Semantic Edge Detection with Diverse Deep Supervision

Yun Liu, Ming-Ming Cheng, Deng-Ping Fan, Le Zhang, Jia-Wang Bian, and Dacheng Tao, *Fellow, IEEE*

Abstract—Semantic edge detection (SED), which aims at jointly extracting edges as well as their category information, has far-reaching applications in domains such as semantic segmentation, object proposal generation, and object recognition. SED naturally requires achieving two distinct supervision targets: locating fine detailed edges and identifying high-level semantics. We shed light on how such distracted supervision targets prevent state-of-the-art SED methods from effectively using deep supervision to improve results. In this paper, we propose a novel fully convolutional neural network architecture using diverse deep supervision (**DDS**) within a multi-task framework where lower layers aim at generating category-agnostic edges, while higher layers are responsible for the detection of category-aware semantic edges. To overcome the distracted supervision challenge, a novel information converter unit is introduced, whose effectiveness has been extensively evaluated in several popular benchmark datasets, including SBD, Cityscapes, and PASCAL VOC2012. Source code will be released upon paper acceptance.

Index Terms—Semantic edge detection, diverse deep supervision, information converter.

1 Introduction

The aim of classical edge detection is to detect edges and object boundaries in natural images. It is **category-agnostic**, in that object categories need not be recognized. Classical edge detection can be viewed as a pixel-wise binary classification problem, whose objective is to classify each pixel as belonging to either the class indicating an edge, or the class indicating a non-edge. In this paper, we consider more practical scenarios of semantic edge detection, which jointly achieves edge detection and edge category recognition within an image. Semantic edge detection (**SED**) [1]–[4] is an active computer vision research topic due to its wide-ranging applications, including in object proposal generation [4], occlusion and depth reasoning [5], 3D reconstruction [6], object detection [7], [8], image-based localization [9].

Recently, deep convolutional neural networks (**DCNNs**) reign undisputed as the new de-facto method for category-agnostic edge detection [10]–[14], where near human-level performance has been achieved. However, deep learning for **category-aware** SED, which jointly detects visually salient edges as well as recognizing their categories, has not yet witnessed such vast popularity. Hariharan *et al.* [1] first combined generic object detectors with bottom-up edges to recognize semantic edges. Yang *et al.* [15] proposed a fully convolutional encoder-decoder network to detect object contours but without recognizing specific categories. More recently, CASENet [2] introduces a skip-layer structure to enrich the top-layer category-wise edge activation with bottom-layer features, improving previous state-of-the-art methods with a significant margin.

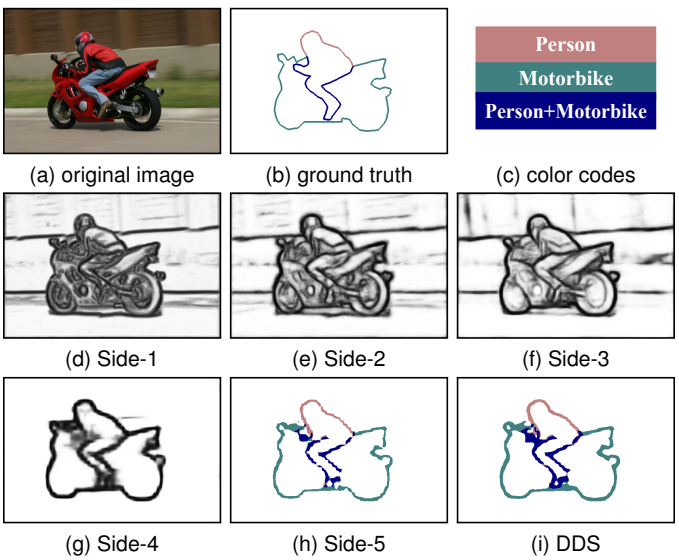


Fig. 1. An example of our DDS algorithm. (a) shows the original image from the SBD dataset [1]. (b)-(c) show its semantic edge map and corresponding color codes. (d)-(g) display category-agnostic edges from Side-1 ∼ Side-4. (h)-(i) show semantic edges of Side-5 and DDS (DDS-R) output, respectively.

Distracted supervision paradox in SED. SED naturally requires achieving two distinct supervision targets: i) locating fine detailed edges by capturing discontinuity among image regions, mainly using low-level features; and ii) identifying abstracted high-level semantics by summarizing different appearance variations of the target categories. This distracted supervision paradox prevents the state-of-the-art SED method, *i.e.* CASENet [2], from successfully applying deep supervision [16], whose effectiveness has been demonstrated in a wide number of other computer vision

- *Y. Liu, M.M. Cheng, D.P. Fan and J.W. Bian are with College of Computer Science, Nankai University. M.M. Cheng is the corresponding author (cmm@nankai.edu.cn).*
- *L. Zhang with Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR).*
- *D. Tao is with the School of Information Technologies, University of Sydney.*

Manuscript received April 19, 2005; revised August 26, 2015.

tasks, *e.g.* image categorization [17], object detection [18], visual tracking [19], and category-agnostic edge detection [11], [12].

In this paper, we propose a **diverse deep supervision** (**DDS**) method, which employs deep supervision with different loss functions for high-level and low-level feature learning as shown in Fig. 2(b). While mainly using high-level convolution (*i.e.* ***conv***) features for semantic classification and low-level *conv* ones for non-semantic edge details may be intuitive and straightforward, directly doing this as in CASENet [2] even degrades the performance compared with directly learning semantic edges without deep supervision or category-agnostic edge guidance. In [2], after unsuccessfully trying various ways of adding deep supervision, Yu *et al.* claimed that deep supervision for lower network layers *is unnecessary*, As illustrated in Fig. 2(b), we propose an ***information converter*** unit to change the backbone DCNN features into different representations, for training category-agnostic or semantic edges respectively. Without such information converters, the low-level (*conv* Side-1 $\sim$ Side-4) and high-level (*conv* Side-5) DCNN features would be optimized towards category-agnostic and semantic edges, respectively, which are difficult to be transformed with simple convolutions between Side-4 and Side-5 features. By introducing the information converter units, a single backbone representation can be effectively learned end-to-end towards different targets. An example of DDS is shown in Fig. 1. The bottom sides of the neural network help Side-5 to find fine details, thus making the final fused semantic edges (Fig. 1 (i)) smoother than those coming from Side-5 (Fig. 1 (h)).

In summary, our main contributions are:

- analyzing the distracted supervision paradox in the context of SED, and why it stops the state-of-the-art SED method [2] from using deep supervision to improve results (Section 3);
- proposing a new SED method, called diverse deep supervision (DDS), which uses *information converters* to avoid the difficulties inherent in learning powerful backbone features with distracted supervision (Section 4); and
- providing detailed ablation studies to further understand the proposed method (Section 5.1).

We extensively evaluate our method on SBD [1], Cityscapes [20], and PASCAL VOC2012 [21] datasets. Our method achieves the new state-of-the-art performance. On the Cityscapes dataset, the mean maximum F-measure of our proposed DDS algorithm at optimal dataset scale (ODS) [22] is 79.3%, comparing with previous state-of-the-art performance of 75.0% [23].

2 Related Work

An exhaustive review of the abundant literature on this topic is out of the scope of this paper. Instead, we first summarize the most important threads of research to solve the problem of classical category-agnostic edge detection, followed by the discussions of deep learning-based approaches, semantic edge detection (SED), and the technique of deep supervision.

Classical category-agnostic edge detection. Edge detection is conventionally solved by designing various filters (*e.g.* Sobel [24] and Canny [25]) or complex models [26], [27] to detect pixels with highest gradients in their local neighborhoods [28]–[30]. To the best of our knowledge, Konishi *et al.* [31] proposed the first data-driven edge detector in which, unlike previous model based approaches, edge detection was posed as statistical inferences. *Pb* features consisting of brightness, color and texture are used in [32] to obtain the posterior probability of each boundary point. *Pb* is further extended to *gPb* [22] by computing local cues from multi-scale and globalizing them through spectral clustering. Sketch tokens are learned from hand-drawn sketches for contour detection [33], while random decision forests are employed in [34] to learn the local structure of edge patches, delivering competitive results among non-deep-learning approaches.

Deep category-agnostic edge detection. The number of success stories of machine learning has seen an all-time rise across many computer vision tasks recently. The unifying idea is deep learning which utilizes neural networks with many hidden layers aimed at learning complex feature representations from raw data [35]–[38]. Motivated by this, deep learning based methods have made vast inroads into edge detection as well [39]–[41]. Ganin *et al.* [42] applied deep neural network for edge detection using a dictionary learning and nearest neighbor algorithm. DeepEdge [43] first extracts candidate contour points and then classifies these candidates. HFL [4] uses SE [34] to generate candidate edge points in contrast to Canny [25] used in DeepEdge. Compared with DeepEdge which has to process input patches for every candidate point, HFL turns out to be more computationally feasible as the input image is only fed into the network once. DeepContour [44] partitions edge data into subclasses and fits each subclass using different model parameters. Xie *et al.* [10], [11] leveraged deeply-supervised nets to build a fully convolutional network for image-to-image prediction. Their deep model, known as HED, fuses the information from the bottom and top *conv* layers. Kokkinos [45] proposed some training strategies to retrain HED. Liu *et al.* [12], [13] introduced the first real-time edge detector, which achieves higher F-measure scores than average human annotators on the popular BSDS500 dataset [22].

Semantic edge detection. By virtue of their strong capacity for semantic representation learning, deep learning based edge detectors tend to generate high responses at object boundary locations, *e.g.* Fig. 1 (d)-(g). This has inspired research on simultaneously detecting edge pixels and classifying them based on associations with one or more object categories. This so-called "category-aware" edge detection is highly beneficial to various vision tasks including object recognition, stereo vision, semantic segmentation, and object proposal generation.

Hariharan *et al.* [1] was the first to propose a principled way of combining generic object detectors with bottom-up contours to detect semantic edges. Yang *et al.* [15] proposed a fully convolutional encoder-decoder network for object contour detection. HFL produces category-agnostic binary

edges and assigns class labels to all boundary points using deep semantic segmentation networks. Maninis *et al.* [3] coupled their convolutional oriented boundaries (COB) with semantic segmentation generated by dilated convolutions [46] to obtain semantic edges. A weakly supervised learning strategy is introduced in [47] in which bounding box annotations alone are sufficient to produce high-quality object boundaries without any object-specific annotations. Yu *et al.* [2] proposed a novel network, CASENet, which has pushed SED performance to a new state-of-the-art. In their architecture, low-level features are only used to augment top classifications. After several failed experiments, they reported that deep supervision on the bottom sides of the lower layers *is unnecessary* for SED. More recently, Yu *et al.* [23] introduced a new training approach, SEAL, to train CASENet [2]. This approach can simultaneously align ground truth edges and learn semantic edge detectors. However, the training of SEAL is very time-consuming due to the heavy CPU computation load, *i.e.* over 16 days to train CASENet on the SBD dataset [1], although we have used a powerful CPU (Intel Xeon(R) CPU E5-2683 v3 @ 2.00GHz $\times$ 56).

Deep supervision. Deep supervision has been demonstrated to be effective in many vision and learning tasks such as image classification [16], [17], object detection [18], [48], [49], visual tracking [19], category-agnostic edge detection [11], [12], salient object detection [50] and so on. Theoretically, the lower layers of deep networks learn discriminative features so that classification/regression at higher layers is easier. In practice, one can explicitly influence the hidden layer weight/filter update process to favor highly discriminative feature maps using deep supervision. However, it may be suboptimal to directly add deep supervision of category-agnostic edges on the bottom sides due to the distracted supervision discussed above. We will introduce a new semantic edge detector with successful diverse deep supervision in the following sections.

3 Discussion of CASENet

Before expounding the proposed method, we first briefly introduce the CASENet architecture [2] as shown in Fig. 2(a).

3.1 CASENet Model

CASENet [2] is built on the well-known backbone network of ResNet [51]. It connects a 1×1 *conv* layer after each of Side-1 $\sim$ Side-3 to produce a single channel feature map $F^{(m)}$. The top Side-5 is connected to a 1×1 *conv* layer to output K-channel class activation map $A^{(5)} = \{A_1^{(5)}, A_2^{(5)}, \cdots, A_K^{(5)}\}$, where K is the number of categories. The *shared concatenation* replicates bottom features $F^{(m)}$ to separately concatenate each channel of the class activation map:

$$A^f = \{F^{(1)}, F^{(2)}, F^{(3)}, A_1^{(5)}, \cdots, F^{(1)}, F^{(2)}, F^{(3)}, A_K^{(5)}\}. \quad (1)$$

Then, a K-grouped 1×1 *conv* is performed on A^f to generate a semantic edge map with K channels, in which the k-th channel represents the edge map for the k-th category.

3.2 Discussion

CASENet [2] only imposes supervision on Side-5 and the final fused activation. In [2], the authors have tried several deeply supervised architectures. They first separately used all of Side-1 $\sim$ Side-5 for SED, with each side connected with a classification loss. The evaluation results are even worse than the basic architecture that directly applies 1×1 convolution at Side-5 to obtain semantic edges. It is widely accepted that the lower layers of neural networks contain low-level, less-semantic features such as local edges, which are unsuitable for semantic classification because semantic category recognition needs abstracted high-level features that appear in the top layers of neural networks. Thus they would obtain poor classification results at bottom sides. Unsurprisingly, simply connecting each low-level feature layer and high-level feature layer with a classification loss and deep supervision for SED task results in a clear performance drop.

Yu *et al.* [2] also attempted to impose deep supervision of binary edges at Side-1 $\sim$ Side-3 in CASENet but observed divergence in the semantic classification at Side-5. With the top supervision of semantic edges, the top layers of the network will be supervised to learn abstracted high-level semantics that can summarize different appearance variations of the target categories. Since the bottom layers are the bases of the top layers for the representation power of the DCNNs, the bottom layers will be supervised to serve the top layers for obtaining high-level semantics through back propagation. Conversely, with bottom supervision of category-agnostic edges, the bottom layers are taught to focus on distinction between edges and non-edges, rather than visual representations for semantic classification. This will cause conflicts in the bottom layers and therefore fail to provide discriminative gradient signals for parameter updating.

Note that Side-4 is not used in CASENet. We believe it is a naive way to **alleviate** the information conflicts by regarding the whole *res4* block as a buffer unit between the bottom and top sides. Indeed, when adding Side-4 to CASENet (see Section 5.1), the new model (*CASENet+S4*) achieves a 70.9% mean F-measure compared with the 71.4% of the original CASENet. This confirms our hypothesis about the buffer function of *res4* block. Moreover, the classical 1×1 *conv* layer after each side [2], [11] is too weak to buffer the conflicts. We therefore propose an information converter unit to solve the conflicts of distracted supervision.

4 Our Approach

Intuitively, by employing different but "appropriate" ground-truths for the bottom and top sides, the learned intermediate representations of the different levels may contain complementary information. However, directly imposing deep supervision does not seem to be beneficial. In this section, we propose a new network architecture for the complementary learning of the bottom and top sides for SED.

4.1 The Proposed DDS Algorithm

Based on above discussion, we hypothesize that the bottom sides of neural networks may not be directly beneficial to

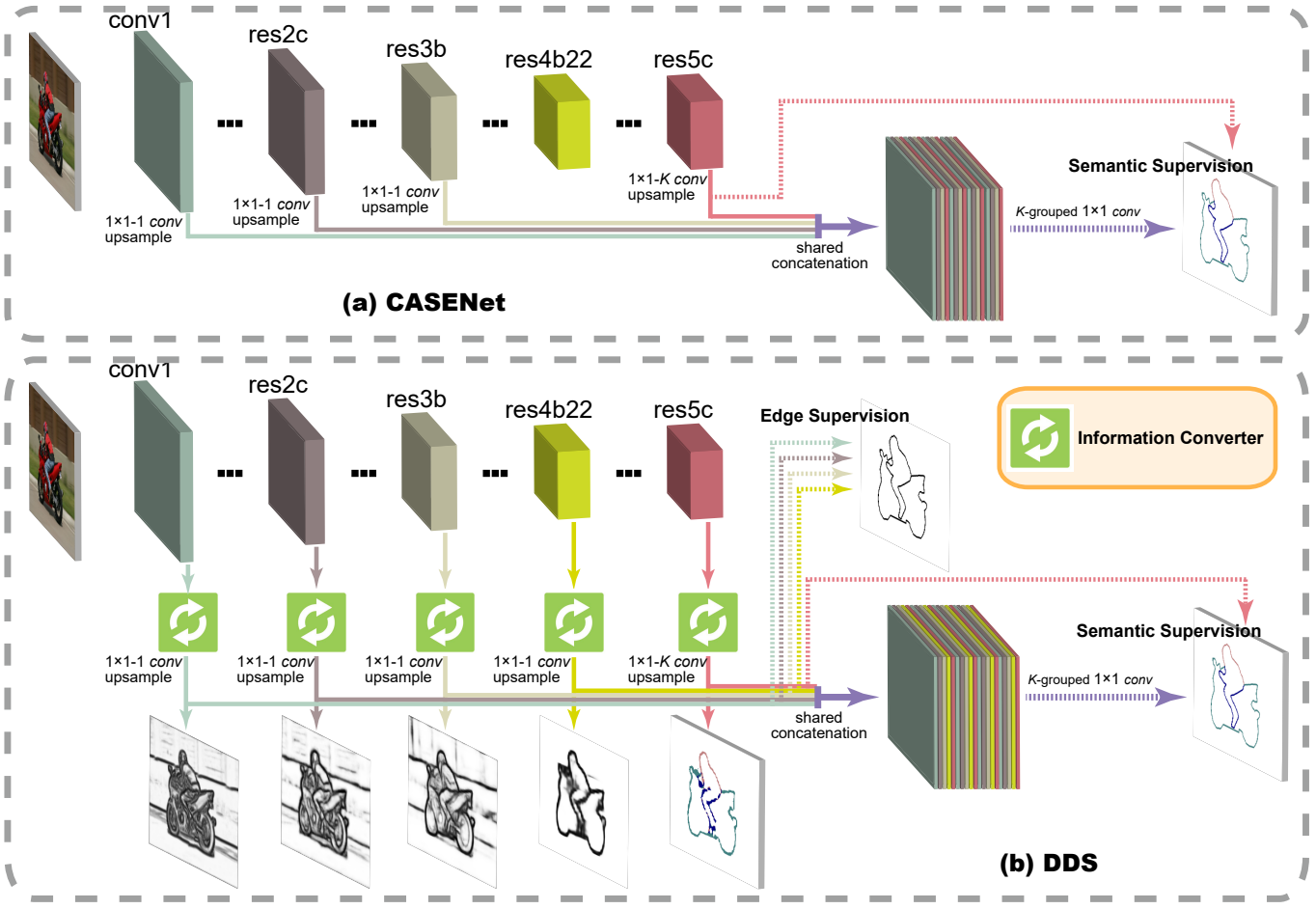


Fig. 2. A comparison between two SED models: CASENet [2] and our DDS. CASENet only adds top supervision on the Side-5 activation, and the authors claimed that deep supervision was not necessary in their architecture. However, our proposed DDS network adds deep supervision on all of the side activation. Note that the *information converters* are crucial for avoiding the distracted supervision among category-agnostic and semantic edge detection.

SED. However, we believe that the bottom sides still encode fine details complementary to the top side (Side-5). With appropriate architecture re-design, we believe that they can be used for category-agnostic edge detection to improve the localization accuracy of semantic edges generated by the top side. To this end, we design a novel *information converter* to assist low-level feature learning and to generate consistent gradient signals from the higher layers. This is essential, as they enable direct influence of the hidden layer weight/filter update process to favor highly discriminative feature maps for correct SED.

Our proposed network architecture is presented in Fig. 2(b). We follow CASENet to use ResNet [51] as our backbone network. After each *information converter* (Section 4.2) in Side-1 $\sim$ Side-4, we connect a 1×1 *conv* layer with a single output channel to produce an edge response map. These predicted maps are then upsampled to the original image size using bilinear interpolation. These side outputs are supervised by binary category-agnostic edges. We perform K-channel 1×1 convolution on Side-5 to obtain semantic edges, where each channel represents the binary edge map of one category. We adopt the same upsampling operation as for Side-1 $\sim$ Side-4. Semantic edges are used to supervise the training of Side-5.

We denote the produced binary edge maps from Side-1 $\sim$ Side-4 as $E = \{E^{(1)}, E^{(2)}, E^{(3)}, E^{(4)}\}$. The semantic edge map from Side-5 is still represented by $A^{(5)}$. A shared concatenation is then performed to obtain the stacked edge activation map:

$$E^f = \{E, A_1^{(5)}, E, A_2^{(5)}, E, A_3^{(5)}, \cdots, E, A_K^{(5)}\}. \quad (2)$$

Note that E^f is a stacked edge activation map, while A^f in CASENet is a stacked feature map. Finally, we apply K-grouped 1×1 convolution on E^f to generate the fused semantic edges. The fused edges are supervised by the ground truth of the semantic edges. As demonstrated in HED [11], the 1×1 convolution fuses the edges from the bottom and top sides well.

4.2 Information Converter

Recently, residual networks have been proved to be easier to optimize than plain networks [51]. The residual learning operation is usually embodied by a shortcut connection and element-wise addition. We describe a residual *conv* block in Fig. 3, which consists of four alternatively connected ReLU and *conv* layers, and the output of the first ReLU layer is added to the output of the last *conv* layer. Our proposed *information converter* combines two residual modules and is

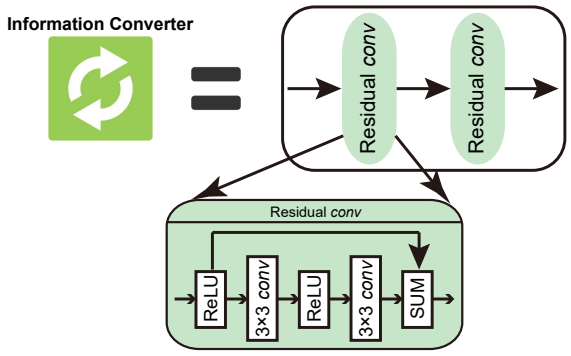


Fig. 3. Schematic of our *information converter* unit (illustrated in the orange box in Fig. 2).

connected to each side of the DDS network to transform the learned representation into the proper form. This operation is expected to avoid the conflicts caused by the discrepancy in different loss functions.

The top supervision of semantic edges will produce gradient signals for learning semantic features, while the bottom supervision of category-agnostic edges will produce category-agnostic gradients. These conflicting gradient signals will confuse the backbone network through back propagation if the distracted supervision is directly imposed. Our information converters can play a buffering role by converting these conflicting signals into a proper representation. In this way, the backbone network will receive consistent update signals and will be optimized towards the same target; furthermore, the different tasks at the bottom and top sides are carried out by the *information converters*.

Our proposed network can successfully combine the fine details from the bottom sides and the semantic classification from the top side. Our experimental results demonstrate that the algorithm solves the problem of conflicts caused by diverse deep supervision. Unlike CASENet, our semantic classification at Side-5 can be well optimized without any divergence. The produced binary edges from the bottom sides help Side-5 to make up the fine details. Thus, the final fused semantic edges can achieve better localization quality.

We use binary edges of single pixel width to supervise Side-1 $\sim$ Side-4 and thick semantic boundaries to supervise Side-5 and the final fused edges. One pixel is viewed as a binary edge if it belongs to the semantic boundaries of any category. We obtain thick semantic boundaries by seeking the difference between a pixel and its neighbors, as in CASENet [2]. A pixel with label k is regarded as a boundary of class k if at least one neighbor with a label k' ($k' \neq k$) exists.

4.3 Multi-task Loss

Two different loss functions, which represent category-agnostic and category-aware edge detection losses, respectively, are employed in our multi-task learning framework. We denote all the layer parameters in the network W. Suppose an image I has a corresponding binary edge map $Y = \{y_i, i = 1, 2, \cdots, |I|\}$. The reweighted *sigmoid* cross-entropy loss function for Side-1 $\sim$ Side-4 can be formulated as

$$\begin{aligned} L_{side}^{(m)}(W) = -\sum_{i\in I}[\beta \cdot (1-y_i) \cdot log(1-P(E_i^{(m)};W)) \\ + (1-\beta) \cdot y_i \cdot log(P(E_i^{(m)};W))], \\ (m = 1, \cdots, 4), \end{aligned} \tag{3}$$

where $\beta = |Y^+|/|Y|$ and $1-\beta = |Y^-|/|Y|$. Y^+ and Y^- represent edge and non-edge ground truth label sets, respectively. $E_i^{(m)}$ is the produced activation value at pixel i for m-th side. $P(\cdot)$ is the standard *sigmoid* function.

For an image I, suppose the semantic ground truth label is $\{\bar{Y}^1, \bar{Y}^2, \cdots, \bar{Y}^K\}$, in which $\bar{Y}^k = \{\bar{y}_i^k, i = 1, 2, \cdots, |I|\}$ is the binary edge map for the k-th category. Note that each pixel may belong to the boundaries of multiple categories. We define the reweighted multi-label loss for Side-5 as

$$\begin{aligned} L_{side}^{(5)}(W) = -\sum_k \sum_{i\in I}[\beta \cdot (1-\bar{y}_i^k) \cdot log(1-P(A_{k,i}^{(5)};W)) \\ + (1-\beta) \cdot \bar{y}_i^k \cdot log(P(A_{k,i}^{(5)};W))], \end{aligned} \tag{4}$$

in which $A_{k,i}^{(5)}$ is the Side-5's activation value for k-th category at pixel i. The loss of the fused semantic activation map is denoted as $L_{fuse}(W)$, which can be similarly defined as

$$\begin{aligned} L_{fuse}(W) = -\sum_k \sum_{i\in I}[\beta \cdot (1-\bar{y}_i^k) \cdot log(1-P(E_{k,i}^{f};W)) \\ + (1-\beta) \cdot \bar{y}_i^k \cdot log(P(E_{k,i}^{f};W))], \end{aligned} \tag{5}$$

where E^f is the fused activation map in Eq. (2). The total loss is formulated as

$$L(W) = \sum_{m=1,\cdots,5} L_{side}^{(m)}(W) + L_{fuse}(W). \tag{6}$$

Using this total loss function, we can use stochastic gradient descent (SGD) to optimize all parameters. We denote DDS trained using the reweighted loss $L(W)$ as **DDS-R**.

Recently, Yu *et al.* [23] proposed to simultaneously align and learn semantic edges. They found that the unweighted (regular) *sigmoid* cross-entropy loss performed better than reweighted loss with their alignment training strategy. Due to the heavy computational load on the CPU, their approach was very time-consuming (over 16 days for SBD dataset [1] with 28 CPU kernels and an NVIDIA TITAN Xp GPU) to train a network. We use their method (SEAL) to align the ground truth edges only once prior to training and apply unweighted *sigmoid* cross-entropy loss to train the aligned edges. The loss function for Side-1 $\sim$ Side-4 can thus be formulated as

$$\begin{aligned} L'^{(m)}_{side}(W) = -\sum_{i\in I}[(1-y_i) \cdot log(1-P(E_i^{(m)};W)) \\ + y_i \cdot log(P(E_i^{(m)};W))], \qquad (m = 1, \cdots, 4). \end{aligned} \tag{7}$$

The unweighted multi-label loss for Side-5 is

$$\begin{aligned} L'^{(5)}_{side}(W) = -\sum_k \sum_{i\in I}[(1-\bar{y}_i^k) \cdot log(1-P(A_{k,i}^{(5)};W)) \\ + \bar{y}_i^k \cdot log(P(A_{k,i}^{(5)};W))]. \end{aligned} \tag{8}$$

TABLE 1
The ODS F-measure (%) of DDS-R/DDS-U and ablation methods on the SBD test set [1]. The best performance of each column is highlighted in **bold**.

Method	aer.	bike	bird	boat	bot.	bus	car	cat	cha.	cow	tab.	dog	hor.	mot.	per.	pot.	she.	sofa	train	tv	mean
Softmax	74.0	64.1	64.8	52.5	52.1	73.2	68.1	73.2	43.1	56.2	37.3	67.4	68.4	67.6	76.7	42.7	64.3	37.5	64.6	56.3	60.2
Basic	82.5	74.2	80.2	62.3	68.0	80.8	74.3	82.9	52.9	73.1	46.1	79.6	78.9	76.0	80.4	52.4	75.4	48.6	75.8	68.0	70.6
DSN	81.6	75.6	78.4	61.3	67.6	82.3	74.6	82.6	52.4	71.9	45.9	79.2	78.3	76.2	80.1	51.9	74.9	48.0	76.5	66.8	70.3
CASENet+S4	84.1	76.4	80.7	63.7	70.3	81.3	73.4	79.4	56.9	70.7	47.6	77.5	81.0	74.5	79.9	54.5	74.8	48.3	72.6	69.4	70.9
DDS\Convt	83.3	77.1	81.7	63.6	70.6	81.2	73.9	79.5	56.8	71.9	48.0	78.3	81.2	75.2	79.7	54.3	76.8	48.9	75.1	68.7	71.3
DDS\DeSup	82.5	77.4	81.5	62.4	70.8	81.6	73.8	80.5	56.9	72.4	46.6	77.9	80.1	73.4	79.9	54.8	76.6	47.5	73.3	67.8	70.9
CASENet [2]	83.3	76.0	80.7	63.4	69.2	81.3	74.9	**83.2**	54.3	74.8	46.4	80.3	80.2	76.6	80.8	53.3	77.2	50.1	75.9	66.8	71.4
DDS-R	85.4	78.3	83.3	65.6	71.4	83.0	75.5	81.3	59.1	75.7	50.7	80.2	82.7	77.0	81.6	58.2	79.5	50.2	76.5	71.2	73.3
DDS-U	**87.2**	**79.7**	**84.7**	**68.3**	**73.0**	**83.7**	**76.7**	82.3	**60.4**	**79.4**	**50.9**	**81.2**	**83.6**	**78.3**	**82.0**	**60.1**	**82.7**	**51.2**	**78.0**	**72.7**	**74.8**

$L'_{fuse}(W)$ can be similarly defined as

$$L'_{fuse}(W) = -\sum_{k}\sum_{i\in I}[(1-\bar{y}_i^k)\cdot log(1-P(E_{k,i}^f;W)) + \bar{y}_i^k \cdot log(P(E_{k,i}^f;W))]. \quad (9)$$

The total loss is the sum across all sides:

$$L'(W) = \sum_{m=1,\cdots,5} L'^{(m)}_{side}(W) + L'_{fuse}(W). \quad (10)$$

We denote DDS trained using the unweighted loss $L'(W)$ as **DDS-U**.

4.4 Implementation Details

We implement our algorithm using the well-known deep learning framework of Caffe [52]. The proposed network is built on ResNet [51]. We follow CASENet [2] to change the strides of the first and fifth convolution blocks from 2 to 1. The *atrous* algorithm is used to keep the receptive field sizes the same as the original ResNet. We also follow CASENet to pre-train the convolution blocks on the COCO dataset [53]. The network is optimized with stochastic gradient descent (SGD). Each SGD iteration chooses 10 images at uniformly random and crops a 352×352 patch from each of them. The weight decay and momentum are set to 0.0005 and 0.9, respectively. We use the learning rate policy of "poly", in which the current learning rate equals the base one multiplying $(1 - curr\_iter/max\_iter)^{power}$. The parameter of $power$ is set to 0.9. We run 25k/80k iterations (max_iter) of SGD for SBD [1] and Cityscapes [20], respectively. For DDS-R training, the base learning rate is set to 5e-7/2.5e-7 for SBD [1] and Cityscapes [20], respectively. For DDS-U training, the loss at the beginning of training is very large. Therefore, for both SBD and Cityscapes, we first pre-train the network with a fixed learning rate of 1e-8 for 3k iterations and then use the base learning rate of 1e-7 to continue training with the same settings as described above. We use the model trained on SBD to test PASCAL VOC2012 without retraining. The side upsampling operation is implemented with deconvolution layers by fixing the parameters to perform bilinear interpolation. All experiments are performed using an NVIDIA TITAN Xp GPU.

5 Experiments

We evaluate our method on several datasets, including SBD [1], Cityscapes [20], and PASCAL VOC2012 [21]. SBD [1] comprises 11,355 images and corresponding labeled semantic edge maps for 20 Pascal VOC classes. It is divided into 8498 training and 2857 test images. We follow [2] to use the training set to train our network and the test set for evaluation. The Cityscapes dataset [20] is a large-scale semantic segmentation dataset with stereo video sequences recorded in street scenarios from 50 different cities. It consists of 5000 images divided into 2975 training, 500 validation, and 1525 test images. The ground truth of the test set has not been published because it is an online competition for semantic segmentation labeling and scene understanding. Hence, we use the training set for training and validation set for testing. The semantic segmentation set of PASCAL VOC2012 [21] consists of 1464 training, 1449 validation, and 1456 test images with the same 20 classes as the SBD dataset. For the same reason as Cityscapes, the semantic labeling of the test set has not been published. We generate a new validation set that excludes the SBD training images, resulting in 904 validation images. We use this new set and the models trained on the SBD training set to test the generalizability of different methods.

For evaluation protocol, we use the standard benchmark with default settings published in [1]. The maximum F-measure (F_m) at the optimal dataset scale (ODS) for each class and mean maximum F-measure across all classes are reported. For Cityscapes and VOC2012, we follow [1] to generate semantic edges of single pixel width, so that the produced boundaries are just the boundaries of *semantic objects or stuff* in semantic segmentation. We also follow [2] to downsample the ground truth and predicted edge maps of Cityscapes to half the original dimensions to speed up evaluation.

5.1 Ablation Studies

We first perform ablation experiments on SBD to investigate various aspects of our proposed DDS algorithm before comparing it with existing state-of-the-art methods. To this end, we propose six DDS variants:

- ***Softmax***, which only adopts the top side (Side-5) with a 21-class *softmax* loss function, such that the ground truth edges of each category do not overlap and thus each pixel has one specific class label.
- ***Basic***, which employs the top side (Side-5) for multi-label classification, meaning that we directly connect the loss function of $L^{(5)}_{side}(W)$ on *res5c* to train the detector.

TABLE 2
The ODS F-measure (%) of DDS-R/DDS-U and competitors on the SBD test set [1]. The best performance of each column is highlighted in **bold**.

Method	aer.	bike	bird	boat	bot.	bus	car	cat	cha.	cow	tab.	dog	hor.	mot.	per.	pot.	she.	sofa	train	tv	mean
InvDet [1]	41.5	46.7	15.6	17.1	36.5	42.6	40.3	22.7	18.9	26.9	12.5	18.2	35.4	29.4	48.2	13.9	26.9	11.1	21.9	31.4	27.9
HFL-FC8 [4]	71.6	59.6	68.0	54.1	57.2	68.0	58.8	69.3	43.3	65.8	33.3	67.9	67.5	62.2	69.0	43.8	68.5	33.9	57.7	54.8	58.7
HFL-CRF [4]	73.9	61.4	74.6	57.2	58.8	70.4	61.6	71.9	46.5	72.3	36.2	71.1	73.0	68.1	70.3	44.4	73.2	42.6	62.4	60.1	62.5
BNF [54]	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
WS [47]	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
DilConv [46]	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
DSN [2]	81.6	75.6	78.4	61.3	67.6	82.3	74.6	82.6	52.4	71.9	45.9	79.2	78.3	76.2	80.1	51.9	74.9	48.0	76.5	66.8	70.3
COB [3]	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
CASENet [2]	83.3	76.0	80.7	63.4	69.2	81.3	74.9	**83.2**	54.3	74.8	46.4	80.3	80.2	76.6	80.8	53.3	77.2	50.1	75.9	66.8	71.4
SEAL [23]	85.2	77.7	83.4	66.3	70.6	82.4	75.2	82.3	58.5	76.5	50.4	80.9	82.2	76.8	**82.2**	57.1	78.9	50.4	75.8	70.1	73.1
DDS-R	85.4	78.3	83.3	65.6	71.4	83.0	75.5	81.3	59.1	75.7	50.7	80.2	82.7	77.0	81.6	58.2	79.5	50.2	76.5	71.2	73.3
DDS-U	**87.2**	**79.7**	**84.7**	**68.3**	**73.0**	**83.7**	**76.7**	82.3	**60.4**	**79.4**	**50.9**	**81.2**	**83.6**	**78.3**	82.0	**60.1**	82.7	51.2	**78.0**	**72.7**	**74.8**

TABLE 3
The ODS F-measure (%) of some competitors on the re-annotated SBD test set [21]. The best performance of each column is highlighted in **bold**.

Method	aer.	bike	bird	boat	bot.	bus	car	cat	cha.	cow	tab.	dog	hor.	mot.	per.	pot.	she.	sofa	train	tv	mean
DSN [2]	83.8	73.6	76.0	61.4	69.2	84.2	74.8	82.0	53.5	73.7	45.3	81.9	79.9	73.0	83.5	55.0	77.2	51.9	80.6	66.7	71.4
CASENet [2]	84.8	72.8	77.9	62.6	70.9	83.5	73.4	81.7	54.7	75.6	44.8	82.6	82.0	74.0	83.0	53.5	77.8	51.7	78.7	63.8	71.5
SEAL [23]	85.5	74.9	80.9	64.7	70.4	85.9	76.5	84.3	58.3	74.2	47.7	84.0	82.4	76.1	**85.7**	59.1	80.1	54.0	81.1	67.1	73.7
DDS-R	86.6	76.4	79.7	65.7	72.7	**86.0**	77.3	83.4	58.5	77.5	51.7	83.4	82.6	76.5	84.9	59.6	80.4	55.2	81.5	69.6	74.5
DDS-U	**88.2**	**77.1**	**82.4**	**67.9**	**73.0**	85.6	**79.2**	**85.2**	**60.6**	**80.5**	**53.2**	**84.2**	**84.0**	**77.5**	85.5	**62.9**	**83.2**	**56.8**	**82.4**	**71.7**	**76.1**

- ***DSN***, which directly applies the deeply supervised network architecture, in which each side of the backbone network is connected to a 1×1 *conv* layer with K channels for SED, and the resulting activation maps from all sides are fused to generate the final semantic edges.
- ***CASENet+S4***, which is similar to CASENet but takes into consideration Side-4 connected to a 1×1 *conv* layer to produce a single-channel feature map, while CASENet only uses Side-1 $\sim$ Side-3 and Side-5.
- ***DDS\Convt***, which removes the *information converters* in DDS, such that deep supervision is directly imposed after each side.
- ***DDS\DeSup***, which removes the deep supervision from Side-1 $\sim$ Side-4 of DDS but retains the *information converters*.

All these variants are trained using the reweighted loss function Eq. (6) (except *Softmax*) and the original SBD data for fair comparison.

We evaluate these variants and the original DDS and CASENet [2] methods on the SBD test dataset. The evaluation results are shown in Table 1. We can see that *Softmax* suffers from significant performance degradation. Because the predicted semantic edges of neural networks are usually thick and overlap with other classes, it is improper to assign a single label to each pixel. Hence, we apply multi-label loss in Eq. (4) and Eq. (5). The *Basic* network achieves an ODS F-measure of 70.6%, which is 0.3% higher than *DSN*. This further verifies our hypothesis presented in Section 3 that features from the bottom layers are not sufficiently discriminative for semantic classification. Furthermore, *CASENet+S4* performs better than *DSN*, demonstrating that the bottom convolutional features are more suited to binary edge detection. Moreover, the F-measure of *CASENet+S4* is lower than the original CASENet. In addition, the improvement from *DDS\DeSup* to DDS-R shows that the success of DDS does not arise due to more parameters (*conv* layers) but instead from the coordination between deep supervision and *information converters*. Adding more *conv* layers but without deep supervision may make network convergence more difficult. Our conclusion is consistent with [2], when comparing *DDS\Convt* with the results of CASENet, namely that there is no value in directly adding binary edge supervision to the bottom sides.

Intuitively, employing different but "appropriate" ground-truths to the bottom and top sides may enhance the feature learning in different layers. Upon this, the learned intermediate representations of different levels will tend to contain complementary information. However, in our case, it may be unnecessary to directly add deep supervision of category-agnostic edges to the bottom sides, because less discriminative gradient signals are likely to arise due to the discrepancy in the loss function of Eq. (6). Instead, we show that with proper architecture re-design, we can employ deep supervision to significantly boost performance. The *information converters* adopted in the proposed method play a central role in guiding lower layers for category-agnostic edge detection. In this way, low-level edges from the bottom layers encode more details, which then assist the top layers to better localize semantic edges. They also serve to generate consistent gradient signals from higher layers. This is essential, as they enable direct influence of the hidden layer weight/filter update process to favor highly discriminative feature maps for correct SED.

The significant performance improvement provided by our proposed DDS-R/DDS-U over *CASENet+S4* and *DDS\Convt* demonstrates the importance of our design, in which different sides use different supervision after the information format conversion. We also note that DDS-U achieves better performance than DDS-R by applying the

TABLE 4
The evaluation results (%) of some competitors on the Cityscapes validation set [20]. The best performance of each column is highlighted in **bold**.

Method	road	sid.	bui.	wall	fen.	pole	light	sign	veg.	ter.	sky	per.	rider	car	tru.	bus	tra.	mot.	bike	mean
DSN [2]	87.8	82.5	83.2	55.2	57.5	81.4	75.9	78.9	86.6	66.1	82.3	87.9	76.2	91.0	55.4	73.2	53.9	61.6	85.4	74.8
CASENet [2]	87.2	82.2	83.0	53.7	57.9	82.9	78.7	79.2	86.0	65.8	82.7	88.0	77.1	90.3	50.6	72.1	56.1	63.5	85.3	74.9
PSPNet [55]	58.7	79.9	73.0	58.4	59.8	79.3	75.3	75.5	76.7	66.0	70.2	80.1	74.6	84.2	**63.1**	76.6	**70.3**	64.5	76.1	71.7
DeepLabv3+ [56]	39.2	32.8	39.5	9.0	7.0	25.2	12.5	19.6	34.6	10.2	23.6	22.7	12.0	22.4	2.3	11.1	9.5	6.0	14.0	18.6
SEAL [23]	88.1	84.5	83.4	55.3	57.2	83.6	78.6	79.7	87.3	69.0	83.5	86.8	77.8	87.2	54.5	73.1	49.0	61.8	85.3	75.0
DDS-R	**90.5**	84.2	86.2	57.7	61.4	85.1	83.8	80.4	88.5	67.6	**88.2**	89.9	80.1	**91.8**	58.6	76.3	56.2	68.8	**87.3**	78.0
DDS-U	90.3	**85.3**	**86.7**	**58.8**	**61.5**	**86.9**	**84.7**	**83.0**	**89.3**	**69.8**	**88.2**	**90.3**	**80.5**	91.7	62.5	**77.4**	61.5	**70.5**	**87.3**	**79.3**

TABLE 5
The ODS F-measure (%) of some competitors on the VOC2012 validation set [21]. The best performance of each column is highlighted in **bold**.

Method	aer.	bike	bird	boat	bot.	bus	car	cat	cha.	cow	tab.	dog	hor.	mot.	per.	pot.	she.	sofa	train	tv	mean
DSN [2]	83.5	**60.5**	81.8	58.0	66.4	82.7	69.9	83.0	49.7	78.6	50.8	78.4	74.7	74.1	82.0	55.0	79.9	55.2	78.3	68.6	70.5
CASENet [2]	84.6	60.1	82.7	59.2	68.1	84.3	69.9	83.5	51.9	81.2	50.4	80.4	76.7	74.4	81.9	55.8	82.0	54.9	77.8	67.0	71.3
SEAL [23]	85.2	60.0	84.4	**61.8**	70.3	85.5	71.7	83.7	53.8	82.1	50.1	**81.4**	76.8	75.4	83.7	59.1	80.9	54.4	78.7	72.2	72.6
DDS-R	86.3	58.2	86.0	60.2	71.6	85.2	72.6	83.0	53.0	82.1	54.0	79.4	77.8	74.9	83.5	57.3	81.7	53.6	79.7	71.0	72.6
DDS-U	**87.1**	60.0	**86.6**	60.8	**72.6**	**87.0**	**73.2**	**85.3**	**56.5**	**83.9**	**55.8**	80.3	**79.6**	**75.9**	**84.5**	**61.7**	**85.1**	**57.0**	**80.5**	**74.0**	**74.4**

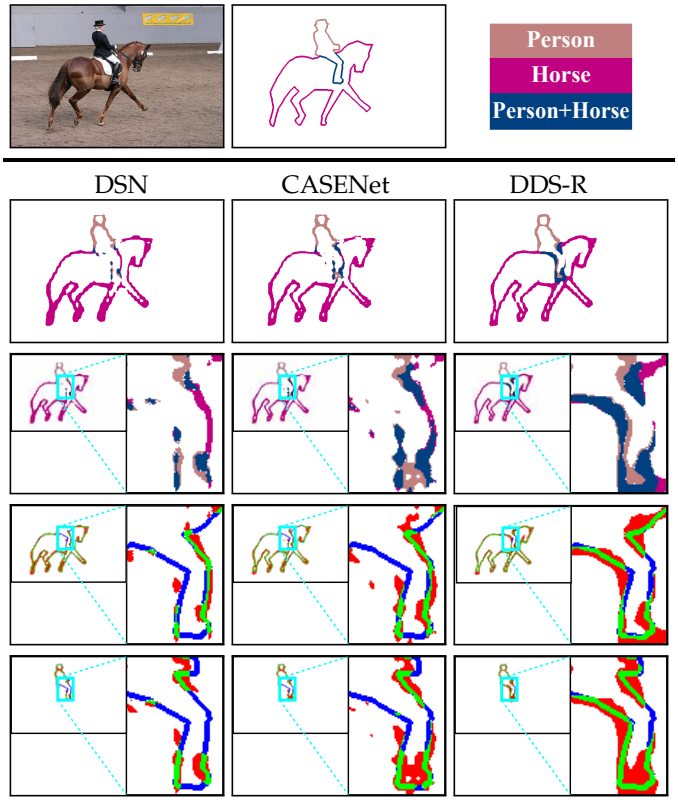


Fig. 4. A qualitative comparison of DSN, CASENet and DDS-R. First row: the original image, ground truth, and category color codes. This image is taken from the SBD [1] test set. Second row: the semantic edges predicted by different methods. Third row: an enlarged area of predicted edges. Fourth row: the predicted horse boundaries only. Last row: the predicted person boundaries only. Green, red, white, and blue pixels represent true positive, false positive, true negative, and false negative points, respectively, at the threshold of 0.5. Best viewed in color.

unweighted loss function and aligned edges [23]. After exploring DDS with several variants and establishing the effectiveness of the approach, we summarize the results obtained by our method and compare it with several state-of-the-art methods.

5.2 Evaluation on SBD

We compare DDS-R/DDS-U on the SBD dataset with several state-of-the-art methods including InvDet [1], HFL-FC8 [4], HFL-CRF [4], BNF [54], WS [47], DilConv [46], DSN [2], COB [3], CASENet [2] and SEAL [23]. Results are summarized in Table 2.

InvDet is a non-deep learning based approach which shows competitive results among other conventional approaches. COB is a state-of-the-art category-agnostic edge detection method, and combining it with semantic segmentation of DilConv produces a competitive semantic edge detector [3]. COB's superiority over DilConv reflects the effectiveness of the fusion algorithm in [3]. The fact that both CASENet and DDS-R/DDS-U outperform COB illustrates the importance of directly learning semantic edges, because the combination of binary edges and semantic segmentation is insufficient for SED. DDS-U achieves the state-of-the-art performance across all competitors, and outperforms other methods on 16 of 20 classes. The ODS F-measure of the proposed DDS-U is 1.7% higher than SEAL and 3.4% higher than CASENet, so delivering a new state-of-the-art. The average runtimes of DSN, CASENet, and DDS are shown in Table 6. Hence DDS can generate state-of-the-art semantic edges with only a slight reduction in speed.

Yu *et al.* [23] discovered that some of the original SBD labels are a little noisy, so they re-annotated 1059 images from the test set to form a new test set. We compare our method with DSN [2], CASENet [2] and SEAL [23] on this new dataset. Both DDS-R and DDS-U achieve better performance than previous methods. Specifically, the mean ODS F-measures of DDS-R and DDS-U are 0.8% and 2.4% higher than recent SEAL [23], respectively. Note that SEAL retrains CASENet with a new training strategy: *i.e.* simultaneous alignment and learning. With the same training strategy, DDS-R obtains a 3.0% higher F-measure than CASENet.

To better visualize the edge prediction results, an example is shown in Fig. 4. We also show the normalized images

Fig. 5. Some examples from SBD dataset [1]. **From top to bottom:** color codes, original images, ground truth, DSN, CASENet [2], SEAL [23], our DDS-R and DDS-U. We follow the color coding protocol in [23], [57].

of side activation in Fig. 6. All activations are obtained before *sigmoid* non-linearization. For simpler arrangement of the figures, we do not display Side-4 activation of DDS-R. From Side-1 to Side-3, one can see that the feature maps of DDS-R are significantly clearer than for DSN and CASENet. Clear category-agnostic edges can be found with DDS-R, while DSN and CASENet suffer from noisy activation. For example, in CASENet, without imposing deep supervision on Side-1 $\sim$ Side-3, edge activation can barely be found. For category classification activation, DDS-R can separate horse and person clearly, while DSN and CASENet can not. Therefore, the *information converters* also help to better optimize Side-5 for category-specific classification. This further verifies the feasibility of the proposed DDS architecture.

TABLE 6
The average runtime per image on the SBD dataset [1].

Method	DSN	CASENet	SEAL	DDS
Time (s)	0.171	0.166	0.166	0.175

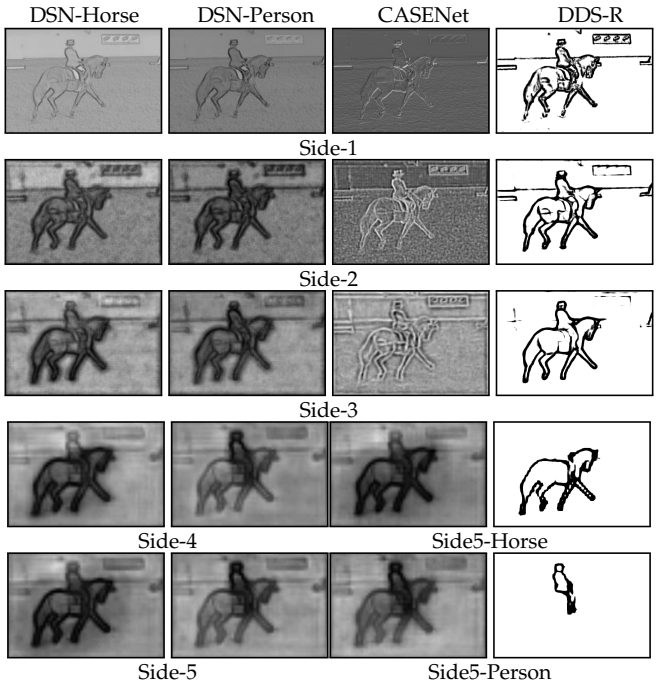


Fig. 6. Side activation maps on the input image of Fig. 4. The first two columns display DSN's side class classification activation for the classes of horse and person, respectively. The last two columns show the side features of Side-1 $\sim$ Side-3 and class classification activation of Side-5 for CASENet and our DDS-R, respectively. These images are obtained by normalizing the activation to [0, 255]. Note that all activation are directly outputted without any non-linearization, *e.g. sigmoid* function.

More qualitative examples are displayed in Fig. 5. DDS-R/DDS-U can produce clearer and smother edges than the other detectors. In the second column, it is interesting to note that most detectors can recognize the boundaries of the objects with missing annotations, *i.e.* , the obscured dining table and human arm. In the third column, DDS-R/DDS-U can generate strong responses at the boundaries of the small cat, while the other detectors only have weak responses. This demonstrates that DDS is more robust for detecting small objects. We also find that DDS-U and SEAL can generate thinner edges, suggesting that training with regular unweighted *sigmoid* cross entropy loss and refined ground truth edges is helpful for accurately locating thin boundaries.

5.3 Evaluation on Cityscapes

The Cityscapes [20] dataset is more challenging than SBD [1]. The images in Cityscapes are captured in more complicated scenes, usually in urban street scenes in different cities. There are more objects, especially overlapping objects, in each image. Thus, Cityscapes may be better for testing semantic edge detectors. We compare DDS not only with three semantic edge detectors including DSN [2], CASENet [2] and SEAL [23], but also with two state-of-the-art semantic segmentation models, PSPNet [55] and DeepLabv3+ [56]. We extract the semantic segment boundaries for PSPNet and DeepLabv3+ to generate their corresponding semantic edges of single pixel width [1]. The evaluation results are reported in Table 4. Both DDS-R and DDS-U significantly outperform the other methods. PSPNet [55] is competitive in terms of SED but performs less well than the edge detectors. Although DeepLabv3+ [56] is much better than PSPNet [55] in terms of semantic segmentation, DeepLabv3+ [56] performs surprisingly worse than its competitors for SED. This suggests that semantic segmentation cannot always generate reliable boundaries and that further SED research is necessary. With the same loss function, the mean ODS F-measure of DDS-R is 3.1% higher than CASENet, and DDS-U is 4.3% higher than SEAL. Some qualitative comparisons are shown in Fig. 7. We can see that DDS-R/DDS-U produces smoother and clearer edges.

5.4 Evaluation on PASCAL VOC2012

VOC2012 [21] contains the same object categories as SBD dataset [1]. For the validation set, we exclude the images that appear in the SBD training set, resulting in a new validation set containing 904 images. We use the new validation set to test some competitors with the model trained on SBD. In this way, we can test the generalizability of the various methods. However, the original annotations of VOC2012 leave a thin unlabeled area near each object boundary, affecting the evaluation. Instead, we follow the methodology in [15] and employ a dense CRF model [58] to fill the uncertain area with the neighboring object labels. We further follow [1] to generate semantic edges of single pixel width. The evaluation results are summarized in Table 5. DDS-U achieves the best performance, as expected, indicating that the DDS network has good generalizability.

6 Conclusion

In this paper, we study the SED problem. Previous methods suggest that deep supervision is not necessary [2], [23] for SED. Here we show that this is false and, with proper architecture re-design, that the network can be deeply supervised to improve detection results. The core of our approach is the introduction of novel *information converters*, which play a central role in guiding the lower layers for category-agnostic edge detection. They also help to generate consistent gradient signals with the ones arising from category-aware edge detection in the higher layers. DDS is essential, as it enables influencing the hidden layer weight/filter update process directly to favor highly discriminative feature maps for improved SED. DDS achieves the state-of-the-art performance on several datasets including SBD [1], Cityscape [20], and PASCAL VOC2012 [21]. Our idea to leverage deep supervision for training a deep network opens a path towards putting more emphasis utilizing rich feature hierarchies from deep networks for SED as well as other high-level tasks such as semantic segmentation [3], [59], object detection [3], [7], and instance segmentation [60], [61].

Future Work. Besides category-agnostic edge detection and SED, relevant tasks are commonly exist in computer vision [62], such as segmentation and saliency detection, object detection and keypoint detection, edge detection and skeleton extraction. Building multi-task networks to solve relevant tasks is a good way to save computation resources in practical applications [63]. However, distracted supervision between different tasks usually prevent this target as shown in this paper. From this point of view, the proposed DDS

Fig. 7. Some examples from Cityscapes dataset [20]. **From top to bottom:** color codes, original images, ground truth, DSN, CASENet [2], SEAL [23], our DDS-R and DDS-U. We follow the color coding protocol in [23], [57]. We can see that the produced edges of DDS are smoother and clearer.

provides a new perspective to multi-task learning. In the future, we plan to leverage the idea of *information converter* for more relevant tasks.

ACKNOWLEDGMENTS

This research was supported by NSFC (NO. 61572264), the national youth talent support program, Tianjin Natural Science Foundation for Distinguished Young Scholars (NO. 17JCJQJC43700), Tianjin key S&T Projects on new generation AI, and Huawei Innovation Research Program.

REFERENCES

[1] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Int. Conf. Comput. Vis.*, 2011, pp. 991–998.

[2] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam, "CASENet: Deep category-aware semantic edge detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 5964–5973.

[3] K.-K. Maninis, J. Pont-Tuset, P. Arbelaez, and L. Van Gool, "Convolutional oriented boundaries: From image segmentation to high-level tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[4] 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 *Int. Conf. Comput. Vis.*, 2015, pp. 504–512.

[5] M. R. Amer, S. Yousefi, R. Raich, and S. Todorovic, "Monocular extraction of 2.1 d sketch using constrained convex optimization," *Int. J. Comput. Vis.*, vol. 112, no. 1, pp. 23–42, 2015.

[6] Q. Shan, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, "Occluding contours for multi-view stereo," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 4002–4009.

[7] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 36–51, 2008.

[8] V. Ferrari, F. Jurie, and C. Schmid, "From images to shape models for object detection," *Int. J. Comput. Vis.*, vol. 87, no. 3, pp. 284–303, 2010.

[9] S. Ramalingam, S. Bouaziz, P. Sturm, and M. Brand, "Skyline2gps: Localization in urban canyons using omni-skylines," in *IEEE\RSJ Int. Conf. Intell. Robot. Syst.*, 2010, pp. 3816–3823.

[10] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Int. Conf. Comput. Vis.*, 2015, pp. 1395–1403.

[11] ——, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, no. 1-3, pp. 3–18, 2017.

[12] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3000–3009.

[13] 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.*, 2018.

[14] X. Hu, Y. Liu, K. Wang, and B. Ren, "Learning hybrid convolutional features for edge detection," *Neurocomputing*, 2018.

[15] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang, "Object contour detection with a fully convolutional encoder-decoder network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 193–202.

[16] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1–9.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[19] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Int. Conf. Comput. Vis.*, 2015, pp. 3119–3127.

[20] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 3213–3223.

[21] 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, 2012.

[22] P. Arbeláez, 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.

[23] Z. Yu, W. Liu, Y. Zou, C. Feng, S. Ramalingam, B. Kumar, and J. Kautz, "Simultaneous edge alignment and learning," in *Eur. Conf. Comput. Vis.*, 2018, pp. 400–417.

[24] I. Sobel, "Camera models and machine perception," Stanford Univ Calif Dept of Computer Science, Tech. Rep., 1970.

[25] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.

[26] M. Mafi, H. Rajaei, M. Cabrerizo, and M. Adjouadi, "A robust edge detection approach in the presence of high impulse noise intensity through switching adaptive median and fixed weighted mean filtering," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5475–5490, 2018.

[27] P.-L. Shui and F.-P. Wang, "Anti-impulse-noise edge detection via anisotropic morphological directional derivatives," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4962–4977, 2017.

[28] P. E. Trahanias and A. N. Venetsanopoulos, "Color edge detection using vector order statistics," *IEEE Trans. Image Process.*, vol. 2, no. 2, pp. 259–264, 1993.

[29] R. C. Hardie and C. G. Boncelet, "Gradient-based edge detection using nonlinear edge enhancing prefilters," *IEEE Trans. Image Process.*, vol. 4, no. 11, pp. 1572–1577, 1995.

[30] P. V. Henstock and D. M. Chelberg, "Automatic gradient threshold determination for edge detection," *IEEE Trans. Image Process.*, vol. 5, no. 5, pp. 784–787, 1996.

[31] 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.*, vol. 25, no. 1, pp. 57–74, 2003.

[32] 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.*, vol. 26, no. 5, pp. 530–549, 2004.

[33] J. J. Lim, C. L. Zitnick, and P. Dollár, "Sketch tokens: A learned mid-level representation for contour and object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3158–3165.

[34] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, 2015.

[35] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?" *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017–5032, 2015.

[36] H. Lee and H. Kwon, "Going deeper with contextual cnn for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, 2017.

[37] P. Tang, X. Wang, B. Feng, and W. Liu, "Learning multi-instance deep discriminative patterns for image classification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3385–3396, 2017.

[38] Y. Liu, P.-T. Jiang, V. Petrosyan, S.-J. Li, J. Bian, L. Zhang, and M.-M. Cheng, "DEL: deep embedding learning for efficient image segmentation." in *Int. Joint Conf. Artif. Intell.*, 2018.

[39] 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.

[40] R. Deng, C. Shen, S. Liu, H. Wang, and X. Liu, "Learning to predict crisp boundaries," in *Eur. Conf. Comput. Vis.*, 2018, pp. 570–586.

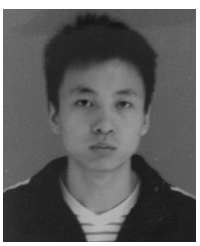
[41] W. Yang, J. Feng, J. Yang, F. Zhao, J. Liu, Z. Guo, and S. Yan, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5895–5907, 2017.

[42] Y. Ganin and V. Lempitsky, "N^4-Fields: Neural network nearest neighbor fields for image transforms," in *Asian Conf. Comput. Vis.*, 2014, pp. 536–551.

[43] G. Bertasius, J. Shi, and L. Torresani, "DeepEdge: A multi-scale bifurcated deep network for top-down contour detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4380–4389.

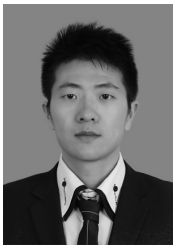
[44] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3982–3991.

[45] I. Kokkinos, "Pushing the boundaries of boundary detection using deep learning," in *Int. Conf. Learn. Represent.*, 2016.
[46] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Int. Conf. Learn. Represent.*, 2016.
[47] A. Khoreva, R. Benenson, M. Omran, M. Hein, and B. Schiele, "Weakly supervised object boundaries," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 183–192.
[48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
[49] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[50] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.
[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.
[52] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[53] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[54] G. Bertasius, J. Shi, and L. Torresani, "Semantic segmentation with boundary neural fields," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 3602–3610.
[55] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2881–2890.
[56] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.
[57] Z. Yu, W. Liu, Y. Zou, C. Feng, S. Ramalingam, B. V. Kumar, and J. Kautz, "Simultaneous edge alignment and learning," *arXiv preprint arXiv:1808.01992*, 2018.
[58] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *Adv. Neural Inform. Process. Syst.*, 2011, pp. 109–117.
[59] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 4545–4554.
[60] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother, "Instancecut: from edges to instances with multicut," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 5008–5017.
[61] Z. Hayder, X. He, and M. Salzmann, "Boundary-aware instance segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 5696–5704.
[62] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3712–3722.
[63] Q. Hou, J. Liu, M.-M. Cheng, A. Borji, and P. H. Torr, "Three birds one stone: A unified framework for salient object segmentation, edge detection and skeleton extraction," *arXiv preprint arXiv:1803.09860*, 2018.



Yun Liu is a PhD candidate at College of Computer Science, Nankai University. He received his bachelor degree from Nankai University in 2016. His research interests include computer vision and machine learning.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. Then he did 2 years research fellow, with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests includes computer graphics, computer vision, and image processing. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, and CCF-Intel Young Faculty Researcher Program.



Deng-Ping Fan received his BS and MS degree from the Fujian Agriculture and Forestry University, Fujian, China, 2010 and Guanxi Normal University, Guanxi, China, 2015, respectively. He currently is a Ph.D candidate in the Nankai University, Tianjin, China. He is working with Prof. Ming-Ming Chen. His research interests includes computer vision and deep learning, especially in salient object detection, metric, and sketch synthesis.

Le Zhang received the B.Eng degree from University of Electronic Science and Technology Of China in 2011. He received his M.Sc and Ph.D.degree form Nanyang Technological University (NTU) in 2012 and 2016, respectively. Currently, he is a scientist at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. His current research interests include deep learning and computer vision.

Jia-Wang Bian received his B.Eng degree from Nankai University in 2016. Then he worked in Singapore University of Technology and Design (SUTD) as a research assistant, and worked in Advanced Digital Sciences Center (ADSC) as a training engineer at the same time. His current research interests lie in the field of computer vision.

Dacheng Tao is Professor of computer science and an ARC Laureate Fellow with the School of Information Technologies and with the Faculty of Engineering and Information Technologies, and the Inaugural Director of the UBTECH Sydney Artificial Intelligence Centre, at the University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science. His research results have expounded in one monograph and 200+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, IJCV, CVPR; and ACM SIGKDD, with several best paper awards. He is a fellow of the Australian Academy of Science, AAAS, IEEE, IAPR, OSA, and SPIE.