

# Unmixing Convolutional Features for Crisp Edge Detection

Linxi Huan, Nan Xue, Xianwei Zheng\*, Wei He, Jianya Gong, Gui-Song Xia

**Abstract**—This paper presents a context-aware tracing strategy (CATS) for crisp edge detection with deep edge detectors, based on an observation that the localization ambiguity of deep edge detectors is mainly caused by the mixing phenomenon of convolutional neural networks: *feature mixing* in edge classification and *side mixing* during fusing side predictions. The CATS consists of two modules: a novel *tracing loss* that performs feature unmixing by tracing boundaries for better side edge learning, and a *context-aware fusion block* that tackles the side mixing by aggregating the complementary merits of learned side edges. Experiments demonstrate that the proposed CATS can be integrated into modern deep edge detectors to improve localization accuracy. With the vanilla VGG16 backbone, in terms of BSDS500 dataset, our CATS improves the F-measure (ODS) of the RCF and BDCN deep edge detectors by 12% and 6% respectively when evaluating without using the morphological non-maximal suppression scheme for edge detection.

**Index Terms**—Edge detection, deep learning, convolutional feature unmixing

## 1 INTRODUCTION

As a fundamental computer vision task of localizing boundaries of perceptually salient objects in natural images, edge detection has been studied with a long history [1]. In the early stage, the low-level features were extensively studied for detecting edges [2], [3], [4], [5]. These approaches can obtain crisp edge maps but leave a challenging problem of suppressing high-frequency texture regions. Later, learning techniques were introduced to classify image patches as edge and non-edge classes [6], [7]. End-to-end deep edge detectors [8], [9], [10] have been recently proposed to learn multi-level side edges with deep supervision, and weight the side predictions according to the learned side importance to obtain the final predictions.

Deep learning solutions dramatically improved the performance of edge detection as the classification on hierarchical deep features generated by large receptive fields can robustly suppress false alarms in texture regions [8], [9], [10], [11], [12], [13]. These methods, however, have a challenging issue of detecting crisp edge maps that are free of localization ambiguity. As shown in Fig. 1, both the network outputs of HED [8] and RCF [9] suffer from localization ambiguity in the regions that contain true positive edge pixels across convolutional stages. To obtain crisp edge maps, it is required to use a morphological non-maximal suppression (NMS) scheme for the network outputs [6], [8], [9], [10], [11], [14], [15].

There have been some works that focused on the crispness of deep edge detection [16], [17] without using the morphological NMS scheme. Wang *et al.* [16] proposed to

use a new refinement architecture to tackle the problem of crispness. In the same period, Deng *et al.* [17] found that the commonly-used weighted cross entropy loss is a key factor of posing the issue of localization ambiguity and adopted a Dice loss [18] for edge detection. Limited by the numerical stability issue, the Dice loss tends to yield weaker edge response than the weighted cross entropy, which is disadvantageous for distinguishing edge and non-edge points, and [17] thus has to add the ResNeXt block [19] to obtain better features than the commonly-used vanilla VGG16 [20] network.

In this paper, we focus on the crispness of deep edge detection. That is to say, we expect the raw edge predictions by the convolution layers are with less localization ambiguity around the true positive predictions while maintaining strong capability of texture suppression. This is important for separating the adjacent edges that are glued together, as such glued edges in the raw predictions can be hard for the post-processing operation to handle. In contrast to the previous works [16], [17], we are going to study if it is possible to learn crisp edge maps with the commonly-used backbone architecture VGG16 [20] for edge detection.

We found that the issue of localization ambiguity for deep edge detectors is mainly caused by the mixing phenomenon in convolutional neural networks (CNNs).

- **Feature Mixing.** Limited by the existence of max-pooling layers (for larger receptive field) and up-sampling operators (for pixel-wise prediction), the higher-level convolutional features are spatially mixed. When training a network to obtain a (higher-level) side edge map at a convolutional stage using the commonly-used weighted cross entropy loss or its variants, it is challenging to correctly classify multiple pixels sourced with the severe class-imbalance issue from the same mixed feature vectors. As a result, the higher convolutional stage is, the blurrier side edge maps are obtained by existing deep edge detectors .
- **Side Mixing.** When the side outputs are obtained, it is

---

- L. Huan and X. Zheng are with the LIESMARS, Wuhan University, Wuhan 430079, China. E-mail: {whu\_hlx, zhengxw}@whu.edu.cn.
- N. Xue and G.-S. Xia are with the School of Computer Science, Wuhan University, Wuhan 430079, China.
- W. He is with the RIKEN Center, Tokyo 1030027, Japan.
- J. Gong is with the LIESMARS and the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China.
- Linxi Huan and Nan Xue contributed equally to this work.
- Corresponding author: X. Zheng.

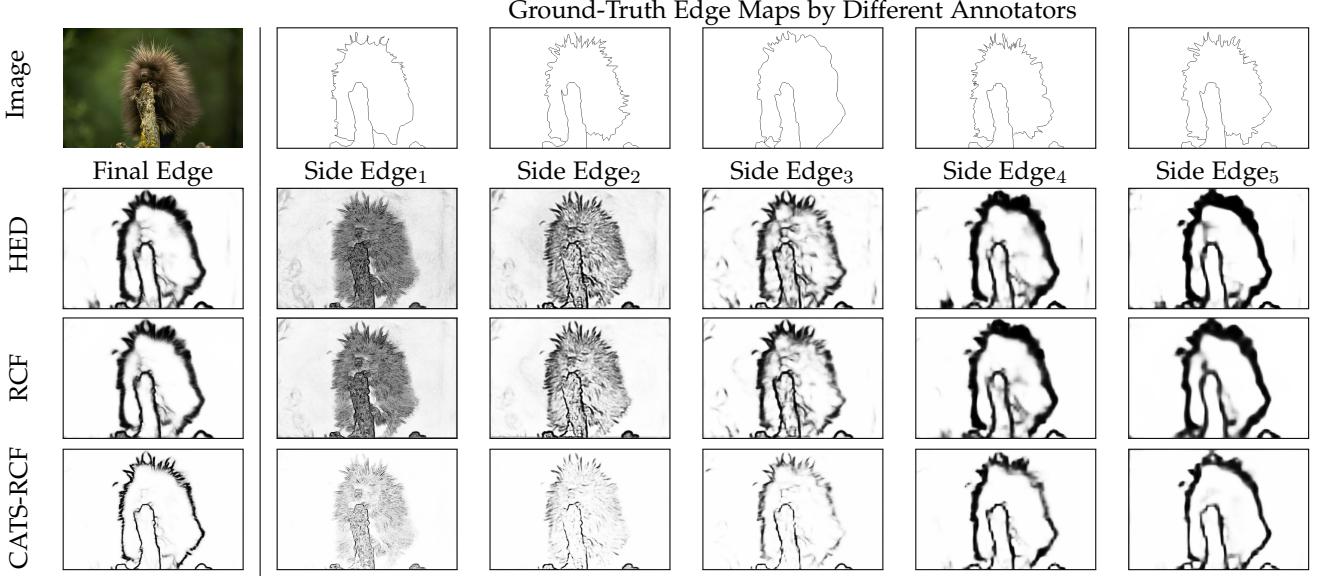


Fig. 1. Qualitative comparison between the prior arts of deep edge detection (HED [8] and RCF [9]) and our proposed CATS. The top row displays an example image and its edge annotations from the BSDS500 dataset [21]. The final edge predictions and the multi-level side edge maps estimated by different edge detection approaches are listed in rows. In the bottom line, we leverage our proposed CATS on RCF [9] to obtain the most precisely-located side edge maps and final edge prediction.

required to fuse them into a final edge prediction. The phenomenon of *side mixing* is easily to be observed in existing detectors as they simply weight the side edge predictions according to the learned side importance. On the one hand, because the final edge prediction should be able to suppress the complicated textures in an image, the high-level side edges will have greater importance than low-level ones. On the other hand, such a fusion strategy equally treats all pixels in the same side output, thus making it hard to selectively preserve the complementary merits of different side edges. As shown in Fig. 1, the final edge predictions of both HED and RCF are dominated by the high-level side edge maps while ignoring the low-level ones that have less localization ambiguities.

Based on the above discussion, we are going to unmix the convolutional features to learn crisp edge maps. First, we address the *feature mixing* by tracing the true positive boundaries during training. We term the false positives caused by the *feature mixing* as the *confusing pixels* (of edges). To this end, we present a novel tracing loss with a specific focus on confusing pixels. We use a boundary tracing function to trace the actual position of an edge by enlarging the response difference between an edge pixel and the neighboring confusing pixels. Once the traced pixels are delineated, all the remained pixels belong to the background category and we only need to suppress the textures. Different from the existing loss functions for edge detection, we propose a texture suppression function to robustly handle texture regions by holistically suppressing the non-edge pixels that lie in the same image patch rather than treating these pixels as independent individuals. For further alleviation of localization ambiguity, a pixel-wise fusion mechanism is required to handle the side mixing caused by image-level weighted average. We present a simple context-aware fusion (CoFusion) block that dynamically learns filters [22] with contextual information across side predictions under the

general self-attention framework [23], [24], [25]. A similar design of leveraging dynamic filters for edge perception task was previously presented in DFF [26], it, however, focuses on fusing side predictions with high-level information for semantic edge detection. By contrast, our CoFusion block aims at addressing side mixing with the clues provided by the side edges themselves. The CoFusion block handles the side mixing that is hard for the tracing loss to alleviate with image-level weighted average; while the tracing loss offers necessary guidance to exert the CoFusion block.

The tracing loss and the CoFusion block together form a context-aware tracing strategy (CATS) that is specifically designed for crisp edge detection. As displayed at the bottom of Fig. 1, with the proposed tracing function available, the side edge maps are reliably learned with less localization ambiguity, and the CoFusion fusion block makes the final edge predictions better in the aspect of crispness.

In experiments, we leverage our proposed CATS on the existing state-of-the-art deep edge detectors HED [8], RCF [9] and BDCN [10] to demonstrate that the CATS can fully exploit the multi-level side outputs in terms of supervision and fusion to obtain precisely-located edge predictions. In the standard evaluation of edge detectors, our proposed CATS consistently improves the accuracy of HED, RCF and BDCN on three common edge detection datasets. We achieve the state-of-the-art performance with a F-measure (ODS) of 0.812, 0.752, 0.897 on BSDS500 [21], NYUDv2 [27] and MultiCue [28] for RGB data, respectively.

Furthermore, we quantitatively evaluate the aforementioned edge detectors with CATS on BSDS500 and NYUDv2 datasets without using the standard morphological non-maximal suppression scheme. The results validate that the CATS improves the F-measure (ODS) performance by large margins on BSDS500 and NYUDv2 with at most 12% and 7.6% respectively with robust edge localization ability.

Our main contributions are summarized as follow: (1) We explicitly treat the false positive pixels as the confusing

pixels, and propose a tracing loss to address the localization ambiguity for deep edge detectors. (2) We present a context-aware tracing strategy to learn precisely-located edge maps from natural images in an end-to-end manner. (3) The proposed CATS consistently improves the performance of HED, RCF and BDCN on the BSDS500, the NYUDv2 and the Multicue datasets, and dramatically improves the edge localization accuracy of these detectors with no post-processing conditions.

## 2 METHODOLOGY

### 2.1 Overview

In this section, we elaborate the context-aware tracing strategy (CATS) that alleviates edge localization ambiguity by a *tracing loss* and a *context-aware fusion (CoFusion)* block for crisp edge learning. For feature unmixing with the weighted cross entropy, the tracing loss introduces a boundary tracing function to specifically separate the confusing pixels from edges, and a texture suppression function to perform holistic texture-region smoothing. Supervised by the tracing loss, the CoFusion block learns to selectively aggregate the different delineation merits of side edges in a pixel-wise fashion, which tackles the side mixing during the fusion.

### 2.2 The tracing loss

#### 2.2.1 Weighted cross entropy

Given an edge prediction  $\hat{Y} = \{\hat{y}_i\}_{i=1}^N$  and the corresponding edge label  $Y = \{y_i\}_{i=1}^N$ , the weighted cross entropy is formulated as

$$L_{ce}(\hat{Y}, Y) = -\lambda \alpha \sum_{i \in Y^+} \log \hat{y}_i - (1 - \alpha) \sum_{i \in Y^-} \log (1 - \hat{y}_i), \quad (1)$$

where  $Y^+ = \{i | y_i \in Y, y_i > \delta\}$ ,  $Y^- = \{i | y_i \in Y, y_i = 0\}$  denote the edge and non-edge sample sets respectively, and  $\alpha$  is the proportion of the negative pixels in the set  $Y^+ \cup Y^-$ . The hyper-parameter  $\lambda$  serves for importance balancing between edge and non-edge samples, and  $\delta$  is a threshold to remove semantically controversial pixels when there are several annotators [9], [10].

The weighted cross entropy can effectively supervise the network to learn reasonable edge maps, but it has highly imbalanced attention between edge and non-edge samples, making it hard to distinguish the confusing pixels that share features with edges and consistently smooth high-frequency regions. Consequently, as illustrated in Fig. 2, the detected edge maps usually contain misclassified confusing pixels that incur thick edges around true positive edge pixels, and get false positives that form dark shadows in texture regions.

#### 2.2.2 Feature unmixing by tracing boundaries

Different from the non-edge pixels that lie beyond edges in a complete texture regions, the confusing pixels are intertwined with edges by the shared features. For a clear separation from edges, the confusing pixels therefore require an edge-aware suppression that is capable of feature unmixing in the high-level side edges, rather than being simply smoothed as texture points.

Based on the above discussion, we devised a boundary tracing function to weaken the interaction between edges

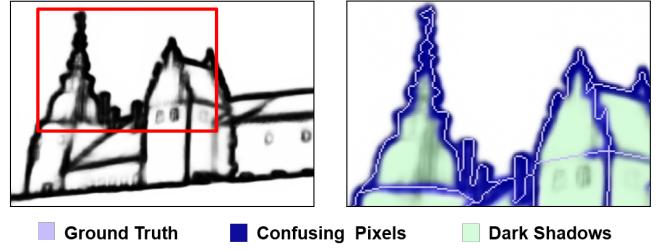


Fig. 2. An illustration of confusing pixels. The left image displays edges predicted by RCF [9] and the right image is the enlarged rectangle region (red box) in the left image. Pixels in different colors in the right image indicate the ground truth edge pixels (purple), the confusing pixels of edges (blue), and the dark shadows in texture regions (light green).

and their confusing neighbors by enlarging the response difference for crisp edge delineation. The boundary tracing function is formulated as

$$L_{bdry}(\hat{Y}, Y) = - \sum_{p \in E} \log \left( \sum_{i \in L_p} \hat{y}_i / \left( \sum_{i \in R_p^e \setminus L_p} \hat{y}_i + \sum_{i \in L_p} \hat{y}_i \right) \right), \quad (2)$$

In Eq. (2),  $E$  is the set of all the edge points in the edge label  $Y$ .  $R_p^e$  denotes a small image patch (e.g., a  $7 \times 7$  rectangle patch) that contains edge fragments, and the patch centers at an edge point  $p$ . The set of the edge points in  $R_p^e$  is represented as  $L_p$ .

Minimizing  $L_{bdry}(\hat{Y}, Y)$  will force  $\sum_{i \in R_p^e \setminus L_p} \hat{y}_i$  to be 0 while increasing  $\sum_{i \in L_p} \hat{y}_i$ , which enlarges the response difference between edge fragments and accompanying confusing pixels. The boundary tracing function  $L_{bdry}$  therefore smooths confusing pixels with awareness to the predicted response strength of neighboring edge fragments, and traces the actual edge position with less of the ambiguity caused by confusing pixels. In practice,  $L_{bdry}$  is computed via three convolution layers that are applied on the prediction and the label. The time complexity is about  $\mathcal{O}((2 + 3k^2) \times M \times N)$  with an  $M \times N$  image and a  $k \times k$  patch size.

With edges and confusing pixels handled by the boundary tracing function, the remained texture regions can be consistently suppressed by a texture suppression function defined as

$$L_{tex}(\hat{Y}, Y) = - \sum_{p \in Y \setminus \hat{E}} \log(1 - \sum_{i \in R_p^t} \hat{y}_i / |R_p^t|), \quad (3)$$

where  $R_p^t$  represents an image patch that centers at a non-edge point  $p$  (e.g., a  $3 \times 3$  rectangle area), and  $\hat{E}$  is a set that includes all edges and their confusing pixels that have been used in the boundary tracing function.  $\hat{E}$  serves as a buffer zone to weaken the negative interaction between edge fragments and texture areas that require a stronger suppression than confusing pixels.

As revealed by Eq. (3), the texture suppression function groups the non-edge points in the same texture region patch, and holistically suppresses these pixels rather than treating them as independent individuals. Focusing on complete texture regions beyond edges, the texture suppression function actually works in a complementary way with the boundary tracing function.

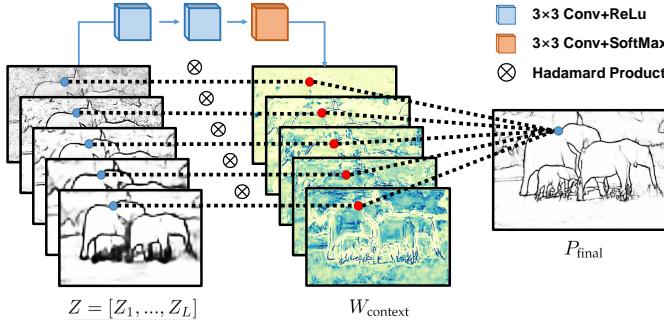


Fig. 3. The mechanism of the CoFusion Block.

With the boundary tracing function and the texture suppression function, the tracing loss follows as

$$\text{TracingLoss}(\hat{Y}, Y) = L_{\text{ce}} + \lambda_1 L_{\text{bdry}} + \lambda_2 L_{\text{tex}}, \quad (4)$$

where  $\hat{Y}$  and  $Y$  respectively denote the edge prediction and the edge label,  $\lambda_1$  and  $\lambda_2$  are hyper-parameters to balance each element in the tracing loss.

During model training,  $L_{\text{ce}}$  performs a rough edge learning,  $L_{\text{bdry}}$  tackles edge localization refinement by feature unmixing, and  $L_{\text{tex}}$  imposes a strong holistic suppression on texture regions. With  $L_{\text{bdry}}$  and  $L_{\text{tex}}$ , the tracing loss processes the non-edge points that are gathered according to their surrounding context with target-specific suppression, and achieves crisp edge generation with less localization ambiguity than single weighted cross entropy.

### 2.3 Context-aware fusion block

To leverage the edge details in low-level edge maps and the global context with robust texture suppression in the high-level counterparts, weighted average is a common operation adopted in previous works for obtaining a unified edge prediction from multi-level side edges [8], [9], [10]. Albeit effective, this average operation incurs the issue of side mixing as all the pixels in a side edge map share the same fixed weight and have equal importance during fusion.

For harnessing the multi-level side maps without side mixing, as shown in Fig. 3, we designed a context-aware fusion (CoFusion) block within a self-attention framework to absorb the merits and bypass the limitations of different side maps by a position-wise aggregation of side edges.

Let  $Z = [Z_1, \dots, Z_L] \in \mathbb{R}^{H \times W \times L}$  denote  $L$  side edge heatmaps. The CoFusion block first learns a weight map  $W_{\text{context}} \in \mathbb{R}^{H \times W \times L}$  from  $Z$  by an attention block. To derive a score map  $A_{\text{score}} = [a_{ijl}] \in \mathbb{R}^{H \times W \times L}$  from  $Z$ , the attention block adopts three  $3 \times 3$  convolution layers to capture contextual information for inferring  $W_{\text{context}}$ .  $A_{\text{score}}$  is subsequently normalized through softmax activation to get a weight map  $W_{\text{context}} = [w_{ijl}]$ , where

$$w_{ijl} = e^{a_{ijl}} / \sum_{k=1}^L e^{a_{ijk}}. \quad (5)$$

A vector  $v_{ij} \in \mathbb{R}^{1 \times 1 \times L}$  in  $W_{\text{context}}$  determines how much each side map contributes to the pixel  $p_{ij}$  of the final edge prediction  $P_{\text{final}}$ . Because  $v_{ij}$  is calculated with the features around  $p_{ij}$  across side maps,  $W_{\text{context}}$  changes according to the context information in the multi-level side maps and

leverages their different delineation preferences for the final prediction. With the context-aware weights  $W_{\text{context}}$ ,  $P_{\text{final}}$  can be calculated as follows.

$$\text{CoFusion}(Z) = \sum_{l=1}^L W_{\text{context}} \otimes Z; \quad (6)$$

$$P_{\text{final}} = \text{sigmoid}(\text{CoFusion}(Z)), \quad (7)$$

where  $\otimes$  denotes Hadamard product.

## 3 EXPERIMENTS AND ANALYSIS

To validate the effectiveness and generality of the CATS, we implement it in Pytorch framework [29] with three VGG16-based [20] edge detectors, including HED [8], RCF [9] and BDCN [10]. In this section, we evaluate the CATS on three challenging edge detection benchmarks, including NYUDv2 [27], BSDS500 [21] and Multicue [28] datasets. An ablation study is subsequently presented for a more comprehensive performance analysis of the CATS.

### 3.1 Datasets

**BSDS500** [21]. This dataset is a challenging edge detection benchmark that is composed of 200 training, 100 validation and 200 test images. Each image in this dataset is annotated by several annotators. The training and validation data are jointly augmented for model training as in [8] and [9].

**NYUDv2** [27]. This is a challenging dataset for indoor scene parsing and is also a commonly used benchmark for edge detection evaluation [8], [9], [10], [17]. It contains 1449 densely annotated RGB-D images, and is divided into 381 training, 414 validation and 654 testing images. We use this dataset to compare our proposed CATS with the state-of-the-art edge detectors, and also perform a comprehensive analysis on the NYUDv2 dataset in the ablation study.

**Multicue dataset** [28]. This dataset strictly distinguishes the definitions of boundary and edge, and it thus consists of two sub datasets: Multicue Boundary and Multicue Edge. This dataset regards semantically meaningful pixels as object boundary points, and treats pixels with abrupt perceptual changes as low-level edge points. We testify the CATS on both sub-datasets. Same with [8], [9], [10], [28], we randomly split the Multicue dataset into 80 training and 20 testing samples, and conducted three independent trials on both sub-datasets.

### 3.2 Implementation details

Experiments in this paper are conducted on a single GeForce RTX 2080 Ti GPU. The model parameters are updated by stochastic gradient descent (SGD) optimizer with momentum 0.9 and weight decay 0.0002, and the batch size is set to 10. The initial learning rate is 1e-6 and is multiplied with 0.1 after every given period shown in Tab. 1. The hyper-parameters in the tracing loss are also specified in Tab. 1.

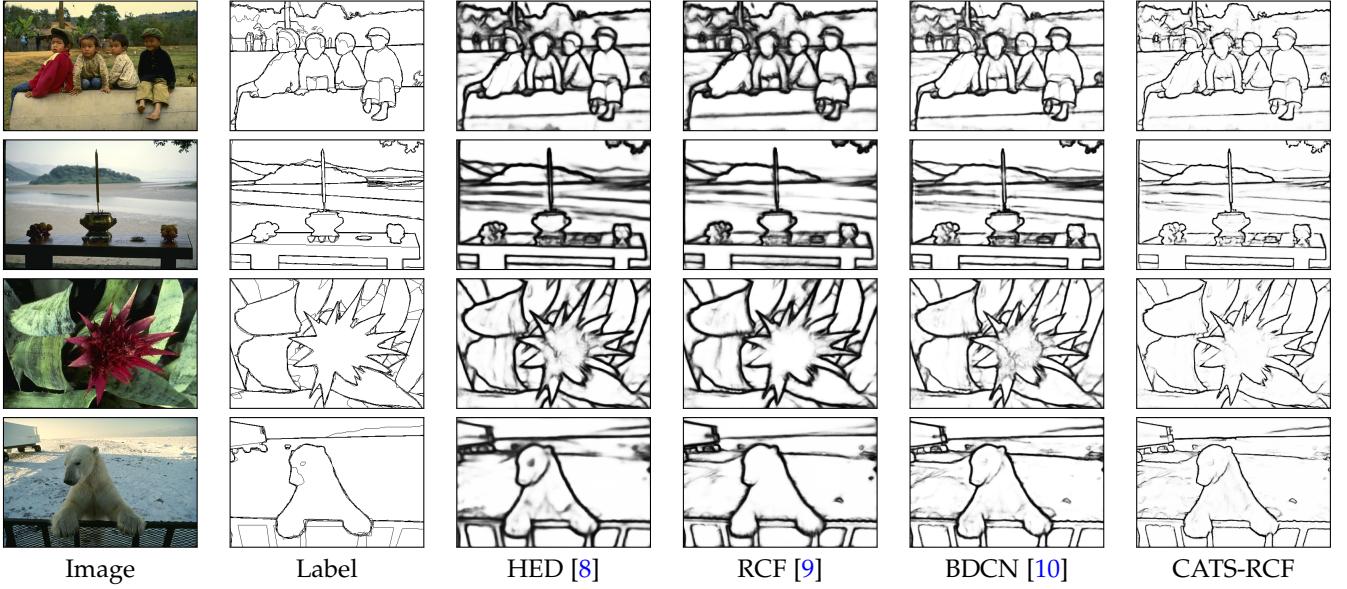


Fig. 4. Qualitative comparison between the prior arts of deep edge detection and our proposed CATS on the BSDS500 dataset [21].

TABLE 1  
Parameter setting of the tracing loss in experiments.

Param. Data	period/epoch	$\delta$	$\lambda$	$\lambda_1/\lambda_2$		
				L1-L3	L4-L5	Final
NYUDv2	20/60	-	1.2	4/0.05	2/0.1	6/0.05
BSDS500	10/40	0.3	1.1	2/0.05	1/0.1	4/0.05
Multicue_bdr	20/60	0.3	1.2	2/0.05	1/0.1	4/0.03
Multicue_edge	20/60	0.2	1.1	4/0.01	2/0.01	6/0.01

### 3.3 Evaluation protocols

The core step of evaluating an edge detector is to pixel-wisely match the ground truth of edges to the binarized edge prediction under a specified maximum allowed distance tolerance. Following the previous works, a parametric curve of the precision and recall will be drawn to evaluate the overall performance of an edge detector with different thresholds of binarization. The best performance of an edge detector will be reported by the F-measure scores with the optimal thresholds at both dataset scale (ODS) and image scale (OIS). The maximum allowed distance tolerance for correct matches between the edge predictions and the annotations is conventionally set to 0.0075 for the BSDS500 and Multicue datasets, and 0.011 for the NYUDv2 dataset [8], [9], [10], [12].

**Standard evaluation protocol.** Before binarizing a given edge prediction, the general evaluation procedure in previous researches will first apply a standard post-processing scheme. The post-processing scheme includes a non-maximum suppression (NMS) step and a mathematical morphology operation to obtain a thinned edge map, and the standard evaluation protocol then uses the thinned edge prediction for matching with the ground truth of edge maps.

**Crispness-emphasized evaluation protocol.** Although the post-processing scheme can partially remove the falsely-alarm edge pixels around the true positive ones, it would be interesting to explore whether the raw outputs of the deep edge detectors have a better localization ability of edges. Accordingly, we remove the standard post-processing scheme to evaluate the performance of a trained edge detector in the aspect of crispness.

### 3.4 Comparison with state-of-the-art methods

In this section, we report the statistic comparison of our method with existing state-of-the-art models on BSDS500, NYUDv2 and Multicue datasets. We tested and validated the proposed CATS on the VGG16-based HED [8], RCF [9] and BDCN [10]. Models in experiments were fine-tuned with a VGG16 [20] model pre-trained on ImageNet [30].

#### 3.4.1 BSDS500 Dataset

The numerical results on BSDS500 dataset are listed in Tab. 2, showing that the application of CATS achieves consistent performance boosting with HED, RCF and BDCN, by at most 0.7% in both ODS and OIS under the standard evaluation protocol.

TABLE 2  
Quantitative analysis on the BSDS500 dataset. SEval denotes the standard evaluation, and CEval is the crispness-emphasized evaluation. For fair comparison, we only list the single-scale results generated by models trained with only BSDS500 data.

Methods	SEval		CEval	
	ODS	OIS	ODS	OIS
Human	0.803	0.803	-	-
OEF [31]	0.746	0.770	-	-
N <sup>4</sup> -Fields [15]	0.753	0.769	-	-
DeepContour [6]	0.757	0.776	-	-
HFL [32]	0.767	0.788	-	-
CEDN [11]	0.788	0.804	-	-
DeepBoundary [14]	0.789	0.811	-	-
COB [12]	0.793	0.820	-	-
CED [16]	0.794	0.811	0.642	0.656
AMH-Net [33]	0.798	0.829	-	-
DCD [34]	0.799	0.817	-	-
LPCB [17]	0.800	0.816	0.693	0.700
HED [8]	0.793	0.811	0.588	0.608
CATS-HED	<b>0.800</b>	<b>0.816</b>	<b>0.666</b>	<b>0.676</b>
RCF [9]	0.799	0.815	0.585	0.604
CATS-RCF	<b>0.805</b>	<b>0.822</b>	<b>0.705</b>	<b>0.716</b>
BDCN [10]	0.807	0.822	0.636	0.650
CATS-BDCN	<b>0.812</b>	<b>0.828</b>	<b>0.696</b>	<b>0.705</b>

In Tab. 2, the CATS-applied models gain over the original counterparts by a dramatic margin in terms of the ODS and OIS scores of crispness-emphasized evaluation. The significant improvement indicates that the CATS can

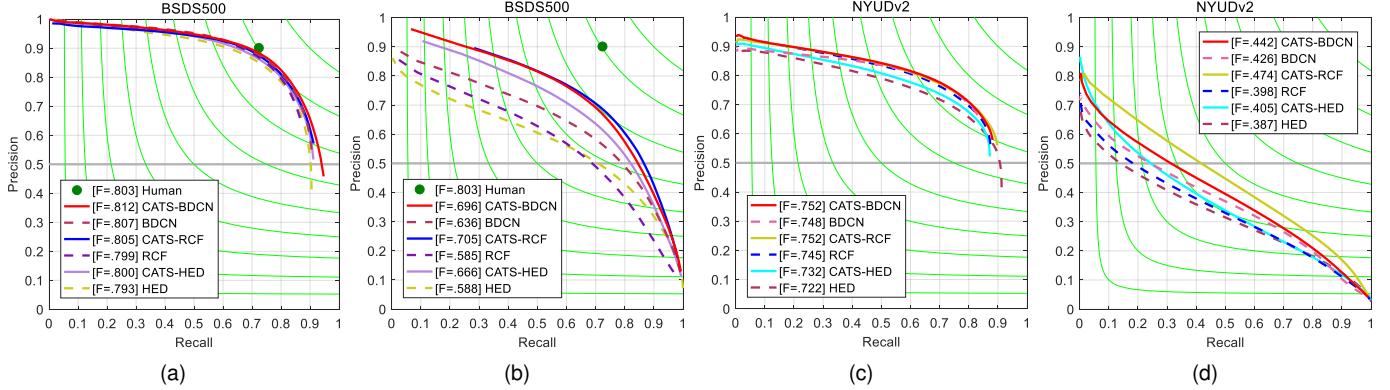


Fig. 5. The precision-recall curves of the CATS-based and the compared original models on BSDS500 and NYUDv2 datasets. (a) and (c) depict the results under the standard evaluation, while (b) and (d) are the results of the crispness-emphasized evaluation.

facilitate accurate edge localization with a robust removal of false positives along edges (*i.e.*, the confusing pixels). One may notice that the standard evaluation scores do not always synchronize with the crispness-emphasized evaluation counterparts, although they show simultaneous performance improvement when using the CATS on the same model. The post-processing operation will help mitigate the impact of the false positives around true positive edges, and the standard evaluation is therefore more influenced by the misclassified texture points than the confusing pixels. Despite of the different focuses of the two evaluation protocols, the results still indicate that the standard evaluation can benefit from less localization ambiguity.

Visualization results on BSDS500 dataset can be found in Fig. 4. In the given examples, the CATS successfully tackles the confusing pixels caused by the issue of feature mixing and side mixing during fusion to generate crisper edges than other networks, which in turn makes adjacent boundaries easier to distinguish. More detailed comparison is supplied in Fig. 5, which gives the precision-recall curves of the CATS-based models and the corresponding originals on BSDS500 regarding the two evaluation schemes.

### 3.4.2 NYUDv2 Dataset

We independently conducted experiments on the RGB and HHA data provided in the NYUDv2 dataset, and averaged the RGB- and HHA-based predictions to generate merged edge predictions like previous work [8], [9], [10]. The quantitative comparison with several existing methods is reported in Tab. 3.

Compared to the original HED, RCF and BDCN, their CATS versions all show steady accuracy growth regarding the standard evaluation, where the gain in ODS and OIS can respectively reach as much as 1.0% and 0.9% with respect to the RGB data. The crispness-emphasized evaluation also reveals the remarkable improvement in edge localization accuracy by the proposed CATS, which brings 7.6% and 7.5% progress at most in the ODS and OIS scores respectively for the RGB data. In Fig. 6, we present the qualitative results on the NYUDv2 dataset, where the proposed CATS generates crisper edges than other models, illustrating the capability of CATS for robust texture suppression in complex circumstances. For a comprehensive comparison, Fig. 5 plots the precision-recall curves of the CATS-based models and the original networks.

TABLE 3  
Quantitative analysis on the NYUDv2 dataset. SEval and CEval respectively refer to the standard evaluation and the crispness-emphasized evaluation.

Methods	Data	SEval		CEval	
		ODS	OIS	ODS	OIS
LPCB [17]	RGB	0.739	0.754	-	-
	HHA	0.707	0.719	-	-
	RGB-HHA	0.762	0.778	-	-
HED [8]	RGB	0.722	0.737	0.387	0.404
	HHA	0.691	0.704	0.335	0.350
	RGB-HHA	0.746	0.764	0.368	0.384
CATS-HED	RGB	<b>0.732</b>	<b>0.746</b>	<b>0.405</b>	<b>0.418</b>
	HHA	<b>0.693</b>	<b>0.703</b>	<b>0.349</b>	<b>0.362</b>
	RGB-HHA	<b>0.755</b>	<b>0.770</b>	<b>0.382</b>	<b>0.397</b>
RCF [9]	RGB	0.745	0.759	0.398	0.413
	HHA	0.701	0.712	0.333	0.348
	RGB-HHA	0.764	0.778	0.374	0.385
CATS-RCF	RGB	<b>0.752</b>	<b>0.765</b>	<b>0.474</b>	<b>0.488</b>
	HHA	<b>0.710</b>	<b>0.721</b>	<b>0.433</b>	<b>0.445</b>
	RGB-HHA	<b>0.768</b>	<b>0.782</b>	<b>0.439</b>	<b>0.452</b>
BDCN [10]	RGB	0.748	0.762	0.426	0.450
	HHA	0.704	0.716	0.347	0.367
	RGB-HHA	0.766	0.779	0.375	0.392
CATS-BDCN	RGB	<b>0.752</b>	<b>0.765</b>	<b>0.442</b>	<b>0.462</b>
	HHA	<b>0.712</b>	<b>0.724</b>	<b>0.422</b>	<b>0.439</b>
	RGB-HHA	<b>0.770</b>	<b>0.783</b>	<b>0.418</b>	<b>0.435</b>

### 3.4.3 Multicue Dataset

The CATS-applied models are further compared with HED, RCF and BDCN on the edge and boundary data of Multicue via the standard evaluation procedure. In evaluation, the performance of a model is measured by the average score of three independent trials, and the performance fluctuation is appraised by the mean standard deviation. The quantitative results with only single-scale input are presented in Tab. 4.

In Tab. 4, the CATS-applied approaches obtain an apparent detection accuracy increase against the corresponding original models on both the Multicue Boundary and Multicue Edge datasets.

### 3.5 Ablation Study

The ablation study was conducted on NYUDv2 dataset to avoid the impact of the label threshold, and the NYUDv2 dataset was only augmented by flipping. RCF [9] with weighted cross entropy was chosen as the baseline for comparison. We verified the functionality of each component in the tracing loss, and investigated the side unmixing effect of

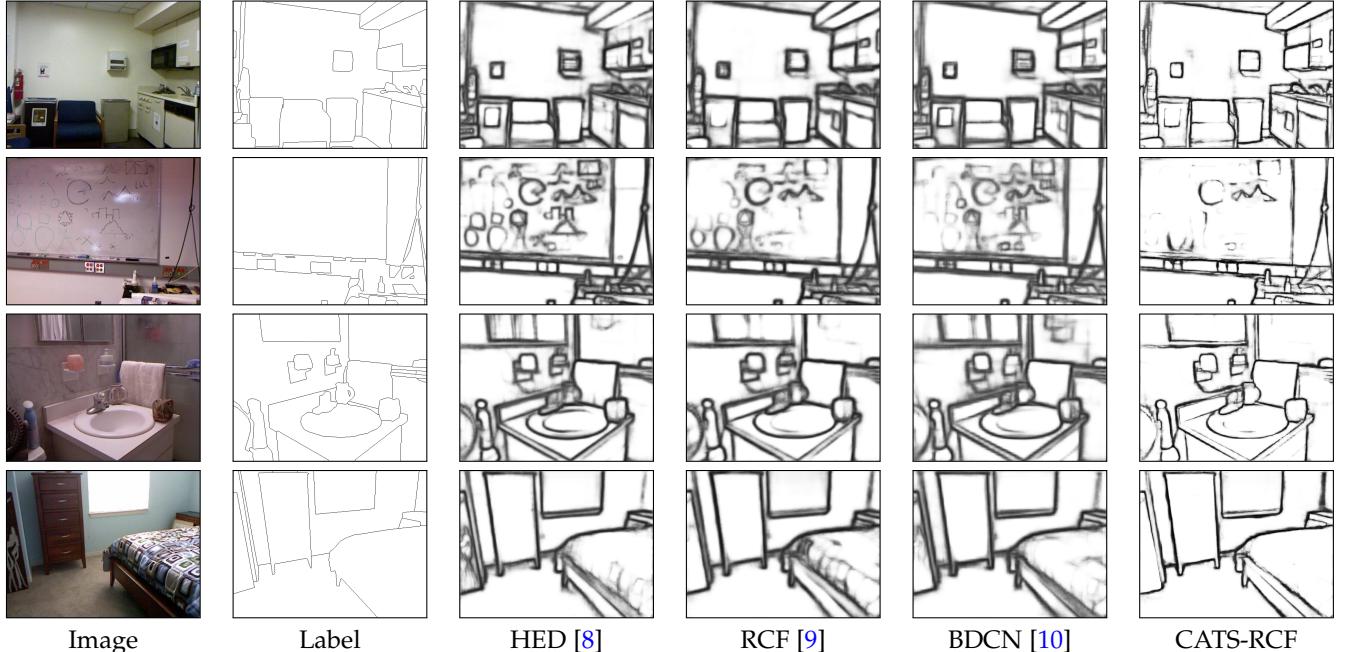


Fig. 6. Qualitative comparison between the prior arts of deep edge detection and our proposed CATS on the NYUDv2 dataset [27].

TABLE 4

Evaluation results on Multicue dataset. The mean standard deviations are given in the parentheses.

Category	Method	ODS	OIS
Boundary	Human [28]	0.760(0.017)	-
	Multicue [28]	0.720(0.014)	-
	HED [8]	0.814(0.011)	0.822(0.008)
	CATS-HED	<b>0.827(0.007)</b>	<b>0.830(0.008)</b>
	RCF [9]	0.817(0.004)	0.825(0.005)
	CATS-RCF	<b>0.841(0.001)</b>	<b>0.846(0.002)</b>
	BDCN [10]	0.836(0.001)	0.846(0.003)
Edge	Human [28]	0.750(0.024)	-
	Multicue [28]	0.830(0.002)	-
	HED [8]	0.851(0.014)	0.864(0.011)
	CATS-HED	<b>0.885(0.004)</b>	<b>0.893(0.003)</b>
	RCF [9]	0.857(0.004)	0.862(0.004)
	CATS-RCF	<b>0.892(0.001)</b>	<b>0.895(0.001)</b>
	BDCN [10]	0.891(0.001)	0.898(0.002)
	CATS-BDCN	<b>0.897(0.001)</b>	<b>0.904(0.001)</b>

the CoFusion block by substituting it for the weighted average layer in RCF. All models were conventionally trained from scratch, where the weights in convolution layers were initialized by zero-mean Gaussian distributions with a standard deviation 0.01 and the biases were set to 0. Tab. 5 presents the details of quantitative analysis.

TABLE 5

The numerical results for ablation study.  $L_{bdry}$  and  $L_{tex}$  denote the boundary tracing function and texture suppression function. SEval and CEval respectively denote the standard evaluation and the crispness-emphasized evaluation.

$L_{bdry}$	$L_{tex}$	CoFusion	CATS			SEval			CEval		
			ODS	OIS	ODS	OIS	ODS	OIS	ODS	OIS	ODS
RCF			0.720	0.736	0.381	0.397					
RCF	✓		0.722	0.738	0.382	0.399					
RCF		✓	0.726	0.743	0.383	0.396					
RCF	✓	✓	0.729	0.744	0.411	0.424					
RCF		✓	0.725	0.739	0.381	0.394					
RCF	✓		0.727	0.740	0.427	0.441					
RCF		✓	0.728	0.743	0.409	0.427					
CATS-RCF	✓	✓	0.731	0.744	0.451	0.468					

In Tab. 5, the complete CATS improves the performance of RCF by 1.1% in ODS and 0.8% in OIS, while using the

tracing loss or the CoFusion block alone gains over the original RCF by at most 0.9% in ODS and 0.8% in OIS.

In terms of the crispness-emphasized evaluation displayed in Tab. 5, the proposed CATS gains over the RCF by 7% in ODS and 7.1% in OIS. The results in the first three rows show that, although  $L_{bdry}$  or  $L_{tex}$  alone can promote the performance of RCF with post-processing, they still have similar crispness scores with the weighted cross entropy. This is because that  $L_{bdry}$  and  $L_{tex}$  focus on different types of non-edge points. Only applying one of them is insufficient for a comprehensive refinement of side edges. In contrast, with the effective feature unmixing and texture suppression offered by combining  $L_{bdry}$  and  $L_{tex}$ , the tracing loss brings a remarkable crispness enhancement to RCF. Similarly, the CoFusion block guided by the weighted cross entropy also increases the scores of the standard evaluation measures, but with little crispness improvement. When working with  $L_{bdry}$  or  $L_{tex}$ , however, the CoFusion block shows its power in addressing side mixing for crisp edge detection. With the complete CATS where the CoFusion is guided by the tracing loss, CATS-RCF achieves the most crispness improvement, which demonstrates that the tracing loss and the CoFusion block work complementarily for crisp edge detection.

### 3.6 Deeper into the CoFusion block

For a perceptual understanding of the side unmixing effect in the CoFusion block, we visualized the weight maps generated by the CoFusion block in Fig. 7.

It can be found that the CoFusion block gives more attention to the two highest-level outputs on the finely smoothed non-edge regions and higher weights to the lower-level side predictions at boundaries. By concentrating on the different merits of side maps for edge delineation, the CoFusion block solves the issue of side mixing in the fusion procedure.

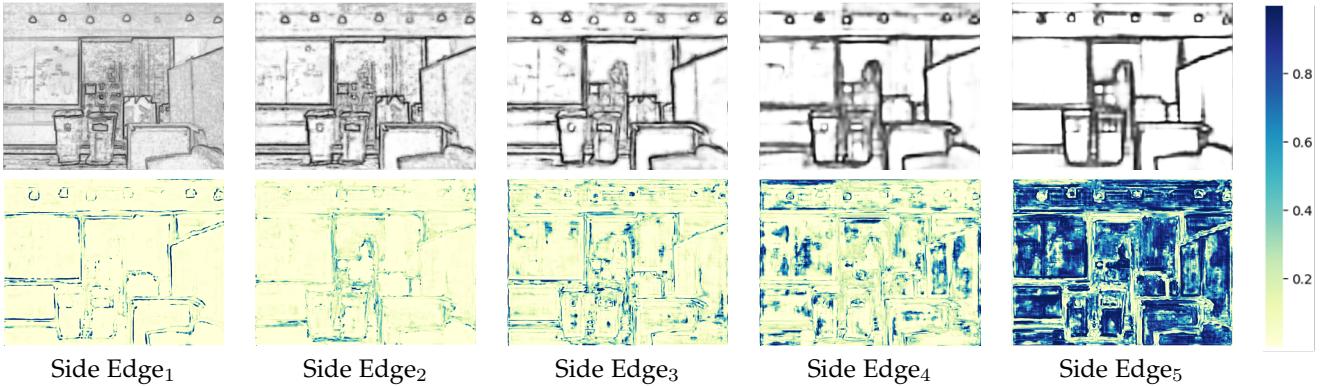


Fig. 7. The weight maps generated by the CoFusion block in CATS-RCF for multi-level side edges on the NYUDv2 dataset [27]. The results show the different attention preference of multi-level side edges for non-edge region smoothing and edge details.

## 4 CONCLUSION

In this paper, we introduce a simple yet effective context-aware tracing strategy (CATS) to tackle the problem of localization ambiguity for modern deep edge detectors. The CATS addresses this issue by unmixing the edge features with a tracing loss and the side predictions with a CoFusion block during the fusion procedure. The effectiveness of CATS was validated in three VGG16-based edge detectors, *i.e.*, HED, RCF, and BDCN. Extensive experiments demonstrate that the CATS can bring consistent performance improvement to the three detectors. Especially, the qualitative and the crispness-emphasized evaluation results show that the CATS facilitates modern edge detectors to obtain crisp edges from raw edge predictions, with the localization accuracy dramatically improved. In the future, it would be interesting to explore the effectiveness of CATS for other down-stream vision tasks. Source code and more results are available at <https://github.com/WHUHLX/CATS>.

## ACKNOWLEDGMENT

This research was funded by the National Key Research and Development Program of China under Grant 2018YFB0505401, and the National Natural Science Foundation of China Project under Grant 42071370.

## REFERENCES

- [1] L. G. Roberts, "Machine perception of three-dimensional solids," *PhD thesis, Massachusetts Institute of Technology*, 1963.
- [2] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.
- [3] J. Kittler, "On the accuracy of the sobel edge detector," *Image and Vision Computing*, vol. 1, no. 1, pp. 37–42, 1983.
- [4] D. 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.*, vol. 26, no. 5, pp. 530–549, 2004.
- [5] X. Ren and L. Bo, "Discriminatively trained sparse code gradients for contour detection," in *NeurIPS*, 2012, pp. 584–592.
- [6] 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, pp. 3982–3991.
- [7] G. Bertasius, J. Shi, and L. Torresani, "Deepedge: A multi-scale bifurcated deep network for top-down contour detection," in *CVPR*, 2015.
- [8] S. Xie and Z. Tu, "Holistically-nested edge detection," *International Journal of Computer Vision*, vol. 125, no. 1, pp. 3–18, 2017.
- [9] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939–1946, 2019.
- [10] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *CVPR*, 2019, pp. 3828–3837.
- [11] 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, pp. 193–202.
- [12] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool, "Convolutional oriented boundaries: From image segmentation to high-level tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 819 – 833, 2018.
- [13] Y. Liu and M. S. Lew, "Learning relaxed deep supervision for better edge detection," in *CVPR*, 2016, pp. 231–240.
- [14] I. Kokkinos, "Pushing the boundaries of boundary detection using deep learning," *arXiv preprint arXiv:1511.07386*, 2015.
- [15] Y. Ganin and V. Lempitsky, "N<sup>4</sup>-fields: Neural network nearest neighbor fields for image transforms," in *ACCV*, 2014, pp. 536–551.
- [16] Y. Wang, X. Zhao, Y. Li, and K. Huang, "Deep crisp boundaries: From boundaries to higher-level tasks," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1285–1298, 2019.
- [17] R. Deng, C. Shen, S. Liu, H. Wang, and X. Liu, "Learning to predict crisp boundaries," in *ECCV*, 2018, pp. 570–586.
- [18] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*. IEEE, 2016, pp. 565–571.
- [19] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017, pp. 1492–1500.
- [20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015, pp. 1–14.
- [21] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2011.
- [22] X. Jia, B. Brabandere, T. Tuytelaars, and L. Van Gool, "Dynamic filter networks," *NeurIPS*, 01 2016.
- [23] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*. PMLR, 2015, pp. 2048–2057.
- [24] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *CVPR*, 2016, pp. 3640–3649.
- [25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, September 2018.
- [26] Y. Hu, Y. Chen, X. Li, and J. Feng, "Dynamic feature fusion for semantic edge detection," in *IJCAI*, 08 2019, pp. 782–788.
- [27] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012, pp. 746–760.
- [28] D. A. Mély, J. Kim, M. McGill, Y. Guo, and T. Serre, "A systematic comparison between visual cues for boundary detection," *Vision Research*, vol. 120, pp. 93–107, 2016.
- [29] B. Steiner, Z. Devito, S. Chintala, S. Gross, A. Paszke, F. Massa, A. Lerer, G. Chanan, Z. Lin, E. Yang *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019, pp. 8026–8037.

- [30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
- [31] S. Hallman and C. C. Fowlkes, "Oriented edge forests for boundary detection," in *CVPR*, 2015, pp. 1732–1740.
- [32] 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, pp. 504–512.
- [33] 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 *NeurIPS*, 2017, pp. 3961–3970.
- [34] 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, pp. 859–864.