# Multi-Oriented Open-Set Text Recognition via Direction-Aware Channel-Adaptive Fusion

1st Mengdie Han
*School of Cyber Security and Computer*
*HeBei University*
Baoding, China
0009-0001-9544-6691

2nd Fang Yang*
*School of Cyber Security and Computer*
*HeBei University*
Baoding, China
0000-0001-5948-3965

3rd XiuFen Miao
*School of Cyber Security and Computer*
*HeBei University*
Baoding, China
0009-0000-4211-7723

***Abstract*—The open-set text recognition task aims to recognize unknown text categories in complex scenes, but large variations in character sets, text lengths, and orientations increase difficulty. Existing methods, such as MOoSE, use identical-structure experts for different orientations, but still lack sufficient directional modeling and effective multi-scale feature fusion. To address this, we propose a Direction-Aware and Channel-Adaptive (DACA) framework, integrating a Direction-Aware Attention (DAA) module to enhance directional feature extraction for multi-oriented text. Each expert is specialized for a different text orientation pattern. We also design a Channel-Adaptive BiFPN (CA-BiFPN) for improved multi-scale feature fusion. Experiments show that DACA improves line-level accuracy by 4.99% over MOoSE on a multi-directional open-set benchmark, increases F-score for rejected samples by 3%–7%, and achieves 0.9% higher accuracy on horizontal text, demonstrating significant advantages in multi-directional text recognition.**

***Index Terms*—Open-set text recognition, multi-oriented text recognition, Direction-aware, BiFPN.**

## I. Introduction

Open-set text recognition (OSTR) [1] is an emerging paradigm in scene text understanding, which requires a model to not only recognize characters from known classes but also identify and classify novel characters without retraining. This task setting is particularly suitable for scene text recognition (STR) [2, 3] across multiple languages such as Chinese, Korean, Japanese, and English.

In real-world scenarios, scene text often exhibits irregular layouts and diverse orientations. However, existing open-set text recognition [4] approaches rarely address the challenge of multi-orientation text, which significantly limits their practical applicability. Conventional techniques, such as text rotation or aspect-ratio normalization, fail to enhance orientation robustness while introducing substantial computational overhead. Although methods like [5] attempt to mitigate directional sensitivity, they suffer from poor convergence due to the scarcity of vertically oriented samples.

Open-set multi-oriented scene text recognition remains challenging because changes in text orientation alter character structures, while open-set settings require reliance on fine-grained cues rather than memorized classes. This combination amplifies domain shifts across orientations and increases ambiguity when recognizing unseen characters.

MOoSE [6] introduced a mixture-of-experts framework for multi-oriented scene text, reducing orientation discrepancies and partially alleviating data scarcity via shared substructures. Inspired by prior methods [7, 8, 9, 10], it uses a rule-based routing mechanism to classify text images into three categories, which are processed by three word-level experts. However, the feature extractors themselves lack explicit mechanisms to model directional dependencies, limiting their ability to capture long-range structural information along different spatial axes. Moreover, the feature fusion strategy remains coarse, treating all channels equally across scales and restricting effective interaction between low-level stroke details and high-level semantic representations. These limitations hinder robust generalization to irregular layouts and unseen characters.

To address these issues, we propose a Direction-Aware and Channel-Adaptive (DACA) framework for open-set multi-oriented text recognition. Our key insight is that robust open-set recognition under orientation variation requires explicit directional modeling at the feature extraction stage and adaptive channel-wise integration during multi-scale feature fusion. To this end, we introduce a Direction-aware Attention (DAA) module that encodes horizontal and vertical spatial dependencies directly into the feature maps, enhancing orientation sensitivity without increasing architectural complexity. In addition, we design a Channel-Adaptive BiFPN (CA-BiFPN) that dynamically adjusts channel contributions during bidirectional feature fusion, enabling fine-grained integration of structural cues critical for unseen character recognition.

The main contributions of this paper are summarized as follows:

1.We propose a Direction-aware Attention (DAA) module that enhances directional sensitivity and cross-orientation generalization by explicitly encoding spatial directionality and positional cues via adaptive directional pooling.

2.We integrate the CA-BiFPN module to achieve bidirectional fusion of low-level (strokes, edges) and high-level (semantic structures) features, dynamically adjusting feature representations text orientation and layout, thereby improving cross-orientation generalization and character discrimination.

* Corresponding author

3.Extensive experiments demonstrate that our method significantly improves open-set multi-orientation scene text recognition performance.

## II. Related work

Close-set text recognition assumes that training and testing share the same character set. Existing methods [11, 12, 13] achieve high accuracy under this fixed setting, but struggle in real-world scenarios where character sets are dynamic. When encountering unseen characters, these models tend to misclassify them, requiring costly retraining to maintain performance.

Zero-shot text recognition methods [14, 15] address unseen characters by leveraging auxiliary information. Some approaches model characters with compositional structures such as strokes and radicals [15, 16], while others encode glyph shapes to generate prototype embeddings and perform similarity-based recognition. Recent advances include SideNet [17], which integrates radicals, glyphs, and strokes, and HierCode [18], which further improves hierarchical encoding efficiency and generalization.

Open-set text recognition (OSTR) extends zero-shot recognition by enabling both recognition and rejection of known and unknown characters using side information [1, 19, 20]. Recent methods improve OSTR through label-to-prototype learning [1], shape-aware visual reconstruction [20], character-context decoupling [21], and context-free learning [4]. While these approaches provide important foundations for OSTR, they mainly focus on horizontal text and struggle with vertically oriented text.

Multi-oriented text recognition [5] is challenging. Existing methods either rotate vertical text to horizontal, risking loss of directional information, or normalize images to square shapes [22], increasing computation and potential distortion. Recent approaches decouple content and orientation to obtain direction-agnostic features, but complex architectures struggle with unseen characters or extremely long vertical text [23]. Methods like SAN [24] choose to resize the vertical samples and horizontal samples into different sizes for cotraining. However, the vast varying sample aspect ratios within the horizontal samples causes these methods also to face excessive computational costs from over-stretching like AON [22] and can limit their performance.

To address these challenges, MOoSE introduced a mixture-of-experts framework to reduce orientation-related domain gaps and alleviate data scarcity via shared knowledge among experts. However, limitations remain: traditional CNNs may lose directional and spatial information during scaling and padding; single-directional feature fusion restricts reverse information flow; and pre-defined fusion rules produce imbalanced contributions across scales and lack adaptability.

To overcome these limitations, we propose a lightweight attention mechanism that enhances the model's ability to capture text orientation and spatial positional cues, and a bidirectional, learnable-weight feature pyramid network to achieve adaptive, hierarchical feature fusion. Together, these designs significantly improve the performance of open-set multi-oriented text recognition.

## III. Method

We propose a Direction-Aware and Channel-Adaptive (DACA) framework (Fig. 1) that takes preprocessed text images and rendered prototype images as inputs. Features are first extracted by a DSBN-ResNet45 backbone with a Direction-aware Attention (DAA) module, then enhanced by a word-level aggregator. CA-BiFPN performs channel-adaptive fusion to obtain refined word-level representations $F$, while character-level features are processed by a character-level aggregator to generate prototype features $P$. Finally, both $F$ and $P$ are input to an open-set classifier for visual prediction.

During feature extraction, DAA enhances the network's sensitivity to orientation and spatial relationships, preserving positional cues along each axis while highlighting informative regions, thus enabling each expert to focus on different directional features, improving the capture of multi-oriented text. In the word feature aggregator, CA-BiFPN performs bidirectional channel-wise fusion, dynamically adjusting each channel's contribution to enable precise feature selection and suppression.

### A. Direction-aware Attention(DAA) Module

To enhance orientation awareness, we introduce a Direction-aware Attention (DAA) module into the feature extraction stage, which can be flexibly inserted into a convolutional backbone. Unlike conventional channel or spatial attention mechanisms that compress spatial features via global pooling, DAA preserves positional information along horizontal and vertical axes using adaptive directional pooling. This allows the network to capture long-range dependencies in a direction-aware manner while retaining fine-grained spatial details, which is particularly important for open-set multi-oriented scene text recognition where distinguishing unseen characters relies on subtle structural variations.

The proposed DAA is conceptually related to coordinate-based attention mechanisms [25], but is specifically tailored for open-set multi-oriented text recognition, where directional cues play a critical role in modeling orientation-induced structural differences across text layouts.

*a) Directional Global Pooling:* To encourage the attention block to capture long-range interactions spatially while preserving precise positional information, we decompose the global pooling into a pair of one-dimensional feature encoding operations.Specifically, Given an input feature map $X$, directional context descriptors along height and width are first computed via adaptive average pooling:

$$\begin{aligned} X_h &= \text{AdaptiveAvgPool}_{H\times 1}(X), \\ X_w &= \text{AdaptiveAvgPool}_{1\times W}(X), \end{aligned} \tag{1}$$

which produces direction-aware features while preserving spatial information along the non-pooled dimension. In practice, adaptive average pooling divides the input feature map into

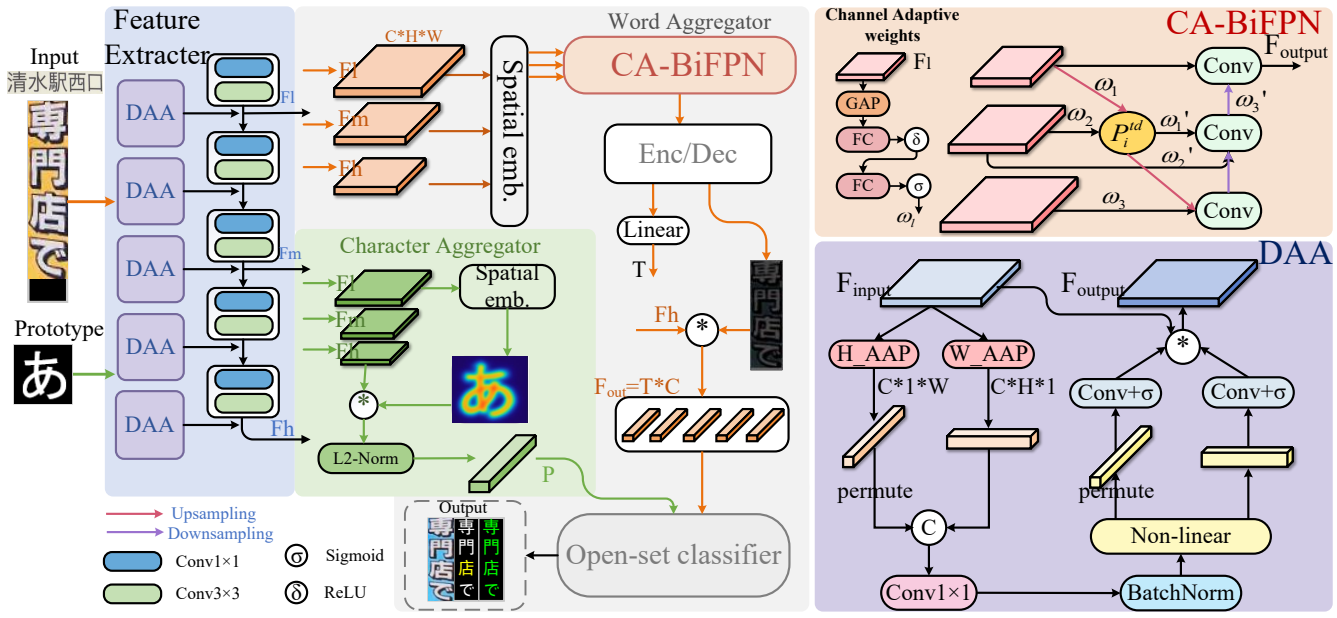


Fig. 1. The DACA framework.The blue region represents the feature extraction stage, the gray region corresponds to the word-level feature aggregator and the open-set classifier, the green region denotes the character-level feature aggregator, the purple region indicates the DAA module, and the peach region corresponds to the CA-BiFPN module.$\sigma$ and $\delta$ denote activation functions, and C denotes concatenation.

evenly spaced regions corresponding to the target output size and computes the average value of each region:

$$Y[c,i,j] = \frac{1}{|R_{i,j}|} \sum_{(h,w)\in R_{i,j}} X[c,h,w], \tag{2}$$

where $R_{i,j}$ denotes the set of input positions mapped to output location $(i,j)$, and $c$ indexes the channel. Unlike global average pooling (GAP), which outputs a single value per channel, adaptive pooling allows the output to retain directional spatial resolution, making it suitable for horizontal or vertical context extraction. By aggregating features separately along these two axes, DAA generates direction-aware and position-sensitive attention maps, explicitly encoding both directional and positional cues.

*b) Directional Attention Generation:* Combining the outputs of (1), we first concatenate them and then feed them into a shared $1\times1$ convolutional transformation function to produce $Y'$:

$$Y' = h\_swish(\text{Conv}([z^h, z^w])), \tag{3}$$

where $[\cdot,\cdot]$ denotes the concatenation operation along the spatial dimension.$Y' \in \mathbb{R}^{\frac{C}{r}\times(H+W)}$ is an intermediate feature map that encodes spatial information in the horizontal and vertical directions. Here, $r$ is the reduction ratio used to control the block size(32 is optimal). Hard Swish ($h\_swish$) is an efficient, differentiable, and gated smooth non-linear activation function, defined as:

$$h\_swish(x) = x \cdot \frac{\text{ReLU6}(x+3)}{6}. \tag{4}$$

It exhibits a smooth quadratic non-linearity in the interval $[-3,3]$, outputs 0 for $x \le -3$, and is approximately linear for $x \ge 3$. Compared to Swish, it is computationally more efficient, requiring only ReLU6 and division without exponential operations.

The reduced feature $Y'$ is split along the original height and width dimensions: $Y'_h \in \mathbb{R}^{C/r\times H}$ , $Y'_w \in \mathbb{R}^{C/r\times W}$. where $C$ denotes the reduced channel number. The vertical component $Y'_w$ is transposed back to its original orientation. Two independent $1\times1$ convolutions generate attention maps along each direction:

$$\begin{aligned} A_h &= \sigma(\text{Conv}_{1\times1}(Y'_h)), \\ A_w &= \sigma(\text{Conv}_{1\times1}(Y'_w)) \end{aligned} \tag{5}$$

where $\sigma(\cdot)$ denotes the Sigmoid function.

Finally, the original feature map $X$ is reweighted by the directional attention maps as :

$$y_c(i,j) = x_c(i,j) \times A^h_c(i) \times A^w_c(j). \tag{6}$$

Direction-aware Attention (DAA) captures spatial directional information and long-range dependencies with minimal cost using only $1\times1$ convolutions. It improves feature quality, object localization, and recognition performance. Unlike existing attention mechanisms for closed-set recognition, DAA explicitly handles orientation-induced shifts in open-set text recognition with unseen characters and directions.

### *B. CA-BiFPN Module*

To improve channel-wise representation in multi-scale feature fusion, we enhance the original Bidirectional Feature

Pyramid Network (BiFPN) [26] with a channel-adaptive weighting mechanism, forming the CA-BiFPN module. This mechanism adjusts the contribution of each channel based on its response in the feature map, enabling more precise feature selection and suppression.

In the original BiFPN, each input feature has a single scalar weight, treating all channels equally. We introduce a channel-wise adaptive weight vector, allowing each channel's importance to be modulated independently during feature fusion.

Let the input feature be $F_l \in \mathbb{R}^{C\times H\times W}$ with a corresponding channel weight vector $\omega_l \in \mathbb{R}^C$. The fused output is computed as:

$$O_c = \sum_l \frac{\omega_{l,c}}{\epsilon + \sum_m \omega_{m,c}} \odot F_{l,c}, \tag{7}$$

where $\odot$ denotes element-wise multiplication along channels, and $\epsilon$ is a small constant to prevent numerical instability.All summations and normalizations are performed independently for each channel $c$. Here $\epsilon$ is a small constant to avoid division by zero. The channel weight $\omega_i$ is generated by a lightweight channel attention module:

$$\omega_l = \sigma(\mathrm{MLP}(\mathrm{GAP}(F_l))), \tag{8}$$

here, $\mathrm{GAP}(\cdot)$ denotes global average pooling, $\mathrm{MLP}(\cdot)$ is a two-layer multi-layer perceptron, and $\sigma(\cdot)$ is the Sigmoid activation function. Let $z \in \mathbb{R}^C$ denote the global average pooled feature vector obtained from an input feature map. The channel-adaptive weighting vector is computed as:

$$\mathbf{w} = \sigma(W_2\,\delta(W_1 z + b_1) + b_2)\,, \tag{9}$$

where $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are the weight matrices of the first and second fully connected layers, respectively, and $b_1$, $b_2$ are the corresponding bias terms. The reduction ratio $r$ controls the intermediate dimesionality and introduces a bottleneck structure to reduce computational cost.

The first fully connected layer performs a channel reduction, projecting the $C$-dimensional descriptor into a lower-dimensional embedding:

$$h = \delta(W_1 z + b_1), \quad h \in \mathbb{R}^{C/r}, \tag{10}$$

thereby learning a compact representation of inter-channel dependencies. The second fully connected layer restores the channel dimension:

$$\tilde{\mathbf{w}} = W_2 h + b_2, \quad \tilde{\mathbf{w}} \in \mathbb{R}^C, \tag{11}$$

which enables the model to assign an independent importance weight to each channel. Finally, a sigmoid activation $\sigma(\cdot)$ is applied to normalize the weights into the range $(0, 1)$, making them suitable for element-wise channel re-weighting in subsequent feature fusion operations.

For each input feature $F_i$, channel-adaptive weights $w_i$ and $w_i'$ are computed separately for the top-down and bottom-up paths. Although the input features are identical, the network allows independent weights at different fusion stages to accommodate the distinct semantic requirements at each stage. The top-down and bottom-up fusion stages are expressed as:

$$P_i^{td} = \mathrm{Conv}\left(\frac{\omega_1\odot P_i^{in} + \omega_2 \odot \mathrm{Resize}(P_{i+1}^{in})}{\epsilon+\omega_1+\omega_2}\right),$$
$$P_i^{out} = \mathrm{Conv}\left(\frac{\omega_1'\odot P_i^{in} + \omega_2' \odot P_i^{td} + \omega_3' \odot \mathrm{Resize}(P_{i-1}^{out})}{\epsilon+\omega_1'+\omega_2'+\omega_3'}\right) \tag{12}$$

CA-BiFPN adopts the bidirectional flow of BiFPN while enabling fine-grained, channel-adaptive feature fusion. Experiments show it improves feature representation and detection accuracy with minimal overhead, highlighting important channels to enhance selectivity, being lightweight with negligible computation, and offering better interpretability through channel attention maps.

### C. Optimization

The model is optimized by minimizing the sum of the classification loss and the length prediction loss for each expert,

$$L = \sum_{a\in\mathcal{E}} \mathrm{CrossEntropy}(\hat{Y}_a, Y_a^*) + \sum_{a\in\mathcal{E}} \mathrm{CrossEntropy}(\hat{t}_a, t_a^*) \tag{13}$$

In the training stage, each expert is associated with a dedicated data stream generated by the dispatcher [6] during preprocessing. During training, each expert dynamically generates prototypes for learning, while during evaluation, cached prototypes are used to speed up computation and support incremental recognition of unseen characters by generating new prototypes features.

## IV. Experiment

In this study,we first perform ablation studies on the MOoSE datasets [6] to assess each module's effectiveness under the GZSL protocol, focusing on multi-oriented open-set text recognition. Then, the full model is evaluated with multiple open-set protocols to verify its ability to recognize novel characters and improved performance. Finally, we compare our method with state-of-the-art approaches on the OSTR task to show overall performance.

### A. Implementation Details

The implementation is based on the MOoSE codebase, which is built upon the **PyTorch** framework. We follow the same training and evaluation data splits as defined in the original implementation. The training samples are composed of data from the ART [27], RCTW [28], CTW [29], and LSVT [30] datasets, as well as the Latin and Chinese subsets of the MLT [31]dataset. The test set consists of the Japanese subset of the MLT dataset.

This paper employs four evaluation protocols, each corresponding to different scenarios and requirements in open-set text recognition:Generalized Zero-Shot Learning (GZSL) [32] , Open-Set Recognition (OSR) [33] , Generalized Open-Set Recognition (GOSR) [34] ,and Open-Set Text Recognition

TABLE I
DETAILED ABLATION STUDY OF DAA AND CA-BIFPN UNDER FOUR SPLITS OF THE MOOSTR PROTOCOL.

| Split | Module | DAA | CA-BiFPN | LA(all)↑ | LA(hori)↑ | LA(vert)↑ | Recall↑ | Precision↑ | F-score↑ |
|---|---|---|---|---|---|---|---|---|---|
| GZSL | DAA | ✓ | | 38.21 | 38.23 | **32.65** | - | - | - |
| | CA-BiFPN | | ✓ | 37.74 | 39.31 | 31.87 | - | - | - |
| | Full Model | ✓ | ✓ | **39.26** | **39.65** | 32.27 | - | - | - |
| OSR | DAA | ✓ | | 71.70 | - | - | **77.18** | 94.25 | 84.86 |
| | CA-BiFPN | | ✓ | **72.01** | - | - | 70.96 | **95.27** | 81.34 |
| | Full Model | ✓ | ✓ | 70.16 | - | - | 77.15 | 94.43 | **84.92** |
| GOSR | DAA | ✓ | | 60.13 | - | - | 71.58 | 76.63 | 74.02 |
| | CA-BiFPN | | ✓ | **62.92** | - | - | 63.34 | 80.97 | 71.08 |
| | Full Model | ✓ | ✓ | 62.46 | - | - | **72.83** | **80.98** | **76.69** |
| OSTR | DAA | ✓ | | 62.97 | - | - | 84.72 | 84.43 | 84.57 |
| | CA-BiFPN | | ✓ | 65.16 | - | - | 79.01 | 87.56 | 83.07 |
| | Full Model | ✓ | ✓ | **65.90** | - | - | **85.68** | **87.93** | **86.79** |

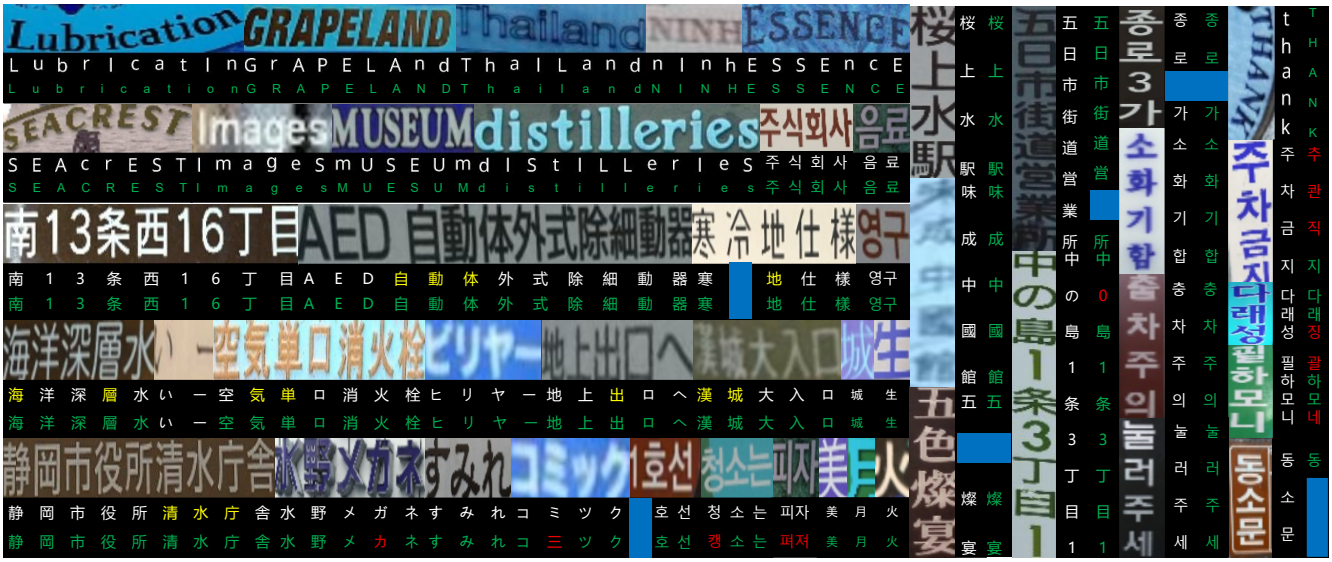

Fig. 2. The visualization of Recognition Results.They include text recognition in three languages: English, Japanese, and Korean. Yellow characters indicate new characters, white characters represent previously seen characters, blue rectangles mark new characters that should be rejected, green indicates correct recognition, and red indicates incorrect recognition.

(OSTR) [1] .The recognition performance of the model is primarily evaluated using line-level accuracy (LA),The rejection performance is measured by recall (RE) and precision (PR), and are summarized with the F-score (FM).

### B. Ablative Experiments

Ablation experiments are conducted on the GZSL split to evaluate recognition of both seen and novel characters. Results in Table I show that both the Direction-aware Attention (DAA) and CA-BiFPN modules consistently improve performance, especially on novel characters. Specifically, DAA enhances spatial modeling and increases line accuracy by 3.94%. CA-BiFPN further improves performance by 3.49% through adaptive multi-level feature fusion. These results confirm the effectiveness of the proposed modules for open-set text recognition.

### C. Multi-Oriented Open-Set Text Recognition

We systematically evaluate the MOoSE framework in multi-orientation scenarios. Trained for 200,000 iterations as described in the reference, qualitative results on the GZSL dataset (Fig. 2) demonstrate robustness to text style and orientation variations. DACA generalizes well to unseen characters, performing effectively on both horizontal and vertical samples, though line accuracy is slightly lower for styles like Japanese kana. Overall line accuracy improves by 4.99% over the baseline. On GOSR and OSTR datasets, average accuracy increases by around 6%, showing better generalization to unseen kanji with similar structures, while OSR accuracy for seen characters rises by about 1%. Our model achieves over 70% recall and 80% precision in rejecting anomalous characters. The smaller recall-precision gap between OSR and GOSR indicates improved recognition of kana, approaching

TABLE II

THE PERFORMANCE ON THE MOOSTR BENCHMARK,THE MULTI-ORIENTATION COUNTERPART, AND ON THE JAPANESE AND KOREAN SUBSETS OF THE MLT DATASET AFTER TRAINING ON CHINESE AND ENGLISH DATA. BOLD INDICATES THE BEST, UNDERLINE INDICATES THE SECOND BEST, AND ITALIC INDICATES UNSEEN NEW CHARACTERS.TABLES I AND III FOLLOW THE SAME CONVENTION.

| Split | Registered | Out-of-Set | Method | LA(all)↑ | Recall↑ | Precision↑ | F-score↑ |
|---|---|---|---|---|---|---|---|
| GZSL | Latin,Shared Kanji,*Unique Kanji*,*Kana* | ∅ | MOoSE | 34.27 | - | - | - |
| | | | MOoSE-XL | 37.39 | - | - | - |
| | | | Ours | **39.26** | - | - | - |
| OSR | Latin,Shared Kanji | *Unique Kanji,Kana* | MOoSE | 69.08 | 71.56 | 94.05 | 81.28 |
| | | | MOoSE-XL | **75.56** | 74.05 | **95.79** | 83.53 |
| | | | Ours | 70.16 | **77.15** | 94.43 | **84.92** |
| GOSR | Latin,Shared Kanji *Unique Kanji* | *Kana* | MOoSE | 56.73 | 62.27 | 76.98 | 68.85 |
| | | | MOoSE-XL | 59.81 | 63.83 | **81.89** | 71.74 |
| | | | Ours | **62.46** | **72.83** | 80.98 | **76.69** |
| OSTR | Shared Kanji, *Unique Kanji* | Latin,*Kana* | MOoSE | 59.96 | 80.42 | 84.91 | 82.61 |
| | | | MOoSE-XL | 61.75 | 80.01 | **88.18** | 83.90 |
| | | | Ours | **65.90** | **85.68** | 87.93 | **86.79** |

TABLE III

PERFORMANCE ON THE OPEN –SET TEXT RECOGNITION TASK.

| Split | Name | LA | R | P | FM |
|---|---|---|---|---|---|
| GZSL | OSOCR-Large | 30.83 | - | - | - |
| | OpenCCD | 36.57 | - | - | - |
| | OpenCCD-Large | 41.31 | - | - | - |
| | OpenSAVR | 40.96 | - | - | - |
| | OpenSAVR-XL | <u>42.58</u> | - | - | - |
| | CFOR-XL | **44.47** | - | - | - |
| | MOoSE-XL | 39.56 | - | - | - |
| | Ours | 40.43 | - | - | - |
| OSTR | OSOCR-Large | 58.57 | 24.46 | **93.78** | 38.80 |
| | OpenSAVR | <u>71.86</u> | 69.72 | 90.86 | 78.90 |
| | OpenSAVR-XL | **72.33** | 72.96 | <u>92.62</u> | 81.62 |
| | CFOR | 71.80 | **86.36** | 89.52 | **87.91** |
| | MOoSE-XL | 64.80 | 80.49 | 89.12 | 84.59 |
| | Ours | 70.61 | <u>84.50</u> | 89.90 | <u>87.12</u> |

performance on new kanji. Kana samples mostly lie within known kanji boundaries, while new kanji have higher recall. Compared to MOoSE, our model generalizes better to kana, and the recall gap between OSTR and GOSR shows that even without user information, known characters can be reliably rejected, allowing rare characters to be removed.

### D. Open-Set Text Recognition

Under the OSTR protocol, we evaluate the proposed Direction-aware Attention (DAA) framework for comparison with prior methods. Since the test set contains only horizontal samples, the vertical expert is excluded. The model is trained following [6] and evaluated after 200,000 iterations. As reported in Table III, the proposed method achieves a higher FM score than existing approaches.

Although the gain in horizontal line accuracy is limited, the model shows improved recognition across multiple orientations. This compensates for the modest horizontal improvement and is crucial for robust real-world text recognition.

Compared to MOoSE, the proposed method brings larger gains on vertically oriented and mixed-orientation text, while maintaining comparable performance on horizontal samples.

## V. CONCLUSION

To improve directional feature modeling and multi-level feature fusion in multi-oriented open-set text recognition, we propose the Direction-aware Attention (DAA) and CA-BiFPN modules. DAA encodes global information along horizontal and vertical directions, while CA-BiFPN enables channel-specific bidirectional feature fusion. Ablation studies confirm their effectiveness and complementarity.

The proposed model achieves strong recognition and rejection performance, notably improving unseen Kana recognition, and handles both isolated characters and text lines with robust generalization. Limitations include occasional instability in distinguishing known from novel characters and sensitivity to scaling or font variations. Future work will focus on enhancing rejection of ambiguous known characters.

## REFERENCES


[1] Chang Liu, Chun Yang, Hai-Bo Qin, Xiaobin Zhu, Cheng-Lin Liu, and Xu-Cheng Yin. Towards open-set text recognition via label-to-prototype learning. *Pattern Recognition*, 134:109109, 2023.

[2] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, pages 9989–9999, 2020.

[3] Chang Liu et al. Gccnet: Grouped channel composition network for scene text detection. *Neurocomputing*, 454:135–151, 2021.

[4] Chang Liu, Chun Yang, Zhiyu Fang, Hai-Bo Qin, and Xu-Cheng Yin. Cfor: Character-first open-set text recognition via context-free learning. *IEEE Transactions on Image Processing*, 33:6497–6507, 2024.

[5] Haiyang Yu, Xiaocong Wang, Bin Li, and Xiangyang Xue. Orientation-independent chinese text recognition in scene images. In *IJCAI*, pages 1667–1675, 8 2023.

[6] Chang Liu, Simon Corbillé, and Elisa Hope Barney Smith. Moose: Multi-orientation sharing experts for open-set scene text recognition. In *ICDAR*, pages 93–110, 2024.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern*, 39(6):1137–1149, 2017.

[8] Albert Q Jiang et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[9] Isaac Ong et al. Routellm: Learning to route llms from preference data. In *ICLR*, 2025.

[10] Shanghaoran Quan. DMoERM: Recipes of mixture-of-experts for effective reward modeling. In *ACL*, pages 7006–7028, August 2024.

[11] Shuai Zhao, Ruijie Quan, Linchao Zhu, and Yi Yang. Clip4str: A simple baseline for scene text recognition with pre-trained vision-language model. *IEEE Transactions on Image Processing*, 33:6893–6904, 2024.

[12] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Syst. Appl.*, 41(18):8027–8048, 2014.

[13] Wenchao Wang, Jianshu Zhang, Jun Du, Zi-Rui Wang, and Yixing Zhu. Denseran for offline handwritten chinese character recognition. In *ICFHR*, pages 104–109, 2018.

[14] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Adaptive text recognition through visual matching. In *ECCV*, pages 51–67, 2020.

[15] Jianshu Zhang, Jun Du, and Lirong Dai. Radical analysis network for learning hierarchies of chinese characters. *Pattern Recognition*, 103:107305, 2020.

[16] Tianwei Wang et al. Decoupled attention network for text recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12216–12224, Apr. 2020.

[17] Ziyan Li, Yuhao Huang, Dezhi Peng, Mengchao He, and Lianwen Jin. Sidenet: Learning representations from interactive side information for zero-shot chinese character recognition. *Pattern Recognition*, 148:110208, 2024.

[18] Yuyi Zhang et al. Hiercode: A lightweight hierarchical codebook for zero-shot chinese text recognition. *Pattern Recognition*, 158:110963, 2025.

[19] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *CVPR*, pages 16901–16911, June 2024.

[20] Chang Liu, Chun Yang, and Xu-Cheng Yin. Open-set text recognition via shape-awareness visual reconstruction. In Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi, editors, *ICDAR*, pages 89–105, 2023.

[21] Chang Liu, Chun Yang, and Xu-Cheng Yin. Open-set text recognition via character-context decoupling. In *CVPR*, pages 4513–4522, June 2022.

[22] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Aon: Towards arbitrarily-oriented text recognition. In *CVPR*, pages 5571–5579, 2018.

[23] Jie-Bo Hou et al. Ham: Hidden anchor mechanism for scene text detection. *IEEE Transactions on Image Processing*, 29:7904–7916, 2020.

[24] Chankyu Choi, Youngmin Yoon, Junsu Lee, and Junseok Kim. Simultaneous recognition of horizontal and vertical text in natural images. In *ACCV Workshops*, pages 202–212, 2019.

[25] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13713–13722, June 2021.

[26] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, pages 10778–10787, 2020.

[27] Chee Kheng Chng et al. Icdar2019 robust reading challenge on arbitrary-shaped text - rrc-art. In *ICDAR*, pages 1571–1576, 2019.

[28] Baoguang Shi et al. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *ICDAR*, volume 01, pages 1429–1434, 2017.

[29] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34(3):509–521, 2019.

[30] Yipeng Sun et al. Icdar 2019 competition on large-scale street view text with partial labeling - rrc-lsvt. In *ICDAR*, pages 1557–1562, 2019.

[31] Nibal Nayef et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition — rrc-mlt-2019. In *ICDAR*, pages 1582–1587, 2019.

[32] Farhad Pourpanah et al. A review of generalized zero-shot learning methods. *IEEE Trans. Pattern*, 45(4):4051–4070, 2023.

[33] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Trans. Pattern*, 35(7):1757–1772, 2013.

[34] Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Trans. Pattern*, 43(10):3614–3631, 2021.