CPA-GNN: Contextual-Based Pattern-Aware Graph Neural Network for Text Spotting

Anant Sinha[1], Shivakumara Palaiahnakote2, 3, Umapada Pal[1] and Mo Saraee2, 3

1.Veermata Jijabai Technological Institute, Mumbai, India

2.Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India

3.School of Science, Engineering and Environment, University of Salford, Manchester, UK

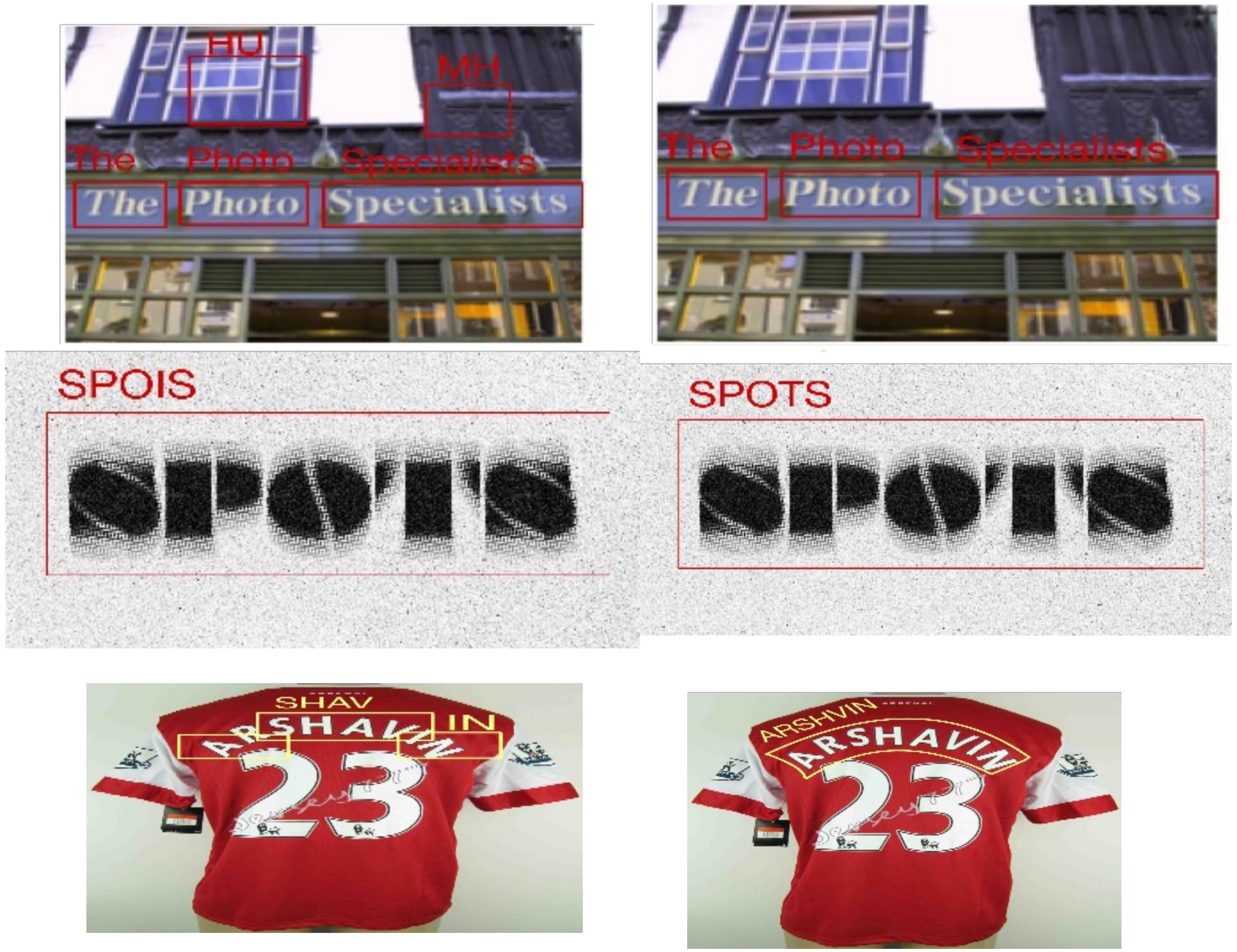
4.Data Science and Artificial Intelligence Hub, University of Salford, Manchester, UK

Submitted to International Conference on Pattern Recognition (ICPR), 2026

Slide Structure Plan

1. Introduction
2. Limitations of Existing Methods
3. Contributions
4. Proposed Method Architecture
   - Contextual Detection Refinement GNN
   - Pattern-Aware Recognition GNN
   - End-to-End Text Spotting Pipeline
5. Experimental Results
6. Conclusion and Future Work

Introduction: The Text Spotting Challenge



Text and non-text in scene images are strongly correlated. Existing models (left) fail on adverse text; CPA-GNN (right) fixes bounding boxes and recognizes text correctly across scene, printed and handwritten domains.

Limitations of Existing Methods

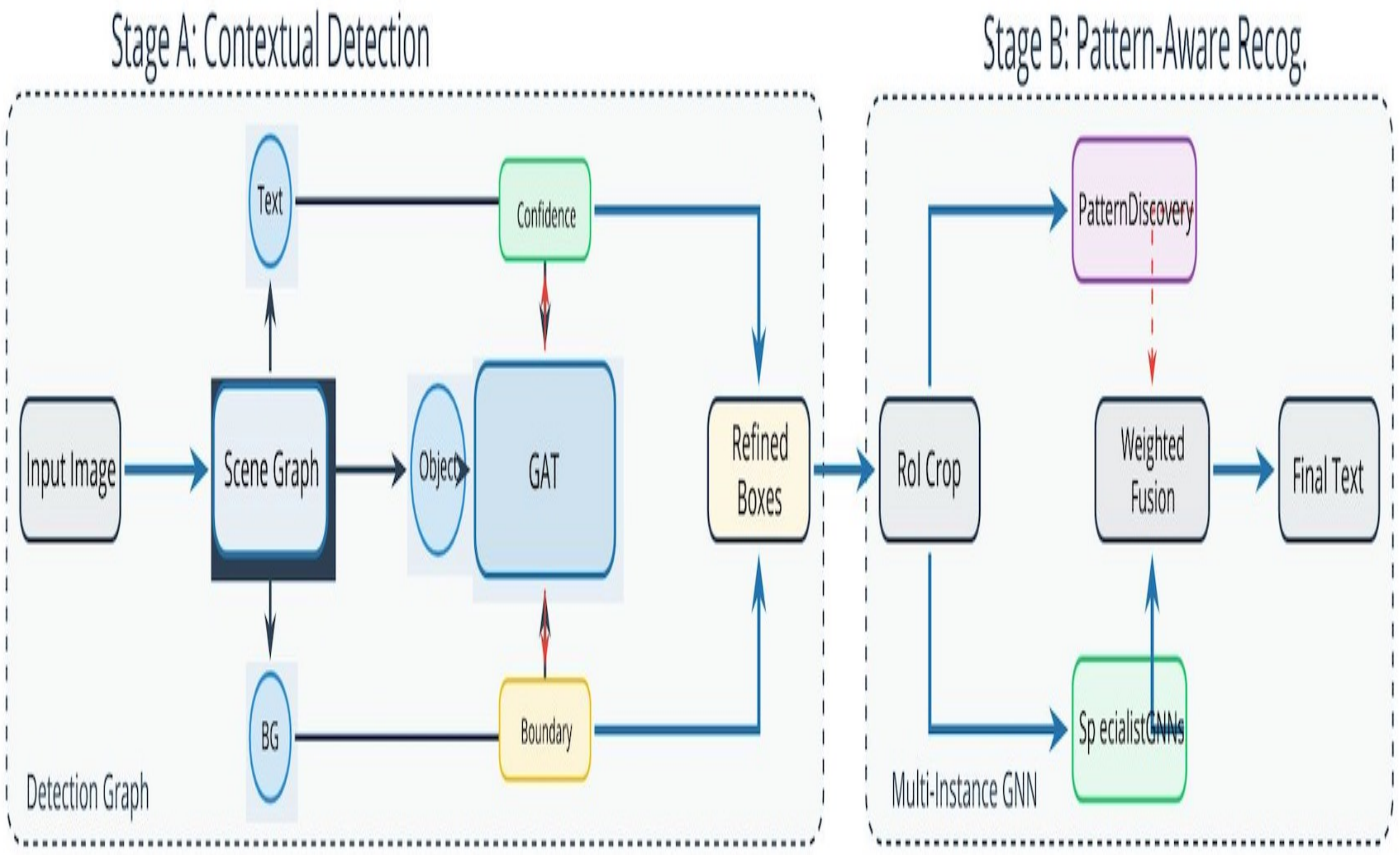
Why Existing Pipelines Fall Short

- **Lack of Context**: Detectors (EAST, CRAFT, DBNet) view text regions in isolation and ignore correlation with surrounding objects and background.
- **Style Variance**: Recognizers (CRNN, ASTER, TrOCR) use one shared encoder, failing on long-tail styles like handwriting or curved text.
- **False Positives**: Local pixel-level features alone cannot distinguish real text from text-like textures in cluttered scenes.
- **Hallucination in LMMs**: Vision-language spotters (Monkey) invent text and give imprecise, loose bounding boxes.
- **Domain Gap**: Jointly trained spotters let the recognition head overfit to the detector's output format, hurting cross-domain generalization.

Contributions

- Implementing a **graph neural network** for extracting **contextual information** between text and non-text in the images.
- Introducing multiple branches to a graph neural network for identifying the pattern of different text styles, called a **"team of specialists."**
- Achieves the **best accuracy** on challenging benchmarks — Total-Text, CDW and ICDAR 2015 — outperforming the state of the art.

Proposed Method Architecture



- Two novel modules: the Contextual Detection Refinement GNN and the Pattern-Aware Recognition GNN.
- Stage A refines text proposals using graph attention over text, object and background nodes.
- Stage B routes recognition features to specialist GNN branches based on the discovered text style.

Contextual Detection Refinement GNN

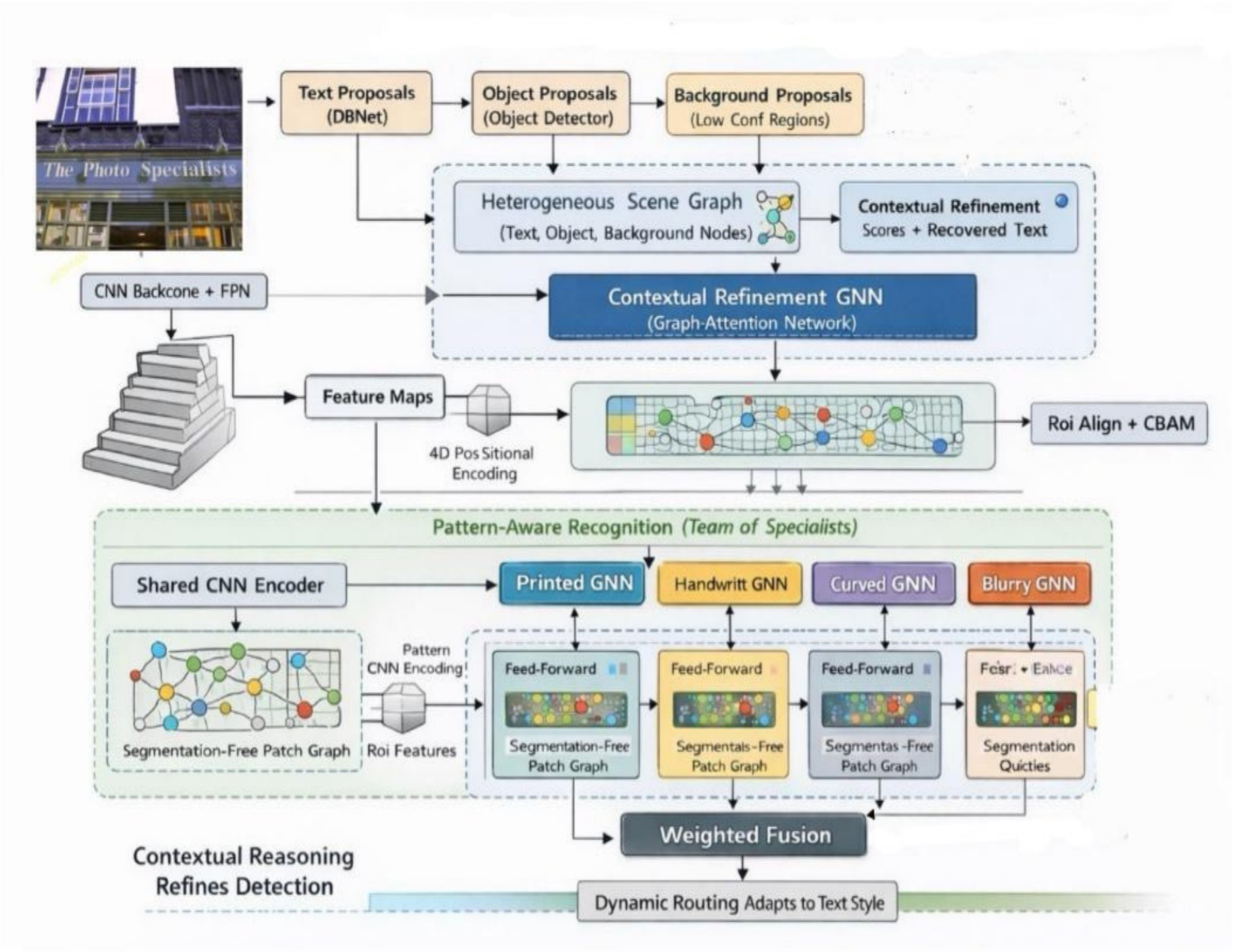
- Builds a heterogeneous scene graph linking Text, Object and Background nodes.
- L layers of Graph Attention Networks (GAT) propagate context between nodes.
- High attention validates weak text near supportive objects; low attention suppresses background noise.

Pattern-Aware Recognition GNN

University of Salford MANCHESTER

- A Pattern Discovery router assigns each text instance to $K = 4$ latent style specialists.
- A segmentation-free patch graph activates only on text structure, filtering out the background.
- Weighted Fusion combines all specialist outputs via a Log-Sum-Exp for the final prediction.

End-to-End Text Spotting Pipeline



- Proposal Generation: a shared ResNet-50 + FPN backbone extracts multi-scale features.
- Contextual Refinement plus RoIAlign and CBAM sharpen character strokes before recognition.
- Pattern-Aware Recognition then routes each proposal to its style-specific specialist branch.

Dataset Summary

Dataset	Scope
ICDAR 2015	Scene text
Total-Text	Curved text
CDW	Irregular/curved
FUNSD	Document forms

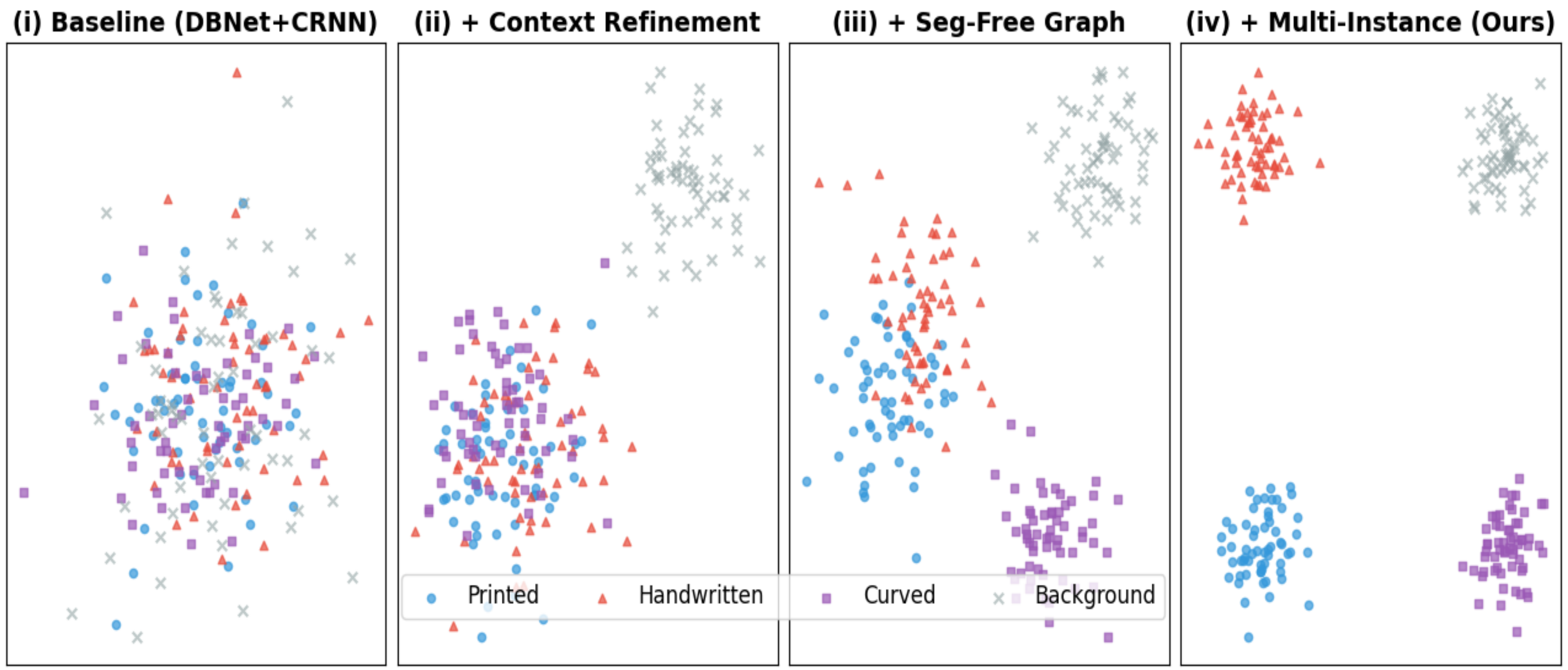
Text Detection Results (F-score)

Method	Detection Datasets (F-score)		
	IC15 (Standard)	Total-Text (Curved)	CDW (Irregular)
EAST (CVPR'17)	80.7	78.4	71.2
CRAFT (CVPR'19)	86.9	82.4	78.5
DBNet (AAAI'20)	87.3	84.7	81.0
DBNet++ (TPAMI'22)	89.1	86.0	83.4
DeepSolo (CVPR'23)	89.6	86.8	85.1
TextDiffuser-2 (ECCV'24)	90.0	87.1	**86.5**
Ours (CPA-GNN)	**90.3**	**87.8**	88.2

Text Spotting Results (H-Mean)

Method	Spotting Datasets (H-Mean)		
	IC15 (Standard)	Total-Text (Curved)	CDW (Irregular)
Mask TextSpotter v3 (ECCV'20)	83.3	71.2	70.5
ABCNet v2 (TPAMI'21)	84.2	78.1	76.9
TESTR (CVPR'22)	86.6	83.3	82.1
DeepSolo (CVPR'23)	87.2	82.4	-
GLASS (ECCV'22)	89.5	83.3	**-**
Ours (CPA-GNN)	**88.2**	**88.0**	88.0

Ablation Study

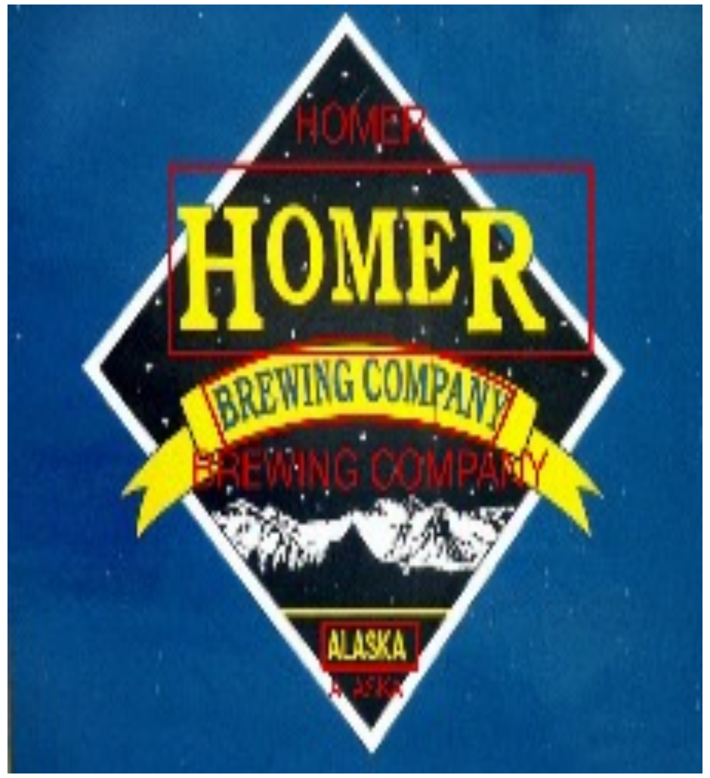


- Baseline (DBNet+CRNN): entangled features with heavy overlap between text styles and background noise.
- +Context Refinement and +Seg-Free Graph progressively separate the curved-text and background clusters.
- +Multi-Instance (full model) yields compact, well-separated style clusters and the best word accuracy across all three datasets.

Cross-Domain Generalisation (Sim-to-Real)

Method	Trained on SynthText, tested on:		
	IC15 (Standard)	Total-Text (Curved)	CDW (Irregular)
CRNN	52.4	43.5	39.1
ASTER	56.8	49.8	44.2
PARSeq	68.4	62.1	59.5
Ours (CPA-GNN)	78.5	74.2	72.0

Qualitative Results



- CPA-GNN accurately detects and recognizes text across Total-Text, CDW and ICDAR 2015.
- Handles printed, stylized and handwritten text within a single unified framework.
- Contextual reasoning recovers text instances that baseline detectors miss entirely.

Limitations

- Dense occlusion severs edges in the patch graph, fragmenting the recognition output.
- Out-of-distribution styles (e.g. graffiti) yield a high-entropy routing vector that confuses specialist fusion.
- Future work: introduce a language model to improve robustness to these failure cases.

Conclusion and Future Works

- To address the computational cost of the multi-branch design, future work will distil CPA-GNN into a smaller, single-branch "Student" network via a Teacher-Student framework.
- Proposed CPA-GNN: a unified graph-based framework combining a Contextual Detection Refinement GNN with a Pattern-Aware "Team of Specialists" Recognition GNN for text spotting.
- Experiments on ICDAR 2015, Total-Text and CDW yield state-of-the-art word accuracy; we plan to integrate a language model to further improve robustness to occlusion and out-of-distribution styles.

References

1. R. Ronen, S. Tsiper, O. Anschel, I. Lavi, A. Markovitz, R. Litman: GLASS: Global to Local Attention for Scene-Text Spotting. In: Proc. ECCV (2022).
2. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio: Graph Attention Networks. In: Proc. ICLR (2018).
3. Y. Baek, B. Lee, D. Han, S. Yun, H. Lee: Character Region Awareness for Text Detection. In: Proc. CVPR (2019).
4. M. Liao, Z. Wan, C. Yao, X. Chen, X. Bai: Real-time Scene Text Detection with Differentiable Binarization. In: Proc. AAAI (2020).
5. J. Cui et al.: TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering. In: Proc. ECCV (2024).
6. M. Ye et al.: DeepSolo: Point Sequence Prediction for Text Detection. In: Proc. CVPR (2023).
7. D. Bautista, R. Atienza: Scene Text Recognition with Permuted Autoregressive Sequence Models. In: Proc. ECCV (2022).
8. D. Karatzas et al.: ICDAR 2015 Competition on Robust Reading. In: Proc. ICDAR (2015).

Thank You

University of Salford MANCHESTER