# CPN: Complementary Proposal Network for Unconstrained Text Detection

**Longhuang Wu, Shangxuan Tian*, Youxin Wang, Pengfei Xiong**

Shopee Pte. Ltd.
{wlonghuang, wyxileroy, xiongpengfei2019}@gmail.com, tianshangxuan@u.nus.edu

## Abstract

Existing methods for scene text detection can be divided into two paradigms: segmentation-based and anchor-based. While Segmentation-based methods are well-suited for irregular shapes, they struggle with compact or overlapping layouts. Conversely, anchor-based approaches excel for complex layouts but suffer from irregular shapes. To strengthen their merits and overcome their respective demerits, we propose a Complementary Proposal Network (CPN) that seamlessly and parallelly integrates semantic and geometric information for superior performance. The CPN comprises two efficient networks for proposal generation: the Deformable Morphology Semantic Network, which generates semantic proposals employing an innovative deformable morphological operator, and the Balanced Region Proposal Network, which produces geometric proposals with pre-defined anchors. To further enhance the complementarity, we introduce an Interleaved Feature Attention module that enables semantic and geometric features to interact deeply before proposal generation. By leveraging both complementary proposals and features, CPN outperforms state-of-the-art approaches with significant margins under comparable computation cost. Specifically, our approach achieves improvements of 3.6%, 1.3% and 1.0% on challenging benchmarks ICDAR19-ArT, IC15, and MSRA-TD500, respectively. Code for our method will be released.

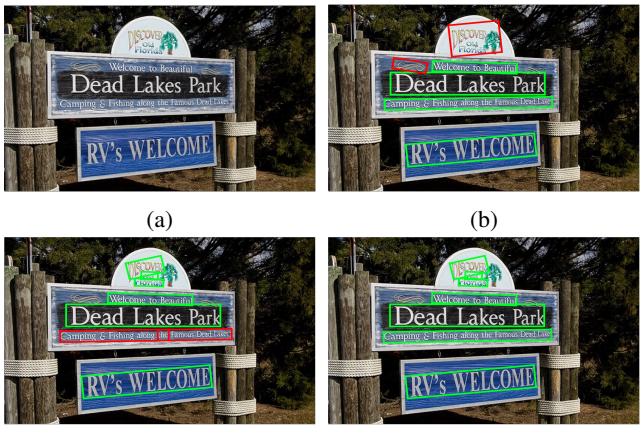


Figure 1: Comparison between existing detectors and our proposed CPN. The input image is shown in (a). Segmentation based methods tend to produce false positives and negatives while facing compact layouts as illustrated in (b). Anchor based approaches struggle for texts of large aspect ratio as shown in (c). Our CPN addresses the above issues by complementary proposals and features, with results given in (d). True positive results are colored in green, while the false boxes are indicated in red.

## Introduction

Automated detection of various texts in scenes has been extensively studied for years due to its applications in many useful tasks such as multi-modal image and video understanding, autonomous indoor and outdoor navigation, etc. However, the detection of text in arbitrary shapes and complex layouts remains a challenge for most existing scene text detectors, leading to a high number of false positives and false negatives as illustrated in Figure 1b and 1c.

Prevalent scene text detection work follows two typical paradigms: segmentation-based approach and anchor-based approach. The segmentation approach has achieved great success thanks to its simplicity and capacity to deal with text of arbitrary shapes (Liao et al. 2020; Wang et al. 2019a). However, this approach often struggles while facing texts of complex layouts and background noise, as illustrated in Figure 1b. This is largely due to the challenges in identification of text boundaries and complicated heuristic post-processing for pixel grouping (Wang et al. 2019a; Deng et al. 2018; Tian et al. 2019). On the other hand, the anchor approach generally achieves higher accuracy due to its two-stage design with proposal generation and verification. However, it is designed for generic object detection tasks and often suffers while dealing with scene texts of arbitrary shapes such as curved texts and long text lines, as illustrated in 1c.

In light of the aforementioned analysis, we contend that the segmentation approach and the anchor approach are inherently complementary to each other, not only in terms of candidate proposals but also in the features used to generate them. By combining the strengths of these two approaches, we could potentially address their respective limitations and lead to improved performance in unconstrained

*Corresponding author

scene text detection. However, directly integrating these two approaches faces two issues. First, the textual representation of them are distinct, how can effective and efficient unification be achieved? Second, segmentation-based methods require complex post-processing to generate candidate boxes on the CPU. This results in low efficiency for both training and inference of the entire network, preventing end-to-end training. Here, we present a **C**omplementary **P**roposal **N**etwork (CPN) that exploits rich semantic and geometric information for detecting text instances. Our CPN consists of two complementary designs as shown in Figure 2. First, we propose two efficient complementary networks for proposal generation, namely, Deformable Morphology Semantic Network (Deformable MSN) and Balanced Region Proposal Network (Balanced RPN). Deformable MSN learns high-level semantic information with deformable morphological operators to obtain semantic proposals, while Balanced RPN captures low-level geometric structure of text instances to predict geometric proposals. Second, we introduce an Interleaved Feature Attention (IFA) module, which not only enhances the mutual information between semantic and geometric features and explicitly supervises each other, but also reduces the computational burden of the pipeline.

We conducted extensive experiments on five text detection benchmarks to evaluate the performance of our proposed CPN. Results show that the CPN approach significantly outperforms state-of-the-art methods, where it achieves an f-measure of 81.7% (+3.6%) on the large-scale arbitrary-shaped IC19-ArT dataset (Chng et al. 2019). Furthermore, our CPN achieves a comparable system efficiency (11.2 fps) to standard two-stage detectors such as Mask R-CNN (9.6 fps) on IC15 dataset. To summarize, the main contributions of this work are as follows:

1. We propose a dual Complementary Proposal Network (CPN) that integrates semantic and geometric proposals, along with the features used to generate them. To the best of our knowledge, this is the first study to explore the integration of two distinct proposal paradigms for text detection.
2. We present a Deformable MSN with a deformable morphology operator to efficiently generate semantic proposals with parallelized GPU computations.
3. We introduce an Interleaved Feature Attention module that enables interaction and supervision between semantic and geometric features for complementary features and computation reduction.
4. Extensive experiments over multiple prevailing benchmarks show our proposed CPN outperforms the state-of-the-art by large margins.

## Related Work

Before the deep learning era, bottom-up pipelines are widely adopted in scene text detection with hand-crafted features such as Maximally Stable Extremal Regions (MSER) (Neumann and Matas 2012) and Stroke Width Transform (SWT) (Epshtein, Ofek, and Wexler 2010). Over the past few years, deep learning based text detectors have become prevalent which can be roughly divided into anchor-based methods and segmentation-based methods.

### Anchor-based Methods

The anchor-based methods treat text instances as common objects and adapt the pipeline of generic object detection, e.g., SSD (Liu et al. 2016) and Faster-RCNN (Ren et al. 2015), into text detection. They utilize offset regression from predefined anchor boxes or points in feature maps to predict text locations. To address the problem of large aspect ratios in scene texts, TextBoxes (Liao et al. 2017) has designed compact anchors with different aspect ratios and scales to cover texts with varied sizes. RRPN (Ma et al. 2018) adjusts horizontal anchors to rotated ones with angle prediction to localize arbitrary-oriented text regions, followed by a rotated RoI pooling layer before text/non-text classification. DMPNet (Liu and Jin 2017) trys to handle multi-oriented text instances with tighter quadrilateral sliding windows instead of horizontal ones. In general, anchor-based methods have two stages with proposal generation in the first stage, followed by a text verification network. Therefore, they tend to achieve better performance when compared with the simplified one-stage segmentation-based approaches.

### Segmentation-based Methods

Methods belonging to this category draw inspiration from semantic segmentation and they aim to gather individual pixels for accurate text detection. PixelLink (Deng et al. 2018) predicts the linkage relationships between pixels to group pixels within the same text instance. PSENet (Wang et al. 2019a) adopts a progressive scale algorithm to expand the text kernels of different scales. In TextField (Xu et al. 2019), the offset field of each pixel in text regions is predicted for connecting neighborhood pixels. To effectively simplify the post-processing step, Liao *et al.* (Liao et al. 2020) propose differentiable binarization and achieve a good balance between speed and accuracy. To conclude, segmentation-based methods predict text boxes in one shot, making the whole pipeline much simpler and more efficient. Besides, it can intrinsically handle text of arbitrary shapes.

Despite their respective superiority, methods in each group may suffer from different weaknesses. Anchor-based methods may fail when handling text instances of irregular shapes and extreme aspect ratios due to their structural limitations. Segmentation-based methods behave poorly for adjacent and overlapping text instances. However, those approaches intrinsically compensate each other considering the above merits and demerits. Therefore, we consider proposing a novel pipeline that can combine their merits to overcome their demerits.

## Methodology

The architecture of the proposed text detection pipeline is presented in Figure 2. It consists of three components: the backbone network, the proposed Complementary Proposal Network (CPN), followed by an lightweight RoI head. The CPN is composed of two novel parallel complementary

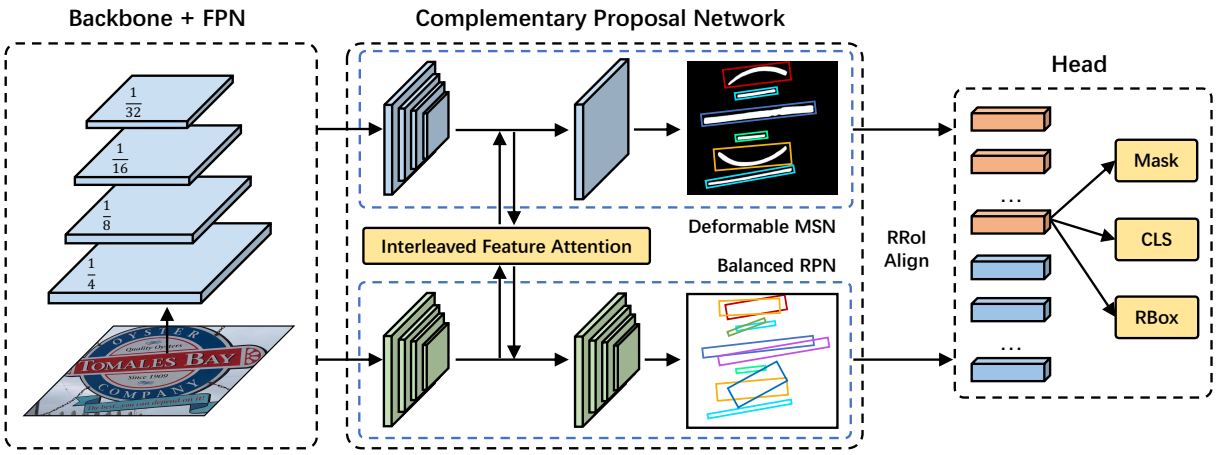


Figure 2: The pipeline of the proposed Complementary Proposal Network (CPN). Given an input image, multi-scale FPN features are extracted and fed to two parallel networks (Deformable MSN and Balanced RPN) for generating complementary semantic and geometric proposals. An Interleaved Feature Attention (IFA) module is designed to promote interaction between branch features, encouraging them to capture more spatial and scale-aware information. Features for the merged proposals are then identified with RRoI align before passing to the RoI head, where final text boxes and masks are generated.

branches designed to efficiently integrate semantic and geometric proposals. Additionally, the Interleaved Feature Attention (IFA) module is integrated into the CPN to enhance the interactions between semantic and geometric features, further improving their complementarity.

The main objective of the CPN is to generate diverse text proposals that can accommodate a wide range of scales, shapes, and orientations under varied scenes. To this end, we have designed two networks that work in tandem to strengthen their merits and overcome their respective demerits. Specifically, the Deformable Morphology Semantic Network (Deformable MSN) generates semantic proposals, while the Balanced Region Proposal Network (Balanced RPN) generates geometric proposals. These proposals are merged as shown in Figure 2 and passed to the RROI align layer. Considering that the Deformable MSN network has already captured rich semantic information, we adopt a lightweight mask head in R-CNN by decreasing the four 3x3 convolution layers to one. We optimize the entire pipeline through end-to-end multi-task learning without complicated post-processing steps.

## Deformable Morphology Semantic Network

The Deformable Morphology Semantic Network (Deformable MSN) in the upper part of the CPN in Figure 2 generates accurate candidate proposals, particularly for text with curved shapes and extreme aspect ratios where the geometric proposal branch often fails. Unlike existing segmentation-based approaches (Liao et al. 2020; Wang et al. 2019a,b) that utilize CPU extensive operations such as the Vatti clipping algorithm (Vatti 1992) for box generation, we propose a Deformable MSN with differentiable morphology operators that can be fully parallelized on GPUs, resulting in superior efficiency. Furthermore, as the Deformable MSN utilizes instance segmentation as supervision, the generated proposals can capture high-level semantic information effectively. The detailed structure of Deformable MSN is presented in Figure 3. It predicts a text erosion map and a structuring kernel map, followed by a deformable dilation layer to produce the text confidence map, and then the corresponding oriented rectangular proposals.

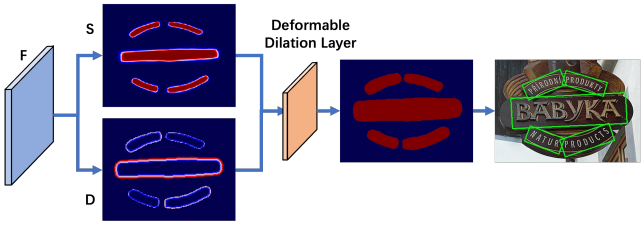


Figure 3: Illustration of the proposed Deformable Morphology Semantic Network (Deformable MSN) for semantic proposals. It predicts a text erosion map $S$ and a structuring kernel map $D$, followed by a deformable dilation layer to produce text candidate regions, and thereafter the corresponding oriented proposals.

**Text Erosion Map Generation** In order to create labels for the text erosion map during training, we erode the text regions by specific sizes and set pixels inside as positive while others as negative. In contrast to previous approaches (Liao et al. 2020; Wang et al. 2019a,b) which utilize Vatti clipping, we design a more efficient approach by using morphological erosion operators with circular structuring kernels of varying sizes to shrink the pixels in original text regions. For a text instance $t$, the circular structuring kernel size is given by

$$e_t = k \times A_t / L_t \tag{1}$$

where $A_t$ and $L_t$ are the area and perimeter of $t$, and $k$ is the scale hyperparameter.

**Structuring Kernel Map Generation** The morphological operator adopts structuring kernel to provide specified size

and shape information while transforming an image. We define the structuring kernel $D(p)$ for a given pixel $p$ as

$$D(p) = \begin{cases} k \times \left|\overrightarrow{pS_p}\right|, & p \in \mathbb{T} \\ 0, & p \notin \mathbb{T} \end{cases} \tag{2}$$

where $p$ and $S_p$ belong to the same text instance and $S_p$ is the nearest pixel for $p$ in the erosion map $S$. $k$ is the normalization factor and $\mathbb{T}$ represents all the text pixels. For pixels in non-text areas ($p \notin \mathbb{T}$), we set the values to 0.

**Semantic Proposal Generation** In most segmentation-based pipelines, post-processing is sophistically designed over text confidence map to rebuild text instances. However, they inevitably introduce some complicated operators, e.g. Vatti-clipping algorithm, which are difficult to be paralleled on GPU and often lead to inferior performance. To build a more efficient and end-to-end trainable network, we propose the deformable morphological dilation layer which is defined as

$$\mathrm{dst}(x, y) = \max_{(\Delta x, \Delta y)} \mathrm{S}\left(x + \Delta x, y + \Delta y\right) \tag{3}$$

where $(\Delta x)^2 + (\Delta y)^2 < \lceil d_{x,y} \rceil^2$. $d_{x,y}$ is the predicted structuring kernel size at pixel $p(x, y)$ and $\lceil d_{x,y} \rceil$ is the corresponding radius.

Given the predicted text erosion map $S$, we binarize it and conduct connected-component labeling to obtain text instances map. Then, the labeled map is fed into the deformable morphological dilation layer with structuring kernel map $D$ to rebuild binary text instances. The deformable morphological dilation layer can be seen as a MaxPooling operator whose kernels are in circular shapes with varying radius at different locations. Finally, we produce oriented rectangular boxes by computing the rotated minimal area rectangles from the previous dilated binary regions. In general, the number of proposals generated by the Deformable MSN is less than 100, which is much smaller than the geometric proposal network.

## Balanced Region Proposal Network

To address multi-oriented problem for anchor-based frameworks, existing works like (Ma et al. 2018) have designed rotated anchors with different angles, scales and aspect ratios, with extensive computation. To efficiently generate oriented text proposals, we explore the midpoint offset representation by oriented Region Proposal Network (RPN) (Xie et al. 2021) to capture the geometric characteristics, which complement the deformable morphology network that deals from the semantic perspective.

**Proposal Number Balance** In most anchor-based detection pipelines, the maximum number of proposals is set to more than 1k for better performance, e.g. 2k proposals in (Xie et al. 2021) and 1k proposals in (He et al. 2017). Considering that the semantic branch normally provides much less ($\sim$100) and tighter proposals covering the majority of text regions, we doubt the redundancy of the geometric proposals by RPN. Moreover, a larger number of proposals from RPN may introduce more false positives and dominate the training loss. In practice, we experiment with different numbers of proposals and the outcome is consistent with our assumption. Hence we balance the number of RPN proposals to 300 and refer to this proposal branch as Balanced Region Proposal branch (Balanced RPN). Detailed settings are described in Section  with corresponding results in Table 4.

Meanwhile, we adjusted the anchor ratios to $\{0.5, 1, 3\}$ and set the base scale to 5, making it more suitable for the scene text detection task.

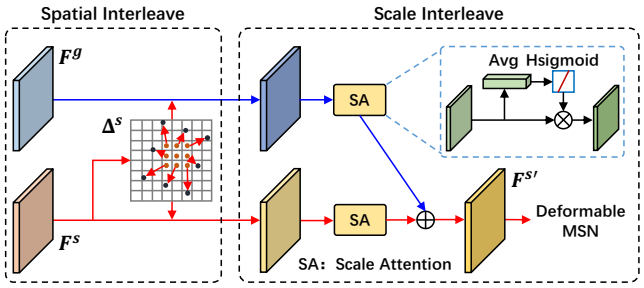


Figure 4: Detailed structure of IFA module with spatial and scale interleaved attentions for the Deformable MSN branch, and the Balanced RPN branch shares the same design.

## Interleaved Feature Attention

To further leverage the complementary information, we propose an Interleaved Feature Attention (IFA) module between different branches for more interactions, as illustrated in Figure 4. By interacting with each other, the semantic branch and geometric branch can encode both spatial and scaled context. Take the semantic proposal branch on the top part of Figure 4 as an example. We first apply spatial aware attention $\mathcal{O}$ among different levels of FPN features $F^s$ to obtain the offset field $\Delta^s$ given by Equation 4. Then $\Delta^s$ will be shared to different deformable convolutions (Dai et al. 2017) $\mathcal{D}_s$ and $\mathcal{D}_g$ to obtain the interleaved features of the semantic branch and geometric branch respectively. As the semantic features $F^s$ and geometric features $F^g$ may be in different feature spaces, we further utilize scale-aware attentions $\mathscr{S}_s$ and $\mathscr{S}_g$ to learn the corresponding weights for both features before fusion. The ultimate features $F^{s\prime}$ for the semantic branch is produced by adding the previous two features after scale attention.

$$\begin{aligned} \Delta^s &= \mathcal{O}\left(F^s\right) \\ F^{s\prime} &= \frac{1}{2} \sum_{i \in \{s,g\}} \left(\mathscr{S}_i\left(\mathcal{D}_i\left(F^i, \Delta^s\right)\right)\right) \end{aligned} \tag{4}$$

The scale aware attention $\mathscr{S}$ is given as

$$\mathscr{S}(F) = \rho\left(f\left(\frac{1}{HW}\sum_{HW} F\right)\right) \cdot F \tag{5}$$

where $H$ and $W$ indicate the spatial scale of corresponding feature map $F$, $f(\cdot)$ is a linear function approximated by a $1 \times 1$ convolutional layer, and $\rho(x) = \max\left(0, \min\left(1, (x+1)/2\right)\right)$ is a hard-sigmoid function.

The proposed IFA is a lightweight module that captures geometric and semantic information to produce complementary text features, leading to an obvious performance lift in both recall and precision. Additionally, IFA deducts the original FPN features from 256 channels to 128 channels, resulting in faster inference speeds and reduced overall model complexity.

### Training Targets

The whole pipeline is jointly trained with the loss given by

$$\mathcal{L} = \mathcal{L}_t + \alpha_1 \mathcal{L}_e + \alpha_2 \mathcal{L}_{geo} + \alpha_3 \mathcal{L}_{rcnn} \tag{6}$$

Where $\mathcal{L}_t$, $\mathcal{L}_e$, $\mathcal{L}_{geo}$, $\mathcal{L}_{rcnn}$ represents the losses for the text erosion map, structuring kernel map, Balanced RPN and Mast R-CNN head, respectively. The $\alpha_1$, $\alpha_2$, $\alpha_3$ are the corresponding weights for each loss.

**Loss for Text Erosion Map** To avoid the network bias to non-text pixels, we adopt dice coefficient loss for shrunk text instances in Deformable MSN. The dice coefficient $\mathcal{L}_t$ is computed as

$$\mathcal{L}_t = \frac{2 \times \sum_p \left( \hat{S}_p \times S_p^* \right)}{\sum_p \hat{S}_p^2 + \sum_p {S_p^*}^2} \tag{7}$$

where $\hat{S}_p$ and $S_p^*$ refer to the value of pixel $p$ in the predicted and ground truth text erosion map, respectively.

**Loss for Structuring Kernel Map** We extend the smoothed $L1$ loss proposed in (Girshick 2015) by adding an extra ratio term. Then the loss function can be defined as

$$\mathcal{L}_e = \frac{1}{\Omega} \sum_{p \in \Omega} SL1(\hat{D}_p, D_p^*) + \beta \log \frac{\max(\hat{D}_p, D_p^*)}{\min(\hat{D}_p, D_p^*)} \tag{8}$$

where $\hat{D}_p$ and $D_p^*$ are the predicted and ground truth values for pixel $p$ in the structuring kernel map. $SL1()$ is the smoothed $L1$ loss and $\beta$ is the weighted factor. $\Omega$ denotes the set of positive elements in $D_p^*$.

**Loss for Balanced RPN** We use the oriented RPN loss (Xie et al. 2021) to optimize the Balanced RPN branch given by

$$\mathcal{L}_{geo} = \frac{1}{N} \left( \sum_{i=1}^{N} F_{cls}\left(y_i^*, y_i\right) + y_i \sum_{i=1}^{N} F_{reg}\left(t_i^*, t_i\right) \right) \tag{9}$$

where $N$ is the number of anchors in a mini-batch and $i$ is the index. $y_i$ is the label and $y_i^*$ denotes the probability. $t_i$ is ground-truth offset while $t_i^*$ is the predicted one. $F_{cls}$ is the cross-entropy loss and $F_{reg}$ is the Smooth $L1$ loss.

## Experiments

We adopts five widely studied datasets IC19-ArT (Chng et al. 2019), CTW1500 (Yuliang et al. 2017), IC17-MLT (Nayef et al. 2017), IC15 (Karatzas et al. 2015), MSRA-TD500 (Yao, Bai, and Liu 2014), which contain a variety of different scenarios, to evaluate the performance of our proposed complementary network.

### Implementation Details

We use ResNet50 (He et al. 2016) as our backbone. All the networks are optimized with AdamW (Loshchilov and Hutter 2017) with batch size set to 16. On the IC17-MLT dataset, we train the model for 75 epochs without using extra data such as SynthText. The initial learning rate is set to $1 \times 10^{-4}$ and divided by 10 at 65 and 70 epochs. For the rest of the datasets, we fine-tune the model with their corresponding train sets on the previous IC17-MLT model. During fine-tuning, the model is trained for 24 epochs with an initial learning rate set to $5 \times 10^{-5}$ and decayed by 0.1 after 20 epochs. Three augmentation schemes are implemented for training: 1) each side of the images is randomly re-scaled within the range of [480, 2560] without maintaining the aspect ratio, 2) each image is randomly flipped horizontally and rotated within the range of $[-10°, 10°]$, 3) $640 \times 640$ random samples are cropped from each transformed image. For straight-text datasets, we directly used the predicted oriented bounding boxes as detection outputs. For curved-text datasets, we use the predicted masks as detection outputs.

### Comparison with the state-of-the-arts

As presented in Table 1 and 2, our proposed approach demonstrates superior performance compared to previous state-of-the-art methods, with a significant margin across all five datasets and in various scenarios.

**Curved text detection** The dataset IC19-ArT has a balanced distribution of all text shapes, providing a comprehensive evaluation that is presented in Table 1. Without whistles and bells, our network outperforms the state-of-the-art by a significant margin (**3.6%**) and achieves an impressive f-measure of 81.7%. It's noteworthy that our approach achieves a considerable recall gain of **6.2%** over the previous SOTA, which adopts a transformer-based pipeline. The experiments demonstrate that our complementary network can effectively boost both recall and precision.

**Multi-oriented text detection** To validate our network on multi-oriented text instances, we evaluate on the TD500 and IC15 datasets, and the results are presented in Table 1. During inference, we resize the shorter side of the images to 960 on the IC15 dataset. Despite the presence of a large number of small and low-resolution text regions in IC15, our approach achieves a promising 90.4% f-measure which surpasses state-of-the-art. For example, it outperforms FSG (Tang et al. 2022) by 1.3% on f-measure. On the TD500 dataset, we resize the longer side of images to 800 while keeping the aspect ratio. Our approach exceeds the best-reported results from TextPMs (Zhang et al. 2022) by 1.0% for f-measure with better fps. As visualized in the second column of Figure 6, our CPN can handle multi-oriented text instances as well as long text lines with ease.

**Multilingual text detection** We demonstrate the capability of our proposed approach in detecting multilingual texts on the IC17-MLT dataset. During inference, we keep the aspect ratio and resize the longer side of the images to 1920. As listed in Table 2, our network achieves an impressive f-measure of 80.0% with a recall of 75.4% on this challenging dataset with multilingual text. Compared to the previous

| Method | Ext | IC15 | | | | MSRA-TD500 | | | | CTW1500 | | | | IC19-ArT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R | P | F | FPS | R | P | F | FPS | R | P | F | FPS | R | P | F |
| ***Sem-based*** | | | | | | | | | | | | | | | | |
| PSENet-1s (Wang et al. 2019a) | MLT | 84.5 | 86.9 | 85.7 | 1.6 | - | - | - | - | 79.7 | 84.8 | 82.2 | 3.9 | 52.2 | 75.9 | 61.9 |
| CRAFT (Baek et al. 2019) | MLT | 84.3 | 89.8 | 86.9 | - | 78.2 | 88.2 | 82.9 | 8.6 | 81.1 | 86.0 | 83.5 | - | 68.9 | 77.3 | 72.9 |
| LOMO (Zhang et al. 2019) | MLT$^+$ | 83.5 | 91.3 | 87.2 | 3.4 | - | - | - | - | 76.5 | 85.7 | 80.8 | 4.4 | - | - | - |
| PCR (Dai et al. 2021) | MLT | - | - | - | - | 82.0 | 88.5 | 85.2 | - | 82.3 | 87.2 | 84.7 | 11.8 | 66.1 | **84.0** | 74.0 |
| DB++ (Liao et al. 2022) | Syn | 83.9 | 90.9 | 87.3 | 10.0 | 83.3 | 91.5 | 87.2 | **29.0** | 82.8 | 87.9 | 85.3 | 26.0 | - | - | - |
| TextBPN (Zhang et al. 2022) | MLT | - | - | - | - | 84.5 | 86.6 | 85.6 | 10.7 | 83.6 | 86.5 | 85.0 | 12.2 | - | - | - |
| TextPMs (Zhang et al. 2022) | MLT | 84.9 | 89.9 | 87.4 | - | 86.9 | 91.0 | 88.9 | 10.6 | 83.8 | 87.8 | 85.8 | 9.1 | - | - | - |
| ***Geo-based*** | | | | | | | | | | | | | | | | |
| DRRG (Zhang et al. 2020) | MLT | 84.7 | 88.5 | 86.6 | - | 82.3 | 88.1 | 85.1 | - | 83.0 | 85.9 | 84.5 | - | - | - | - |
| Contour (Wang et al. 2020) | - | 86.1 | 87.6 | 86.9 | 3.5 | - | - | - | - | 84.1 | 83.7 | 83.9 | 4.5 | 62.1 | 73.2 | 67.2 |
| I3CL (Du et al. 2022) | Syn | - | - | - | - | - | - | - | - | 84.5 | 87.4 | 85.9 | 7.6 | 71.3 | 82.7 | 76.6 |
| FSG (Tang et al. 2022) | Syn$^+$ | 87.3 | 90.9 | 89.1 | **12.9** | 84.7 | 91.4 | 87.9 | - | 82.4 | 88.1 | 85.2 | - | - | - | - |
| DPText-DETR (Ye et al. 2023) | MLT$^+$ | - | - | - | - | - | - | - | - | 86.2 | **91.7** | **88.8** | - | 73.7 | 83.0 | 78.1 |
| ***Ours*** | MLT | **89.2** | **91.7** | **90.4** | 11.2 | **88.3** | **91.6** | **89.9** | 13.3 | **89.6** | 88.0 | **88.8** | 12.0 | **79.9** | 83.6 | **81.7** |

Table 1: Experimental results on IC15, MSRA-TD500, CTW1500 and IC19-ArT. "Sem-based" represents segmentation-based methods and "geo-based" refers to the anchor-based approaches. "Ext" means extra data is used for pre-training. "Syn" and "MLT" denote the SynthText and IC17-MLT datasets, "$^+$" denotes the use of extra data. R, P, F and FPS refer to recall, precision, f-measure and frame per second, respectively. Best results are in bold, while the second ones are underlined.

| Method | R | P | F |
|---|---|---|---|
| PSENet-1s (Wang et al. 2019a) | 68.2 | 73.8 | 70.9 |
| LOMO (Zhang et al. 2019) | 60.6 | 78.8 | 68.5 |
| CRAFT (Baek et al. 2019) | 68.2 | 80.6 | 73.9 |
| SPCNet (Xie et al. 2019) | 68.6 | 80.6 | 74.1 |
| DRRG (Zhang et al. 2020) | 61.0 | 75.0 | 67.3 |
| SD (Xiao et al. 2020) | 72.8 | 84.2 | 78.1 |
| DB++ (Liao et al. 2022) | 67.9 | 83.1 | 74.7 |
| FSG (Tang et al. 2022) | 73.2 | **87.3** | 79.6 |
| **Ours** | **75.4** | 85.3 | **80.0** |

Table 2: Experimental results on IC17-MLT dataset.

| Sem | Geo | IFA | IC17-MLT | | | CTW1500 | | |
|---|---|---|---|---|---|---|---|---|
| | | | R | P | F | R | P | F |
| ✓ | | | 64.7 | 82.9 | 72.7 | 84.3 | **90.1** | 87.1 |
| | ✓ | | 73.0 | 85.1 | 78.6 | 86.4 | 88.3 | 87.3 |
| ✓ | ✓ | | 74.9 | 84.6 | 79.5 | 87.7 | 88.0 | 87.9 |
| ✓ | ✓ | ✓ | **75.4** | **85.3** | **80.0** | **89.6** | 88.0 | **88.8** |

Table 3: Ablation study of different modules on IC17-MLT and CTW1500 datasets. "Sem" represents the Deformable MSN generating semantic proposals and "Geo" refers to the Balanced RPN producing geometrical ones.

state-of-the-art method, FSG (Tang et al. 2022), we achieve a 0.4% gain in f-measure and 2.2% in recall without using additional training data such as SynthText (Gupta, Vedaldi, and Zisserman 2016). To our best knowledge, our approach is the first framework to achieve an f-measure over 80%.

## Complementary Proposal Analysis

Furthermore, we investigate whether the two branches in CPN, Deformable MSN and Balanced RPN, can produce high-recall and high-quality proposals, as well as whether

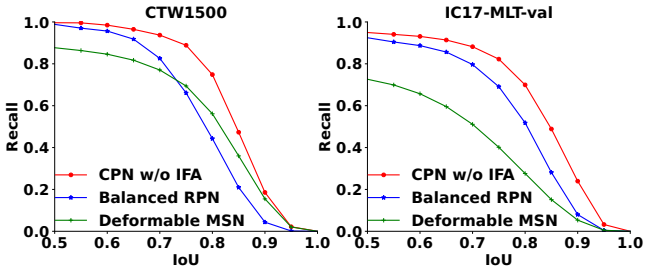


Figure 5: Recall vs. IoU overlap ratio on the CTW1500 test set and IC17-MLT validation set. Rotated bounding boxes are applied while computing IoU on CTW1500.

they complement each other. Experiments are typically conducted on the curved CTW1500 test set and the multilingual IC17-MLT validation set. In Figure 5, we give the Recall-to-IoU curve which is related to the quality of proposals. The left plots on CTW1500 test set in Figure 5 show that the Deformable MSN performs better when IoU gets larger, while Balanced RPN behaves completely opposite. This is intuitively accepted since semantic proposals are more accurate than geometric proposals for curved and long text instances. Regarding to the CPN$^*$ (without the IFA module), the recall is consistently much higher than either one across all IoU thresholds, indicating that the two parallel branches can effectively complement each other. On the IC17-MLT validation set, Balanced RPN gives higher recall than Deformable MSN under all IoU threshold settings, suggesting that anchor-based methods may perform better than segmentation-based methods on challenging multilingual datasets. Similarly, our CPN$^*$ shows a steady increase in recall compared to both approaches. In summary, the proposed two networks can well complement each other with much higher recall.


Figure 6: The qualitative results of our proposed method under various scenes, such as curved text, long text line, multi-oriented text, complex layout and multilingual text.

| Dataset | Number of Geometric proposals | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 100 | 300 | 500 | 1000 | 2000 |
| IC15 | 88.15 | 90.21 | 90.36 | 90.40 | 90.41 | 90.41 |
| CTW1500 | 87.63 | 88.70 | 88.80 | 88.80 | 88.76 | 88.71 |

Table 4: F-meansure on the CTW1500 and IC15 dataset when setting various number of proposals in Balanced RPN.

| Method | | GFLOPs | Params | FPS |
|---|---|---|---|---|
| Mask RCNN | | 142.39 | 43.75M | 9.6 |
| Ours | CPN w/o IFA | 117.74 | 42.54M | 10.1 |
| | CPN | **99.61** | **37.38M** | **11.2** |

Table 5: Comparisons on the FLOPs, number of parameters, and inference speed.

## Ablation Study

**Effectiveness of the two proposal networks** To validate the effectiveness of the proposed Deformable MSN and Balanced RPN branches, we conduct ablation study on the IC17-MLT and CTW1500 datasets. The components studied and corresponding results are summarized in Table 3. With regards to the Deformable MSN, the recall is much lower compared to the Balanced RPN, which validates the proof in Figure 5. The consistent and substantial gains observed on both datasets after integrating the two networks demonstrate the effectiveness of our design.

**Effectiveness of the IFA** Table 3 shows that the network can further improve the f-measure by 0.5% and 0.9% on the IC17-MLT and CTW1500 datasets when incorporating the IFA. This suggests that the interaction between semantic and geometric features is crucial for enhancing performance. Furthermore, the IFA employs a lightweight design with only half the number of feature channels of FPN, which reduces the overall network's GLOPs by about 15% and parameters by 12%, as illustrated in Table 5.

**Influence of the proposal number** Traditionally, anchor-based detectors require more than 1k proposals in RPN for higher recall. However, our semantic network generates proposals after the component grouping operation, which typically results in less than 100 proposals. As we have two complementary proposal networks, we investigated how the number of proposals influences detection performance by varying the number of geometrical proposals. Our results in Table 4 suggest that using a large number of proposals only slightly improves performance. With the aid of Deformable MSN, our CPN becomes less reliant on redundant proposals.

**Complexity of the model** To analyze the model complexity of our network, we compute the FLOPs, number of model parameters, and inference speed. For a fair comparison, we resize the input images to 640 on both sides for all models to calculate the FLOPs. We also use the images from the IC15 test dataset to measure the inference speed by FPS. As shown in Table 5, our proposed architecture has a lower computational cost in terms of FLOPs, model size and a faster inference speed than the standard Mask R-CNN.

## Conclusions

We present the Complementary Proposal Network (CPN), an innovative approach that combines the strengths of segmentation-based and anchor-based methods for scene text detection. The CPN includes two efficient networks, the Deformable MSN and Balanced RPN, which work together to generate semantic and geometric proposals. The Deformable MSN produces semantic proposals based on instance segmentation, while the Balanced RPN generates geometric proposals based on pre-defined anchors. By working in concert, these two proposal branches reinforce each other's strengths and compensate for their individual weaknesses. The CPN has been tested on five popular scene text detection datasets and has achieved impressive results. Our study aims to encourage further exploration of the complementary relationships between anchor-based and segmentation-based methods for scene text detection and object detection in general.

## References


Baek, Y.; Lee, B.; Han, D.; Yun, S.; and Lee, H. 2019. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9365–9374.

Chng, C. K.; Liu, Y.; Sun, Y.; Ng, C. C.; Luo, C.; Ni, Z.; Fang, C.; Zhang, S.; Han, J.; Ding, E.; et al. 2019. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1571–1576. IEEE.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Dai, P.; Zhang, S.; Zhang, H.; and Cao, X. 2021. Progressive contour regression for arbitrary-shape scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7393–7402.

Deng, D.; Liu, H.; Li, X.; and Cai, D. 2018. Pixellink: Detecting scene text via instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Du, B.; Ye, J.; Zhang, J.; Liu, J.; and Tao, D. 2022. I3CL: Intra-and Inter-Instance Collaborative Learning for Arbitrary-shaped Scene Text Detection. *International Journal of Computer Vision*, 1–17.

Epshtein, B.; Ofek, E.; and Wexler, Y. 2010. Detecting text in natural scenes with stroke width transform. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 2963–2970. IEEE.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2315–2324.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, 1156–1160. IEEE.

Liao, M.; Shi, B.; Bai, X.; Wang, X.; and Liu, W. 2017. Textboxes: A fast text detector with a single deep neural network. In *Thirty-first AAAI conference on artificial intelligence*.

Liao, M.; Wan, Z.; Yao, C.; Chen, K.; and Bai, X. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11474–11481.

Liao, M.; Zou, Z.; Wan, Z.; Yao, C.; and Bai, X. 2022. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 919–931.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Liu, Y.; and Jin, L. 2017. Deep matching prior network: Toward tighter multi-oriented text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1962–1969.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; and Xue, X. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11): 3111–3122.

Nayef, N.; Yin, F.; Bizid, I.; Choi, H.; Feng, Y.; Karatzas, D.; Luo, Z.; Pal, U.; Rigaud, C.; Chazalon, J.; et al. 2017. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, 1454–1459. IEEE.

Neumann, L.; and Matas, J. 2012. Real-time scene text localization and recognition. In *2012 IEEE conference on computer vision and pattern recognition*, 3538–3545. IEEE.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Tang, J.; Zhang, W.; Liu, H.; Yang, M.; Jiang, B.; Hu, G.; and Bai, X. 2022. Few Could Be Better Than All: Feature Sampling and Grouping for Scene Text Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4563–4572.

Tian, Z.; Shu, M.; Lyu, P.; Li, R.; Zhou, C.; Shen, X.; and Jia, J. 2019. Learning shape-aware embedding for scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4234–4243.

Vatti, B. R. 1992. A generic solution to polygon clipping. *Communications of the ACM*, 35(7): 56–63.

Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; and Shao, S. 2019a. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9336–9345.

Wang, W.; Xie, E.; Song, X.; Zang, Y.; Wang, W.; Lu, T.; Yu, G.; and Shen, C. 2019b. Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network. In *Proceedings of the IEEE International Conference on Computer Vision*, 8440–8449.

Wang, Y.; Xie, H.; Zha, Z.-J.; Xing, M.; Fu, Z.; and Zhang, Y. 2020. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11753–11762.

Xiao, S.; Peng, L.; Yan, R.; An, K.; Yao, G.; and Min, J. 2020. Sequential deformation for accurate scene text detection. In *European Conference on Computer Vision*, 108–124. Springer.

Xie, E.; Zang, Y.; Shao, S.; Yu, G.; Yao, C.; and Li, G. 2019. Scene text detection with supervised pyramid context network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 9038–9045.

Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021. Oriented R-CNN for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3520–3529.

Xu, Y.; Wang, Y.; Zhou, W.; Wang, Y.; Yang, Z.; and Bai, X. 2019. Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Transactions on Image Processing*, 28(11): 5566–5579.

Yao, C.; Bai, X.; and Liu, W. 2014. A unified framework for multioriented text detection and recognition. *IEEE Transactions on Image Processing*, 23(11): 4737–4749.

Ye, M.; Zhang, J.; Zhao, S.; Liu, J.; Du, B.; and Tao, D. 2023. Dptext-detr: Towards better scene text detection with dynamic points in transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3241–3249.

Yuliang, L.; Lianwen, J.; Shuaitao, Z.; and Sheng, Z. 2017. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*.

Zhang, C.; Liang, B.; Huang, Z.; En, M.; Han, J.; Ding, E.; and Ding, X. 2019. Look more than once: An accurate detector for text of arbitrary shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10552–10561.

Zhang, S.-X.; Zhu, X.; Hou, J.-B.; Liu, C.; Yang, C.; Wang, H.; and Yin, X.-C. 2020. Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9699–9708.

Zhang, S.-X.; Zhu, X.; Yang, C.; and Yin, X.-C. 2022. Arbitrary Shape Text Detection via Boundary Transformer. *arXiv preprint arXiv:2205.05320*.