Dynamic Relation Transformer for Contextual Text Block Detection

Jiawei Wang[1,3], Shunchi Zhang[2,3], Kai Hu[1,3], Chixiang Ma[3], Zhuoyao Zhong[3], Lei Sun[3], Qiang Huo[3]

[1]University of Science and Technology of China, Hefei, China

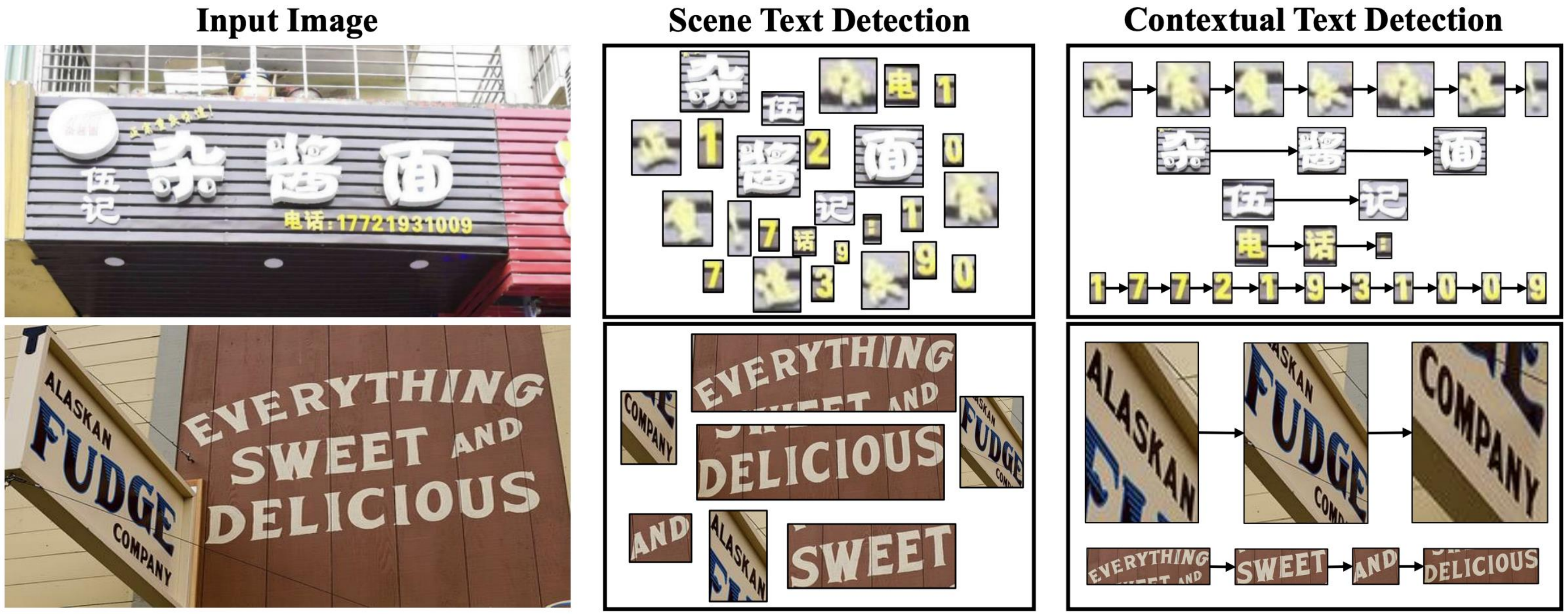
[2]Xi'an Jiaotong University, Xi'an, China

[3]Microsoft Research Asia, Beijing, China

ICDAR-2024, Athens, Greece, 30 August – 4 September, 2024

Goal of Contextual Text Block Detection

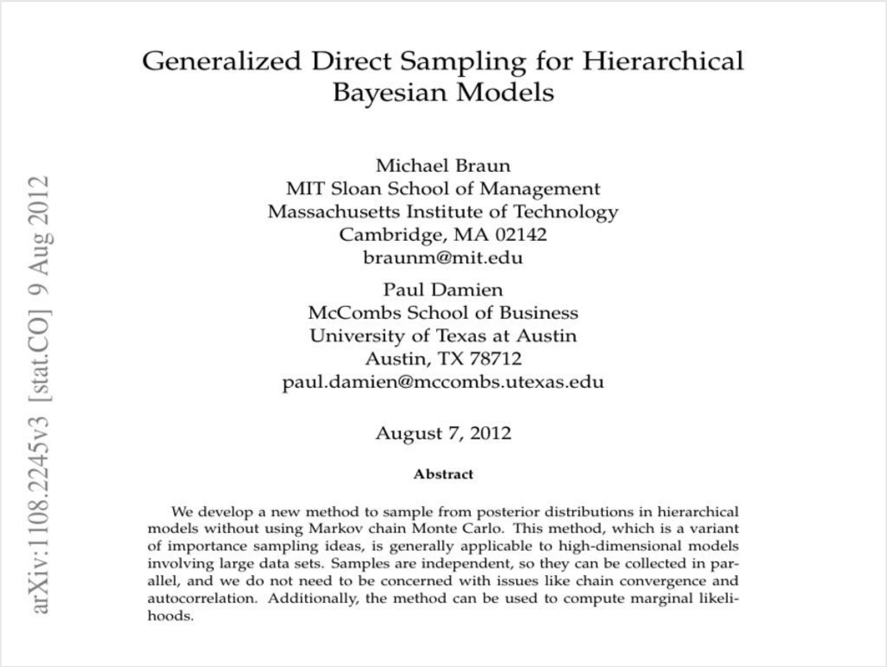
- Contextual Text Block Detection[1] (CTBD) aims to detect contextual text blocks within natural scenes, which are aggregates of one or more integral text units, such as characters, words, or text-lines, arranged in their natural reading order.



[1] Xue, Chuhui, et al. "Contextual text block detection towards scene text understanding." *European Conference on Computer Vision*. 2022.

Challenges of Contextual Text Block Detection

Within Document Images



Generalized Direct Sampling for Hierarchical Bayesian Models

Michael Braun
MIT Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02142
braunm@mit.edu

Paul Damien
McCombs School of Business
University of Texas at Austin
Austin, TX 78712
paul.damien@mccombs.utexas.edu

August 7, 2012

Abstract

We develop a new method to sample from posterior distributions in hierarchical models without using Markov chain Monte Carlo. This method, which is a variant of importance sampling ideas, is generally applicable to high-dimensional models involving large data sets. Samples are independent, so they can be collected in parallel, and we do not need to be concerned with issues like chain convergence and autocorrelation. Additionally, the method can be used to compute marginal likelihoods.

arXiv:1108.2245v3 [stat.CO] 9 Aug 2012

- Consistent font styles and sizes
- Clear spatial alignment
- Lack of background noises

Within Natural Scenes

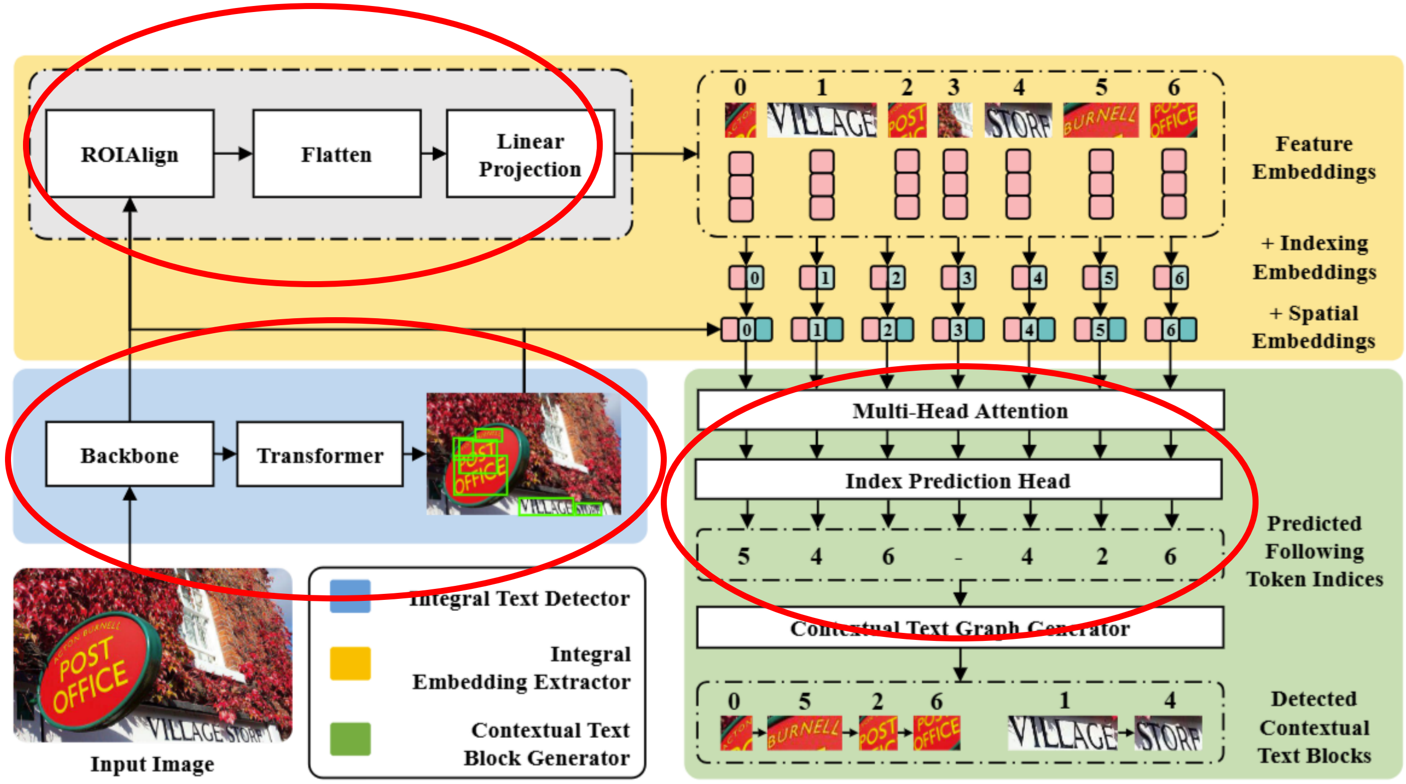
- Diversity in text font styles and sizes
- Unclear spatial alignment among text units
- Background noises that obscure text

Prior Arts

- Top-Down Methods
  - Adopt box-regression based object detection frameworks to identify text blocks
    - E.g., R-CNN, Fast R-CNN, Faster R-CNN, YOLOv5, Deformable DETR, …
  - Leverage instance segmentation frameworks to segment text blocks
    - E.g., Mask R-CNN, Mask2Former, SOLO, TransDLANet, Mask DINO, …
  - Facing challenges in accurately detecting contextual text blocks in complex natural scenes and obtaining the reading order among the text units

- Bottom-Up Methods
  - Detect the text units first, and then group them into text blocks arranged in their natural reading order
    - E.g., Post-OCR Paragraph Recognition, Unified Line and Paragraph Detection, Hybrid POD, HierText, CUTE, …

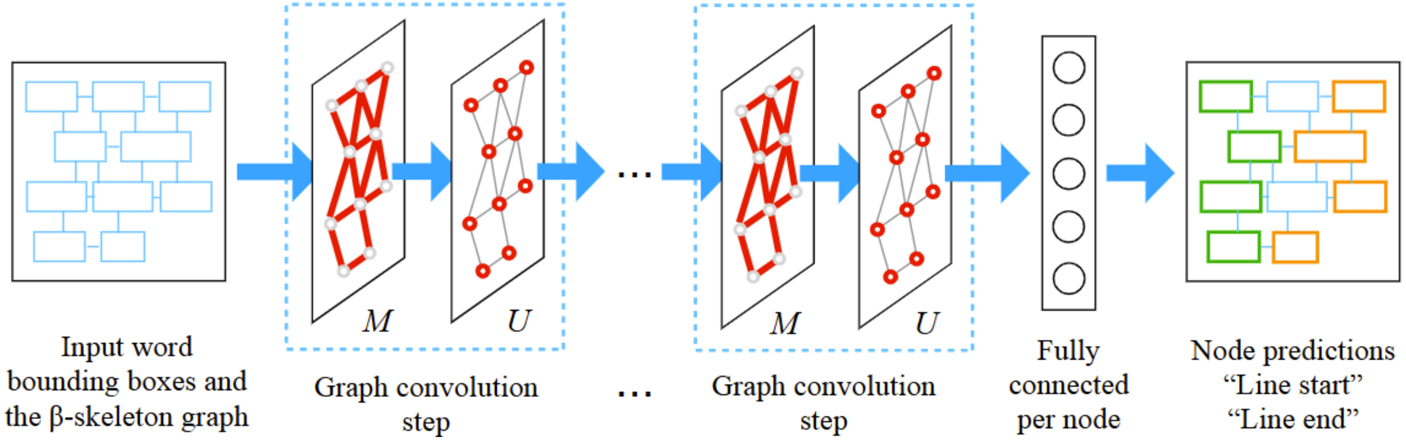
CUTE[1]: An NLP Perspective



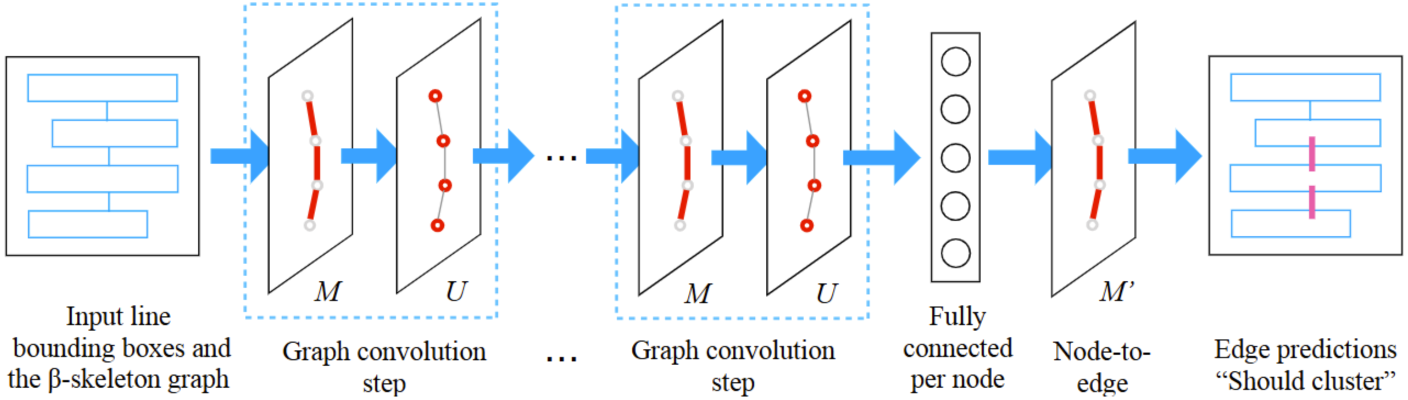
- First to define the task of contextual text block detection
- Establish two benchmark datasets
- Frame it as a sequence modeling problem
- Inefficient prediction in vast index space
- Challenges in modeling more complex relationships
- Limited in leveraging broader visual features for CTBD

[1] Xue, Chuhui, et al. "Contextual text block detection towards scene text understanding." *European Conference on Computer Vision*. 2022.

Post-OCR Paragraph Recognition[1]: Introduce Graph Structure into Paragraph Recognition



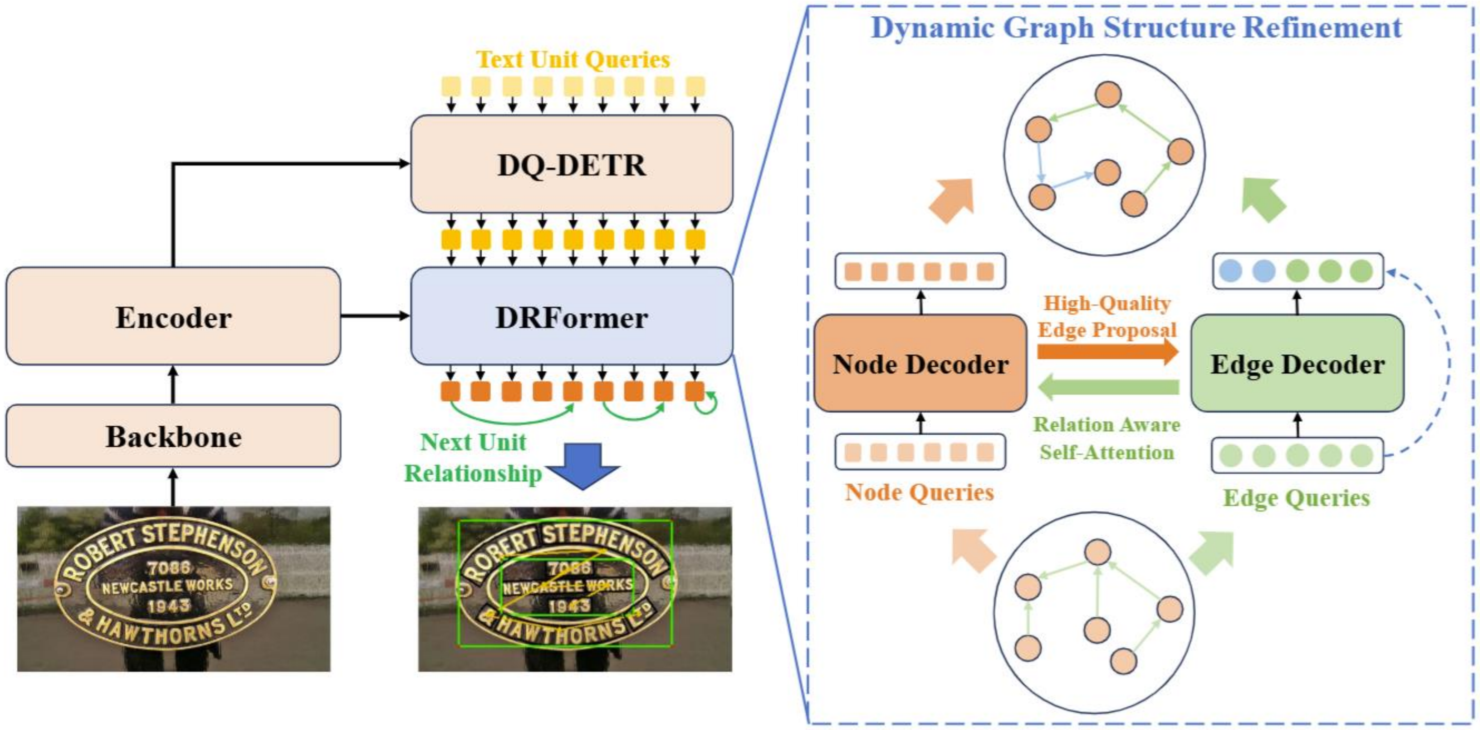
- Frame paragraph recognition as a relation prediction problem
- Leverage a GCN to model the relationships.



- Limit the capability to capture complex relationships due to the "static" graph
- Focus remains primarily on physical paragraphs in printed text scenarios.

[1] Wang, Renshen, Yasuhisa Fujii, and Ashok C. Popat. "Post-ocr paragraph recognition by graph convolutional networks." *Winter Conference on Applications of Computer Vision*. 2022.

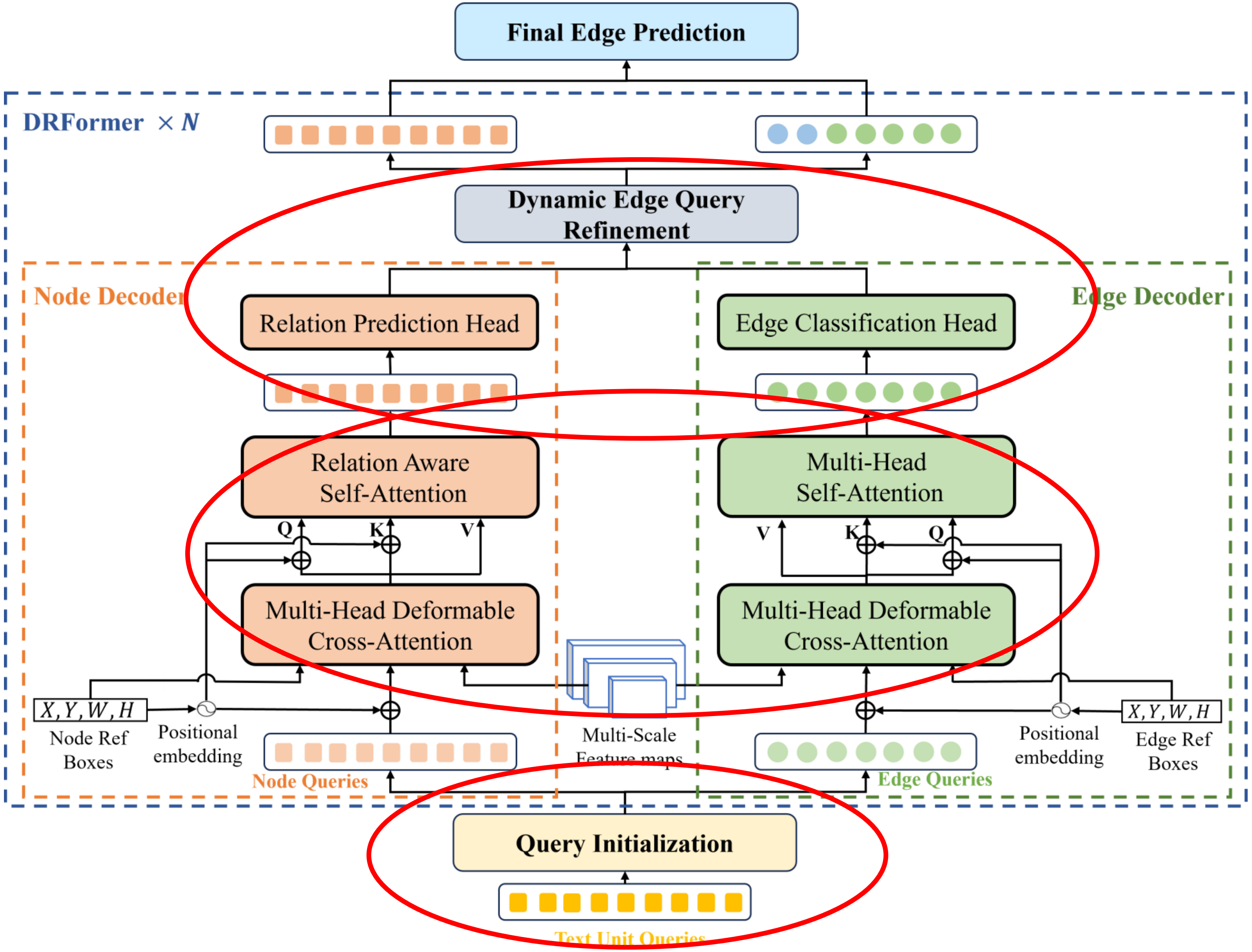
Core Idea of Our Approach:
Introduce Dynamic Graph Structure to CTBD

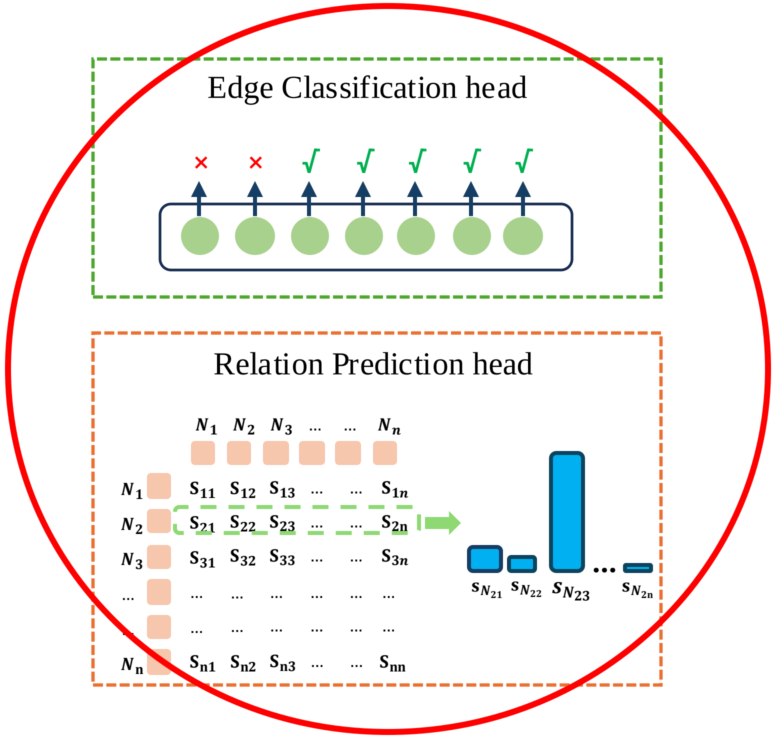


- Propose to frame contextual text block detection as *a graph generation problem*.
- Introduce *a dynamic graph structure refinement process* to progressively improve the quality of generated graphs.
- Introduce a dual-interactive transformer decoder, *Dynamic Relation Transformer (DRFormer)*, to support the iterative refinement process:
  - Node Decoder generates *high-quality edge proposals*
  - Edge Decoder facilitates *relation-aware self-attention* and *prunes incorrect edges*

[1] Ma, Chixiang, et al. "DQ-DETR: Dynamic Queries Enhanced Detection Transformer for Arbitrary Shape Text Detection." International Conference on Document Analysis and Recognition, 2023.

DRFormer: Dynamic Relation TransFormer





- ***Dynamic Graph Structure Refinement***
- ***Relation Aware Self-Attention***
- ***Multi-Head Deformable Cross-Attention***

Benchmark Datasets

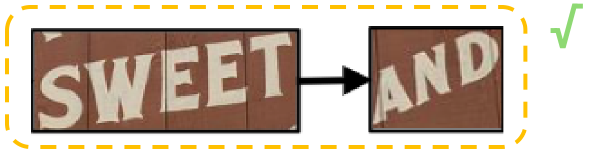
- ReCTS-Context
  - Includes a corpus of 15,000 training images and 5,000 test images.
  - The majority of text units are characters, presenting a unique challenge in predicting reading order relationships.
- SCUT-CTW-Context
  - Contains a corpus of 940 training images and 498 test images.
  - The majority of text units are words, offering rich contextual information across various scenes.

The statistics of the ReCTS-Context and SCUT-CTW-Context datasets: ‘integral’: Integral Text Units; ‘block’: Contextual Text Blocks; ‘#’: Number.

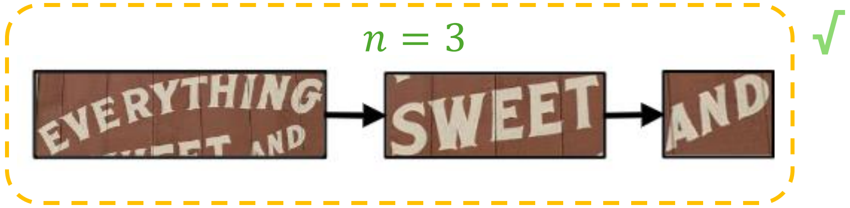
Dataset	Integral Text	# integral	# block	# image	# integral per block	# integral per image	# block per image
ReCTS-Context	Character	440,027	107,754	20,000	4.08	22.00	5.39
SCUT-CTW-Context	Word	25,208	4,512	1,438	5.56	17.65	3.17

Evaluation Metrics

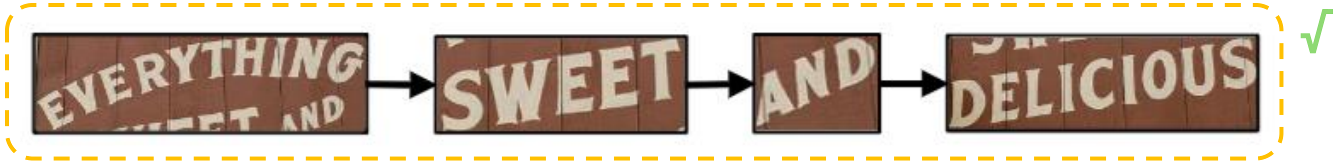
- Local Accuracy (LA)
  - Evaluate the accuracy of order prediction for neighboring text units.



- Local Continuity (LC)
  - Evaluate the continuity of text units by computing a modified n-gram precision score as inspired by BLEU, where n varies from 1 to 5.



- Global Accuracy (GA)
  - Evaluate the detection accuracy of complete contextual text blocks.



Comparisons with Prior Arts

- Performance comparison on SCUT-CTW-Context

Models	IoU=0.5			IoU=0.75			IoU=0.5:0.05:0.95		
	LA	LC	GA	LA	LC	GA	LA	LC	GA
LINK-R50 [40]	25.5	3.3	18.9	20.3	3.2	14.7	19.3	2.9	14.3
CUTE-R50 [39]	54.0	39.2	30.7	41.6	31.2	23.7	39.4	29.0	22.1
LINK-R101 [40]	25.7	3.4	19.2	20.0	2.9	14.7	19.6	2.7	14.4
CUTE-R101 [39]	55.7	39.4	32.6	40.6	29.0	22.8	40.0	28.3	22.7
Baseline-R50	67.6	55.7	45.8	56.5	43.6	37.3	47.4	37.1	31.9
DRFormer-R50	**69.6**	**59.0**	**47.8**	**58.1**	**46.0**	**39.3**	**48.9**	**39.3**	**33.3**

- Performance comparison on ReCTS-Context

Models	IoU=0.5			IoU=0.75			IoU=0.5:0.05:0.95		
	LA	LC	GA	LA	LC	GA	LA	LC	GA
LINK-R50 [40]	68.2	57.5	48.4	53.8	50.2	38.4	53.0	47.7	37.3
CUTE-R50 [39]	70.4	64.7	51.6	54.4	56.6	39.5	53.9	53.6	38.9
LINK-R101 [40]	70.8	59.1	49.9	54.5	51.0	39.0	53.4	48.3	37.9
CUTE-R101 [39]	72.4	67.3	53.8	55.1	**57.0**	40.2	54.6	**53.9**	39.4
Baseline-R50	82.2	71.4	69.6	63.2	50.0	52.8	56.4	46.0	47.6
DRFormer-R50	**83.3**	**74.6**	**71.8**	**67.6**	55.9	**56.9**	**59.4**	50.0	**50.6**

- Upper Bound Evaluation with GT Text Units

Models	SCUT-CTW-Context			ReCTS-Context		
	LA	LC	GA	LA	LC	GA
LINK-R50	30.2	4.5	22.8	83.8	68.4	61.1
CUTE-R50	71.5	58.5	49.7	92.1	82.8	76.0
LINK-R101	45.5	6.3	31.7	86.7	75.0	69.6
CUTE-R101	71.5	58.7	52.6	**93.1**	83.7	77.8
Baseline-R50	80.3	71.0	58.7	90.9	81.8	82.8
DRFormer-R50	**83.9**	**76.0**	**60.5**	92.8	**85.9**	**85.5**

Effectiveness of Various Components

- Key components:
  - *Dynamic Graph Structure Refinement (DGSR)*
  - *Cross-Attention First (CAF)*
  - *Relation-Aware Self-Attention (RASA)*

- Ablation studies on SCUT-CTW-Context dataset.

#	Method	DGSR	CAF	RASA	LA	LC	GA
1	Baseline				80.3	71.0	58.7
2		✓			82.3	72.6	58.8
3		✓	✓		83.4	75.3	60.2
4	DRFormer	✓	✓	✓	**83.9**	**76.0**	**60.5**

Comparison Examples

Baseline:
Node Decoder
only

DRFormer

Conclusion and Future Work

- Conclusion
  - Framing contextual text block detection as a graph generation problem is an effective problem formulation for CTBD.
  - DRFormer provides a promising avenue for integrating dynamic graph structures into the relation prediction process.

- Future work
  - Integrate text embeddings to enhance relation prediction accuracy.
  - Explore applying dynamic graph structure refinement to related tasks like Scene Graph Generation and Graph Structure Learning.

Thanks for your listening!

Personal Website

Linkedin

X (Twitter)

WeChat