employing the BDCN structure to train each CNN layer with layer-specific supervision.

2. Related Work

This work is related to edge detection, multi-scale representation learning, and network cascade structure. We briefly review these three lines of works, respectively.

Edge Detection: Most edge detection methods can be categorized into three groups, *i.e.*, traditional edge operators, learning based methods, and the recent deep learning, respectively. Traditional edge operators [22, 4, 45, 34] detect edges by finding sudden changes in intensity, color, texture, *etc.* Learning based methods spot edges by utilizing supervised models and hand-crafted features. For example, Dollár *et al.* [10] propose structured edge which jointly learns the clustering of groundtruth edges and the mapping of image patch to clustered token. Deep learning based methods use CNN to extract multi-level hierarchical features. Bertasius *et al.* [2] employ CNN to generate features of candidate contour points. Xie *et al.* [49] propose an end-to-end detection model that leverages the outputs from different intermediate layers with skip-connections. Liu *et al.* [30] further learn richer deep representations by concatenating features derived from all convolutional layers. Xu *et al.* [51] introduce a hierarchical deep model to extract multi-scale features and a gated conditional random field to fuse them.

Multi-Scale Representation Learning: Extraction and fusion of multi-scale features are fundamental and critical for many vision tasks, *e.g.*, [19, 52, 6]. Multi-scale representations can be constructed from multiple re-scaled images [12, 38, 11], *i.e.*, an image pyramid, either by computing features independently at each scale [12] or using the output from one scale as the input to the next scale [38, 11]. Recently, innovative works DeepLab [5] and PSPNet [55] use dilated convolutions and pooling to achieve multi-scale feature learning in image segmentation. Chen *et al.* [6] propose an attention mechanism to softly weight the multi-scale features at each pixel location.

Like other image patterns, edges vary dramatically in scales. Ren *et al.* [39] show that considering multi-scale cues does improve performance of edge detection. Multiple scale cues are also used in many approaches [48, 39, 24, 50, 30, 34, 51]. Most of those approaches explore the scale-

space of edges, *e.g.*, using Gaussian smoothing at multiple scales [48] or extracting features from different scaled images [1]. Recent deep based methods employ image pyramid and hierarchal features. For example, Liu *et al.* [30] forward multiple re-scaled images to a CNN independently, then average the results. Our approaches follow a similar intuition, nevertheless, we build SEM to learn multi-scale representations in an efficient way, which avoids repetitive computation on multiple input images.

Network Cascade: Network cascade [21, 37, 25, 46, 26] is an effective scheme for many vision applications like classification [37], detection [25], pose estimation [46] and semantic segmentation [26]. For example, Murthy *et al.* [37] treat easy and hard samples with different networks to improve classification accuracy. Yuan *et al.* [54] ensemble a set of models with different complexities to process samples with different difficulties. Li *et al.* [26] propose to classify easy regions in a shallow network and train deeper networks to deal with hard regions. Lin *et al.* [29] propose a top-down architecture with lateral connections to propagate deep semantic features to shallow layers. Different from previous network cascade, BDCN is a bi-directional pseudo-cascade structure, which allows an innovative way to supervise each layer individually for layer-specific edge detection. To our best knowledge, this is an early and original attempt to adopt a cascade architecture in edge detection.

3. Proposed Methods

3.1. Formulation

Let (X, Y) denote one sample in the training set \mathbb{T} , where $X = \{x_j, j = 1, \dots, |X|\}$ is a raw input image and $Y = \{y_j, j = 1, \dots, |X|\}, y_j \in \{0, 1\}$ is the corresponding groundtruth edge map. Considering the scale of edges may vary considerably in one image, we decompose edges in Y into S binary edge maps according to the scale of their depicted objects, *i.e.*,

$$Y = \sum_{s=1}^S Y_s, \quad (1)$$

where Y_s contains annotated edges corresponding to a scale s . Note that, we assume the scale of edges is in proportion to the size of their depicted objects.

Our goal is to learn an edge detector $D(\cdot)$ capable of detecting edges at different scales. A natural way to design $D(\cdot)$ is to train a deep neural network, where different layers correspond to different sizes of receptive field. Specifically, we can build a neural network N with S convolutional layers. The pooling layers make adjacent convolutional layers depict image patterns at different scales.

For one training image X , suppose the feature map generated by the s -th convolutional layer is $N_s(X) \in \mathbb{R}^{w \times h \times c}$.

Using $N_s(X)$ as input, we design a detector $D_s(\cdot)$ to spot edges at scale s . The training loss for $D_s(\cdot)$ is formulated as

$$\mathcal{L}_s = \sum_{X \in \mathbb{T}} |P_s - Y_s|, \quad (2)$$

where $P_s = D_s(N_s(X))$ is the edge prediction at scale s . The final detector $D(\cdot)$ hence is derived as the ensemble of detectors learned from scale 1 to S .

To make the training with Eq. (2) possible, Y_s is required. It is not easy to decompose the groundtruth edge map Y manually into different scales, making it hard to obtain the layer-specific supervision Y_s for the s -th layer. A possible solution is to approximate Y_s based on ground truth label Y and edges predicted at other layers, *i.e.*,

$$Y_s \sim Y - \sum_{i \neq s} P_i. \quad (3)$$

However, Y_s computed in Eq. (3) is not an appropriate layer-specific supervision. In the following paragraph, we briefly explain the reason.

According to Eq. (3), for a training image, its predicted edges P_s at layer s should approximate Y_s , *i.e.*, $P_s \sim Y - \sum_{i \neq s} P_i$. In other words, we can pass the other layers' predictions to layer s for training, resulting in an equivalent formulation, *i.e.*, $Y \sim \sum_i P_i$. The training objective can thus become $\mathcal{L} = \mathcal{L}(\hat{Y}, Y)$, where $\hat{Y} = \sum_i P_i$. The gradient *w.r.t* the prediction P_s of layer s is

$$\frac{\partial(L)}{\partial(P_s)} = \frac{\partial(L(\hat{Y}, Y))}{\partial(P_s)} = \frac{\partial(L(\hat{Y}, Y))}{\partial(\hat{Y})} \cdot \frac{\partial(\hat{Y})}{\partial(P_s)}. \quad (4)$$

According to Eq. (4), for edge predictions P_s, P_i at any two layers s and $i, s \neq i$, their loss gradients are equal because $\frac{\partial(\hat{Y})}{\partial(P_s)} = \frac{\partial(\hat{Y})}{\partial(P_i)} = 1$. This implies that, with Eq. (3), the training process dose not necessarily differentiate the scales depicted by different layers, making it not appropriate for our layer-specific scale learning task.

To address the above issue, we approximate Y_s with two complementary supervisions. One ignores the edges with scales smaller than s , and the other ignores the edges with larger scales. Those two supervisions train two edge detectors at each scale. We define those two supervisions at scale s as

$$\begin{aligned} Y_s^{s2d} &= Y - \sum_{i < s} P_i^{s2d}, \\ Y_s^{d2s} &= Y - \sum_{i > s} P_i^{d2s}, \end{aligned} \quad (5)$$

where the superscript $s2d$ denotes information propagation from shallow layers to deeper layers, and $d2s$ denotes the prorogation from deep layers to shallower layers.

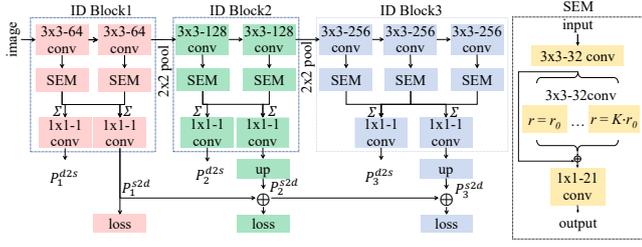


Figure 3. The detailed architecture of BDCN and SEM. For illustration, we only show 3 ID Blocks and the cascade from shallow to deep. The number of ID Blocks in our network can be flexibly set from 2 to 5 (see Fig. 9).

For scale s , the predicted edges P_s^{s2d} and P_s^{d2s} approximate to Y_s^{s2d} and Y_s^{d2s} , respectively. Their combination is a reasonable approximation to Y_s , *i.e.*,

$$P_s^{s2d} + P_s^{d2s} \sim 2Y - \sum_{i < s} P_i^{s2d} - \sum_{i > s} P_i^{d2s}, \quad (6)$$

where the edges predicted at scales $i \neq s$ are depressed. Therefore, we use $P_s^{s2d} + P_s^{d2s}$ to interpolate the edge prediction at scale s .

Because different convolutional layers depict different scales, the depth of a neural network determines the range of scales it could model. A shallow network may not be capable to detect edges at all of the S scales. However, a large number of convolutional layers involves too many parameters and makes the training difficult. To enable edge detection at different scales with a shallow network, we propose to enhance the multi-scale representation learned in each convolutional layer with the Scale Enhancement Module (SEM). The detail of SEM will be presented in Sec. 3.2.

3.2. Architecture of BDCN

Based on Eq. (6), we propose a Bi-Directional Cascade Network (BDCN) architecture to achieve layer-specific training for edge detection. As shown in Fig. 2, our network is composed of multiple ID Blocks, each of which is learned with different supervisions inferred by a bi-directional cascade structure. Specifically, the network is based on the VGG16 [44] by removing its three fully connected layers and last pooling layer. The 13 convolutional layers in VGG16 are then divided into 5 blocks, each follows a pooling layer to progressively enlarge the receptive fields in the next block. The VGG blocks evolve into ID Blocks by inserting several SEMs. We illustrate the detailed architecture of BDCN and SEM in Fig. 3.

ID Block is the basic component of our network. Each ID block produces two edge predictions. As shown in Fig. 3, an ID Block consists of several convolutional layers, each is followed by a SEM. The outputs of multiple SEMs are fused and fed into two 1×1 convolutional layers to generate two edges predictions P^{d2s} and P^{s2d} , respectively. The cascade structure shown in Fig. 3 propagates the

edge predictions from the shallow layers to deep layers. For the s -th block, P_s^{s2d} is trained with supervision Y_s^{s2d} computed in Eq. (5). P_s^{d2s} is trained in a similar way. The final edge prediction is computed by fusing those intermediate edge predictions in a fusion layer using 1×1 convolution.

Scale Enhancement Module is inserted into each ID Block to enrich the multi-scale representations in it. SEM is inspired by the dilated convolution proposed by Chen *et al.* [5] for image segmentation. For an input two-dimensional feature map $\mathbf{x} \in \mathcal{R}^{H \times W}$ with a convolution filter $\mathbf{w} \in \mathcal{R}^{h \times w}$, the output $\mathbf{y} \in \mathcal{R}^{H' \times W'}$ of dilated convolution at location (i, j) is computed by

$$\mathbf{y}_{ij} = \sum_{m,n} \mathbf{x}_{[i+r \cdot m, j+r \cdot n]} \cdot \mathbf{w}_{[m,n]}, \quad (7)$$

where r is the dilation rate, indicating the stride for sampling input feature map. Standard convolution can be treated as a special case with $r = 1$. Eq. (7) shows that dilated convolution enlarges the receptive field of neurons without reducing the resolution of feature maps or increasing the parameters.

As shown on the right side of Fig. 3, for each SEM we apply K dilated convolutions with different dilation rates. For the k -th dilated convolution, we set its dilation rate as $r_k = \max(1, r_0 \times k)$, which involves two parameters in SEM: the dilation rate factor r_0 and the number of convolution layers K . They are evaluated in Sec. 4.3.

3.3. Network Training

Each ID Block in our network is trained with two layer-specific side supervisions. Besides that, we fuse the intermediate edge predictions with a fusion layer as the final result. Therefore, BDCN is trained with three types of loss. We formulate the overall loss \mathcal{L} as,

$$\mathcal{L} = w_{side} \cdot \mathcal{L}_{side} + w_{fuse} \cdot \mathcal{L}_{fuse}(P, Y), \quad (8)$$

$$\mathcal{L}_{side} = \sum_{s=1}^S \mathcal{L}(P_s^{d2s}, Y_s^{d2s}) + \mathcal{L}(P_s^{s2d}, Y_s^{s2d}), \quad (9)$$

where w_{side} and w_{fuse} are weights for the side loss and fusion loss, respectively. P denotes the final edge prediction.

The function $\mathcal{L}(\cdot)$ is computed at each pixel with respect to its edge annotation. Because the distribution of edge/non-edge pixels is heavily biased, we employ a class-balanced cross-entropy loss as $\mathcal{L}(\cdot)$. Because of the inconsistency of annotations among different annotators, we also introduce a threshold γ for loss computation. For a groundtruth $Y = (y_j, j = 1, \dots, |Y|)$, $y_j \in (0, 1)$, we define $Y_+ = \{y_j, y_j > \gamma\}$ and $Y_- = \{y_j, y_j = 0\}$. Only pixels corresponding to Y_+ and Y_- are considered in loss computation. We hence

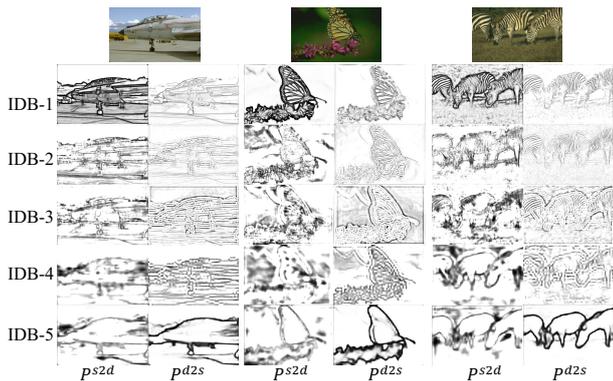


Figure 4. Examples of edges detected by different ID Blocks (IDB for short). Each ID Block generates two edge predictions, P^{s2d} and P^{d2s} , respectively.

define $\mathcal{L}(\cdot)$ as

$$\mathcal{L}(\hat{Y}, Y) = -\alpha \sum_{j \in Y_-} \log(1 - \hat{y}_j) - \beta \sum_{j \in Y_+} \log(\hat{y}_j), \quad (10)$$

where $\hat{Y} = (\hat{y}_j, j = 1, \dots, |\hat{Y}|)$, $\hat{y}_j \in (0, 1)$ denotes a predicted edge map, $\alpha = \lambda \cdot |Y_+| / (|Y_+| + |Y_-|)$, $\beta = |Y_-| / (|Y_+| + |Y_-|)$ balance the edge/non-edge pixels. λ controls the weight of positive over negative samples.

Fig. 4 shows edges detected by different ID blocks. We observe that, edges detected by different ID Blocks correspond to different scales. The shallow ID Blocks produce strong responses on local details and deeper ID Blocks are more sensitive to edges at larger scale. For instance, detailed edges on the body of zebra and butterfly can be detected by shallow ID Block, but are depressed by deeper ID Block. The following section tests the validity of BDCN and SEM.

4. Experiments

4.1. Datasets

We evaluate the proposed approach on three public datasets: *BSDS500* [1], *NYUDv2* [43], and *Multicue* [35].

BSDS500 contains 200 images for training, 100 images for validation, and 200 images for testing. Each image is manually annotated by multiple annotators. The final groundtruth is the averaged annotations by the annotators. We also utilize the strategies in [49, 30, 47] to augment training and validation sets by randomly flipping, scaling and rotating images. Following those works, we also adopt the PASCAL VOC Context dataset [36] as our training set.

NYUDv2 consists of 1449 pairs of aligned RGB and depth images. It is split into 381 training, 414 validation, and 654 testing images. *NYUDv2* is initially used for scene understanding, hence is also used for edge detection in previous works [15, 40, 49, 30]. Following those works, we

augment the training set by randomly flipping, scaling, and rotating training images.

Multicue [35] contains 100 challenging natural scenes. Each scene has two frame sequences taken from left and right view, respectively. The last frame of left-view sequence is annotated with edges and boundaries. Following [35, 30, 50], we randomly split 100 annotated frames into 80 and 20 images for training and testing, respectively. We also augment the training data with the same way in [49].

4.2. Implementation Details

We implement our network using PyTorch. The VG-G16 [44] pretrained on ImageNet [8] is used to initialize the backbone. The threshold γ used for loss computation is set as 0.3 for *BSDS500*. γ is set as 0.3 and 0.4 for *Multicue boundary* and *edges* datasets, respectively. *NYUDv2* provides binary annotations, thus does not need to set γ for loss computation. Following [30], we set the parameter λ as 1.1 for *BSDS500* and *Multicue*, set λ as 1.2 for *NYUDv2*.

SGD optimizer is adopted to train our network. On *BSDS500* and *NYUDv2*, we set the batch size to 10 for all the experiments. The initial learning rate, momentum and weight decay are set to $1e-6$, 0.9, and $2e-4$ respectively. The learning rate decreases by 10 times after every 10k iterations. We train 40k iterations for *BSDS500* and *NYUDv2*, 2k and 4k iterations for *Multicue boundary* and *edge*, respectively. w_{side} and w_{fuse} are set as 0.5, and 1.1, respectively. Since *Multicue* dataset includes high resolution images, we randomly crop 500×500 patches from each image in training. All the experiments are conducted on a NVIDIA GeForce1080Ti GPU with 11GB memory.

We follow previous works [49, 30, 47, 51], and perform standard Non-Maximum Suppression (NMS) to produce the final edge maps. For a fair comparison with other work, we report our edge detection performance with commonly used evaluation metrics, including Average Precision (AP), as well as F-measure at both Optimal Dataset Scale (ODS) and Optimal Image Scale (OIS). The maximum tolerance allowed for correct matches between edge predictions and groundtruth annotations is set to 0.0075 for *BSDS500* and *Multicue* dataset, and is set to 0.011 for *NYUDv2* dataset as in previous works [30, 35, 50].

4.3. Ablation Study

In this section, we conduct experiments on *BSDS500* to study the impact of parameters and verify each component in our network. We train the network on the *BSDS500* training set and evaluate on the validation set. Firstly, we test the impact of the parameters in SEM, *i.e.*, the number of dilated convolutions K and the dilation rate factor r_0 . Experimental results are summarized in Table 1.

Table 1 (a) shows the impact of K with $r_0=4$. Note that,

Table 1. Impact of SEM parameters to the edge detection performance on *BSDS500* validation set. (a) shows the impact of K with $r_0=4$. (b) shows the impact of r_0 with $K=3$.

(a)				(b)				
K	ODS	OIS	AP	r_0	rate	ODS	OIS	AP
0	.7728	.7881	.8093	0	1,1,1	.7720	.7881	.8116
1	.7733	.7845	.8139	1	1,2,3	.7721	.7882	.8124
2	.7738	.7876	.8169	2	2,4,6	.7725	.7875	.8132
3	.7748	.7894	.8170	4	4,8,12	.7748	.7894	.8170
4	.7745	.7896	.8166	8	8,16,24	.7742	.7889	.8169

Table 2. Validity of components in BDCN on *BSDS500* validation set. (a) tests different cascade architectures. (b) shows the validity of SEM and the bi-directional cascade architecture.

(a)				(b)			
Architecture	ODS	OIS	AP	Method	ODS	OIS	AP
baseline	.7681	.7751	.7912	baseline	.7681	.7751	.7912
S2D	.7683	.7802	.7978	SEM	.7748	.7894	.8170
D2S	.7710	.7816	.8049	S2D+D2S	.7762	.7872	.8013
S2D+D2S	.7762	.7872	.8013	(BDCN w/o SEM)			
(BDCN w/o SEM)				BDCN	.7765	.7882	.8091

$K=0$ means directly copying the input as output. The results demonstrate that setting K larger than 1 substantially improves the performance. However, too large K does not constantly boost the performance. The reason might be that, large K produces high dimensional outputs and makes edge extraction from such high dimensional data difficult. Table 1 (b) also shows that larger r_0 improves the performance. But the performance starts to drop with too large r_0 , e.g., $r_0=8$. In our following experiments, we fix $K=3$ and $r_0=4$.

Table 2 (a) shows the comparison among different cascade architectures, i.e., single direction cascade from shallow to deep layers (S2D), from deep to shallow layers (D2S), and the bi-directional cascade (S2D+D2S), i.e., the BDCN w/o SEM. Note that, we use the VGG16 network without fully connected layer as baseline. It can be observed that, both S2D and D2S structures outperform the baseline. This shows the validity of the cascade structure in network training. The combination of these two cascade structures, i.e., S2D+D2S, results in the best performance. We further test the performance of combining SEM and S2D+D2S and summarize the results in Table 2 (b), which shows that SEM and bi-directional cascade structure consistently improve the performance of baseline, e.g., improve the ODS F-measure by 0.7% and 0.8% respectively. Combining SEM and S2D+D2S results in the best performance. We can conclude that, the components introduced in our method are valid in boosting edge detection performance.

4.4. Comparison with Other Works

Performance on BSDS500: We compare our approach with recent deep learning based methods including CED [47], RCF [30], DeepBoundary [23], DCD [27], COB [32], HED [49], HFL [3], DeepEdge [2] and Deep-

Table 3. Comparison with other methods on *BSDS500* test set. † indicates trained with additional PASCAL-Context data. ‡ indicates the fused result of multi-scale images.

Method	ODS	OIS	AP
Human	.803	.803	-
SCG [40]	.739	.758	.773
PMI [20]	.741	.769	.799
OEF [17]	.746	.770	.820
DeepContour [42]	.757	.776	.800
HFL [3]	.767	.788	.795
HED [49]	.788	.808	.840
CEDN [53] †	.788	.804	-
COB [32]	.793	.820	.859
DCD [27]	.799	.817	.849
AMH-Net [51]	.798	.829	.869
RCF [30]	.798	.815	-
RCF [30] †	.806	.823	-
RCF [30] ‡	.811	.830	-
Deep Boundary [23]	.789	.811	.789
Deep Boundary [23] ‡	.809	.827	.861
Deep Boundary [23] ‡ + Grouping	.813	.831	.866
CED [47]	.794	.811	.847
CED [47] ‡	.815	.833	.889
LPCB [9]	.800	.816	-
LPCB [9] †	.808	.824	-
LPCB [9] ‡	.815	.834	-
BDCN	.806	.826	.847
BDCN †	.820	.838	.888
BDCN ‡	.828	.844	.890

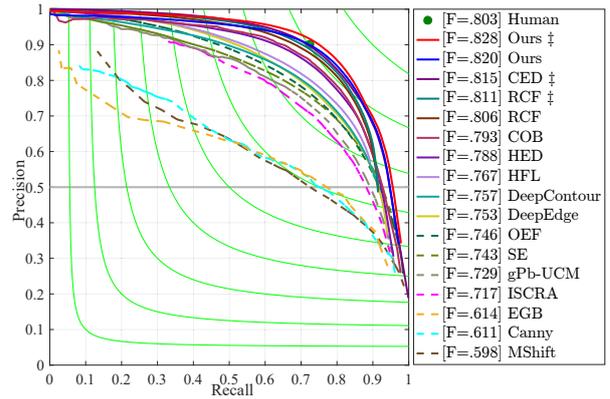


Figure 5. The precision-recall curves of our method and other works on *BSDS500* test set.

Contour [42], and traditional edge detection methods, including SCG [40], PMI [20] and OEF [17]. The comparison on *BSDS500* is summarized in Table 3 and Fig. 5, respectively.

As shown in the results, our method obtains the F-measure ODS of 0.820 using single scale input, and achieves 0.828 with multi-scale inputs, both outperform all of these competing methods. Using a single-scale input, our method still outperforms the recent CED [47] and DeepBoundary [23] that use multi-scale inputs. Our method also outperforms the human perception by 2.5% in F-measure ODS. The F-measure OIS and AP of our approach are also higher than the ones of the other methods.

Performance on NYUDv2: *NYUDv2* has three types of inputs, i.e., RGB, HHA, and RGB-HHA, respectively. Fol-

Table 4. Comparison with recent works on *NYUDv2*.

Method		ODS	OIS	AP
gPb-UCM [1]		.632	.661	.562
gPb+NG [15]		.687	.716	.629
OEF[17]	RGB	.651	.667	—
SE [10]		.695	.708	.679
SE+NG+ [16]		.706	.734	.738
HED [49]	RGB	.720	.734	.734
	HHA	.682	.695	.702
	RGB-HHA	.746	.761	.786
RCF [30]	RGB	.729	.742	—
	HHA	.705	.715	—
	RGB-HHA	.757	.771	—
AMH-Net-ResNet50 [51]	RGB	.744	.758	.765
	HHA	.716	.729	.734
	RGB-HHA	.771	.786	.802
LPCB [9]	RGB	.739	.754	—
	HHA	.707	.719	—
	RGB-HHA	.762	.778	—
COB-ResNet50[33]	RGB-HHA	.784	.805	.825
BDCN	RGB	.748	.763	.770
	HHA	.707	.719	.731
	RGB-HHA	.765	.781	.813

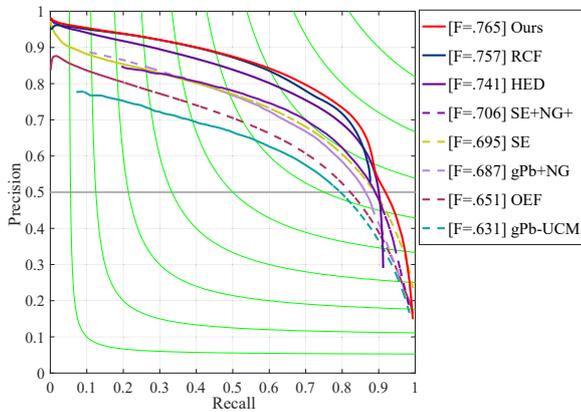


Figure 6. The precision-recall curves of our method and compared works on *NYUDv2*.

lowing previous works [49, 30], we perform experiments on all of them. The results of RGB-HHA are obtained by averaging the edges detected on RGB and HHA. Table 4 shows the comparison of our method with several recent approaches, including gPb-ucm [1], OEF [17], gPb+NG [15], SE+NG+ [16], SE [10], HED [49], RCF [30] and AMH-Net [51]. Fig. 6 shows the precision-recall curves of our method and other competitors. All of the evaluation results are based on a single scale input.

As shown in Table 4 and Fig. 6, our performance is competitive, *i.e.*, outperforms most of the compared works except AMH-Net [51]. Note that, AMH-Net applies the deeper ResNet50 to construct the edge detector. With a shallower network, our method still outperforms AMH-Net on the RGB image, *i.e.*, our 0.748 vs. 0.744 of AMH-Net in F-measure ODS. Compared with previous works, our improvement over existing works is actually more substantial, *e.g.*, on *NYUDv2* our gains over RCF [30] and HED [49] are 0.019 and 0.028 in ODS, higher than the 0.009 gain of

Table 5. Comparison with recent works on *Multicue*. ‡ indicates the fused result of multi-scale images.

Cat.	Method	ODS	OIS	AP
Boundary	Human [35]	.760 (0.017)	—	—
	Multicue [35]	.720 (0.014)	—	—
	HED [50]	.814 (0.011)	.822 (0.008)	.869(0.015)
	RCF [30]	.817 (0.004)	.825 (0.005)	—
	RCF [30] ‡	.825 (0.008)	.836 (0.007)	—
	BDCN	.836 (0.001)	.846(0.003)	.893(0.001)
	BDCN ‡	.838 (0.004)	.853 (0.009)	.906 (0.005)
Edge	Human [35]	.750 (0.024)	—	—
	Multicue [35]	.830 (0.002)	—	—
	HED [50]	.851 (0.014)	.864 (0.011)	—
	RCF [30]	.857 (0.004)	.862 (0.004)	—
	RCF [30] ‡	.860 (0.005)	.864 (0.004)	—
	BDCN	.891 (0.001)	.898 (0.002)	.935(0.002)
	BDCN ‡	.894 (0.002)	.901 (0.004)	.941 (0.005)

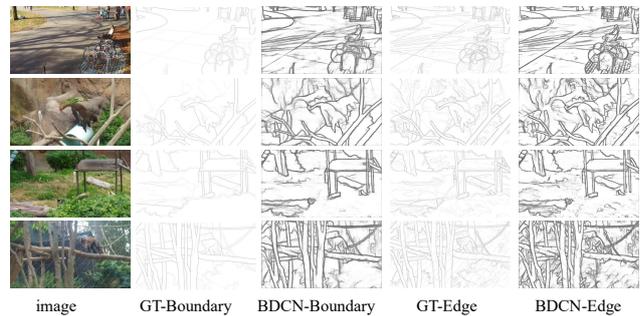


Figure 7. Examples of our edge detection results before Non-Maximum Suppression on *Multicue* dataset.

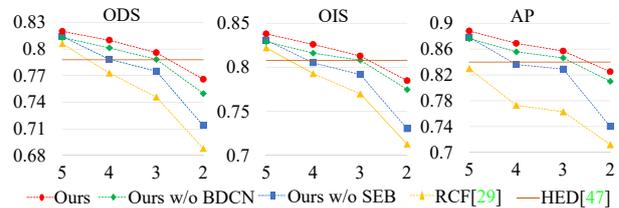


Figure 8. Comparison of edge detection accuracy as we decrease the number of ID Blocks from 5 to 2. HED learned with VGG16 is denoted as the solid line for comparison.

RCF [30] over HED [49].

Performance on Multicue: *Multicue* consists of two sub datasets, *i.e.*, *Multicue boundary* and *Multicue edge*. As done in RCF [30] and the recent version of HED [50], we average the scores of three independent experiments as the final result. We show the comparison with recent works in Table 5, where our method achieves substantially higher performance than RCF [30] and HED [49]. For boundary detection task, we outperform RCF and HED by 1.3% and 2.4%, respectively in F-measure ODS. For edge detection task, our performance is 3.4% and 4.3% higher than the ones of RCF and HED. Moreover, the performance fluctuation of our method is considerably smaller than those two methods, which means our method delivers more stable results. Some edge detection results generated by our approach on *Multicue* are presented in Fig. 7.

Discussions: The above experiments have shown the

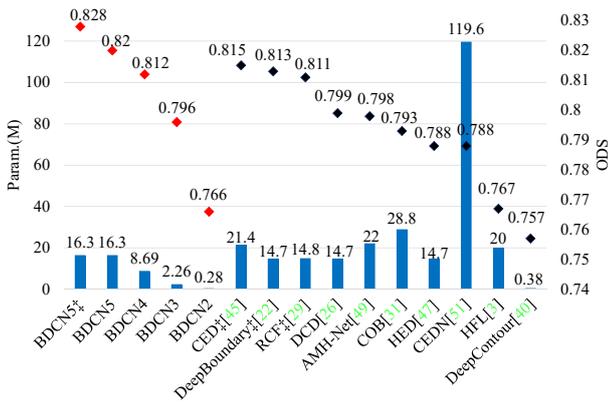


Figure 9. Comparison of parameters and performance with other methods. The number behind “BDCN” indicates the number of ID Block. †means the multiscale results.

competitive performance of our proposed method. We further test the capability of our method in learning multi-scale representations with shallow networks. We test our approach and RCF with different depth of networks, *i.e.*, using different numbers of convolutional block to construct the edge detection model. Fig. 8 presents the results on *BSDS500*. As shown in Fig. 8, the performance of RCF [30] drops more substantially than our method as we decrease the depth of networks. This verifies that our approach is more effective in detecting edges with shallow networks. We also show the performance of our approach without the SEM and the BDCN structure. These ablations show that removing either BDCN or SEM degrades the performance. It is also interesting to observe that, without SEM, the performance of our method drops substantially. This hence verifies the importance of SEM to multi-scale representation learning in shallow networks.

Fig. 9 further shows the comparison of parameters *vs.* performance of our method with other deep net based methods on *BSDS500*. With 5 convolutional blocks in VGG16, HED [49], RCF [30], and our method use similar number of parameters, *i.e.*, about 16M. As we decrease the number of ID Blocks from 5 to 2, our number of parameters decreases dramatically, drops to 8.69M, 2.26M, and 0.28M, respectively. Our method still achieves F-measure ODS of 0.766 using only two ID Blocks with 0.28M parameters. It also outperforms HED and RCF with a more shallow network, *i.e.*, with 3 and 4 ID Blocks respectively. For example, it outperforms HED by 0.8% with 3 ID Blocks and just 1/6 parameters of HED. We thus conclude that, our method can achieve promising edge detection accuracy even with a compact shallow network.

To further show the advantage of our method, we evaluate the performance of edge predictions by different intermediate layers, and show the comparison with HED [49] and RCF [30] in Table 6. It can be observed that, the intermediate predictions of our network also consistently out-

Table 6. The performance (ODS) of each layer in BDCN, RCF [30], and HED [49] on *BSDS500* test set.

Layer ID.	HED [49]	RCF [30]	BDCN
1	0.595	0.595	0.727
2	0.697	0.710	0.762
3	0.750	0.766	0.771
4	0.748	0.761	0.802
5	0.637	0.758	0.815
fuse	0.790	0.805	0.820

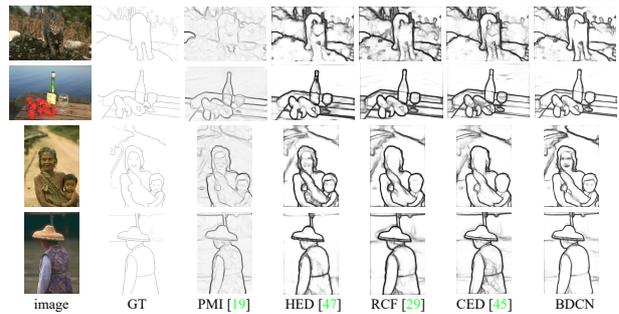


Figure 10. Comparison of edge detection results on *BSDS500* test set. All the results are raw edge maps computed with a single scale input before Non-Maximum Suppression.

perform the ones from HED and RCF, respectively. With 5 ID Blocks, our method runs at about 22fps for edge detection, on par with most DCNN-based methods. With 4, 3 and 2 ID Blocks, it accelerates to 29 fps, 33fps, and 37fps, respectively. Fig. 10 compares some edge detection results generated by our approach and several recent ones.

5. Conclusions

This paper proposes a Bi-Directional Cascade Network for edge detection. By introducing a bi-directional cascade structure to enforce each layer to focus on a specific scale, BDCN trains each network layer with a layer-specific supervision. To enrich the multi-scale representations learned with a shallow network, we further introduce a Scale Enhancement Module (SEM). Our method compares favorably with over 10 edge detection methods on three datasets, achieving ODS F-measure of 0.828, 1.3% higher than current state-of-art on *BSDS500*. Our experiments also show that learning scale dedicated layers results in compact networks with a fraction of parameters, *e.g.*, our approach outperforms HED [49] with only 1/6 of its parameters.

6. Acknowledgement

This work is supported in part by Beijing Natural Science Foundation under Grant No. JQ18012, in part by Natural Science Foundation of China under Grant No. 61620106009, 61572050, 91538111. We also thank NVIDIA’s generosity for providing DGX-1 super-computer and support through the NVAAIL program.

References

- [1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, 2011. [1](#), [3](#), [5](#), [7](#)
- [2] G. Bertasius, J. Shi, and L. Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *CVPR*, pages 4380–4389, 2015. [1](#), [2](#), [6](#)
- [3] G. Bertasius, J. Shi, and L. Torresani. High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In *ICCV*, 2015. [1](#), [2](#), [6](#)
- [4] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, June 1986. [1](#), [2](#)
- [5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. [2](#), [4](#)
- [6] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, pages 3640–3649, 2016. [2](#)
- [7] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002. [1](#)
- [8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. [5](#)
- [9] R. Deng, C. Shen, S. Liu, H. Wang, and X. Liu. Learning to predict crisp boundaries. In *ECCV*, pages 562–578, 2018. [6](#), [7](#)
- [10] P. Dollár and C. L. Zitnick. Fast edge detection using structured forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1558–1570, 2015. [1](#), [2](#), [7](#)
- [11] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015. [2](#)
- [12] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1915–1929, 2013. [2](#)
- [13] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(1):36–51, 2008. [1](#)
- [14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. [1](#)
- [15] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, 2013. [5](#), [7](#)
- [16] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, pages 345–360. Springer, 2014. [7](#)
- [17] S. Hallman and C. C. Fowlkes. Oriented edge forests for boundary detection. In *CVPR*, pages 1732–1740, 2015. [6](#), [7](#)
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [1](#)
- [19] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger. Multi-scale dense convolutional networks for efficient prediction. *arXiv preprint arXiv:1703.09844*, 2017. [2](#)
- [20] P. Isola, D. Zoran, D. Krishnan, and E. H. Adelson. Crisp boundary detection using pointwise mutual information. In *ECCV*. Springer, 2014. [6](#)
- [21] W. Ke, J. Chen, J. Jiao, G. Zhao, and Q. Ye. Srn: Side-output residual network for object symmetry detection in the wild. In *CVPR*, pages 1068–1076, 2017. [3](#)
- [22] J. Kittler. On the accuracy of the sobel edge detector. *Image and Vision Computing*, 1(1):37–42, 1983. [1](#), [2](#)
- [23] I. Kokkinos. Pushing the boundaries of boundary detection using deep learning. *arXiv preprint arXiv:1511.07386*, 2015. [1](#), [6](#)
- [24] S. Konishi, A. L. Yuille, J. M. Coughlan, and S. C. Zhu. Statistical edge detection: Learning and evaluating edge cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(1):57–74, 2003. [2](#)
- [25] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *CVPR*, pages 5325–5334, 2015. [3](#)
- [26] X. Li, Z. Liu, P. Luo, C. Change Loy, and X. Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *CVPR*, 2017. [3](#)
- [27] Y. Liao, S. Fu, X. Lu, C. Zhang, and Z. Tang. Deep-learning-based object-level contour detection with ccg and crf optimization. In *ICME*, 2017. [6](#)
- [28] J. J. Lim, C. L. Zitnick, and P. Dollár. Sketch tokens: A learned mid-level representation for contour and object detection. In *CVPR*, 2013. [1](#)
- [29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017. [3](#)
- [30] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai. Richer convolutional features for edge detection. In *CVPR*, 2017. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [31] Y. Liu and M. S. Lew. Learning relaxed deep supervision for better edge detection. In *CVPR*, 2016. [1](#)
- [32] K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool. Convolutional oriented boundaries. In *ECCV*, 2016. [6](#)
- [33] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool. Convolutional oriented boundaries: From image segmentation to high-level tasks. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 40(4):819–833, 2018. [7](#)
- [34] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(5):530–549, 2004. [2](#)
- [35] D. A. Mély, J. Kim, M. McGill, Y. Guo, and T. Serre. A systematic comparison between visual cues for boundary detection. *Vision research*, 120:93–107, 2016. [5](#), [7](#)
- [36] 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*. [5](#)

- [37] V. N. Murthy, V. Singh, T. Chen, R. Manmatha, and D. Comaniciu. Deep decision network for multi-class image classification. In *CVPR*, pages 2240–2248. IEEE, 2016. 3
- [38] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *ICML*, pages 82–90, 2014. 2
- [39] X. Ren. Multi-scale improves boundary detection in natural images. In *ECCV*, pages 533–545, 2008. 2
- [40] X. Ren and L. Bo. Discriminatively trained sparse code gradients for contour detection. In *NIPS*, 2012. 5, 6
- [41] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *IEEE Trans. Graph.*, volume 23, pages 309–314. ACM, 2004. 1
- [42] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. Deep-contour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *CVPR*, 2015. 1, 2, 6
- [43] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. 5
- [44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 5
- [45] V. Torre and T. A. Poggio. On edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, (2):147–163, 1986. 2
- [46] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 3
- [47] Y. Wang, X. Zhao, and K. Huang. Deep crisp boundaries. In *CVPR*, 2017. 1, 2, 5, 6
- [48] A. P. Witkin. Scale-space filtering. In *Readings in Computer Vision*, pages 329–332. Elsevier, 1987. 2, 3
- [49] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015. 1, 2, 5, 6, 7, 8
- [50] S. Xie and Z. Tu. Holistically-nested edge detection. *International journal of computer vision*, 2017. 1, 2, 5, 7
- [51] D. Xu, W. Ouyang, X. Alameda-Pineda, E. Ricci, X. Wang, and N. Sebe. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In *NIPS*, pages 3964–3973, 2017. 2, 5, 6, 7
- [52] S. Yan, J. S. Smith, W. Lu, and B. Zhang. Hierarchical multi-scale attention networks for action recognition. *Signal Processing: Image Communication*, 61:73–84, 2018. 2
- [53] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang. Object contour detection with a fully convolutional encoder-decoder network. In *CVPR*, pages 193–202, 2016. 6
- [54] Y. Yuan, K. Yang, and C. Zhang. Hard-aware deeply cascaded embedding. *CoRR*, abs/1611.05720, 1, 2016. 3
- [55] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 2