ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection

Yuxin Wang, Hongtao Xie*, Zhengjun Zha, Mengting Xing, Zilong Fu and Yongdong Zhang
University of Science and Technology of China
{wangyx58,metingx,JeromeF}@mail.ustc.edu.cn,{htxie,zhazj,zhyd73}@ustc.edu.cn

Abstract

Scene text detection has witnessed rapid development in recent years. However, there still exists two main challenges: 1) many methods suffer from false positives in their text representations; 2) the large scale variance of scene texts makes it hard for network to learn samples. In this paper, we propose the ContourNet, which effectively handles these two problems taking a further step toward accurate arbitrary-shaped text detection. At first, a scale-insensitive Adaptive Region Proposal Network (Adaptive-RPN) is proposed to generate text proposals by only focusing on the Intersection over Union (IoU) values between predicted and ground-truth bounding boxes. Then a novel Local Orthogonal Texture-aware Module (LOTM) models the local texture information of proposal features in two orthogonal directions and represents text region with a set of contour points. Considering that the strong unidirectional or weakly orthogonal activation is usually caused by the monotonous texture characteristic of false-positive patterns (e.g. streaks.), our method effectively suppresses these false positives by only outputting predictions with high response value in both orthogonal directions. This gives more accurate description of text regions. Extensive experiments on three challenging datasets (Total-Text, CTW1500 and ICDAR2015) verify that our method achieves the state-of-the-art performance. Code is available at https://github.com/wangyuxin87/ContourNet.

1. Introduction

Scene text detection is a task to detect text regions in the complex background and label them with bounding boxes. Accurate detection result benefits a wide scope of real-world applications and is the fundamental step for end-to-end text recognition [36, 5, 39, 24].

Benefiting from the development of deep learning, recent methods have achieved significant improvement in scene text detection task. Meanwhile, the research focus has shifted from horizontal texts [48, 14] to multi-oriented texts [25, 49] and more challenging arbitrary-shaped texts [34, 35] (*e.g.* curved texts). However, due to the specific properties of scene text such as large variance in color, texture, scale, etc., there are still two challenges to be addressed in arbitrary-shaped scene text detection.

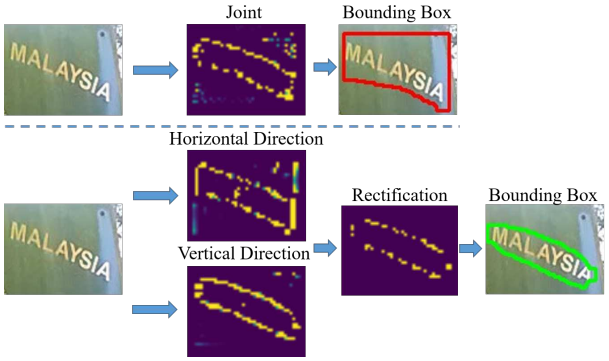


Figure 1. Comparison between jointly modeling the texture information in arbitrary orientations and respectively modeling texture information in two orthogonal directions. We visualize the heatmap of predicted contour points. NMS is implemented to each heatmap for better visualization. FPs can effectively be suppressed by considering the response in two orthogonal directions simultaneously.

The first challenge is false positives (FPs), which is not received enough attention in recent researches [38] and is regarded as one of the key bottlenecks for more accurate arbitrary-shaped scene text detection in this paper. Recent CNN-based methods jointly model the texture information in arbitrary orientations by using $k \times k$ convolutional kernel [46, 43]. However, this operation is sensitive to some specific cases containing similar texture characteristics to text regions, and tends to perform identically high response to these cases (see in top of Fig.1). SPCNET [38] attributes this problem to the lack of context information clues and inaccurate classification scores, thus a text context module is proposed to compensate global semantic feature and bounding boxes are further rectified by the segmentation map. Li-

*Corresponding author

u *et al.* [21] re-score the detection results with the confidence of four vertexes to supervise the compactness of the bounding boxes. Different from these methods, we handle the FPs by using only local texture information, which is a more straightforward approach and contains less computation. As shown in bottom of Fig.1, our motivation mainly comes from two observations: 1) FPs with strong unidirectional texture characteristics are weakly activated in its orthogonal direction (*e.g*. some vertical streaks); 2) FPs can be effectively suppressed by considering the responses in both orthogonal directions simultaneously. Thus, it is reasonable to model the texture information along two orthogonal directions. Inspired by traditional edge detection operators (*e.g*. Sobel, etc.), we heuristically use horizontal and vertical directions in our approach.

The second challenge is the large scale variance of scene texts. Compared with general objects, the scale variation is much larger in scene texts, which makes it hard for CNN-based methods to learn samples. To address this problem, MSR [43] uses a multi-scale network to obtian powerful representation of texts with various scales. DSRN [36] attributes this problem to the inconsistent activation of multi-scale texts, thus a bi-directional operation is proposed to map the convolutional features to a scale-invariant space. Different from these methods solving the large scale variance problem through aggregation of multi-scale features, we pay attention to the shape information and use a scale-invariant metric to optimize our network.

In this paper, we propose a novel text detector to effectively solve these two problems achieving accurate arbitrary-shaped scene text detection, which is called ContourNet. As shown in Fig.2, given an input image, *Adaptive Region Proposal Network* (Adaptive-RPN) first generates text proposals by automatically learning a set of boundary points over the text region that indicate the spatial extend of text instance. The training object of Adaptive-RPN is driven by *IoU* values between the predicted and ground-truth bounding boxes, which is invariant to the scale [27, 49]. Thus, Adaptive-RPN is insensitive to the large scale variance of scene texts and can automatically account for shape information of text regions to achieve finer localization compared with conventional RPN approaches [26, 8]. To capture the distinct texture characteristics in contour regions of texts, we propose a *Local Orthogonal Texture-aware Module* (LOTM) to model the local texture information of proposal features in two orthogonal directions, and represent text region with contour points in two different heatmaps, either of which only responds to the texture characteristics in a certain direction. Finally, *Point Rescoring Algorithm* effectively filters predictions with strong unidirectional or weakly orthogonal activation by considering the response in both orthogonal directions simultaneously. In this way, text regions are detected and represented with a set of high-quality contour points.

The contributions of this paper are three-fold: 1) We propose a novel approach for FP suppression by modeling the local texture information in two orthogonal directions, which is a more straightforward approach and contains less computation compared with previous methods. 2) The proposed Adaptive-RPN effectively handles the large scale variance problem and achieves finer localization of text regions, which can be easily embedded into existing approaches. 3) Without external data for training, the proposed method achieves **85.4%** and **83.9%** in F-measure on Total-Text and CTW1500 dataset with 3.8 FPS and 4.5 FPS respectively, which outperforms recent counterparts by a large margin.

2. Related Works

Scene text detection has been a popular research topic for a long time with many approaches proposed [48, 30, 44, 46, 25, 42, 34, 32, 33]. Conventionally, connected component (CC) based and sliding window based methods have been widely used in text localization [48, 30, 44]. As deep learning becomes the most promising machine learning tool [40, 17, 18, 47], scene text detection has achieved remarkable improvement in recent years. These methods can be divided into two categories: regression based methods and segmentation based methods.

Regression based methods [29, 49], inspired by general object detection methods [7, 19, 8], localize text boxes by predicting the offsets from anchors or pixels. Lyu *et al.* [25] adopt a similar architecture as SSD and rebuild text instance with predicted corner points. Wang *et al.* [35] use recurrent neural network (RNN) for text region refinement and adaptively predict several pairs of points to represent arbitrary-shaped text. Different from these methods localizing text regions by implementing refinement on pre-defined anchors, EAST [49] and DDR [10] propose a new approach for accurate and efficient text detection, which directly regress the offsets from boundaries or vertexes to current point. Based on these direct regression methods, LOMO [46] proposes an iterative refinement module to iteratively refine bounding box proposals for extremely long texts, and then predicts center line, text region, and border offsets to rebuild text instance.

Segmentation based methods [23, 34] are mainly inspired by FCN [22]. Recent segmentation based methods usually use different representation to describe text region, and then rebuild text instance through specific post-procession. PixelLink [4] predicts connections between pixels and localizes text region by separating the links belonging to different text instances. To handle the adjacent texts, Tian *et al.* [32] design a two-step clustering to split dense text instances from segmentation map. PSENet [34] gradually expands kernels at certain scale to split the close

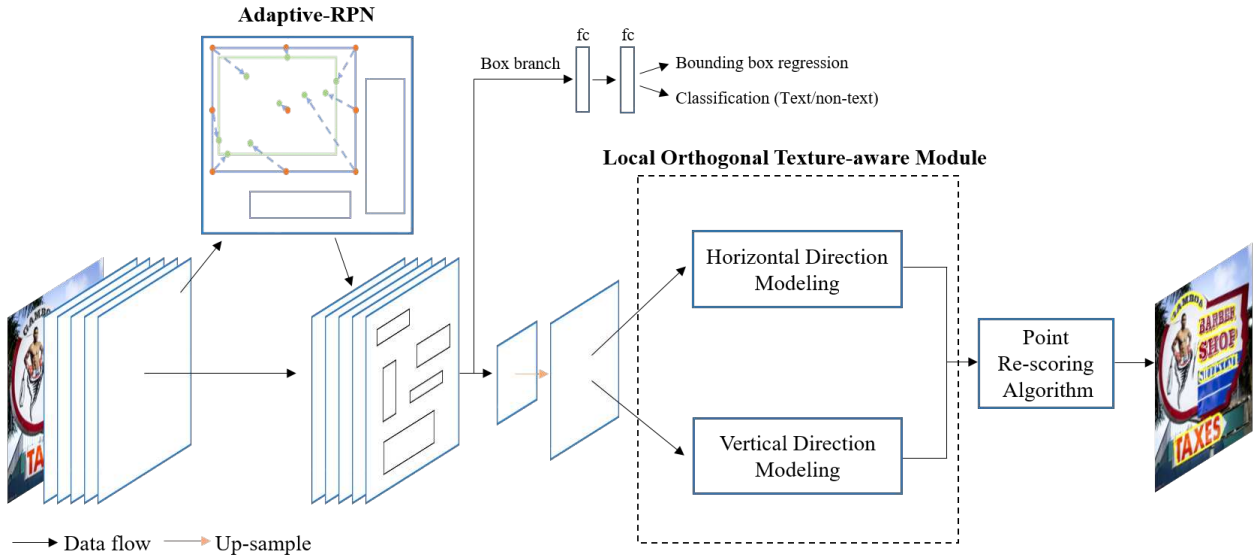


Figure 2. The pipeline of ContourNet. It mainly contains three parts: Adaptive Region Proposal Network (Adaptive-RPN), Local Orthogonal Texture-aware Module (LOTM) and *Point Re-scoring Algorithm*. The box branch is similar to other 2-stage methods.

text instances.

Our method integrates the advantages of regression based methods and segmentation based methods, which adopts a two-stage architecture and represents text region with contour points. Benefiting from Adaptive-RPN and FP suppression, our method effectively handles the large scale variance problem and gives more accurate description of text regions compared with previous methods.

3. Proposed Method

The proposed method mainly consists of three parts: Adaptive-RPN, LOTM and *Point Re-scoring Algorithm*. In this section, we first briefly describe the overall pipeline of the proposed method, and then detail the motivation and implementation of these three parts respectively.

3.1. Overall pipeline

The architecture of our ContourNet is illustrated in Fig.2. First, a backbone network is constructed to generate shared feature maps. Inspired by FPN [16] which can obtain strong semantic features for multi-scale targets, we construct a backbone with FPN-like architecture by implementing lateral connections in the decoding layer. Next, we propose an Adaptive-RPN described in Sec.3.2 for proposal generation by bounding spatial extent of several refined points. The input of LOTM are proposal features obtained by using Deformable RoI pooling [50] and bilinear interpolation to the shared feature maps. Then, LOTM decodes the contour points from proposal features by modeling the local texture information in horizontal and vertical directions respectively. Finally, a *Point Re-scoring Algorithm* is used to filter FPs by considering the responses in both directions simultaneously. The details of LOTM and *Point Re-scoring Algorithm* are presented at Sec.3.3 and 3.4 respectively. Bounding box regression and classification (text/non-text) in box branch are similar to other 2-stage methods, which are used to further refine bounding boxes.

3.2. Adaptive Region Proposal Network

Region Proposal Network is wildly used in existing object detection methods. It aims to predict a 4-d regression vector $\{\Delta x, \Delta y, \Delta w, \Delta h\}$ to refine current bounding box proposal $B_c = \{x_c, y_c, w_c, h_c\}$ to a predicted bounding box $B_t = \{x_c + w_c\Delta x_c, y_c + h_c\Delta y_c, w_c e^{\Delta w_c}, h_c e^{\Delta h_c}\}$, and the training objective is to optimize the smooth l_1 loss [26].

As an approach proposed to improve the *IoU* value between predicted and ground-truth bounding boxes, this aforementioned $4-d$ representation optimized by l_n-norm loss is sensitive to the scale variation. In general, positive bounding boxes are selected through an *IoU* metric (*e.g.* $IoU > 0.5$). However, several pairs of bounding boxes in different scales with the same *IoU* value may have different l_n-norm distances. As there is not a powerful correlation between optimizing l_n-norm loss and improving their *IoU* values [27], we infer that this gap makes it hard for CNN-based methods to learn samples with large scale variance in scene text detection.

To handle this problem, we propose a new Adaptive-RPN to focus on only *IoU* values between predicted and ground-truth bounding boxes which is a scale-invariant metric, and use a set of pre-defined points $P = \{(x_l, y_l)\}_{l=1}^{n}$ (1 center point and $n-1$ boundary points) instead of the 4-d vector for the proposal representation. The refinement can be expressed as:

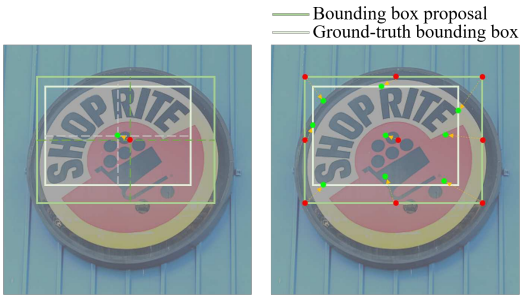


Figure 3. The comparison between conventional RPN (left) and Adaptive-RPN (right). The proposed Adaptive-RPN adaptively regresses the offsets to pre-defined points. Predicted bounding box is generated by bounding the spatial extend of refined points. Red points are pre-defined points in current bounding box proposal (*e.g.* center point in conventional RPN and pre-defined points *P* in Adaptive-RPN), and green points are refined points. The yellow dotted lines indicate the regressed offsets.

$$R = \{x_r, y_r\}_{r=1}^{n} = \{(x_l + w_c \Delta x_l, y_l + h_c \Delta y_l)\}_{l=1}^{n} \quad (1)$$

Where $\{\Delta x_l, \Delta y_l\}_{l=1}^{n}$ are the predicted offsets to pre-defined points, w_c and h_c are width and height of current bounding box proposal. As shown in Fig.3, the predicted offsets are used to process a local refinement on n pre-defined points in current bounding box proposal. Then, we use a max-min function in eq.(2) to bound these refined points with 4 extreme points for the representation of predicted bounding box. Specially, the center point $\{x', y'\}$ is used to normalize the bounding box (*e.g.* if $x_{tl} > x'$, then $x_{tl} = x'$).

$$\begin{aligned} Proposal =& \{x_{tl}, y_{tl}, x_{rb}, y_{rb}\} \\ =& \{min\{x_r\}_{r=1}^{n}, min\{y_r\}_{r=1}^{n}, \\ & max\{x_r\}_{r=1}^{n}, max\{y_r\}_{r=1}^{n}\} \end{aligned} \quad (2)$$

Compared with conventional RPN that considers only rectangular spatial scope, the proposed Adaptive-RPN automatically accounts for shape and semantically important local areas for finer localization of text regions. Without additional supervision, we optimize the regression loss in Adaptive-RPN through an *IoU* loss (see in eq.(4)) by calculating the overlapping between the predicted and ground-truth bounding boxes.

3.3. Local Orthogonal Texture-aware Module

Inspired by traditional edge detection operators (*e.g.* Sobel, etc.) which have achieved remarkable performance before deep learning becomes the most promising machine learning tool, we skillfully incorporate the idea of traditional edge detection operators into LOTM and represent text region with a set of contour points. These points containing strong texture characteristics can accurately localize

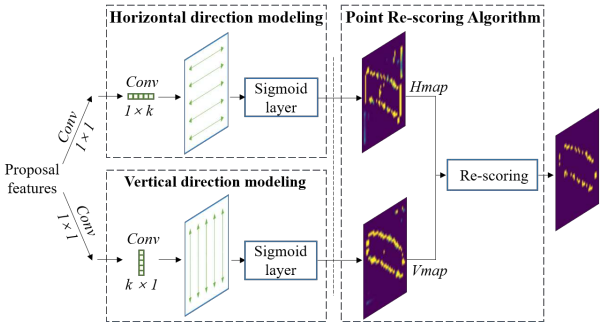


Figure 4. The visualization of LOTM (left). *Point Re-scoring Algorithm* (right) is only used in testing stage.

texts with arbitrary shapes (rectangular and irregular shapes shown in Fig. 5).

As shown in Fig.4, LOTM contains two parallel branches. In the top branch, we slide a convolutional kernel with size $1 \times k$ over the feature maps to model the local texture information in horizontal direction, which only focuses on the texture characteristics in a k-range region. This local operation is proved to be powerful in our experiments, and because of the small amount of computation, it also keeps the efficiency of our method. In a like manner, the bottom branch is constructed to model the texture characteristics in vertical direction through a convolutional kernel with size $k \times 1$. k is a hyper-parameter to control the size of receptive field of texture characteristics, which is discussed in ablation experiments in Sec.4. Finally, two sigmoid layers are implemented to normalize the heatmaps to $[0, 1]$ in both directions. In this way, text regions can be detected in two orthogonal directions and represented with contour points in two differen heatmaps, either of which only responds to texture characteristics in a certain direction.

3.4. Point Re-scoring Algorithm

As false-positive predictions can be effectively suppressed by considering the response value in both orthogonal directions, two heatmaps from LOTM are further processed through *Point Re-scoring Algorithm*. As shown in Algorithm 1, points in different heatmaps are first processed through *Non-Maximum Suppression* (NMS) to achieve a tight representation. Then, to suppress the predictions with strong unidirectional or weakly orthogonal response, we only select the points with distinct response in both heatmaps as candidates. Finally, text region can be represented with polygon made up by these high-quality contour points. NMS_H and NMS_V mean NMS operation in horizontal and vertical direction respectively. We set θ to 0.5 for better trade-off between recall and precision.

3.5. Training Objective

For learning ContourNet, the loss function is formulated as:

$$
\begin{aligned}
L =& L_{Arpn_{cls}} + \lambda_{A_{reg}} L_{Arpn_{reg}} + \lambda_{Hcp} L_{Hcp} \\
&+ \lambda_{Vcp} L_{Vcp} + \lambda_{box_{cls}} L_{box_{cls}} \\
&+ \lambda_{box_{reg}} L_{box_{reg}}
\end{aligned} \tag{3}
$$

Where $L_{Arpn_{cls}}$, $L_{Arpn_{reg}}$, L_{Hcp}, L_{Vcp}, $L_{box_{cls}}$ and $L_{box_{reg}}$ denote Adaptive-RPN classification loss, Adaptive-RPN regression loss, contour point loss in horizontal direction, contour point loss in vertical direction, bounding box classification loss and bounding box regression loss respectively. We use balance weights $\lambda_{A_{reg}}$, λ_{Hcp}, λ_{Vcp}, $\lambda_{box_{cls}}$ and λ_{boxreg} to represent the importance among six losses. We simply balance $\lambda_{A_{reg}}$ and set others to 1 in our experiment.

Adaptive-RPN: Adaptive-RPN is optimized with an *IoU* loss to achieve robust performance on scale variance. The loss function is formulated as:

$$
L_{Arpn_{reg}} = -\log \frac{Intersection + 1}{Union + 1} \tag{4}
$$

Where $Intersection$ and $Union$ are calculated between the predicted and ground-truth bounding boxes. For $L_{Arpn_{cls}}$, we simply use the cross-entropy loss.

LOTM: To solve the unbalanced problem between the size of background and foreground, we use the class-balanced cross-entropy loss for contour point learning. The loss function is formulated as:

$$
L_{cp} = -\frac{N_{neg}}{N} y_i \log p_i - \frac{N_{pos}}{N} (1 - y_i) \log(1 - p_i) \tag{5}
$$

Where y_i and p_i denote ground-truth and prediction. N_{neg} and N_{pos} are numbers of negatives and positives respectively. N is the sum of N_{neg} and N_{pos}. Loss for horizontal prediction (L_{Hcp}) and vertical prediction (L_{Vcp}) have the identical form as L_{cp}.

Algorithm 1 Point Re-scoring Algorithm

Require: Heatmaps in two orthogonal directions: $Hmap$, $Vmap$.
Ensure: Contour point candidates: $Contourmap$.

```
Contourmap = zeros_like(Hmap)
Hmap = NMS_H(Hmap)
Vmap = NMS_V(Vmap)
for (i, j) in Hmap do
    if Confidence[Hmap[i, j]] > θ then
        if Confidence[Vmap[i, j]] > θ then
            Contourmap[i, j] = 1
        end if
    end if
end for
return Contourmap
```

For $L_{boxclass}$ and L_{boxreg} in box branch , we choose the similar forms in [26, 7].

4. Experiments

4.1. Datasets

ICDAR2015 [12] is a dataset proposed in the Challenge 4 of ICDAR 2015 Robust Reading Competition. It contains totally 1500 images (1000 training images and 500 testing images) with annotations labeled as 4 vertices at word level. Different from previous datasets containing horizontal texts only, texts in this benchmark have arbitrary orientations.

CTW1500 [45] is a dataset for curve text detection. It contains 1000 images for training and 500 images for testing. The texts are labeled with 14 boundary points at text-line level.

Total-Text [3] is a recent challenging dataset. Different from CTW1500, the annotations in this dataset are labelled in word-level. This dataset includes horizontal, multi-oriented, and curved texts. It contains 1255 images for training and 300 images for testing.

4.2. Implementation Details

We use the ResNet50 [9] pre-trained on ImageNet as our backbone. The model is implemented in Pytorch and trained on 1 NVIDIA TITANX GPU using Adam optimizer [13]. We only use the official training images of each dataset to train our model. Data augmentation includes random rotation, random horizontal flip and random crop. The models are trained 180k iterations in total. Learning rates start from $2.5 \times 1e-3$, and are multiplied by 0.1 after 120k and 160k iterations. We use 0.9 momentum and 0.0001weight decay. Multi-scale training is used in our training stage. The short side of images is set to {400, 600, 720, 1000, 1200}, and the long side is maintained to 2000. Blurred texts labeled as DO NOT CARE are ignored during training.

As all the datasets use polygon annotations, which are feasible to rebuild texts with arbitrary shapes, we use $distance\_transform\_edt$ in *Scipy* to obtain the two-points wide edge. All the points on the edge are regarded as contour points and used to train our model. The label in Adaptive-RPN can be obtained by using a similar *max-min function* in eq.(2) on ground-truth polygons. During training, we optimize both heatmaps in LOTM with the same supervision.

In testing stage, we use the single scale image as input and evaluate our results through official evaluation protocols. Due to the different scales of test images have a great impact on the detection performance[35, 20], we scale the images in Total-Text and CTW1500 datasets to 720×1280, and fix the resolution to 1200×2000 for ICDAR 2015. Alpha-Shape Algorithm [1] is used to generate bounding

boxes based on contour point candidates.

4.3. Ablation Study

We conduct several ablation studies on CTW1500 and Total-Text datasets to verify the effectiveness of Adaptive-RPN and LOTM. All the models are trained using only official training images.

Adaptive-RPN: We first study the relationship between the performance of Adaptive-RPN and number of pre-defined points. As shown in Tab.1, Adaptive-RPN implemented with 9 pre-defined points obtains 0.6 % improvement in F-measure. We set n to 9 in the remaining experiments.

n-points	Recall	Precision	F-measure
5-points	**85.7**	81.2	83.3
9-points	84.1	**83.7**	**83.9**

Table 1. The relationship between performance and the number of pre-defined points used in Adaptive-RPN. 5-points means top-left, top-right, bottom-right, bottom-left and center points.

Method	Recall	Precision	F-measure
RPN†	83.8	85.1	84.5
Adaptive-RPN†	**83.9**	**86.9**	**85.4**
RPN*	**85.6**	80.8	83.1
Adaptive-RPN*	84.1	**83.7**	**83.9**
	Small	Middle	Large
Gain †	1.4	0.3	1.1
Gain *	1.0	0.7	0.8

Table 2. The performance gain of Adaptive-RPN. * and † are results from CTW1500 and Total-Text respectively. Small, Middle and Large is short for small-size texts, middle-size texts and large-size texts.

size	Recall	Precision	F-measure
3	**83.9**	**86.9**	**85.4**
5	83.6	85.7	84.7
7	83.4	85.4	84.4

Table 3. The relationship between the performance and size of receptive field of texture characteristics in LOTM on Total-Text.

Method	Recall	Precision	F-measure
S-direction	80.5	80.6	80.6
Jointly	82.7	85.3	84.0
LOTM	**83.9**	**86.9**	**85.4**

Table 4. The performance gain of LOTM on Total-Text. S-direction means the texture information is only modeled along a single direction (horizontal direction is implemented here). Jointly means the method jointly models the texture information in a 3×3 convolutional kernel.

To verify the performance gain of the proposed Adaptive-RPN, we conduct several ablation experiments on CTW1500 and Total-Text. LOTM is implemented in all the models. As shown in the top of Tab.2, Adaptive-RPN obtains 0.9% and 0.8% improvement in F-measure on Total-Text and CTW1500 respectively. To further demonstrate the improvement of detecting texts in large variance scale, we further divide the results into three parts according to the size distribution on these two datasets. We consider only the pairs belonging to the same category to be valid for better comparison (*e.g.* small-size predicted bounding box matches small-size ground-truth bounding box. Note that the number of ignored pairs is almost identical in both methods, which affects little to the results.). As shown in the bottom of Tab.2, Adaptive-RPN outperforms conventional RPN in F-measure by a large margin in detecting varying-size texts (*e.g.* 1.4%, 0.3% and 1.1% improvement in F-measure in small-size, middle-size and large-size texts respectively on Total-Text).

LOTM: To evaluate the effectiveness of the proposed LOTM, we conduct several experiments on Total-Text. Firstly, we conduct several experiments to study the relationship between the performance and size of convolutional kernels in LOTM. As shown in Tab.3, model implemented with 1×3 and 3×1 sizes achieves the highest performance (85.4 % in F-measure). When we further increase the size of receptive field, the performance declines. We infer that the larger receptive field containing more noise is harmful to the performance, which further demonstrates the effectiveness of local texture information modeling. We set the size of convolutional kernels to 3 in the remaining experiments.

Secondly, we evaluate the effectiveness of orthogonal modeling. As shown in Tab.4, modeling texture information along only a single direction is a less powerful approach (85.4 % vs 80.6 % in F-measure). Compared with jointly modeling the texture information in arbitrary orientations, LOTM obtains a significant improvement in recall, precision and F-measure with 1.2%, 1.6% and 1.4% respectively.

4.4. Comparisons with State-of-the-Art Methods

We compare our methods with recent state-of-the-art methods on Total-Text, CTW1500 and ICDAR2015 to demonstrate its effectiveness for arbitrary shape text detection.

4.4.1 Evaluation on Curved Text Benchmark

We evaluate the proposed method on Total-Text to test its performance for curved texts.

As shown in Tab.5, with the help of Adaptive-RPN and false-positive suppression, the proposed method achieves a new state-of-the-art result of 83.9%, 86.9% and 85.4% in recall, precision and F-measure respectively without external data, and outperforms existing state-of-the-art methods

Method	Ext	R	P	F	FPS
SegLink* [28]	-	23.8	30.3	26.7	-
EAST* [49]	-	36.2	50.0	42.0	-
Lyu *et al.*[24]	✓	55.0	69.0	61.3	-
TextSnake [23]	✓	74.5	82.7	78.4	-
MSR [43]	✓	74.8	83.8	79.0	4.3
PSENet [33]	-	75.1	81.8	78.3	3.9
PSENet [33]	✓	78.0	84.0	80.9	3.9
Wang *et al.*[35]	-	76.2	80.9	78.5	-
TextDragon [6]	✓	74.2	84.5	79.0	-
TextField [41]	✓	79.9	81.2	80.6	6
PAN [34]	-	79.4	88.0	83.5	**39.6**
LOMO [46]	✓	75.7	**88.6**	81.6	4.4
LOMO† [46]	✓	79.3	87.6	83.3	-
CRAFT [2]	✓	79.9	87.6	83.6	-
Ours	-	**83.9**	86.9	**85.4**	3.8

Table 5. The single-scale results on Total-Text. * indicates the results from [23]. Ext is the short for external data used in training stage. † means testing at multi-scale setting. The evaluation protocol is DetEval [37].

Method	Ext	R	P	F	FPS
CTPN* [31]	-	53.8	60.4	56.9	7.1
SegLink* [28]	-	40.0	42.3	40.8	10.7
EAST* [49]	-	49.1	78.7	60.4	21.2
CTD+TLOC [45]	-	69.8	77.4	73.4	13.3
TextSnake [23]	✓	**85.3**	67.9	75.6	-
PSENet [33]	-	75.6	80.6	78.0	3.9
PSENet [33]	✓	79.7	84.8	82.2	3.9
Tian *et al.*[32]	✓	77.8	82.7	80.1	3
Wang *et al.*[35]	-	80.2	80.1	80.1	-
TextDragon [6]	✓	81.0	79.5	80.2	-
PAN [34]	-	77.7	84.6	81.0	**39.8**
LOMO [46]	✓	69.6	**89.2**	78.4	4.4
LOMO† [46]	✓	76.5	85.7	80.8	-
CRAFT [2]	✓	81.1	86.0	83.5	-
TextField [41]	✓	79.8	83.0	81.4	6
MSR [43]	✓	78.3	85.0	81.5	4.3
Ours	-	84.1	83.7	**83.9**	4.5

Table 6. The single-scale results on CTW1500. * indicates the results from [45]. Ext is the short for external data used in training stage. † means testing at multi-scale setting.

(*e.g.* LOMO [46], PAN [34], PSE[33]) by a large margin. Meanwhile, it also achieves impressive speed (3.8 FPS). Though CRAFT [2] use additional character-level annotations to train their model, our method trained with only original annotations outperforms CRAFT [2] by 1.8 % in F-measure. Besides, LOMO [46] uses external images to train their model and further tests their results at multi-scale level. Our method, which is trained with only official data and tested at single scale, outperforms LOMO [46] by 2.1% in F-measure. The visualization of curved text detection results are shown in Fig.5(a).

Method	Ext	R	P	F	FPS
EAST [49]	✓	73.5	83.6	78.2	13.2
Liao et al. [15]	✓	79.0	85.6	82.2	6.5
Lyu et al. [25]	✓	70.7	**94.1**	80.7	3.6
FOTS [20]	✓	82.0	88.8	85.3	7.8
PixelLink [4]	-	81.7	82.9	82.3	7.3
MSR [43]	✓	78.4	86.6	82.3	4.3
PSENet [33]	-	79.7	81.5	80.6	1.6
PSENet [33]	✓	84.5	86.9	85.7	1.6
PAN [34]	-	77.8	82.9	80.3	**26.1**
TextDragon [6]	✓	81.8	84.8	83.1	-
LOMO [46]	✓	83.5	91.3	87.2	3.4
TextField* [41]	✓	83.9	84.3	84.1	1.8
Liu *et al.* [21]	✓	83.8	89.4	86.5	-
Tian *et al.* [32]	✓	85.0	88.3	86.6	3
CRAFT [2]	✓	84.3	89.8	86.9	-
Wang *et al.* [35]	-	83.3	90.4	86.8	-
Wang *et al.*† [35]	-	86.0	89.2	**87.6**	-
Ours	-	**86.1**	87.6	86.9	3.5

Table 7. The single-scale results on ICDAR2015. * means testing at multi-scale setting. † means SE blocks [11] implemented in their backbone.

4.4.2 Evaluation on Long Curved Text Benchmark

To show the performance of our ContourNet for long curved texts, we compare its performance with state-of-the-arts on CTW1500 dataset, which is annotated at text-line level.

As shown in Tab.6, the proposed method is much better than other counterparts including CTD+TLOC [45], MSR [43], TextSnake [23], which are designed for curved texts. Though text region refinement in LOMO [46] achieves promising results on representing long texts, our ContourNet, which benefits from Adaptive-RPN, achieves much higher performance (83.9 % vs 80.8% in F-measure). Compared with MSR[43] which also uses contour points to describe text regions, ours shows advantages in both recall and F-measure without external data for training, where the relative improvement reaches 5.8% and 2.4% respectively. In addition, the proposed method runs at 4.5 FPS on this dataset, which is faster than most recent methods. The visualization of long curved text detection results are shown in Fig.5(b).

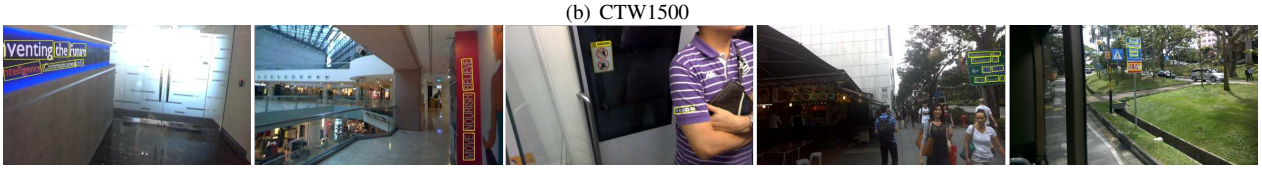
4.4.3 Evaluation on Multi-oriented Text Benchmark

We evaluate our method on ICDAR 2015 to test its performance for multi-oriented texts. RoIAlign [8] is used for the generation of proposal features on this dataset.

Several experimental results are shown in Tab.7. Our

(a) Total-Text

(b) CTW1500



(c) ICDAR2015

Figure 5. Results on different datasets. (a) results on Total-Text; (b) results on CTW1500; (c) results on ICDAR2015.

method achieves 86.9 % in F-measure, which is only a little lower than Wang *et al.*[35] (87.6% in F-measure). However, they implement Squeeze-and-Excitation (SE) blocks [11] in their backbone, which is more powerful to recalibrate channel-wise feature responses. When implemented without SE blocks, their method achieves 86.8 % in F-measure, which is lower than our method. The visualization of multi-oriented text detection results are shown in Fig.5(c).

4.5. Effectiveness of ContourNet

We further demonstrate the effectiveness of our method in the following two aspects. More discussions about this part are shown in the supplementary.

Effectiveness of Adaptive-RPN. As the large scale variance problem exists in scene text detection, conventional RPN obtains a coarse localization of text region when the regression distance is large or the target box has quite different ratio to default box. Benefiting from the awareness of shape information and the scale-invariant training object, the proposed Adaptive-RPN performs better in these cases and achieves finer localization of text regions. Some qualitative examples of conventional RPN and proposed Adaptive-RPN are shown in the supplementary.

Effectiveness of false-positive suppression. 1) Quantification. The value of θ in Point Re-scoring Algorithm affects the ratio of suppressed FPs to caused false negetives (FNs). The value of ratio is considerable when θ goes from 0.1 to 0.9 (an elaborated chart is shown in the supplementary). Thus, our method is much more effective in suppressing FPs than in causing FNs. **2) Qualitative analysis.** Though few FNs are caused, it is worth mentioning that the retained positive points with strong texture information in both orthogonal directions are able to accurately represent texts (see in Fig.1). **3)** Implemented with conventional RPN, our method can achieve 84.5% and 83.1% in F-measure on Total-Text and CTW1500 respectively, surpassing most methods in Tab.5 and Tab.6. Though it is hard to verify which representation is better for arbitrary-shaped text detection (*e.g.* region predictions [35, 34], contour points [43], adaptive points [35], etc.), FP problem is the uniform challenge in each method. In this regard, our method obtains a significant improvement compared with the previous.

5. Conclusion

In this paper, we propose a novel scene text detection method (ContourNet) to handle the false positives in text representation and the large scale variance problem. ContourNet mainly consists of three parts including Adaptive-RPN, LOTM and *Point Re-scoring Algorithm*. Adaptive-RPN localizes the preliminary proposals of texts by bounding the spatial extend of several semantic points. LOTM models the local texture information in two orthogonal directions and represents text region with contour points. *Point Re-scoring Algorithm* filters FPs by considering the response values in both orthogonal directions simultaneously. The effectiveness of our approach has been demonstrated on several public benchmarks including long, curved and oriented text cases. In future works, we prefer to develop an end-to-end text reading system.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2017YFC0820600), the National Nature Science Foundation of China (61525206, U1936210), the Youth Innovation Promotion Association Chinese Academy of Sciences (2017209), the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

References

[1] Nataraj Akkiraju, Herbert Edelsbrunner, Michael Facello, Ping Fu, EP Mucke, and Carlos Varela. Alpha shapes: definition and software. In *Proceedings of the 1st International Computational Geometry Software Workshop*, volume 63, page 66, 1995.

[2] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.

[3] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 935–942. IEEE, 2017.

[4] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *AAAI*, pages 6773–6780, 2018.

[5] Shancheng Fang, Hongtao Xie, Zheng-Jun Zha, Nannan Sun, Jianlong Tan, and Yongdong Zhang. Attention and language ensemble for scene text recognition with convolutional sequence modeling. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 248–256. ACM, 2018.

[6] Wei Feng, Wenhao He, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. Textdragon: An end-to-end framework for arbitrary shaped text spotting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9076–9085, 2019.

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Deep direct regression for multi-oriented scene text detection. In *ICCV*, pages 745–753, 2017.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[12] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th ICDAR*, pages 1156–1160. IEEE, 2015.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[15] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-Song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *CVPR*, pages 5909–5918, 2018.

[16] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017.

[17] An-An Liu, Yu-Ting Su, Wei-Zhi Nie, and Mohan Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):102–114, 2016.

[18] C. Liu, H. Xie, Z. Zha, L. Yu, Z. Chen, and Y. Zhang. Bidirectional attention-recognition model for fine-grained object classification. *IEEE Transactions on Multimedia*, pages 1–1, 2019.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[20] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018.

[21] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3052–3058, 2019.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[23] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *ECCV*, pages 19–35. Springer, 2018.

[24] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018.

[25] Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai. Multi-oriented scene text detection via corner localization and region segmentation. In *CVPR*, pages 7553–7563, 2018.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[27] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[28] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2550–2558, 2017.

[29] Shangxuan Tian, Shijian Lu, and Chongshou Li. Wetext: Scene text detection under weak supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1492–1500, 2017.

[30] Shangxuan Tian, Yifeng Pan, Chang Huang, Shijian Lu, Kai Yu, and Chew Lim Tan. Text flow: A unified text detection system in natural scene images. In *Proceedings of the IEEE international conference on computer vision*, pages 4651–4659, 2015.

[31] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *European conference on computer vision*, pages 56–72. Springer, 2016.

[32] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiaya Jia. Learning shape-aware embedding for scene text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4234–4243, 2019.

[33] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[34] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8440–8449, 2019.

[35] Xiaobing Wang, Yingying Jiang, Zhenbo Luo, Cheng-Lin Liu, Hyunsoo Choi, and Sungjin Kim. Arbitrary shape scene text detection with adaptive text region representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6449–6458, 2019.

[36] Yuxin Wang, Hongtao Xie, Zilong Fu, and Yongdong Zhang. Dsrn: a deep scale relationship network for scene text detection. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 947–953. AAAI Press, 2019.

[37] Christian Wolf and Jean-Michel Jolion. Object count/area graphs for the evaluation of object detection and segmentation algorithms. *International Journal of Document Analysis and Recognition (IJDAR)*, 8(4):280–296, 2006.

[38] Enze Xie, Yuhang Zang, Shuai Shao, Gang Yu, Cong Yao, and Guangyao Li. Scene text detection with supervised pyramid context network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9038–9045, 2019.

[39] Hongtao Xie, Shancheng Fang, Zheng-Jun Zha, Yating Yang, Yan Li, and Yongdong Zhang. Convolutional attention networks for scene text recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1s):1–17, 2019.

[40] Ning Xu, An-An Liu, Yongkang Wong, Yongdong Zhang, Weizhi Nie, Yuting Su, and Mohan Kankanhalli. Dual-stream recurrent neural network for video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2482–2493, 2018.

[41] Yongchao Xu, Yukang Wang, Wei Zhou, Yongpan Wang, Zhibo Yang, and Xiang Bai. Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Transactions on Image Processing*, 2019.

[42] Chuhui Xue, Shijian Lu, and Fangneng Zhan. Accurate scene text detection through border semantics awareness and bootstrapping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 355–372, 2018.

[43] Chuhui Xue, Shijian Lu, and Wei Zhang. Msr: multi-scale shape regression for scene text detection. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 989–995, 2019.

[44] Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao. Robust text detection in natural scene images. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):970–983, 2013.

[45] Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017.

[46] Chengquan Zhang, Borong Liang, Zuming Huang, Mengyi En, Junyu Han, Errui Ding, and Xinghao Ding. Look more than once: An accurate detector for text of arbitrary shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10552–10561, 2019.

[47] Hanwang Zhang, Zheng-Jun Zha, Shuicheng Yan, Jingwen Bian, and Tat-Seng Chua. Attribute feedback. In *Proceedings of the 20th ACM international conference on Multimedia*, 2012.

[48] Zheng Zhang, Wei Shen, Cong Yao, and Xiang Bai. Symmetry-based text line detection in natural scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2558–2567, 2015.

[49] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. EAST: an efficient and accurate scene text detector. In *CVPR*, pages 2642–2651, 2017.

[50] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019.