FPT Education
FPT UNIVERSITY

TextFocus: Efficient Multi-Scale Detection for Arbitrary Scene Text

Author: Do Quang Manh, Tran Minh Khoi, Duong Minh Hieu
Thesis Supervisor: Phan Duy Hung

Table of Contents

01

Introduction

Efficient multi-scale text detection for every resolution

Background

Scene Text Detection represents a pivotal and pervasive computer vision task characterized by its significance in diverse domains.

The challenge pertains to the nuanced realm of arbitrary-text instances

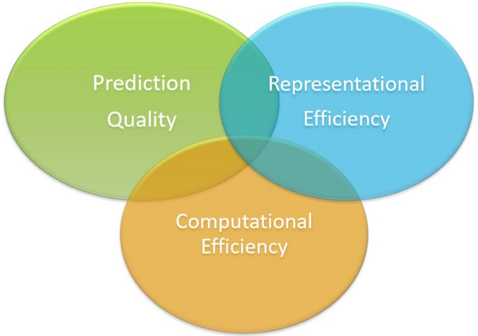
Background

Trade-off between speed and accuracy

The complexity of scene text detection is increased by high-resolution images with tiny textual content.

Objectives

Investigate the current state-of-the-art in arbitrary shape text detection algorithms: identify the limitations of existing methods, and propose an improved approach using multi resolution technique.



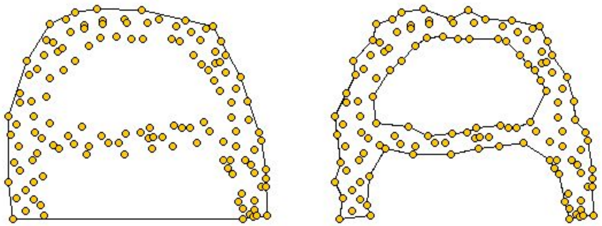
Develop a novel text detection algorithm that has ability to accurately and expediently detect instances of text within images characterized by multi-resolution attributes.

Efficient in terms of speed, resources, and computing usage.

Contributions

Propose TextFocus: levering the power of multiple resolution techniques, precisely the trade-off between accuracy and resource consumption concerns intrinsically linked with high-resolution sample training.

Applying the Alpha-shape algorithm to generate new annotations for the text dataset.



⇒ Provide a promising avenue for further study in text detection.

02

Theoretical foundation

Articles related to the problem and foundation theoris

Literature Review

1. Related works

2. Foundation theories

1. Related works

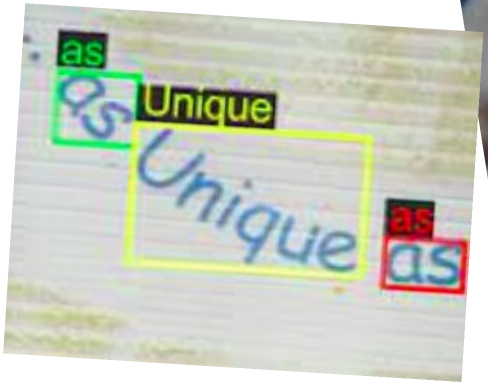
1.1. History of arbitrary shape text detection methods

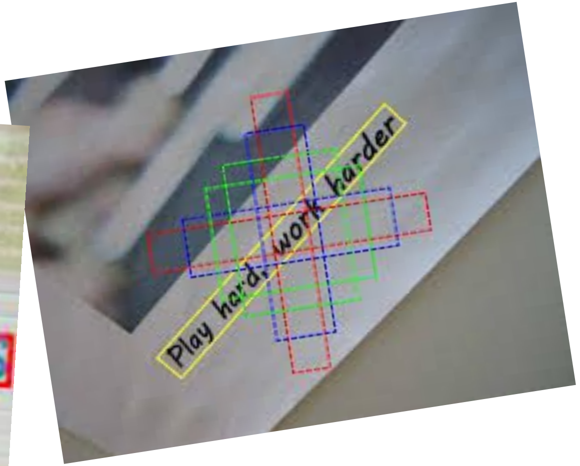
- The problem of arbitrary-shape text detection is challenging due to the variety of shapes and appearances text can take.
-
- Methods can be divided into different groups: regression-based methods, segmentation-based methods, and contour-based methods

1. Related works

1.1. History of arbitrary shape text detection methods

- **Regression-based methods**

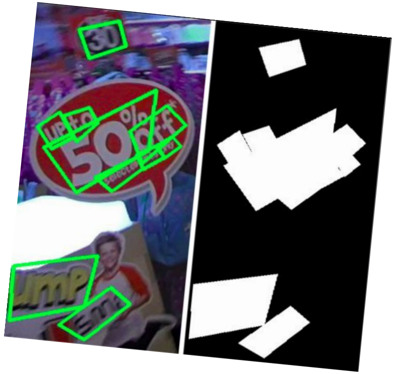




1. Related works

1.1. History of arbitrary shape text detection methods

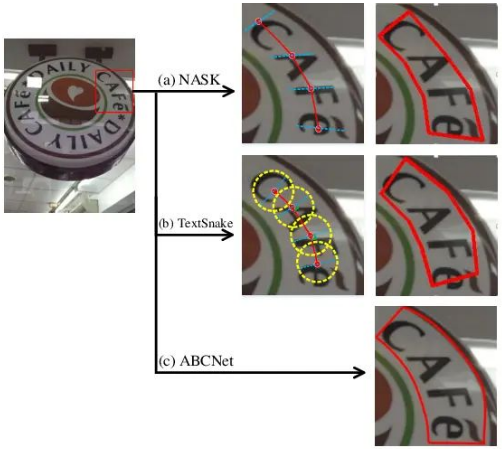
- Segmentation-based methods



1. Related works

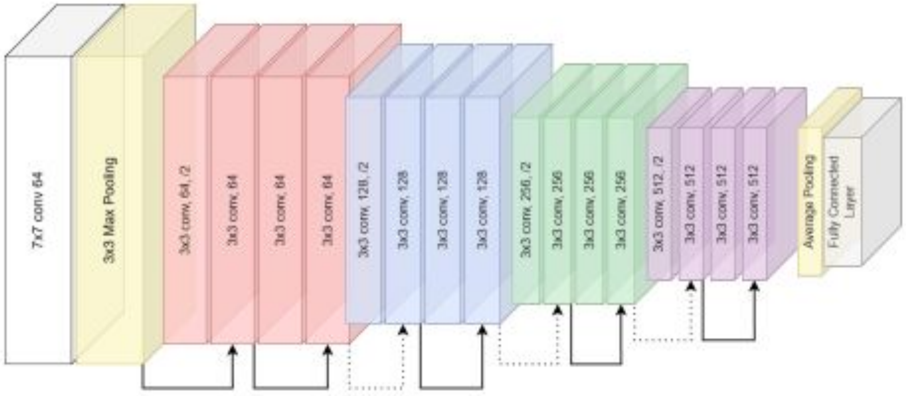
1.1. History of arbitrary shape text detection methods

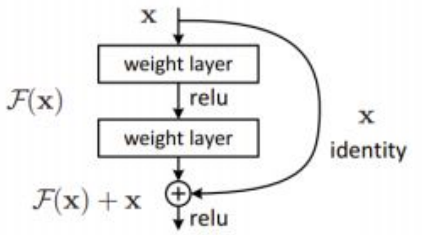
- **Contour-based methods**



2. Foundational theories

2.1. ResNet-18





- ResNet-18 is a powerful and versatile tool for deep learning that has been used to achieve good results specific on speed and resource consumption on many different tasks.

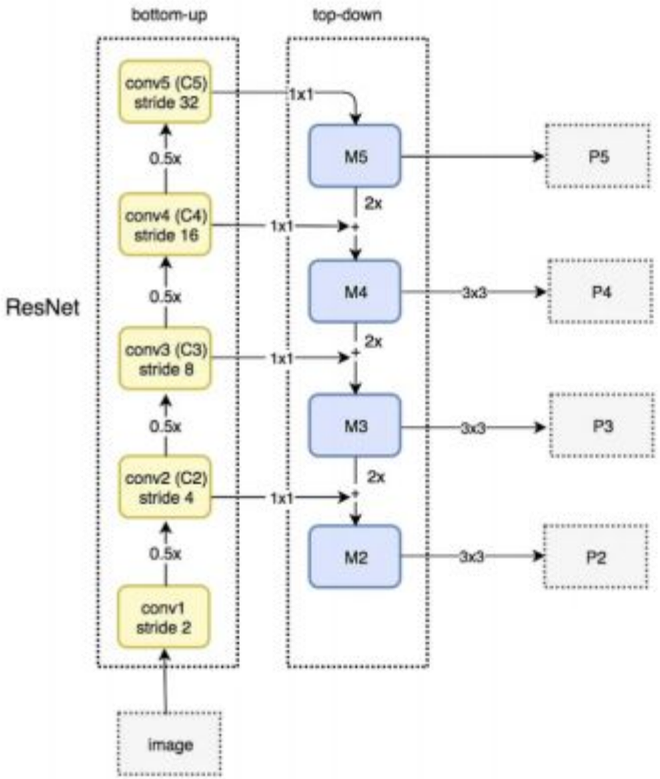
- The architecture achieves commendable accuracy metrics while concurrently adhering to the reasonable parameter count, thereby eclipsing antecedent architectures in efficiency.

2. Foundational theories

2.2. Feature Pyramid Network

- FPN, characterized by its bottom-up and top-down pathways, has a complex interaction between feature extraction and semantic comprehension.

- Greatly benefits tasks such as object detection by facilitating the discernment of objects within images of varying scales and complexities.

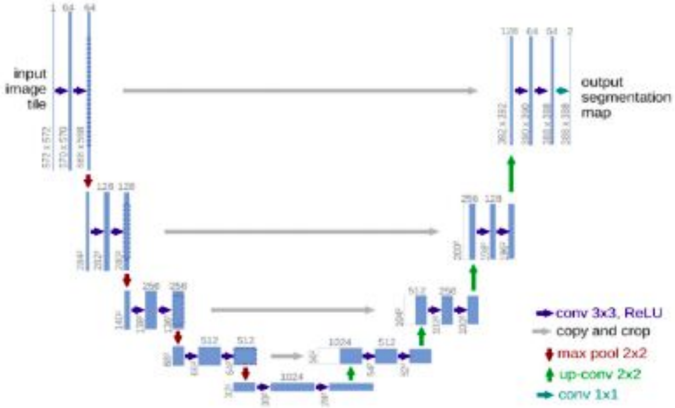


2. Foundational theories

2.3. Encoder-Decoder architecture

- Encoder-Decoder is a widely used deep learning technique successfully applied to various tasks in computer vision and natural language processing.

- The architecture consists of two main components:
  - The encoder takes the input data and generates a lower-dimensional feature representation that captures the most critical information in the data.
  - The decoder produces the output sequence or image by mapping the features back to the original input space.



03

Methodology

Methodology

1. Data enhancement and preprocessing

2. Baseline architecture

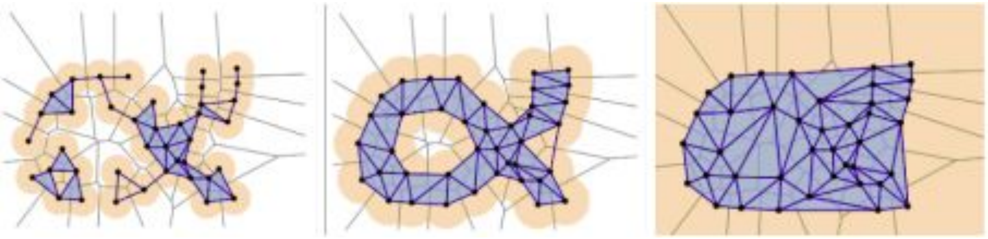
3. Pixel Aggregation Network - PAN

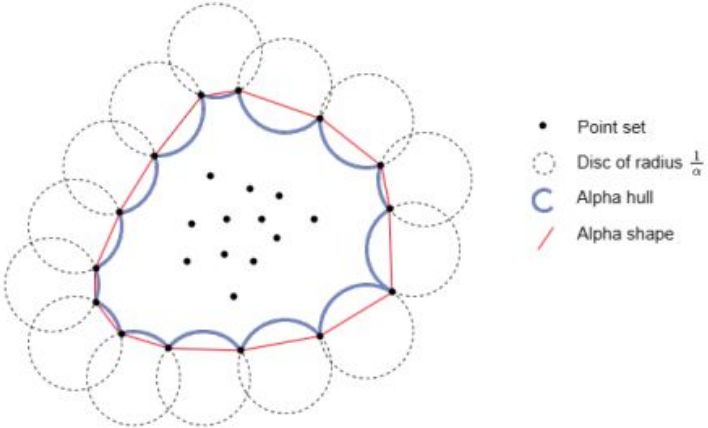
4. Focus branch

5. Implementing TextFocus

Data enhancement and preprocessing

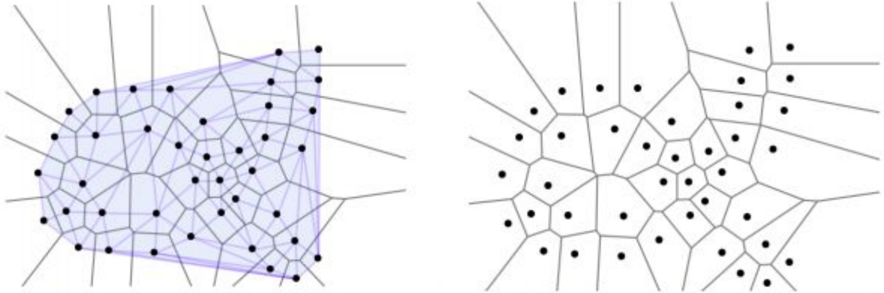
Generate new annotations for CTW dataset with alpha shape :



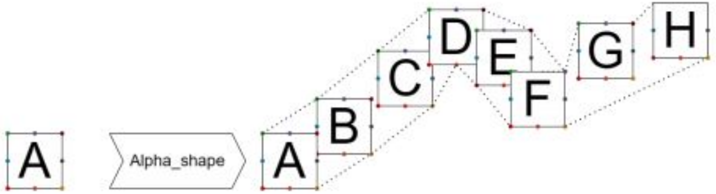


Data enhancement and preprocessing

Generate new annotations for CTW dataset with alpha shape :



Voronoi diagram of a set of points   Voronoi diagram and Delaunay triangulation

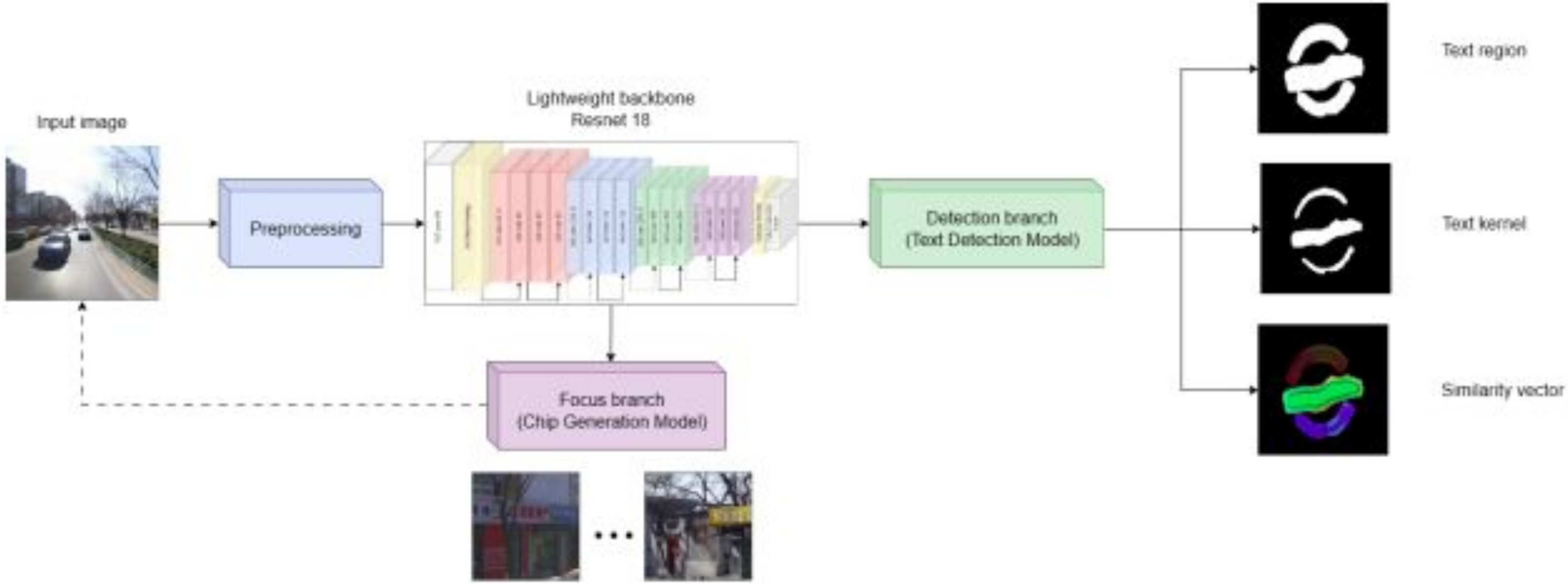


- CTW Dataset exclusively encompassed annotations corresponding to instances of Chinese characters discernible within each individual image.

- Alpha-Shape algorithm was instrumental in circumventing the aforementioned constraint and served as the mechanism through which delineations of text instance boundaries endowed with arbitrary geometries were synthesized.

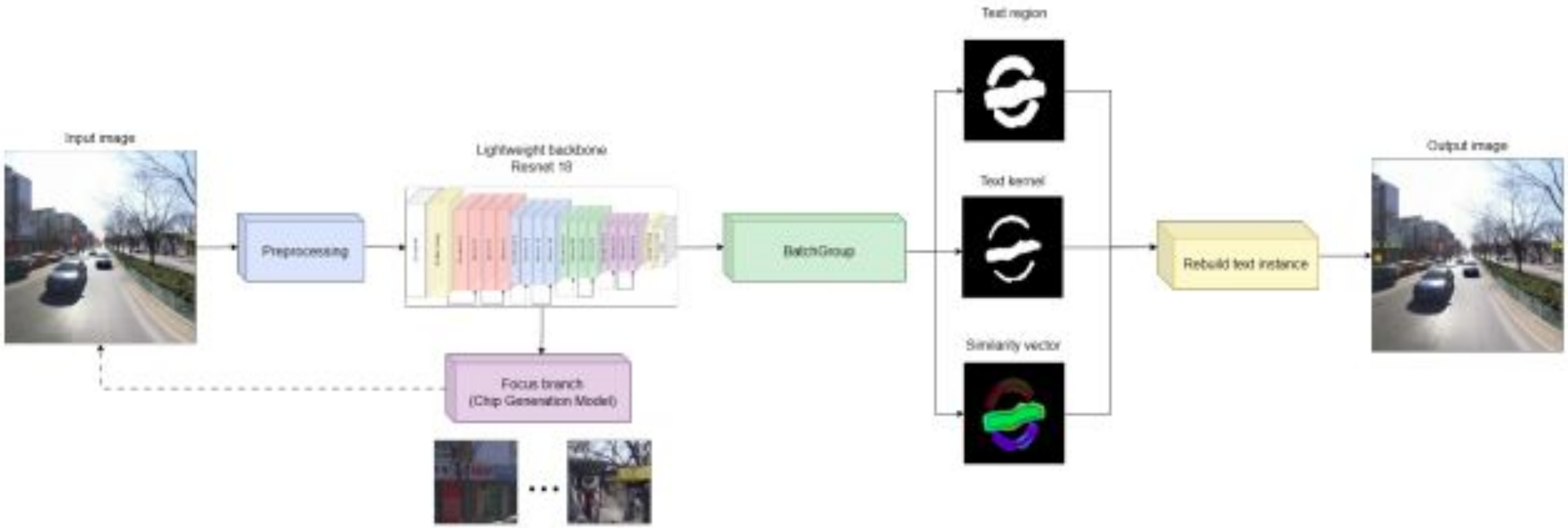
Baseline architecture

1. Training pipeline :

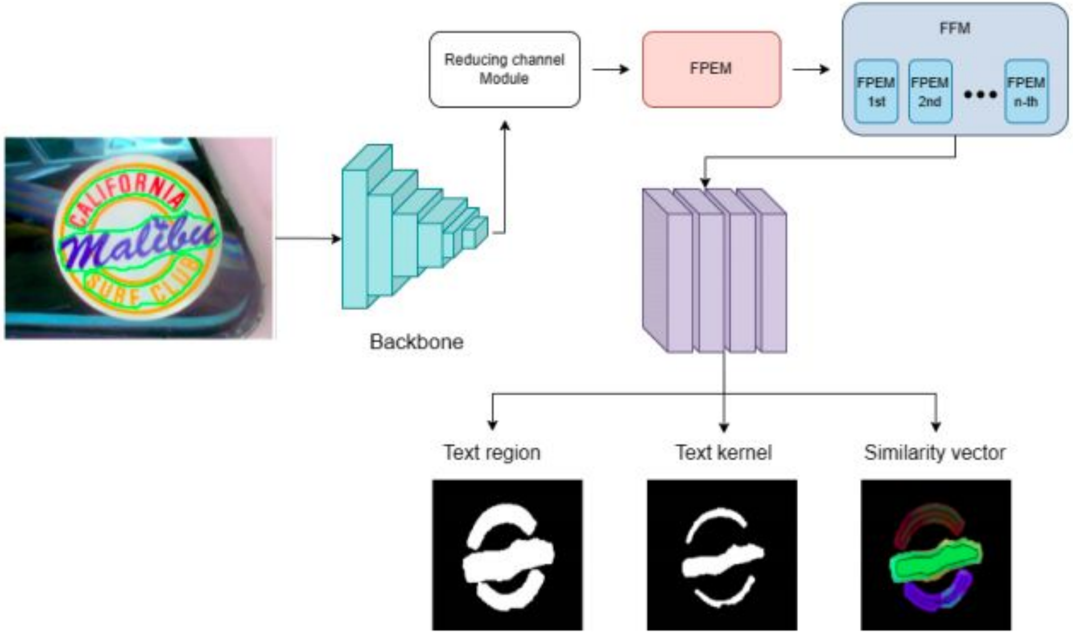


Baseline architecture

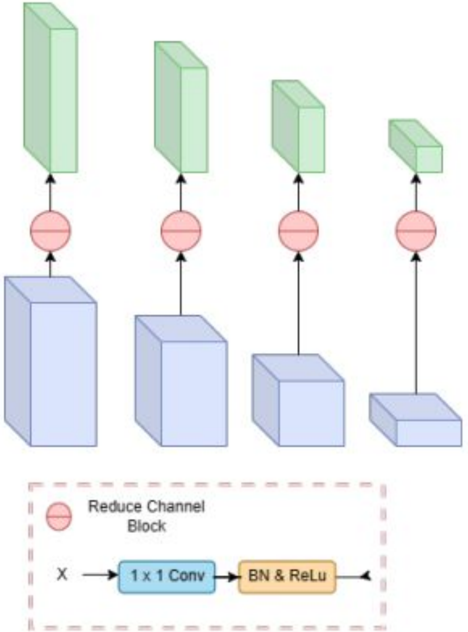
2. Inference pipeline :



Pixel Aggregation Network - PAN

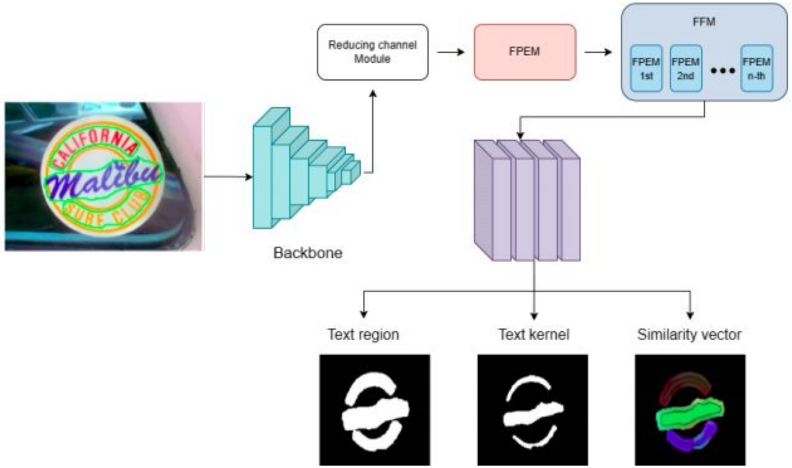


PAN architecture Overview

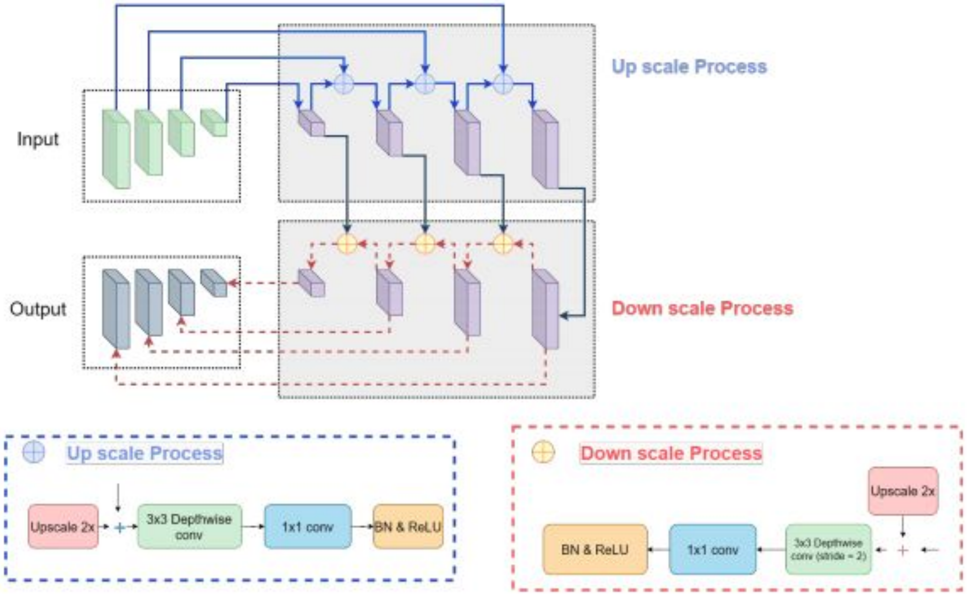


Reducing Channel Block

Pixel Aggregation Network - PAN

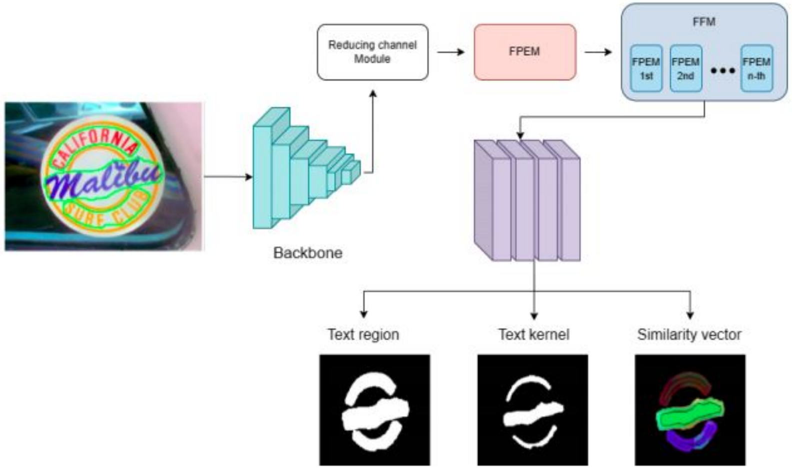


PAN architecture Overview

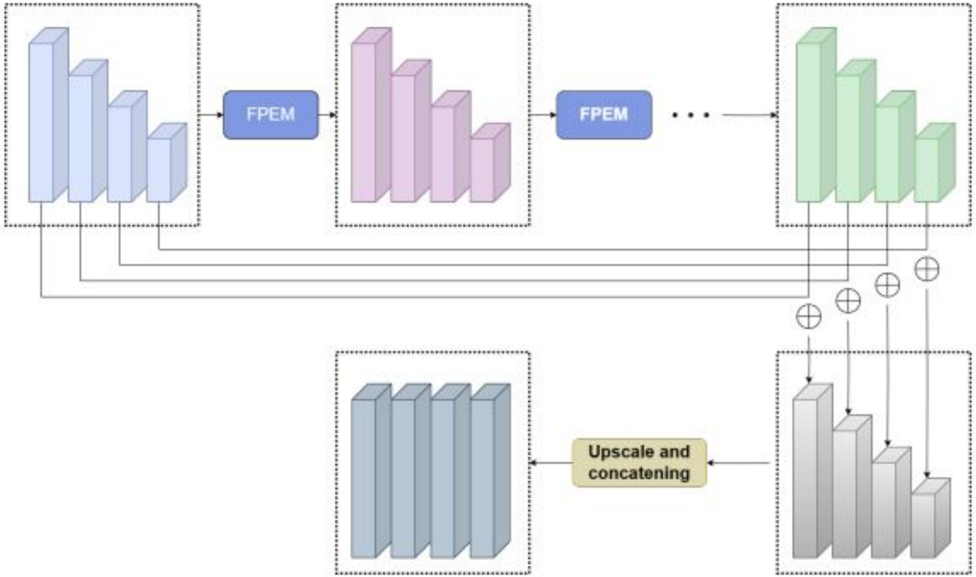


Reducing Channel Block

Pixel Aggregation Network - PAN



PAN architecture Overview



FFM
Feature Fusion Model

Pixel Aggregation Network - PAN

Pixel aggregation loss

$$L_{agg} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|T_i|}\sum_{p\in T_i} ln(\mathcal{D}(p, K_i) + 1)$$

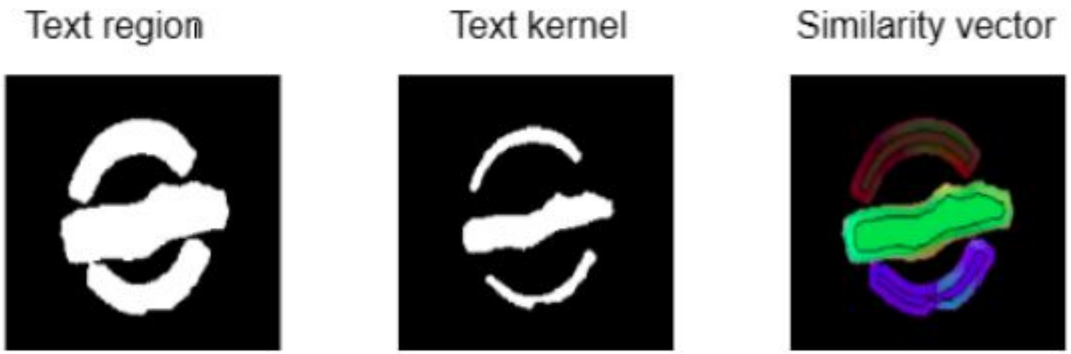
with $\mathcal{D}(p, K_i) = max(||\mathcal{F}(p) - \mathcal{G}(K_i)|| - \theta_{agg}, 0)^2$

T_i is the i_{th} text instance

$F(p)$ is the similarity vector of the pixel p

$\mathcal{G}(\cdot) = \sum_{q\in K_i}\mathcal{F}(q)/|K_i|$

Pixel Aggregation Network - PAN

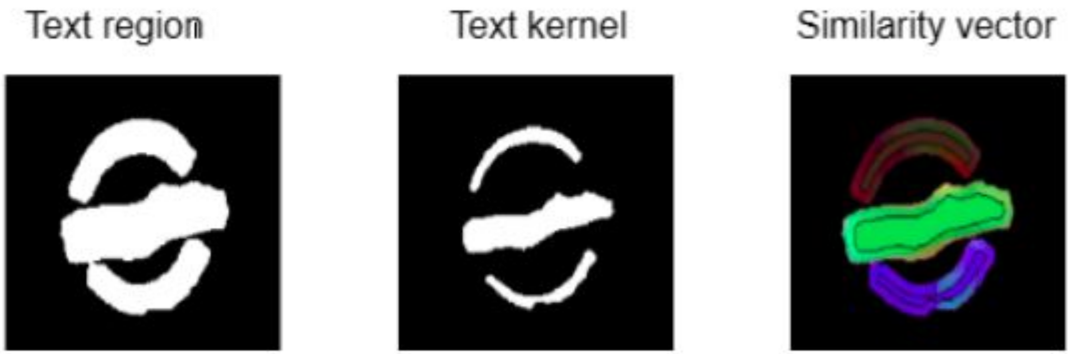


Pixel distance loss

$$L_{dis} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} ln(\mathcal{D}(K_i, K_j) + 1)$$

with $\mathcal{D}(K_i, K_j) = max(\theta_{dis} - ||\mathcal{G}(K_i) - \mathcal{G}(K_j)||, 0)^2$

Pixel Aggregation Network - PAN



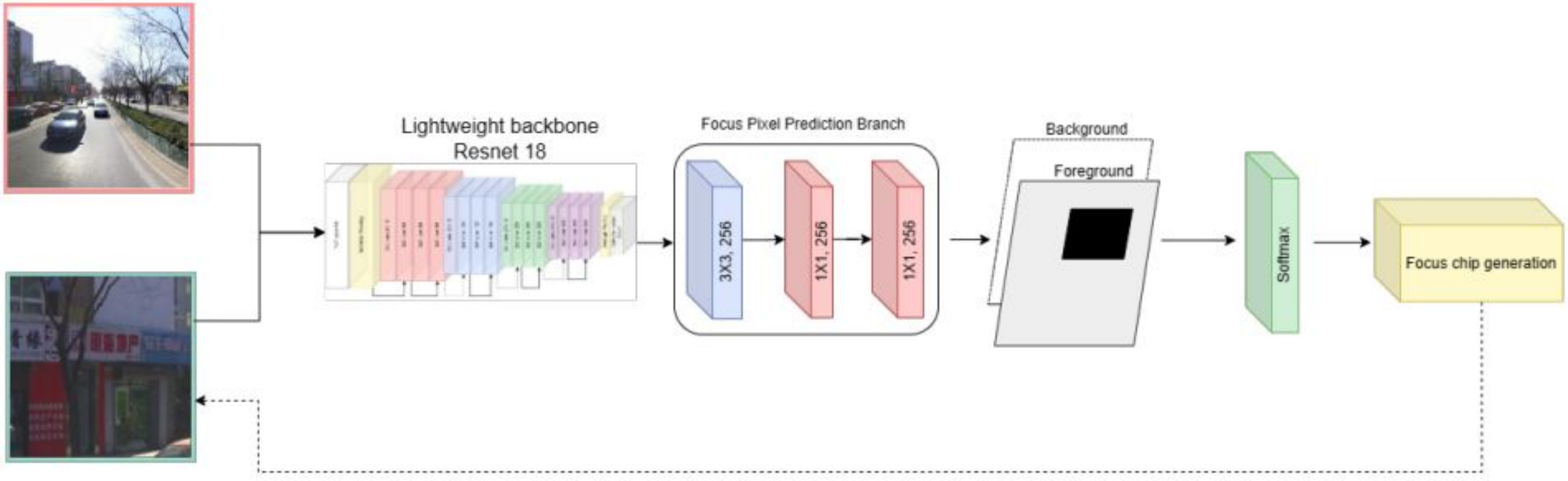
Text regions loss and Kernels loss

$$\mathcal{L}_{tex} = 1 - \frac{2\sum_i P_{tex}(i) G_{tex}(i)}{\sum_i P_{tex}(i)^2 + \sum_i G_{tex}(i)^2},$$

$$\mathcal{L}_{ker} = 1 - \frac{2\sum_i P_{ker}(i) G_{ker}(i)}{\sum_i P_{ker}(i)^2 + \sum_i G_{ker}(i)^2},$$

Focus branch

1. Focus Pixels Finding:



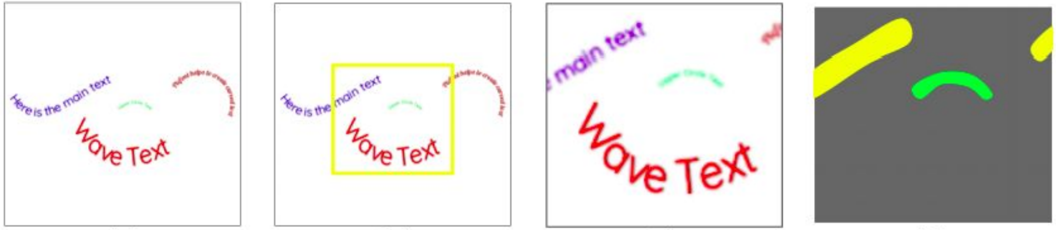
Focus branch

$$F_{conv5}, F_{conv6}, F_{conv7}, F_{conv8} = F_b(I_n)$$

$$P = Sigmoid(F_{Focus}(F_{conv5}))$$

Focus branch

1. Focus Pixels Finding:



(a) (b) (c) (d)

Specifically, given an image of size $W \times H$, and the whole backbone with stride is s, the labels L will be of size $W' \times H'$, where $W' = \lceil \frac{W}{s} \rceil$ and $H' = \lceil \frac{H}{s} \rceil$. Since the stride is s , the rach label $l \in L$ corresponds to $s \times s$ pixels in the image. The label l is defined as follows,

$$l = \begin{cases} 1, & IoU(GT, l) > 0, a < \sqrt{GTArea} < b \\ -1, & IoU(GT, l) > 0, \sqrt{GTArea} < a \\ -1, & IoU(GT, l) > 0, b < \sqrt{GTArea} < c \\ 0, & otherwise \end{cases} \tag{3.8}$$

Focus branch

1. Focus Pixels Finding:

Focus loss

$$\mathcal{L}_{Focus} = -\sum_{i}^{W'}\sum_{j}^{H'}\sum_{c}^{C} t_{i,j,c}k_{i,j}log(p_{i,j,c}) / \sum_{i}^{W'}\sum_{j}^{H'} k_{i,j}$$

$k_{i,j} = 0$ if pixel at position i, j of groundtruth focus map is ignored; $k_{i,j} = 1$ otherwise. $c \in (0, 1)$. $t_{i,j,c} = 1$ if pixel at position i, j of grountruth focus map is c; $t_{i,j,c} = 0$ otherwise. $p_{i,j,c}$ is probability prediction of pixel i, j classified as c.

Focus branch

2. Focus Chips Generation:

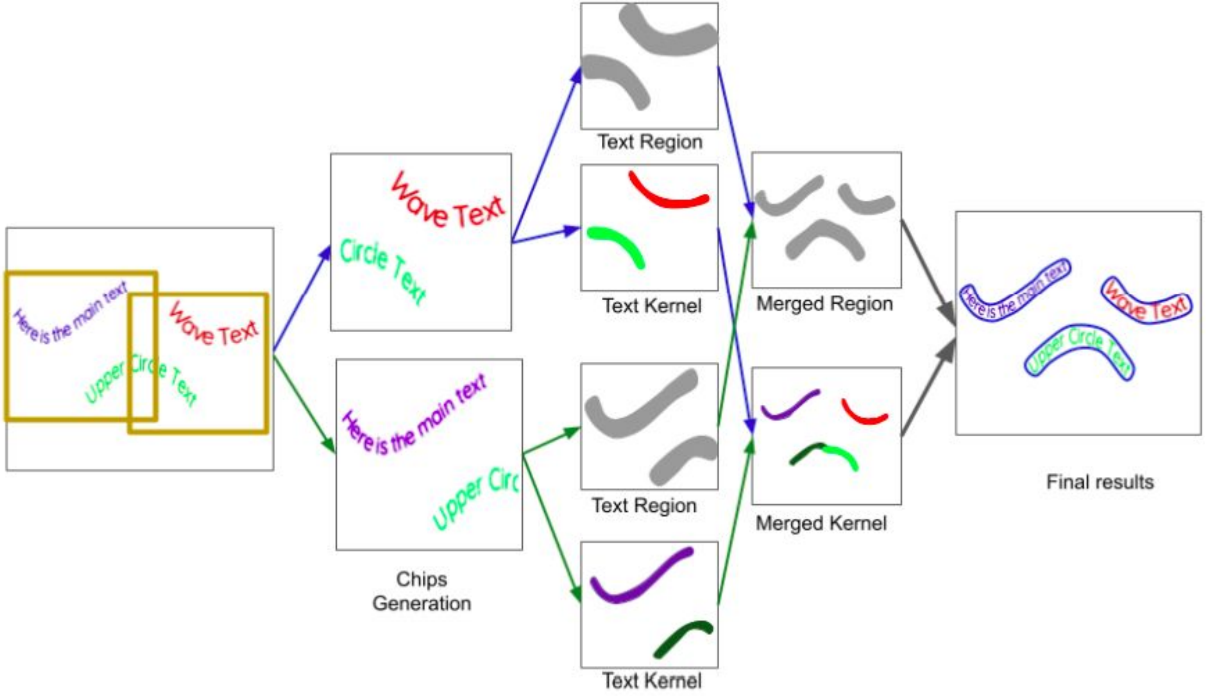
Algorithm 1: Focus Chips Generation

Input: Focus map P, threshold t, dilation constant d, minimum size k
Output: Chips C

```
Transform P into the binary map using threshold t
Dilate binary map with a dxd filter
Obtain the list of connected components S
Generate enclosing chips C for each component in S if the component size is
  larger than k
If chips C overlap, merge these chips
return Chips C
```

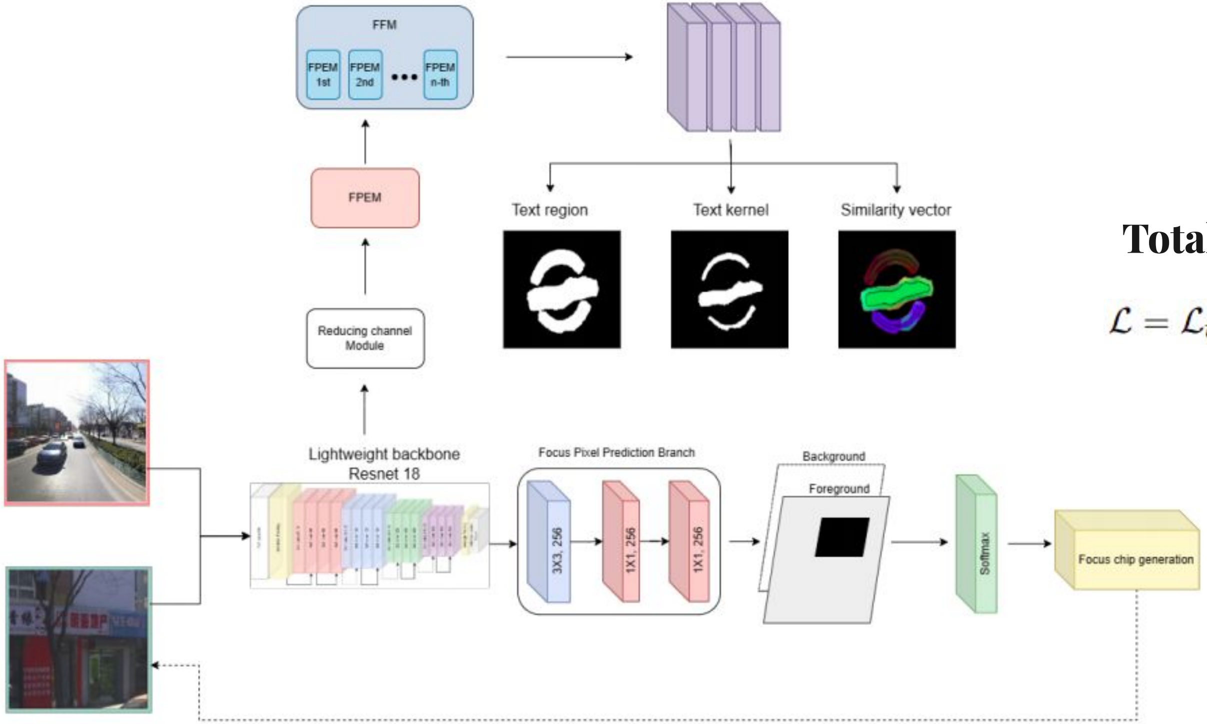
Focus branch

3. Focus Combination for final results:



Description for focus branch process

Implementing TextFocus



The complete architecture of TextFocus

Total loss

$$\mathcal{L} = \mathcal{L}_{tex} + \alpha\mathcal{L}_{ker} + \beta(\mathcal{L}_{agg} + \mathcal{L}_{dis}) + \gamma\mathcal{L}_{Focus}$$

04

Experiments and Results

Datasets

Datasets	Range of resolution(pixel)	Real/synthetic	Annotation level	Language in image
Large CTW [2]	320x240 - 3840x3200	real	word/line	Chinese
Total Text [43]	640x480 - 1920x1080	real	word/line	English
ICDAR15 [42]	300x300 - 2400x2400	real	word/line	Various
Scut-CTW1500 [1]	640x480 - 1920x1080	real	word/line	Chinese
Synth Text [49]	320x240 - 1920x1080	synthetic	word/line	Various

The detail information of datasets

Results and Analysis

Method	ICDAR2015 [42]				Total-Text [43]				SCUT-CTW1500 [1]			
	P	R	F1	FPS	P	R	F1	FPS	P	R	F1	FPS
CTPN [51]	74.2	51.6	60.9	3.55	-	-	-	-	60.4	53.8	56.9	3.57
SegLink [52]	73.1	76.8	75.0	-	30.3	23.8	26.7	-	42.3	40.0	40.8	1.35
EAST [20]	83.6	73.5	78.2	-	50.0	36.2	42.0	-	78.7	49.1	60.4	2.52
RRPN [53]	82.0	73.0	77.0	-	-	-	-	-	-	-	-	-
PSENet [26]	84.5	**86.9**	**85.7**	0.8	78.0	**84.0**	80.9	1.95	79.7	84.8	82.2	0.9
TextSnake [54]	84.9	80.4	8.6	0.55	82.7	74.5	78.4	-	67.9	**85.3**	75.6	-
PAN [27]	84.0	81.9	82.9	**12.29**	**83.6**	78.5	80.1	10.11	**86.4**	81.2	83.7	13.11
Ours (320)	**86.1**	74.5	79.9	8.51	82.7	74.1	78.1	**11.13**	84.8	80.9	82.8	**14.21**
Ours (640)	84.3	85.1	84.7	1.92	82.6	81.5	**82.05**	2.12	84.4	83.8	**84.9**	2.45

Results on ICDAR2015 , Total-Text, SCUT-CTW1500. ”P”, “R”, “F” and “FPS” represent the precision, recall, F-measure, and frame per second, respectively.

Results and Analysis

Method	Large CTW [H-7]			
	R	P	F	FPS
Ours (448)	54.6	52.3	53.4	5.12
Ours (640)	**62.1**	**60.1**	**61.1**	**1.71**

Results on Large CTW. ”P”, “R”, “F” and “FPS” represent the precision, recall, F-measure, and frame per second, respectively.

Results and Analysis

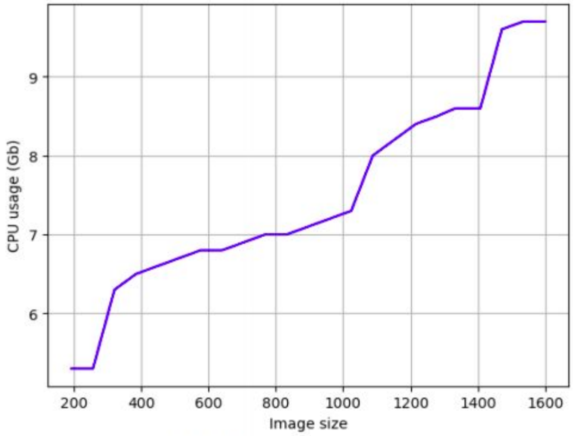
The effectiveness and influence of the backbone and Detection branch

- The model is very effective and can be used in real-time thanks to its lightweight backbone, cascaded pipeline strategy, and segment-based text detection.

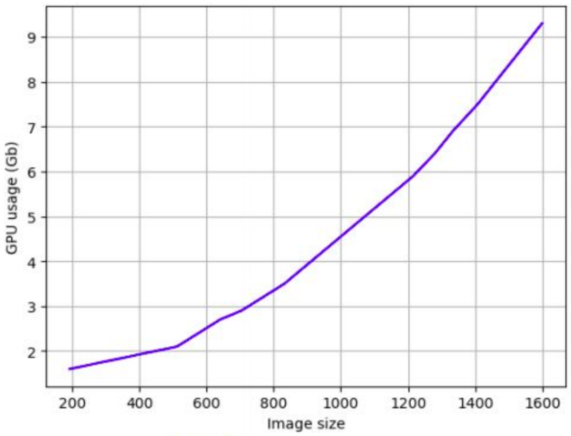
Results and Analysis

The effectiveness of Focus branch

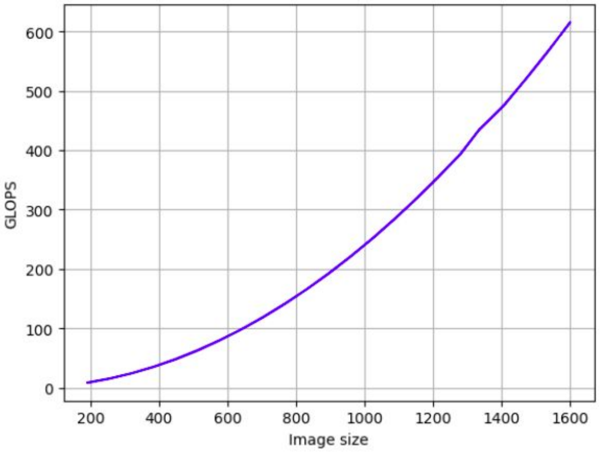
- Saving on resources and computation



CPUconsumption

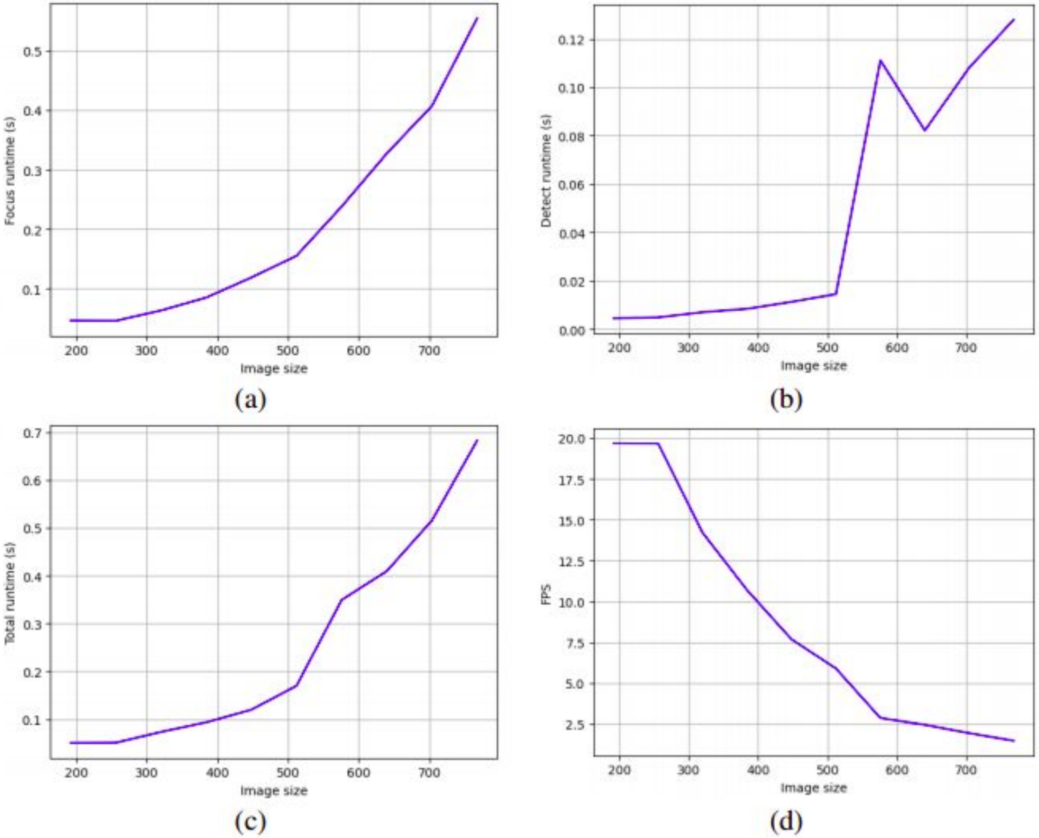


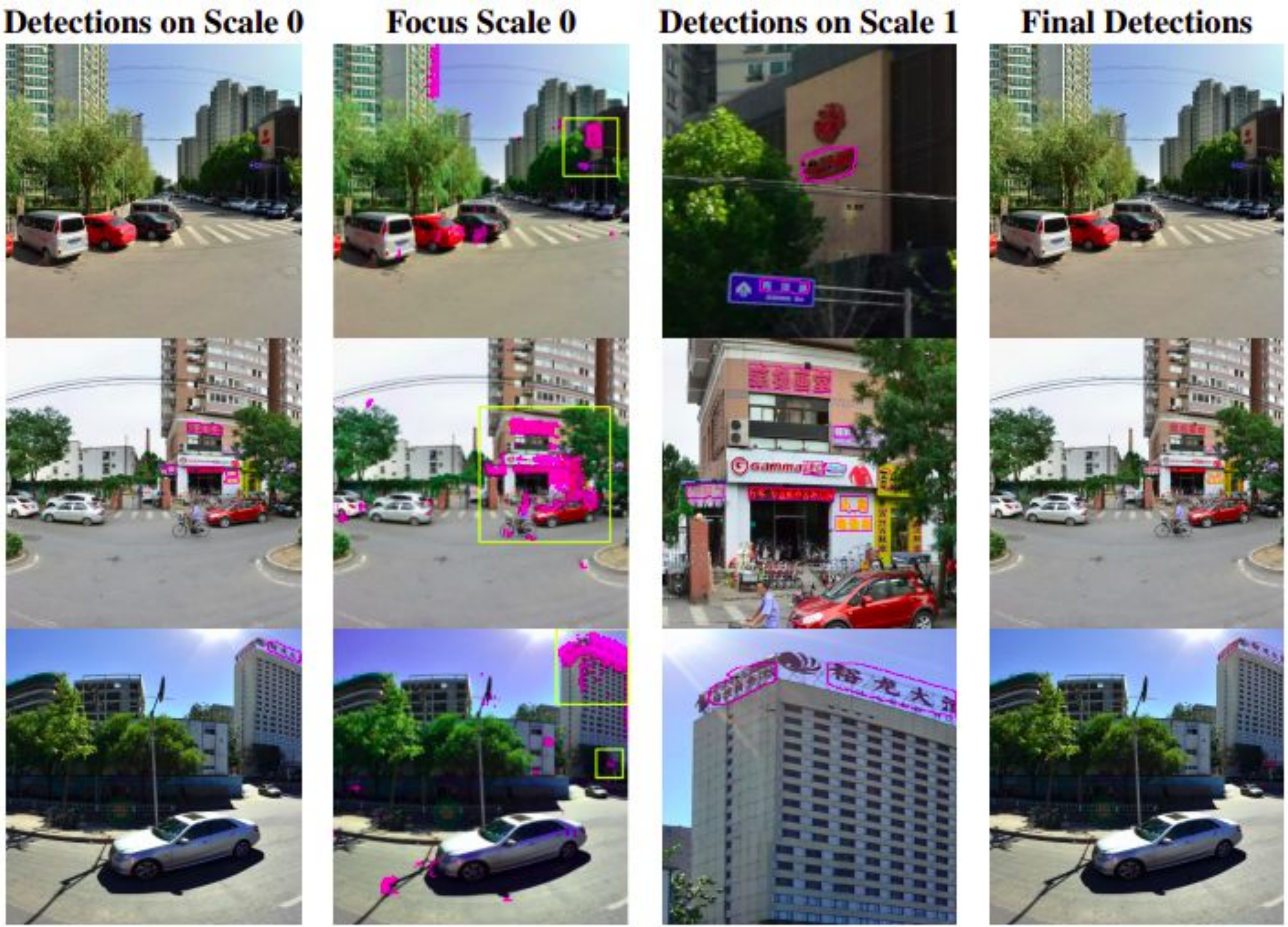
GPU consumption



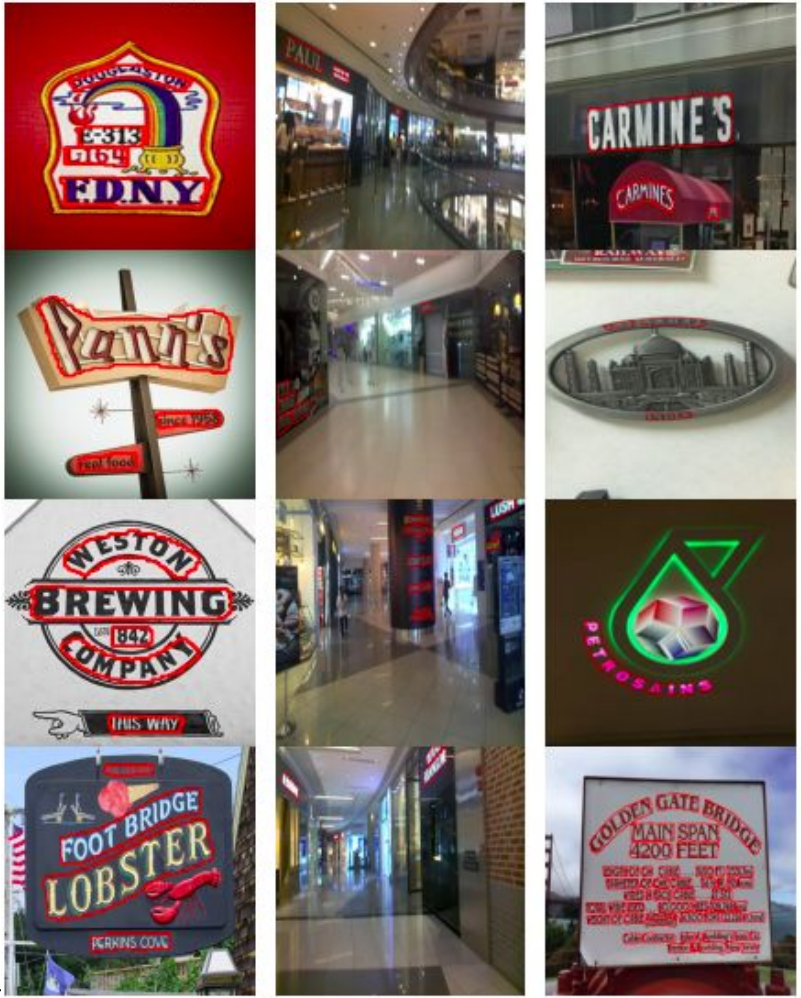
Results and Analysis

- Real-time text detection




Detections on Scale 0
Focus Scale 0
Detections on Scale 1
Final Detections

CTW1500 [1] ICDAR [42] TotalText [43]



05

Conclusion and Future Works

Conclusion and Future Works

1. Conclusion

2. Future Works

Conclusion and Future Works

1. Conclusion

- Proposed a novel method for arbitrary shape text detection using multiple resolutions.

- Designed the TextFocus model with a multi-resolution strategy to enhance text detection accuracy in real-world images.

- Conducted extensive experiments that showed significant improvements over the baseline model in terms of FPS and TIoU-metric.

- Acknowledged limitations, including the need to improve performance on low-resolution images, enhance the focus branch head, and reduce computational complexity for practical applications.

Conclusion and Future Works

2. Future works

- Incorporate advanced deep learning techniques like transfer learning and attention mechanisms.
- Develop a more robust model for effective handling of multiresolution images.
- Revise synthetic data generation for better text placement and segmentation accuracy.
- Extend the method to other image detection applications, such as mini object detection and OCR.
- Explore real-time integration with edge devices, like traffic cameras, for intelligent traffic systems.

Thanks!

QS STARS RATED FOR EXCELLENCE 2012

FPT Education

FPT UNIVERSITY

Demo and Q&A

STARS
RATED FOR EXCELLENCE
2012

FPT Education

FPT UNIVERSITY