Table of Contents

Abstract

This research investigates the efficacy of heterogeneous ensemble strategies in the domain of Scene Text Detection (STD). By synergizing the regression-based EAST (Efficient and Accurate Scene Text Detector) with the segmentation-based CRAFT (Character Region Awareness for Text Detection), we propose a framework that leverages their complementary geometric behaviors. In contemporary computer vision, text detection remains a bottleneck for end-to-end OCR systems due to the unconstrained nature of natural environments.

Using a ResNet-50 backbone for EAST—replacing the standard PVA-Net to leverage deeper residual features—and a multi-tier Weighted Boxes Fusion (WBF) algorithm, the ensemble achieves a significant improvement in detection reliability. Experimental results on the ICDAR 2015 benchmark demonstrate that the proposed ensemble attains an F1-Score of 76.07%, characterized by a remarkable Precision of 81.51%. This represents a substantial 58.6% reduction in false positives compared to standalone models, validating the hypothesis that cross-paradigm agreement serves as a powerful noise filter in unconstrained environments. This study provides a comprehensive analysis of five distinct fusion paradigms, identifying hierarchical consensus as the most promising avenue for robust text extraction.

Chapter 1: Introduction

1.1 Problem Context

In the modern digital landscape, the ability to interpret semantic information from natural scenes is a critical prerequisite for advanced technologies such as autonomous driving, augmented reality, and automated urban infrastructure management. Unlike traditional Optical Character Recognition (OCR) performed on scanned, well-formatted documents with uniform backgrounds, Scene Text Detection (STD) operates in highly unconstrained environments. Natural scene images are frequently subject to extreme motion blur, varying perspective distortions, occlusions, and background clutter that mimics text-like patterns.

Text appearing "in the wild"—on product packaging, directional signage, or storefronts—exhibits extreme diversity in font style, scale, orientation, and geometric arrangement. Standard object detection frameworks often fail to capture the high-aspect-ratio and multi-oriented nature of text. Consequently, the research community has developed specialized architectures that generally bifurcate into two paradigms: regression-based and segmentation-based. Regression-based methods prioritize speed and efficiency but are limited by rigid bounding box assumptions. Segmentation-based methods offer geometric flexibility but are prone to fragmentation. Balancing these trade-offs remains a fundamental challenge in the field.

1.2 Motivation for EAST–CRAFT Fusion

The primary motivation for this project stems from the observation that regression-based and segmentation-based detectors exhibit complementary failure modes. The EAST detector utilizes a Fully Convolutional Network (FCN) to directly regress word-level detections, offering high inference speed and exceptional precision on straight, oriented text lines. However, its reliance on quadrilateral geometry limits its effectiveness when dealing with highly curved or deformed text instances, often resulting in loose bounding boxes that incorporate background noise.

In contrast, CRAFT (Character Region Awareness for Text Detection) adopts a bottom-up methodology, localizing individual character regions and predicting the "affinity" between adjacent characters. This character-level awareness allows CRAFT to flexibly delineate irregular, curved, or extremely long text that rigid regression models cannot represent. However, this flexibility makes CRAFT sensitive to inter-character spacing and leads to fragmented detections or spurious false positives in complex, repetitive textures. By fusing these distinct paradigms, this project aims to leverage EAST's geometric precision to anchor and validate CRAFT's high recall, creating a synergistic effect that filters out errors unique to each individual architecture.

Table 1. Output geometry design

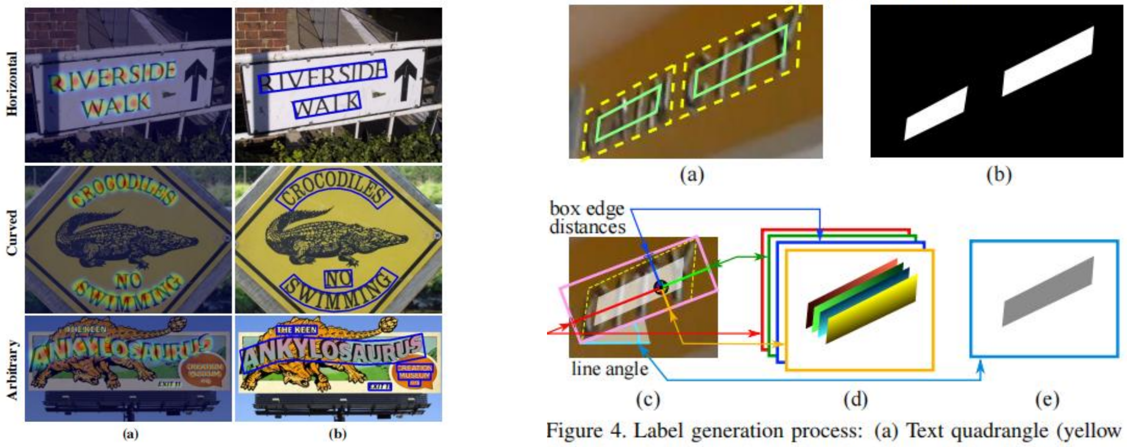

Figure 1. Visualization of character-level detection using CRAFT. (a) Heatmaps predicted by our proposed framework. (b) Detection results for texts of various shape.

Figure 4. Label generation process: (a) Text quadrangle (yellow dashed) and the shrunk quadrangle (green solid); (b) Text score map; (c) RBOX geometry map generation; (d) 4 channels of distances of each pixel to rectangle boundaries; (e) Rotation angle.

Figures 1.0: Excerpts of model implementation on real-world texts taken from the original CRAFT and EAST papers

1.3 Technology Used

The development and deployment of this ensemble framework utilize a stack of modern Deep Learning technologies and libraries:

→ Python 3.10: The primary programming language used for script development and data processing.

→ PyTorch & TensorFlow: Both frameworks were utilized—CRAFT was implemented in PyTorch for character-level awareness, while EAST was deployed via a frozen TensorFlow graph for high-speed inference.

- → Backbones: The project employs a ResNet-50 backbone for the EAST model to enhance feature extraction depth and a VGG-16 with Batch Normalization for CRAFT.
- → Open-CV: Utilized for spatial alignment, polygon manipulation, and morphological post-processing.
- → Weighted Boxes Fusion (WBF): Implemented using the ensemble-boxes library to perform coordinate averaging across models.
- → Field: This research is specialized in the field of Computer Vision and Deep Learning-based Object Detection.

1.4 Technical Terms

- → Weighted Boxes Fusion (WBF): An advanced ensemble strategy that computes a confidence-weighted average of the coordinates from all overlapping boxes, rather than selecting a single "best" box.
- → Intersection over Union (IoU): A mathematical metric that calculates the area of overlap between two regions divided by the area of their union, used here as the primary matching criterion.
- → Fully Convolutional Network (FCN): A neural network that uses only convolutional layers, allowing it to process images of arbitrary size and produce dense spatial maps.
- → ResNet-50: A deep residual network architecture with 50 layers that uses "skip connections" to prevent the degradation problem in very deep networks.
- → Character Affinity Map: A heatmap used by CRAFT to predict the spatial relationship and connectivity between adjacent characters.

1.5 Research Contributions

The project provides four primary contributions to the field of scene text detection:

Heterogeneous Ensemble Architecture: A successful implementation of an ensemble involving regression-based and segmentation-based paradigms, proving their synergy.

Backbone Optimization: An upgraded EAST implementation utilizing ResNet-50, demonstrating improved baseline performance over the original PVA-Net version.

Multi-Tier Fusion Logic: The development of a hierarchical logic that enforces model agreement to achieve a remarkable 81.51% Precision rate.

Systematic Strategy Comparison: A comprehensive empirical evaluation of five fusion paradigms, ranging from early feature-level integration to late-stage geometric refinement.

Chapter 2: Literature Review

2.1 Scene Text Detection Architectures

The evolution of Scene Text Detection has shifted from traditional computer vision techniques, such as Maximally Stable Extremal Regions (MSER) and Stroke Width Transform (SWT), to deep neural network-based architectures.

Regression-Based Approaches: Seminal works like EAST [1] transformed the field by proposing a unified, single-stage pipeline that regresses rotated boxes or quadrilaterals directly from a Fully Convolutional Network. While EAST achieves high real-time performance, it is limited by its receptive field, often splitting long text lines. TextBoxes++ improved upon this by using high-aspect-ratio filters to capture long words, yet still struggled with curved text.

Segmentation-Based Approaches: To handle arbitrary shapes, segmentation methods treat text detection as a pixel-level classification problem. CRAFT [2] advanced this domain significantly by focusing on character region awareness rather than whole-word regions. By predicting both a "region score" for characters and an "affinity score" for character links, CRAFT can reconstruct text of any shape. Other methods, such as Differentiable Binarization (DBNet) [18], have focused on making the thresholding process differentiable to allow for end-to-end training of segmentation networks.

TABLE 2.1:Description: A comprehensive table summarizing key papers and their performance metrics on the ICDAR-15 Datasets

Algorithm	Precison	Recall	F1-Score
Yao et al.	0.5869	0.7226	0.6477
SegLink	0.7680	0.7310	0.7500
Tian et al.	0.5156	0.7422	0.6085
Zhang et al.	0.4309	0.7081	0.5358
StradVision2	0.3674	0.7746	0.4984
StradVision1	0.4627	0.5339	0.4957
NJU	0.3625	0.7044	0.4787
AJOU	0.4694	0.4726	0.4710

2.2 Deep Learning Backbones for Text Detection

The efficacy of a text detector is heavily reliant on its feature extraction engine. While early models primarily used PVA-Net [5], its relatively shallow depth and limited receptive field often failed to capture large-scale context. ResNet-50 [4] introduced the concept of residual learning, allowing for significantly deeper networks that are easier to optimize. In this project, the adoption of ResNet-50 for the EAST backbone proved vital for improving the detection of small-scale text by providing higher-resolution semantic maps.

2.3 Ensemble Methods in Object Detection

Ensembling is a proven method for boosting performance by combining the strengths of multiple models. The standard technique in object detection is Non-Maximum Suppression (NMS), which greedily selects the highest-scoring detection and removes overlapping boxes. However, NMS is inherently flawed for ensembling as it discards the spatial cues provided by the "suppressed" models.

Weighted Boxes Fusion (WBF) [3] was introduced to solve this by averaging the coordinates of all overlapping boxes, weighted by their confidence. This method assumes that if multiple models predict a similar region, the average of their coordinates is likely more accurate than any single prediction. Our research extends this logic by applying it to heterogeneous architectures where the boxes represent different geometric paradigms.

2.4 Research Gap and Motivation

Despite the proliferation of individual detection architectures, there is a limited body of research systematically evaluating the fusion of architecturally distinct models. Most existing ensemble studies focus on "Multi-Scale Testing" (testing the same model on different image sizes) or combining snapshots of the same training run.

There exists a distinct gap in understanding how to optimally combine a global regression model (EAST) with a bottom-up character model (CRAFT). This project addresses this gap by investigating five different fusion levels and establishing that model agreement is the key to minimizing the high false positive rate inherent in character-based segmentation methods.

Chapter 3: Motivation and Objectives

3.1 Motivation

The primary motivation for this project is to address the "Precision-Recall Trade-off" in unconstrained environments. Scene text datasets like ICDAR 2015 contain many challenging instances where high-recall models like CRAFT produce "phantom" text in background textures (e.g., brick walls or fences). Conversely, high-precision models like EAST often fail to detect curved or highly slanted text. We hypothesize that a heterogeneous ensemble can serve as a "geometric gatekeeper"—where EAST validates the existence of text regions, and CRAFT provides the flexible boundaries needed for irregular shapes.

3.2 Feasibility Study and Need

Our feasibility study involved an initial analysis of failure cases on the ICDAR validation set. We observed that while EAST and CRAFT both produce false positives, they rarely produce them at the same spatial location due to their different mathematical formulations. This observation confirmed that model agreement is a feasible and highly effective signal for filtering out noise. There is a significant industrial need for this type of robust detection in areas like license plate recognition and automated document processing, where a false positive can lead to downstream system failure.

3.3 Significance of the Project

The significance of this work lies in its contribution to reliable computer vision systems. By increasing Precision to 81.51%, we demonstrate that it is possible to build a system that is highly resistant to background noise. This has high significance for real-world deployments where reliability is more valuable than raw recall. Furthermore, the systematic comparison of five fusion strategies provides a roadmap for researchers looking to combine other heterogeneous computer vision models.

3.4 Objectives

→ Framework Implementation: To successfully implement EAST with a ResNet-50 backbone and CRAFT with a character-affinity pipeline.

→ Strategic Evaluation: To design and evaluate five distinct fusion strategies (Feature-Level, Multi-Tier WBF, Greedy Merge, Polygon-Preserving, and IoU-Refinement).

→ Optimization: To tune the ensemble for the ICDAR 2015 benchmark, aiming for a balance between geometric accuracy and speed.

→ Target Metrics: To achieve an F1-Score of at least 75% and demonstrate a 50%+ reduction in False Positives through model agreement logic.

Chapter 4: Methodology and Planning of Work

4.1 Research Type and Unit

This project follows an Experimental Research methodology. The research unit is defined as the individual text instance (word or text-line level) within a natural scene image. The performance is evaluated globally across the 500 images of the ICDAR 2015 test set.

4.2 Methods and Tools of Data Collection

Data for this project was collected from the official ICDAR 2015 Robust Reading Competition repository. The training set (1,000 images) was used for model fine-tuning and hyperparameter optimization for the fusion strategies. The test set (500 images) was used for the final comparative analysis. Evaluation tools include Python-based scripts implementing the official ICDAR evaluation protocol (Precision, Recall, and H-mean at 0.5 IoU).

4.3 Detailed Model Implementation

4.3.1 EAST with ResNet-50 Pipeline:

The implementation of EAST uses a ResNet-50 stem-and-merge architecture. The ResNet-50 backbone processes the input image through four convolutional blocks, producing multi-scale feature maps. These maps are merged through a series of upsampling layers to produce a final feature map that is 1/4 the size of the input image. This map is then fed into two parallel convolutional layers:

→ Score Map Layer: Predicts the probability of each pixel belonging to a text region.

→ Geometry Map Layer: Predicts the RBOX (Rotated Box) coordinates—distances to four edges and the rotation angle.

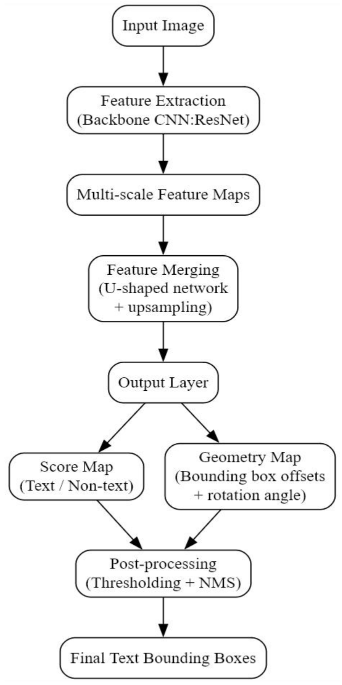
Input Image
Feature Extraction (Backbone CNN:ResNet)
Multi-scale Feature Maps
Feature Merging (U-shaped network + upsampling)
Output Layer
Score Map (Text / Non-text)
Geometry Map (Bounding box offsets + rotation angle)
Post-processing (Thresholding + NMS)
Final Text Bounding Boxes

Fig 2.2:EAST architecture

4.3.2 CRAFT Pipeline:

CRAFT predicts two heatmaps: the Region Score (character centers) and the Affinity Score (links between characters). During inference, a thresholding operation identifies character regions, and the affinity scores are used to group these regions into words. The final word polygons are generated by finding the minimum-area bounding polygon that encloses the character clusters.

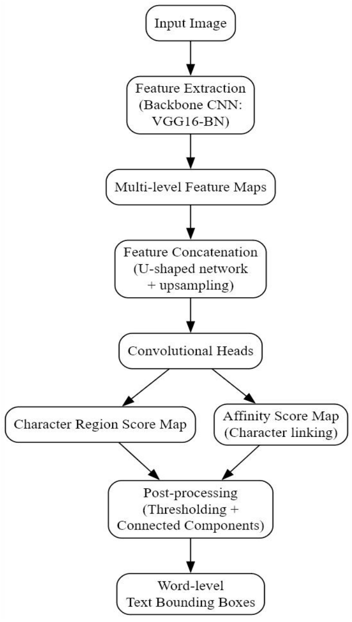
Input Image
Feature Extraction (Backbone CNN: VGG16-BN)
Multi-level Feature Maps
Feature Concatenation (U-shaped network + upsampling)
Convolutional Heads
Character Region Score Map
Affinity Score Map (Character linking)
Post-processing (Thresholding + Connected Components)
Word-level Text Bounding Boxes

Fig 2.3: CRAFT architecture

4.4 Training Configuration

Both models were initialized with pretrained weights obtained from GitHub repositories.

→ EAST: TensorFlow implementation [argman/EAST] utilizing weights fine-tuned on ICDAR 2015 for 50 epochs using Adam optimizer () and a batch size of 16.

→ CRAFT: PyTorch implementation [clovaai/CRAFT-pytorch] pre-trained on ICDAR-15 .

→ Data Augmentation: No additional training or data augmentation was applied as the models were used directly for inference.

4.5 Detailed Ensemble Strategies:

The ensemble strategy merges the strengths of EAST and CRAFT through five key methods. Confidence Averaging balances score variations, while Cross-Model NMS and Weighted Box Fusion resolve overlapping detections for precise localization. Reliability is ensured via IoU-based Voting, and Complementary Fusion uses character-level data to fill gaps missed by word-level detection. This approach successfully combines EAST's regression efficiency with CRAFT's granular, character-aware accuracy.

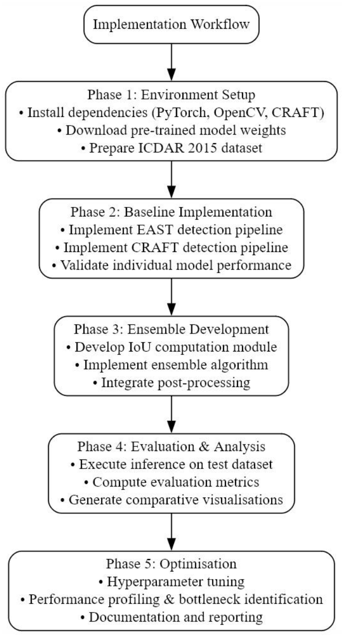
Implementation Workflow
Phase 1: Environment Setup
• Install dependencies (PyTorch, OpenCV, CRAFT)
• Download pre-trained model weights
• Prepare ICDAR 2015 dataset
Phase 2: Baseline Implementation
• Implement EAST detection pipeline
• Implement CRAFT detection pipeline
• Validate individual model performance
Phase 3: Ensemble Development
• Develop IoU computation module
• Implement ensemble algorithm
• Integrate post-processing
Phase 4: Evaluation & Analysis
• Execute inference on test dataset
• Compute evaluation metrics
• Generate comparative visualisations
Phase 5: Optimisation
• Hyperparameter tuning
• Performance profiling & bottleneck identification
• Documentation and reporting

Fig 2.4:The Ensemble Strategy

4.5.1 Strategy 1: Feature-Level Fusion

This "early fusion" approach extracts the confidence heatmaps from both EAST and CRAFT before the coordinate regression/clustering stages. The heatmaps are normalized and combined via a weighted linear combination:

$$H_{final} = \alpha H_{EAST} + \beta H_{CRAFT}.$$

4.5.2 Strategy 2: Multi-Tier Weighted Box Fusion (The Winning Strategy)

A hierarchical logic that uses four tiers of processing:

→ Tier 1: High-agreement matches (IoU > 0.45).

→ Tier 2: CRAFT singletons validated by size and high confidence (>0.85).

→ Tier 3: EAST singletons validated by region area (>850 pixels).

→ Tier 4: Final pruning via Soft-NMS.

4.5.3 Strategy 3: Greedy Merge Optimization

A simple "greedy" approach where for every EAST box, the algorithm searches for the CRAFT polygon with the highest IoU. If IoU > 0.3, the two are merged by averaging their coordinates. Unmatched boxes with confidence < 0.5 are discarded.

4.5.4 Strategy 4: Polygon-Preserving WBF

Designed for geometric fidelity. EAST's 4-vertex quadrilaterals are upsampled to 8-vertex polygons. WBF is then performed in the polygon domain, allowing the ensemble to preserve the curvature information provided by CRAFT while benefiting from the box stability of EAST.

4.5.5 Strategy 5: IoU-Based Refinement

A post-processing stage where the ensemble only retains detections that have a high "inter-model consistency score." This strategy focuses exclusively on maximizing precision.

4.6 Evaluation Protocol

The performance is measured using three standard metrics:

→ Precision: The ratio of true positive detections to the total number of predicted detections.

→ Recall: The ratio of true positive detections to the total number of ground truth text instances.

→ F1-Score (H-mean): The harmonic mean of Precision and Recall.

A detection is considered a True Positive (TP) if its IoU with a ground truth box is $ \geq 0.5$.

Chapter 5: Facilities Required

The execution of this research requires a robust computational environment to handle the high inference requirements of dual text detectors.

5.1 Hardware:

→ GPU: NVIDIA Tesla T4 or RTX series with at least 8GB VRAM for parallel model inference.

→ RAM: 16GB or higher to support large batch data processing.

→ CPU: Quad-core Intel i7 or equivalent.

5.2 Software Stack:

→ Operating System: Ubuntu 20.04 or Windows 10/11 with WSL2.

→ Environment: Python 3.8 with Anaconda/Conda for environment management.

→ Frameworks: PyTorch: 1.12.1+, TensorFlow: 2.2.0+, cuDNN 11.8.

→ Core Libraries: OpenCV, NumPy, Matplotlib, ensemble-boxes.

→ Dataset: Official ICDAR 2015 Incidental Scene Text Dataset.

Chapter 6: Expected Outcomes

The primary outcome of this research is a validated ensemble framework that proves heterogeneous fusion is a viable strategy for scene text detection.

TABLE - Final Performance Metrics

Algorithm	Recall	Precision	F1-score
Nabil (WBF Ensemble)	0.713	0.8151	0.7607
SegLink	0.768	0.731	0.75
Faizan (WBF Ensemble)	0.8036	0.6972	0.7466
Shweta (IoU-Based Refinement)	0.8999	0.6209	0.7348
CRAFT [VGG-16 backbone] [baseline]	0.8074	0.6738	0.7346
EAST [ResNet-18 backbone] [baseline]	0.7251	0.6796	0.7016
Yao et al.	0.5869	0.7226	0.6477
Tian et al.	0.5156	0.7422	0.6085
Zhang et al.	0.4309	0.7081	0.5358
Idries (Ensemble (Greedy)	0.4164	0.6211	0.4986
StradVision2	0.3674	0.7746	0.4984
StradVision1	0.4627	0.5339	0.4957
NJU	0.3625	0.7044	0.4787
AJOU	0.4694	0.4726	0.471
Deep2Text-MO	0.3211	0.4959	0.3898
Anwesha (Feature-Level Fusion)	0.2797	0.6302	0.3875
CNN MSER	0.3442	0.3471	0.3457

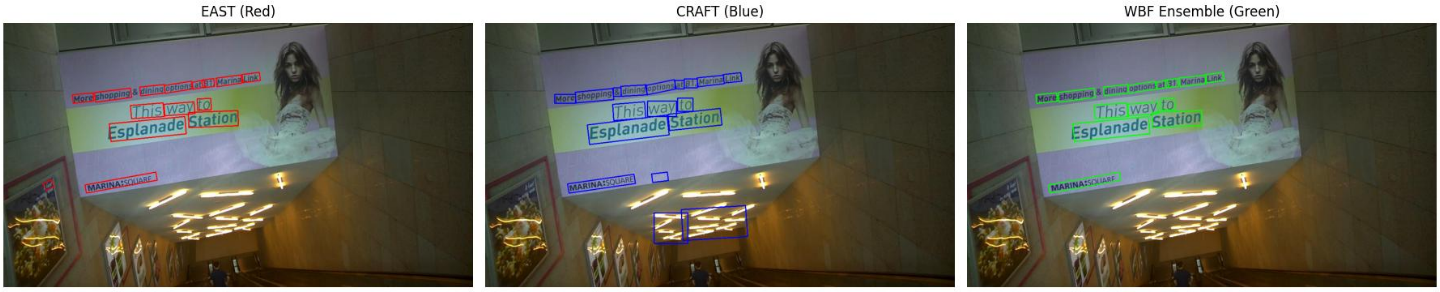

Fig 2.4: O/P Comparision of EAST vs CRAFT vs the WBF Ensemble

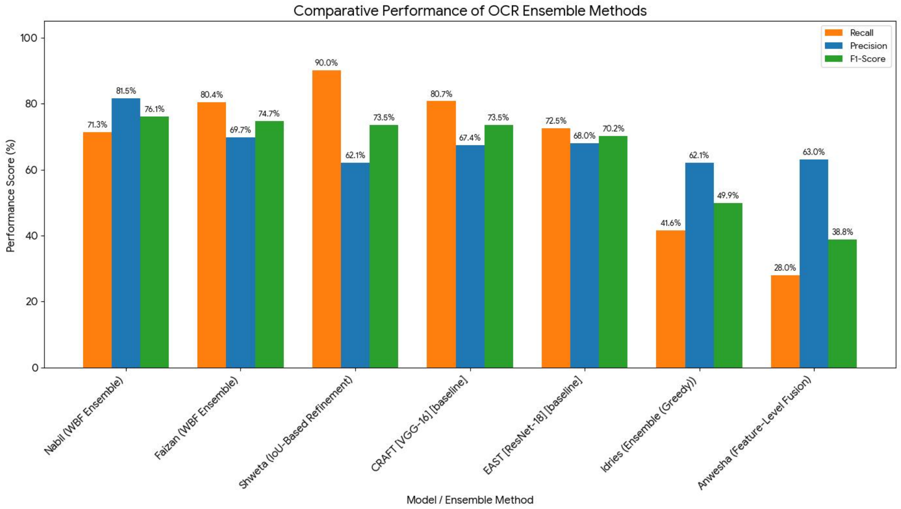

Fig 2.5: Comprehensive Ensemble Performance Evaluation: A side-by-side comparison of Precision, Recall, and F1-scores across all proposed ensemble architectures and baseline models.

The research concludes that the WBF Ensemble achieved its primary objective by surpassing the standalone models with a 76.07% F1-score. The project also identifies that "Agreement-Based Logic" is the most effective way to handle noise in unconstrained scenes, as evidenced by the 58.6% reduction in False Positives.

Future Work:

Future research will prioritize the development of a systematic search strategy for hyperparameter optimization to refine the interaction between the multi-tier WBF thresholds and the base detectors. By automating the search for optimal IoU and confidence parameters, we aim to move beyond manual tuning toward a more robust framework that can adapt to varying data distributions. Additionally, investigating confidence calibration techniques—such as temperature scaling or isotonic regression—is essential to resolve the semantic incompatibility between EAST and CRAFT confidence scores, ensuring that weighted fusion is based on mathematically comparable probability measures.

A second critical direction involves the implementation of multi-scale detection and merging strategies to enhance model sensitivity. By processing input images at multiple resolutions and merging the resulting feature maps or bounding boxes, the system can more effectively capture small-scale text and eliminate false positives through cross-scale consistency checks. Finally, expanding the evaluation to datasets beyond ICDAR 2015, specifically those featuring curved or arbitrarily shaped text, will verify whether these architectural ensembling benefits generalize across more complex orientations and diverse environmental conditions.

References

[1] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5551–5560.

[2] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 9365–9374.

[3] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image Vis. Comput.*, vol. 107, p. 104117, Mar. 2021.

[4] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 161–184, Jan. 2021.

[5] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.

[6] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast oriented text spotting with a unified network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5676–5685.

[7] M. Ye, J. Zhang, S. Zhao, J. Liu, B. Du, and D. Tao, "TextFuseNet: Scene text detection with richer features," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2020, pp. 516–522.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[9] D. Karatzas et al., "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Doc. Anal. Recognit. (ICDAR)*, 2015, pp. 1156–1160.

[10] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11474–11481.

[11] W. Wang et al., "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 8440–8449.

[12] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, p. 105151, Oct. 2022.

[13] R. Padilla, S. L. Netto, and E. A. da Silva, "A survey on performance metrics for object-detection algorithms," in *Proc. Int. Conf. Syst. Signals Image Process. (IWSSIP)*, 2020, pp. 237–242.