IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1139-1153
Journal homepage: http://iieta.org/journals/ts

English Scene Character Detection via Multi-Scale Contextual Feature Fusion and Attention Mechanisms

Gaimin Jin*, Jingan Hu

Department of Public Foreign Languages, Shijiazhuang College of Applied Technology, Shijiazhuang 050081, China

Corresponding Author Email: jingm_123@126.com

https://doi.org/10.18280/ts.430306

Received: 12 January 2026
Revised: 18 June 2026
Accepted: 22 June 2026
Available online: 30 June 2026

Keywords:
scene character detection, multi-scale context awareness, dual-path attention mechanism, dynamic upsampling, progressive boundary refinement, criss-cross attention

ABSTRACT

The robust detection of English characters in natural scenes is persistently challenged by substantial scale variations, intricate background interferences, narrow character contour distortions, and insufficient boundary localization precision. Conventional feature pyramid architectures and fixed upsampling strategies inadequately model long-range contextual dependencies, leading to missed detections of small-scale characters and significant fitting errors for irregular character outlines. In this work, an end-to-end high-accuracy detection framework for English scene characters was constructed. The detection optimization was achieved through a progressive technical system encompassing multi-scale contextual perception feature extraction, dual-path adaptive attention enhancement, dynamic upsampling with contour preservation, and gradual boundary refinement. Long-range feature modeling was realized in a lightweight manner via multi-dilation atrous convolutions and criss-cross attention mechanisms. Feature representation was strengthened through a gated fusion of channel-wise and spatial attention. Contour distortion was suppressed via context-driven coordinate offset dynamic upsampling. Ultimately, iterative boundary optimization was accomplished through curvature-adaptive sampling constrained by prior knowledge of character aspect ratios. The proposed method effectively accommodates complex real-world scenarios, thereby providing reliable technical support for industrial tasks such as English optical character recognition and nameplate inspection on industrial equipment.

1. INTRODUCTION

Character detection in natural scenes constitutes a fundamental and core research direction within the field of machine vision-based intelligent perception, and also serves as an indispensable preliminary step for optical character recognition systems, with its detection performance directly determining the overall accuracy and stability of text recognition tasks [1, 2]. With the rapid iteration of intelligent industries, the demand for natural-scene text detection technologies has been continuously escalating across various deployed scenarios, including autonomous driving environment perception, intelligent inspection of industrial equipment, cross-border commodity verification, and mobile real-time translation [3, 4]. As the most globally prevalent language, English is widely distributed across diverse open and complex scenes. Compared with laboratory-standard printed texts, English characters in real-world scenarios possess distinctive inherent attributes that substantially elevate detection difficulty [5, 6]. Under natural imaging conditions, English characters exhibit vast scale discrepancies, with pixel-size spans potentially reaching two orders of magnitude, thereby imposing stringent requirements on the multi-scale adaptive modeling capacity of detection models [7, 8]. Concurrently, a substantial number of narrow-bodied English characters occupy extremely low pixel proportions and possess fragile edge details, rendering them highly susceptible to feature loss and contour distortion during network sampling procedures [9, 10]. Furthermore, English word characters adhere to fixed spatial arrangements and morphological ratio regularities, encapsulating rich contextual prior information that remains underutilized by existing detection frameworks [11, 12]. On this basis, the core challenges of English character detection in complex scenes are addressed in this work, with targeted breakthroughs in key technologies such as multi-scale feature modeling, contour-preserving sampling, and character-level boundary refinement, thereby effectively enhancing the robustness and accuracy of English character detection under complex interference and severe scale variation, and providing reliable technical support for high-performance industrial-grade English optical character recognition systems [13, 14].

Current mainstream scene character detection methods are predominantly architected upon convolutional feature pyramid structures, coupled with feature regression or segmentation strategies for text localization, and are capable of accommodating text detection tasks with conventional scales and clear backgrounds; however, numerous inherent deficiencies persist in their adaptability to complex English scenes [15, 16]. In terms of multi-scale feature extraction, traditional feature pyramids rely solely on fixed upsampling

and downsampling operations for hierarchical feature transmission, resulting in constrained receptive field coverage and an inability to effectively model long-range contextual dependencies among characters. Meanwhile, classical atrous spatial pyramid pooling methods focus exclusively on local multi-scale feature aggregation, lacking global pixel-wise relational modeling capacity, which renders them highly prone to missed detections of small targets and insufficient feature representation for large targets when confronted with English characters of extreme scales [17, 18]. At the feature fusion level, existing methods predominantly employ fixed weights for the concatenation and integration of high-level and low-level features, failing to dynamically adjust the fusion proportions between semantic information and spatial details according to scene-specific textures, scales, and interference characteristics. Moreover, single-dimensional attention mechanisms can only realize unidirectional feature enhancement, making it difficult to simultaneously accomplish channel-wise semantic calibration and spatially precise reinforcement, thereby causing persistent bottlenecks in small-scale character localization accuracy [19, 20]. During the feature upsampling reconstruction stage, the fixed sampling kernels of conventional transposed convolutions introduce regularized artifacts, which induce edge structure distortions for English characters with slender contours, directly introducing boundary regression errors and substantially degrading the detection quality of narrow characters [21, 22]. At the boundary optimization level, conventional single-stage regression methods are inadequate for achieving sub-pixel-level high-precision localization. Generic post-processing techniques fail to incorporate the morphological-structural priors specific to English characters, and fixed-density boundary sampling strategies cannot adapt to the curvature variations of character contours—resulting in redundant sampling in flat regions and insufficient sampling in curved regions, thus failing to balance detection accuracy with computational efficiency [23, 24].

To address the aforementioned common technical bottlenecks in the field, an end-to-end high-accuracy detection framework for English scene characters is constructed in this work, incorporating a series of targeted technical innovations. A lightweight multi-scale contextual perception feature extraction module is designed, integrating multi-dilation convolutions with criss-cross attention mechanisms, whereby multi-scale contextual modeling capacity is substantially enhanced while computational overhead is significantly reduced. A dual-path attention adaptive fusion module is constructed to simultaneously accomplish channel-wise semantic calibration and spatial position enhancement, with adaptive feature fusion achieved through a dynamic gating mechanism. A context-driven dynamic upsampling strategy is proposed to replace conventional fixed sampling approaches, effectively preserving the contour details of narrow characters. A progressive boundary refinement strategy incorporating character morphological priors is established to achieve efficient and high-precision character contour fitting. The effectiveness of each module is thoroughly validated through experiments and ablation studies on multiple standard datasets, with the overall performance of the proposed method surpassing that of contemporaneous mainstream detection algorithms.

The complete research logic of this work follows a structured progression from problem formulation, through methodology design and experimental validation, to concluding remarks and future outlook, and is organized into six sections. In Section 1, the research background and technical significance are systematically elaborated, the core deficiencies of existing studies are analyzed, and the principal innovations of this work are synthesized. In Section 2, the overall architecture of the proposed detection framework, along with the design principles and implementation methodologies of each core module, is described in detail. In Section 3, multi-dimensional comparative experiments, ablation studies, and targeted robustness experiments are designed, through which comprehensive validation and analysis of the method's performance are conducted. In Section 4, the technical advantages and existing limitations of the proposed method are objectively analyzed, with future directions for subsequent optimization and extended research outlined. In Section 5, the overall research contributions are summarized, and the core conclusions and application value are distilled.

2. METHODOLOGY

2.1 Overview of the overall detection framework

An end-to-end deep learning framework for character detection in complex English scenes is proposed in this work, with the overall network following a progressively optimized feature processing logic to accomplish character detection tasks. The model takes raw scene images as input, whereupon basic multi-scale feature extraction is first performed via a backbone network. ResNet-50 is selected as the default backbone, while the Swin Transformer architecture is also accommodated to ensure feature modeling flexibility. Deep features at three different resolutions are extracted from the backbone network as fundamental representations, with each level of features corresponding to a distinct downsampling scale, thereby effectively covering multi-dimensional feature information of scene characters and balancing shallow spatial detail features with deep semantic features, thus providing stable foundational inputs for subsequent fine-grained feature optimization and detection inference.

The overall network architecture is composed of four core functional modules that are sequentially arranged to form a complete inference pipeline, with each module being orderly connected according to the logic progressing from coarse-grained feature extraction to fine-grained result optimization. The foundational features are first fed into the multi-scale contextual perception feature extraction module, where multi-scale semantic mining and long-range contextual dependency modeling are accomplished, compensating for the scale-adaptability deficiencies inherent in conventional feature representations. Subsequently, collaborative enhancement and dynamic fusion across feature dimensions are realized through the dual-path attention adaptive fusion module, thereby improving the specificity and effectiveness of feature representations. The model further introduces a dynamic upsampling mechanism for feature map resolution reconstruction, whereby contour distortions introduced by conventional sampling methods are avoided while feature dimensions are restored. Finally, iterative optimization of instance contours and morphological constraint correction are achieved through the character-level progressive boundary refinement strategy, with high-accuracy scene English character detection results being output. Each module provides

complementary optimization in a progressive manner, enabling simultaneous enhancement of both detection accuracy and robustness for English characters in complex scenes.

2.2 Multi-scale contextual perception feature extraction module

To address the challenges posed by the large-scale span of English characters in natural scenes and the insufficient receptive field coverage of conventional feature extraction structures, a multi-scale contextual perception feature extraction module is designed in this work, whereby multi-dimensional semantic features and long-range contextual information are simultaneously captured within a controllable computational budget. The architecture of the multi-scale contextual perception feature extraction module is illustrated in Figure 1. The module takes the multi-level features output by the backbone network as input, whereupon a 1 × 1 convolutional operator is first applied to compress the feature channels, reducing the original channel dimension to one-quarter and thereby effectively decreasing the parameter count and computational complexity of subsequent convolutional operations. The compressed features are then fed into three parallel 3 × 3 atrous convolutional branches, through which receptive field representations of varying extents are obtained via the configuration of differentiated dilation rates, thus accommodating the feature extraction requirements of both extremely small-scale and extremely large-scale characters in scenes. The dilation rates of each branch are sequentially set to 3, 6, and 12 in this work, ensuring gradient coverage of the receptive field from local regions to large-scale regions, with each branch maintaining uniform input and output channel dimensions to guarantee multi-scale feature dimensional alignment. The branch output features can be computed as follows:

$$F^{(r)}=Conv_{3\times3}^{(r)}(F_{down}),r\in\{3,6,12\} \quad (1)$$

where, F_{down} denotes the channel-compressed foundational features; $Conv^{(r)}{}_{3\times3}$ represents the atrous convolution operation with dilation rate r; and $F^{(r)}$ denotes the multi-scale feature output corresponding to a single convolutional branch.

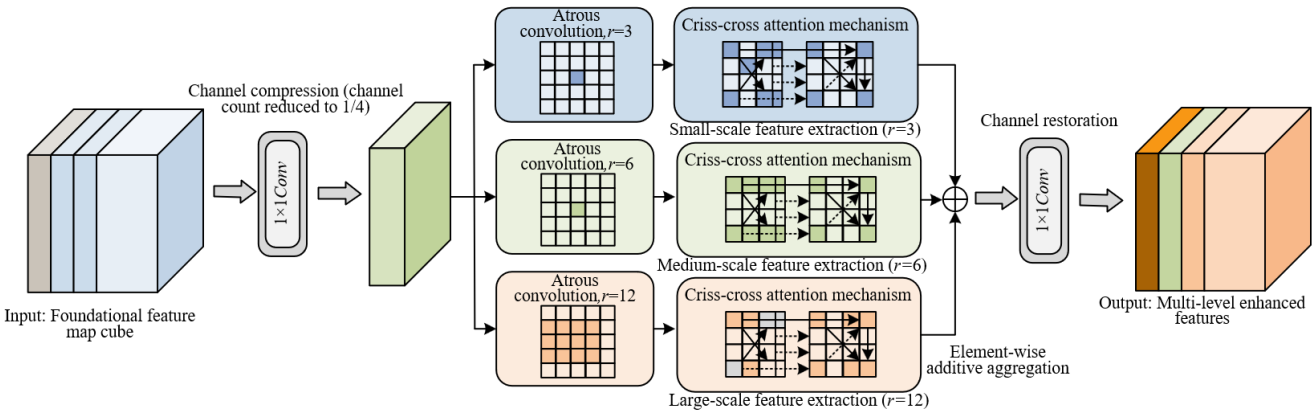

Figure 1. Architecture of the multi-scale contextual perception feature extraction module

To compensate for the deficiency of multi-scale atrous convolutions in modeling pixel-level long-range dependencies, a criss-cross attention mechanism is introduced in this work to perform semantic enhancement on the features of each branch, focusing on pixel-wise associative features along the horizontal and vertical orthogonal directions of the image, which is well-aligned with the horizontally arranged spatial distribution characteristics of English characters. For the output features of each atrous convolution branch, query, key, and value feature maps are generated through linear transformations, with the feature channel dimension being uniformly compressed to one-eighth of the input dimension, thereby preserving core semantic information while reducing the computational cost of attention operations. The attention weights are computed independently along the row and column dimensions of the image, and contextual features from the entire global row and column regions are aggregated pixel-wise, enabling effective modeling of long-range semantic associations within character regions. The feature enhancement process can be defined as:

$$y_{i,j}=\sum_{k=1}^{W}Softmax(q_{i,j}^{\top}k_{i,k})v_{i,k}+\sum_{l=1}^{H}Softmax(q_{i,j}^{\top}k_{l,j})v_{l,j} \quad (2)$$

where, H and W denote the height and width of the feature map, respectively; $q_{i,j}$, $k_{i,k}$, and $v_{i,k}$ denote the query vector, key vector, and value vector corresponding to spatial position (i,j) of the feature map; and $y_{i,j}$ denotes the pixel-wise feature after attention enhancement. Compared with the quadratic complexity of global self-attention, this mechanism reduces the overall computational complexity from $O(H^2W^2)$ to $O(HW(H+W))$, achieving a reduction of over 85% in computational cost for input images at 1280 × 1280 resolution, thereby realizing an effective balance between high-precision contextual modeling and low computational overhead.

After the completion of single-branch multi-scale feature enhancement, the feature information from the three branches with differentiated receptive fields is deeply fused. An element-wise additive strategy is adopted in this work for preliminary aggregation of the multi-branch features, whereby fine-grained texture information and long-range contextual semantics captured by each branch are completely preserved through this fusion approach, while feature redundancy and information offset issues caused by channel concatenation operations are avoided. Cross-channel information interaction and dimension restoration are then performed on the preliminarily aggregated features via a 1 × 1 convolution, with the feature channels being restored to the original dimensions of the module input, thereby achieving complementary

coupling and unified representation of multi-scale semantic features. Following the multi-stage processing, the module ultimately outputs the optimized multi-level enhanced features $\widetilde{F}_3$, $\widetilde{F}_4$, and $\widetilde{F}_5$, which effectively compensate for the deficiencies of conventional feature pyramids in extreme-scale character representation and long-range contextual dependency modeling, providing highly robust foundational feature support for subsequent feature fusion and fine-grained detection tasks.

2.3 Dual-path attention feature enhancement and adaptive fusion module

To address the issues of channel semantic redundancy and weakened spatial localization information inherent in multi-scale contextual features, a dual-path attention feature enhancement and adaptive fusion module is constructed in this work, whereby the quality of feature representations is simultaneously optimized from two dimensions—channel semantic modeling and spatial position reinforcement—enabling fine-grained calibration and scene-adaptive fusion of multi-scale features. The architecture of the dual-path attention feature enhancement and adaptive fusion module is illustrated in Figure 2. The module first performs channel-wise feature selection and dependency modeling through a grouped Transformer structure, with global semantic statistics being computed for the enhanced features at each level. For a hierarchical feature $\widetilde{F}_l$, global spatial information is aggregated via global average pooling, generating a one-dimensional channel descriptor that characterizes the overall response intensity of each channel, which is computed as follows:

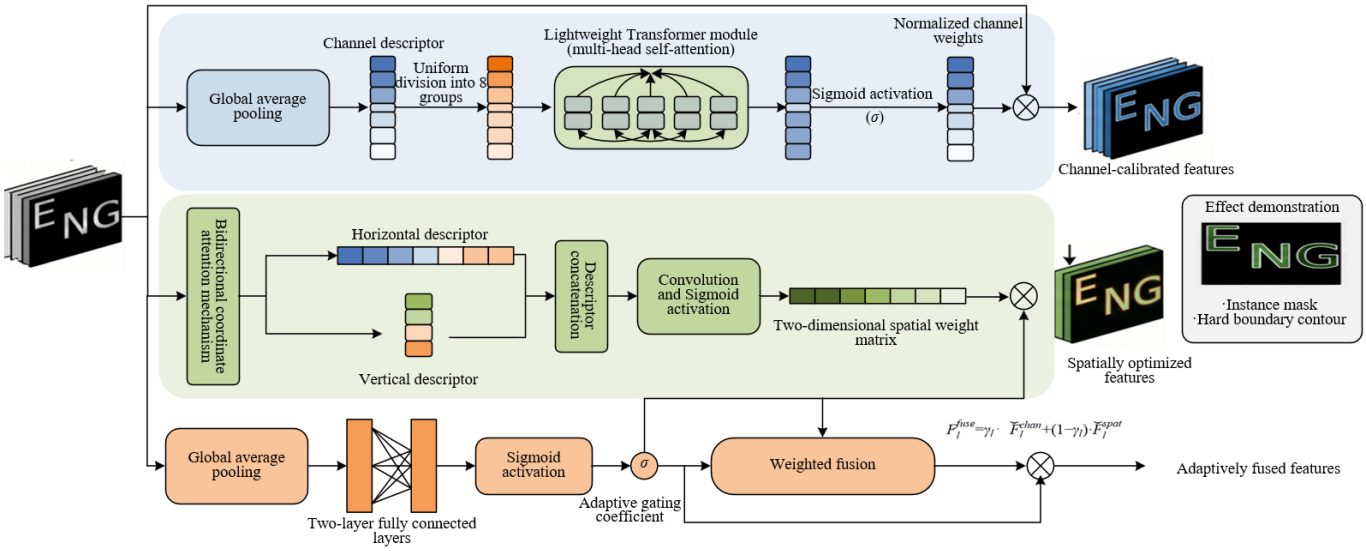

Figure 2. Architecture of the dual-path attention feature enhancement and adaptive fusion module

$$z_l = \frac{1}{H_l W_l} \sum_{i=1}^{H_l} \sum_{j=1}^{W_l} \widetilde{F}_l(i,j), z_l \in \mathrm{R}^{C_l} \tag{3}$$

where, H_l and W_l denote the spatial dimensions of the l-th level feature map, and C_l denotes the channel dimension of the feature. To balance channel interaction modeling capacity with computational efficiency, the channel descriptor is uniformly divided into eight sub-feature groups, within each of which inter-channel relationships are mined through a lightweight Transformer composed of a four-head multi-head self-attention mechanism combined with a two-layer perceptron. After the completion of feature interactions within each group, dimensional concatenation is performed, and normalized channel weight coefficients are generated via a Sigmoid activation function. Finally, semantic calibration of the original features is accomplished through element-wise multiplication, whereby effective feature channels are strengthened and ineffective noise channels are suppressed. The calibration process can be expressed as:

$$\widehat{F}_l = \alpha_l \odot \widetilde{F}_l \tag{4}$$

where, α_l denotes the normalized channel attention weights, $\odot$ denotes the element-wise multiplication operation, and $\widehat{F}_l$ denotes the features after channel-dimensional optimization.

Building upon the channel semantic calibration, a bidirectional coordinate attention mechanism is introduced into the module to achieve spatial-dimensional position information enhancement, whereby character target region features are precisely activated and localization deviations for small-scale and narrow characters are ameliorated. Global pooling sampling is performed on the calibrated features along two orthogonal directions, namely horizontal and vertical, through this mechanism, generating bidirectional spatial descriptors containing precise positional information, thereby effectively overcoming the limitation of conventional spatial attention in encoding long-range positional associations. The bidirectional pooling operations are defined as:

$$p_l^h(i) = \frac{1}{W_l} \sum_{j=1}^{W_l} \widehat{F}_l(i,j), p_l^h \in \mathrm{R}^{C_l \times H_l \times 1} \tag{5}$$

$$p_l^v(j) = \frac{1}{H_l} \sum_{i=1}^{H_l} \widehat{F}_l(i,j), p_l^v \in \mathrm{R}^{C_l \times 1 \times W_l} \tag{6}$$

where, p_l^h and p_l^v denote the positional feature descriptors along the horizontal and vertical directions, respectively. The two types of one-dimensional positional features are expanded to a unified spatial size via dimensional broadcasting, after which channel concatenation is performed. A 2D spatial attention weight matrix is then generated through convolution for dimensionality reduction, nonlinear activation, and

dimensionality restoration. The weight matrix is applied to the input features via element-wise weighting, achieving feature-adaptive enhancement in the image spatial dimension, with the spatially optimized features being ultimately obtained as:

$$\breve{F}_l = \beta_l \odot \hat{F}_l \tag{7}$$

where, β_l denotes the normalized spatial attention weights. To address the limitation that fixed-weight fusion strategies cannot adapt to the feature distributions of complex scenes, an adaptive gating fusion mechanism is designed in this work, whereby the fusion proportions between channel semantic features and spatial positional features are dynamically balanced. Global semantic statistics are performed on the spatially enhanced features through this mechanism, with spatial dimensions being compressed via global average pooling, feature mapping and dimensionality transformation being accomplished through a two-layer fully connected network, and an adaptive gating coefficient in the range of 0 to 1 being ultimately output via a Sigmoid function. The computation process is defined as follows:

$$\gamma_l = Sigmoid(FC(GAP(\breve{F}_l))) \tag{8}$$

where, γ_l denotes the adaptive fusion weight coefficient, and the intermediate fully connected layer dimension is set to one-sixteenth of the original channel dimension, thereby reducing computational overhead while maintaining mapping capacity. Based on this dynamic weight, the features from the channel-enhanced branch and the spatially enhanced branch are fused in a weighted manner, with the fusion formula given as:

$$F_l^{fuse} = \gamma_l \cdot \breve{F}_l^{chan} + (1-\gamma_l) \cdot \breve{F}_l^{spat} \tag{9}$$

Through this fusion approach, the contribution proportions of the two-branch features are automatically adjusted according to the scene complexity, character scale, and background interference level of the input image, with fine spatial details being preferentially preserved in simple scenes and high-level semantic features being emphasized in complex interference scenes. High-quality multi-scale fused features that combine both semantic representation capacity and spatial localization precision are ultimately generated, providing superior feature support for subsequent upsampling reconstruction and boundary detection tasks.

2.4 Dynamic upsampling module with multi-scale context enhancement

During the feature decoding stage of scene character detection networks, conventional upsampling operations commonly rely on transposed convolutions for feature map resolution restoration, with fixed-parameter convolutional kernels inevitably producing periodic grid overlap effects that induce checkerboard artifacts. While such distortion issues have limited impact in generic object detection tasks, they are extremely sensitive to English narrow characters with minimal pixel widths, readily disrupting the continuity and integrity of character edges, causing feature contour distortion, and ultimately resulting in detection boundary regression deviations. To address the detail loss and contour distortion issues caused by fixed sampling patterns, a dynamic upsampling module with multi-scale context enhancement is designed in this work, whereby sampling coordinates are adaptively corrected through the neighborhood semantic information of the features themselves, freeing the operation from the constraints of fixed convolutional kernels. High-fidelity feature reconstruction with both enhanced resolution and complete preservation of character edge details is thereby achieved. A schematic illustration of the context-driven dynamic upsampling process is shown in Figure 3.

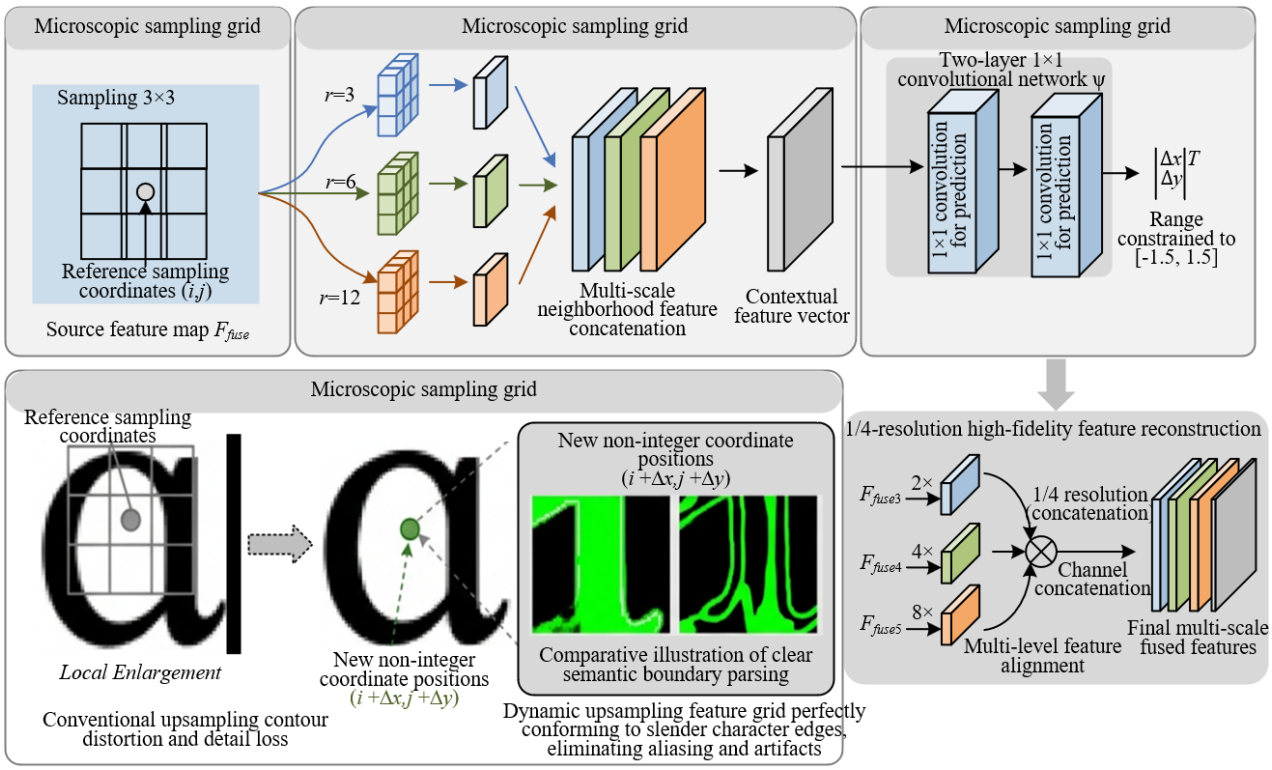

Figure 3. Schematic illustration of the context-driven dynamic upsampling process

The correction accuracy of dynamic sampling coordinates is highly dependent on the quality of feature information in the neighborhood of each sampling point; therefore, multi-scale contextual feature extraction is first performed on each interpolation reference position within the module. With the network upsampling rate being denoted as s, the reference sampling coordinates on the source feature map corresponding to spatial coordinates (i',j') on the target feature map can be mapped as $(i,j) = (i'/s, j'/s)$. The neighborhood sampling range is delineated centered at the reference coordinates, through which multi-scale local feature information is mined in parallel via three groups of atrous convolutions with different dilation rates, capturing fine edge textures and large-range neighborhood association features simultaneously. The aggregated neighborhood contextual feature set is defined as:

$$N_{i,j} = \bigcup_{r\in\{1,2,4\}} \{ Conv_{3\times3}^{(r)}(F^{fuse})_{i+\Delta i, j+\Delta j} : |\Delta i|, |\Delta j| \le k/2 \} \quad (10)$$

where, r denotes the atrous convolution dilation rate, k denotes the neighborhood window size, F^{fuse} denotes the fused features output by the preceding dual-path attention module, and Δi and Δj denote the neighborhood spatial offsets. The multi-group scale features are concatenated and fused along the channel dimension, constructing a feature vector that contains multi-level local semantics and edge information, thereby providing sufficient and fine-grained feature support for subsequent coordinate offset prediction, while sampling deviations caused by insufficient single-scale contextual information are avoided.

Based on the aggregated neighborhood contextual features, a lightweight convolutional network is employed within the module to perform dynamic offset prediction of sampling coordinates. The prediction network is composed of two stacked 1 × 1 convolutional layers, with the first layer reducing the input feature channel dimension to 64 to reduce computational cost, and the second layer outputting two-dimensional coordinate offset parameters corresponding to horizontal and vertical position corrections, respectively. The prediction process can be expressed as:

$$(\Delta i^*, \Delta j^*) = \Phi(N_{i,j}) \quad (11)$$

where, Φ denotes the mapping function of the offset prediction network, and Δi^* and Δj^* denote the final optimized coordinate offsets. To suppress feature disorganization caused by excessive coordinate offsets during training and to enhance model convergence stability, the offset values are constrained within the range of ±1.5 pixels. Based on the corrected precise sampling coordinates, bilinear interpolation is performed on the original fused feature map within the model, enabling adaptive computation of pixel values at the target positions, with sampling positions being dynamically adapted according to local semantics, thereby fundamentally avoiding the regularized artifacts and edge distortions introduced by fixed sampling patterns.

To accommodate the multi-level feature fusion architecture, differential upsampling ratios are applied to the fused features at three different scales. Upsampling ratios of 2×, 4×, and 8× are respectively applied to F^{fuse}_3, F^{fuse}_4, and F^{fuse}_5, with all level features being uniformly mapped to a resolution scale of $H/4 \times W/4$. The multi-level features after size unification are deeply fused via channel concatenation, producing a high-fidelity feature representation that combines both shallow spatial details and deep semantic information, serving as the input for the subsequent character boundary refinement module. This sampling mechanism is fully adapted to the feature distribution characteristics of English narrow characters and small-scale characters, effectively compensating for the detail deficiencies of conventional upsampling methods, and providing a high-quality feature foundation for subsequent high-precision character boundary regression and contour fitting.

2.5 Character-level progressive boundary refinement strategy

To achieve sub-pixel-level boundary fitting and morphological normalization correction for English characters, a coarse-to-fine character-level progressive boundary refinement strategy is designed in this work, whereby detection contour deviations are progressively corrected through a multi-level iterative optimization logic. Coarse-grained character instance prediction is first performed within the model, providing an initial contour basis for subsequent fine-grained optimization. A character region probability map of dimension $H/4 \times W/4$ is output through the segmentation branch within the network, with a precise character foreground mask being generated in conjunction with a differentiable binarization algorithm, thereby effectively distinguishing target regions from complex backgrounds. Simultaneously, a four-channel offset parameter map is output through the boundary regression branch, with pixel-wise distances from the current position to the top, bottom, left, and right boundaries of each character being predicted. Instance partitioning of the prediction results is accomplished through connected component analysis, from which the initial bounding box and complete contour curve of each individual character are obtained, completing the localization output of the coarse detection stage. An iterative flow diagram of the character-level progressive boundary refinement strategy is shown in Figure 4.

To address the issues of feature redundancy and detail deficiency inherent in conventional uniform sampling approaches, a curvature-adaptive keypoint sampling mechanism is adopted in this work for boundary point selection. The initial character contour is constructed as a continuous parametric curve $C(t) = (x(t), y(t))$, with the parameter variable t being normalized to the range of 0 to 1. The curvature magnitude $\|C''(t)\|$ at each position is obtained by solving the second-order differential of the curve, quantitatively characterizing the local bending degree of the contour. Adaptive sampling point selection is achieved based on the cumulative curvature distribution characteristics, with sampling being densified in regions of severe contour deformation and sparsified in smooth regions. The sampling point selection rule is defined as:

$$t_k = inf\left\{ t : \int_0^t \| C''(\tau) \| d\tau \ge \frac{k}{N} \int_0^1 \| C''(\tau) \| d\tau \right\}, k = 1,2,\ldots,N \quad (12)$$

where, t_k denotes the curve parameter corresponding to the k-th sampling point, and N denotes the total number of sampling points per character contour, which is uniformly set to 16 in this work. Through this sampling approach, the geometric features of character corners and curved edges are maximally preserved without increasing computational overhead, thereby enhancing the fineness of contour representation.

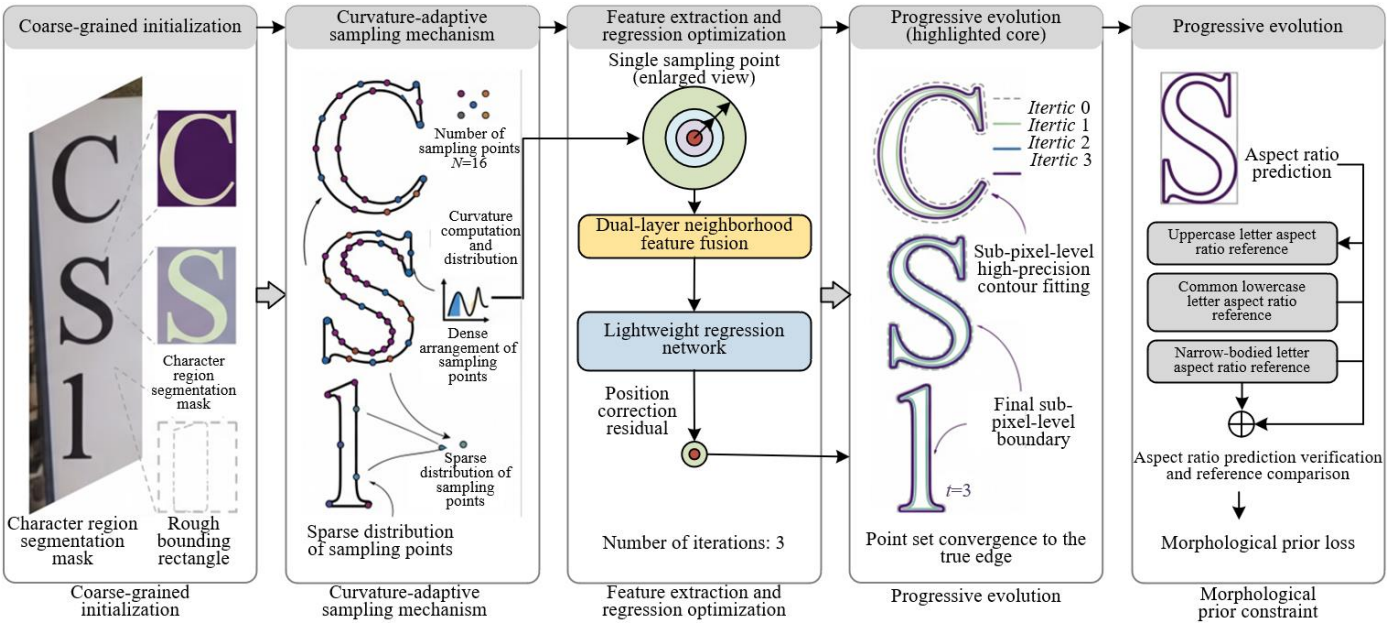

Figure 4. Iterative flow diagram of the character-level progressive boundary refinement strategy

After the completion of boundary keypoint sampling, iterative boundary position optimization is performed within the model based on multi-scale neighborhood contextual features. In each iteration of the update process, dual-layer neighborhood features with radii of 3 and 7 are extracted centered at each boundary keypoint, fusing local fine-grained textures with large-range contextual semantics to compensate for the limitations of single-point feature representation. The multi-layer neighborhood features are concatenated with the original keypoint features and fed into a lightweight regression network, which is composed of two convolutional layers and two fully connected layers in series, for predicting the spatial position residuals of the keypoints. The computation is defined as follows:

$$\Delta p_i^{(t)} = \Psi\left(Concat\left(Feat(p_i), Context(p_i, r_1), Context(p_i, r_2)\right)\right) \quad (13)$$

where, Ψ denotes the mapping function of the boundary regression network, and $\Delta p_i^{(t)}$ denotes the position correction residual at the t-th iteration. Based on the residual results, the coordinate positions are updated point-by-point as $p_i^{(t+1)}=p_i^{(t)}+\Delta p_i^{(t)}$, with high-precision convergence of boundary coordinates being achieved after three iterations of optimization, thereby effectively eliminating localization deviations in the coarse contours.

To address the issue of character morphological distortion that tends to occur during boundary optimization in complex scenes, the inherent aspect ratio prior of English characters is introduced as a regularization constraint in this work, thereby enhancing the physical plausibility of the detection results. In accordance with the morphological distribution characteristics of English characters, the detected targets are categorized into three classes—uppercase letters, common lowercase letters, and narrow-bodied letters—with corresponding standard aspect ratio reference ranges being established. During the training phase, the reference parameters are matched according to the ground-truth categories, while during the inference phase, category adaptation is automatically accomplished through the dimensional characteristics of the coarse segmentation contours. A regularization loss function is constructed by computing the error between the aspect ratio of the predicted contour bounding rectangle and the standard reference value:

$$L_{prior} = \sum_i \| AR(p_i^{(t+1)}) - AR_{ref} \|_2^2 \quad (14)$$

where, $AR(p_i^{(t+1)})$ denotes the actual aspect ratio of the optimized character contour, and AR_{ref} denotes the standard reference value corresponding to the character category. This constraint mechanism effectively suppresses abnormal morphological deviations during the boundary regression process, adapts to the structural characteristics of narrow English characters, and ultimately achieves character boundary detection results that combine both high precision and morphological plausibility.

2.6 Overall loss function and end-to-end training strategy

A composite loss function is constructed in this work based on the principle of multi-task collaborative optimization, through which character region segmentation, coarse-grained boundary regression, fine-grained boundary fitting, and character morphological constraints are jointly supervised, enabling end-to-end parameter optimization of the entire network framework. The overall loss is obtained through the weighted combination of four functionally complementary sub-losses, with the specific expression given as:

$$L_{total} = \lambda_1 L_{seg} + \lambda_2 L_{reg} + \lambda_3 L_{boundary} + \lambda_4 L_{prior} \quad (15)$$

where, λ_1, λ_2, λ_3, and λ_4 denote the balancing weights for each loss term, with the optimal weight configuration of 1.0, 1.0, 0.8, and 0.2 being determined through grid search traversal of the parameter space in this work. The segmentation loss L_{seg} is trained with an equal-weight combination of dice loss and focal loss, whereby the sample imbalance issue between foreground characters and background regions in natural scenes is alleviated, while the model's learning capacity for hard samples such as blurred, occluded, and extremely small

characters is simultaneously enhanced. The coarse boundary regression loss L_{reg} adopts Smooth L1 loss to supervise pixel-level boundary offset parameters, ensuring stable convergence of the initial detection box localization. The boundary refinement loss $L_{boundary}$ measures the deviation between the iteratively optimized keypoints and the ground-truth contours using L1 distance, precisely constraining sub-pixel-level boundary fitting performance. The morphological prior loss L_{prior} incorporates the distribution characteristics of English character aspect ratios to construct a regularization constraint, effectively correcting morphological distortions that occur during the boundary optimization process, thereby improving the plausibility and authenticity of the detection results.

To balance model training stability with scene generalization capacity, a two-stage progressive training strategy is adopted in this work for iterative network parameter updates. The model is first pre-trained on the large-scale SynthText synthetic dataset for 100 epochs, through which general texture features and spatial arrangement regularities of English characters are comprehensively learned, a stable foundational feature representation capacity is established, and the overfitting risk caused by the limited sample size of real-scene datasets is effectively avoided. After the completion of the pre-training phase, targeted fine-tuning is performed on the Total-Text and ICDAR 2015 real-scene datasets, adapting to real-scene characteristics such as complex background interference, severe scale variations, and curved characters. A learning rate warm-up mechanism is introduced during the training process, whereby the learning rate is gradually increased in the initial iterations, thereby avoiding convergence instability caused by parameter update oscillations. Concurrently, a gradient clipping strategy is introduced to constrain the range of extreme gradient values, with the gradient explosion problem during deep network training being suppressed, thereby ensuring stable parameter updates throughout the multi-module collaborative training process and fully unleashing the feature optimization and boundary refinement performance of each core module.

3. EXPERIMENTAL DESIGN AND RESULT ANALYSIS

3.1 Experimental setup

Three publicly available standard datasets were selected in this work for model training and evaluation, encompassing both synthetic and real-scene datasets, through which model generalization capacity training and real-scene performance validation were concurrently addressed. The large-scale SynthText synthetic dataset was employed for model pre-training, providing sufficient foundational feature learning samples for English characters and effectively mitigating the overfitting risk caused by insufficient sample sizes in real-scene datasets. The Total-Text dataset focused on curved and multi-scale English scene texts, serving as an appropriate benchmark for irregular character detection performance testing. The ICDAR 2015 dataset contained extensive natural street-view text with occlusion, low illumination, and complex background interference, and was used to verify the environmental robustness of the model. All datasets were uniformly converted to quadrilateral character annotation formats, with general data augmentation strategies such as random flipping, random scaling, and color space transformation being applied during the training phase, while the original image resolution was preserved during the testing phase to ensure the authenticity of detection results. Detailed information of the datasets is presented in Table 1.

Table 1. Detailed information of experimental datasets

Dataset	Training Set Size	Test Set Size	Scene Characteristics	Evaluation Purpose
Total-Text	1,555 images	293 images	Predominantly English, containing horizontal, multi-directional, and curved text; character height scale range of 8–200 pixels	Primary performance testing for curved text and multi-scale character detection
ICDAR 2015	1,000 images	500 images	Natural street-view scenes with random text orientations and substantial scale variations; extensive occluded and background-interference samples	Robustness testing for detection under complex interference scenarios
SynthText	800,000 images	—	Standardized synthetic English text with precise and complete annotations; diverse scene samples	Model pre-training and enhancement of general feature representation capacity

The experimental hardware platform in this work was equipped with dual NVIDIA A100 80G graphics processing units and an Intel Xeon Gold 6330 central processing unit, with 256 GB of random-access memory being configured to meet the demands of large-scale image training and iterative processing on extensive datasets. The software environment was built upon the PyTorch 2.1 deep learning framework, with CUDA 12.1 and cuDNN 8.9 being employed to accelerate model computations. ResNet-50 was adopted as the default backbone network, with ImageNet pre-trained weights being loaded for transfer learning. During the training phase, the long side of the input images was randomly scaled to 640–2560 pixels, while the short side was constrained not to exceed 1536 pixels. During the testing phase, the long side of the images was fixed at 1280 pixels, with the original aspect ratio being preserved. The AdamW optimizer was employed for parameter updates, with the weight decay coefficient set to 1×10^{-4} and the initial learning rate set to 1×10^{-4}, in conjunction with a cosine annealing learning rate scheduler and a five-epoch learning rate warm-up strategy to ensure training stability. Fine-tuning was performed for 300 epochs on the Total-Text dataset with a batch size of 8, and for 200 epochs on the ICDAR 2015 dataset with a batch size of 16. In the post-processing stage, a differentiable binarization algorithm was adopted, with the segmentation threshold set to 0.3, small connected-component noise below 10 pixels being filtered out, and the number of boundary refinement iterations uniformly set to 3.

3.2 Comparison with mainstream methods

To validate the overall performance advantages of the

proposed method, mainstream scene character detection algorithms published in top computer vision journals and conferences over the past three years were selected as comparison baselines, covering three major technical paradigms: segmentation-based, contour regression-based, and Transformer-based methods. The segmentation-based methods included Progressive Scale Expansion Network v2, Pixel Aggregation Network++, and Differentiable Binarization++; the contour regression-based methods included TextDragon and Complementary Trilateral Decoder Network (CTD-Net); and the Transformer-based methods included TextFormer and Fourier Contour Embedding Network, enabling comprehensive comparison of detection performance across different technical architectures. The quantitative comparison results of all algorithms on the Total-Text and ICDAR 2015 datasets are presented in Tables 2 and 3, respectively.

Table 2. Performance comparison of different algorithms on the Total-Text dataset (%)

Method	Publication Source	Precision	Recall	F-Measure
Progressive Scale Expansion Network v2	CVPR 2022	86.2	83.5	84.8
Pixel Aggregation Network++	IJCV 2022	87.1	84.2	85.6
Differentiable Binarization++	TPAMI 2023	88.3	86.4	87.3
TextDragon	ECCV 2022	87.5	85.1	86.3
Complementary Trilateral Decoder Network	ICCV 2023	88.6	86.8	87.7
TextFormer	CVPR 2023	89.1	87.2	88.1
Fourier Contour Embedding Network	TPAMI 2024	89.7	88.5	89.1
Proposed method	-	90.6	90.4	90.5

Table 3. Performance comparison of different algorithms on the ICDAR 2015 dataset (%)

Method	Publication Source	Precision	Recall	F-Measure
Progressive Scale Expansion Network v2	CVPR 2022	85.3	82.1	83.7
Pixel Aggregation Network++	IJCV 2022	86	83.4	84.7
Differentiable Binarization++	TPAMI 2023	87.2	85.1	86.1
TextDragon	ECCV 2022	86.5	84.3	85.4
Complementary Trilateral Decoder Network	ICCV 2023	87.6	85.7	86.6
TextFormer	CVPR 2023	88	86.2	87.1
Fourier Contour Embedding Network	TPAMI 2024	88.5	87	87.7
Proposed method	-	89.4	89.1	89.3

From the experimental results, it can be observed that the proposed method achieves optimal overall performance on both mainstream English scene character detection datasets. On the Total-Text dataset, an F-measure of 90.5% is attained by the proposed method, which represents a 1.4% improvement over the suboptimal algorithm (Fourier Contour Embedding Network), with a more substantial improvement being observed in recall. This demonstrates that the multi-scale contextual modeling mechanism effectively captures the features of curved and extreme-scale characters, substantially reducing the miss-detection rate of irregular texts. On the ICDAR 2015 complex street-view dataset, an F-measure improvement of 1.2% is achieved by the proposed method over the optimal comparison algorithm, with simultaneous gains in both precision and recall, reflecting the suppression capability of the dual-path attention enhancement mechanism and boundary prior constraints against complex backgrounds and occlusions. The overall experimental results indicate that, compared with algorithms relying on conventional fixed feature fusion and sampling mechanisms, the progressive optimization architecture proposed in this work is better adapted to the inherent characteristics of English scene characters, including substantial scale variations, irregular morphologies, and complex backgrounds, with significant performance advantages being demonstrated in both general and irregular scenarios.

3.3 Module effectiveness ablation experiments

To validate the effectiveness of each of the four core modules in this work, progressive ablation experiments were conducted on the Total-Text dataset based on the ResNet-50+feature pyramid network+differentiable binarization baseline model, with the performance gains and computational costs of each module being comprehensively evaluated from three dimensions: detection accuracy, parameter count, and inference speed. The ablation experiment results obtained by progressively adding the multi-scale contextual perception feature extraction, dual-path attention feature enhancement and adaptive fusion, dynamic upsampling module with multi-scale context enhancement, and character-level progressive boundary refinement modules are presented in Table 4.

The experimental results demonstrate that each module contributes positive performance gains to the model, and the combination of multiple modules exhibits favorable synergistic optimization effects. After the standalone addition of the multi-scale contextual perception feature extraction module, the recall of the model is improved by 1.6%, demonstrating that multi-scale atrous convolutions and criss-cross attention effectively enlarge the receptive field, enhance the modeling capacity for small-scale characters and long-range contextual features, and substantially reduce miss-detected samples. With the addition of the dual-path attention feature enhancement and adaptive fusion module, simultaneous improvements in both precision and recall are observed, confirming that the channel-wise and spatial dual-dimensional adaptive fusion mechanism effectively filters informative features, suppresses background noise, and optimizes feature representation quality. Following the introduction of the dynamic upsampling module with multi-scale context enhancement, the boundary fitting accuracy of the model is substantially enhanced, with notable improvements being achieved in the detection of narrow characters. The character-level progressive boundary

refinement strategy further optimizes character contours, and the aspect ratio prior constraint effectively corrects morphological distortions, ultimately yielding a 0.5% F-measure gain for the model. Compared with the baseline model, the complete algorithm achieves a 4.2% improvement in F-measure, with a parameter increase of only 7.8% and a modest reduction in inference speed, thereby realizing significant detection accuracy enhancement with controlled computational overhead. To validate the necessity of the core components within each module, component-level ablation experiments are conducted, with the results being shown in Figures 5, 6, and 7.

From the component ablation results of the multi-scale contextual perception feature extraction, it can be observed that the multi-branch atrous convolution effectively enhances multi-scale adaptability compared with the single-branch structure, with an F-measure improvement of 0.7% being achieved. The criss-cross attention mechanism achieves 92% of the performance of global self-attention, while reducing the computational cost by 62%, thereby realizing a favorable balance between performance and efficiency.

Table 4. Progressive ablation experiment results of overall modules

Model Configuration	Precision (%)	Recall (%)	F-Measure (%)	Parameters (M)	Frames per Second
Baseline	86.1	84.2	85.3	23.6	28.7
Baseline + multi-scale contextual perception feature extraction	87.2	85.8	86.9	24.3	27.9
Baseline + multi-scale contextual perception feature extraction + dual-path attention feature enhancement and adaptive fusion	87.9	86.9	87.9	24.9	27.3
Baseline + multi-scale contextual perception feature extraction + dual-path attention feature enhancement and adaptive fusion + dynamic upsampling module with multi-scale context enhancement	88.8	87.6	88.7	25.2	26.8
Full model without prior	89.9	88.9	89.8	25.4	26.5
Full model	90.6	90.4	90.5	25.4	24.5

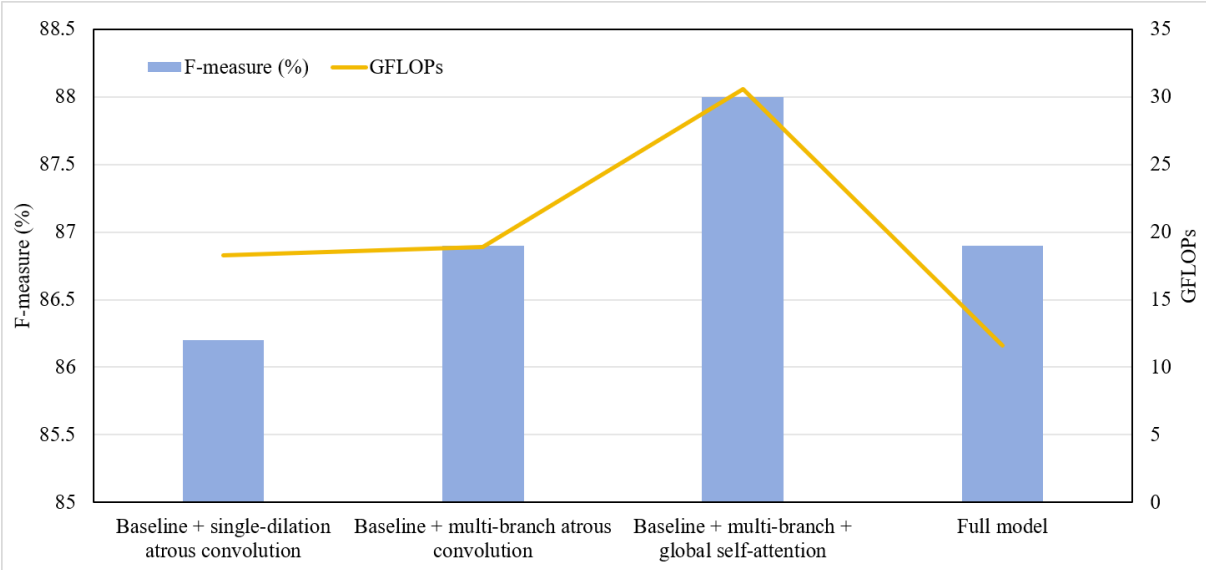

Figure 5. Component ablation experiment results for the multi-scale contextual perception feature extraction

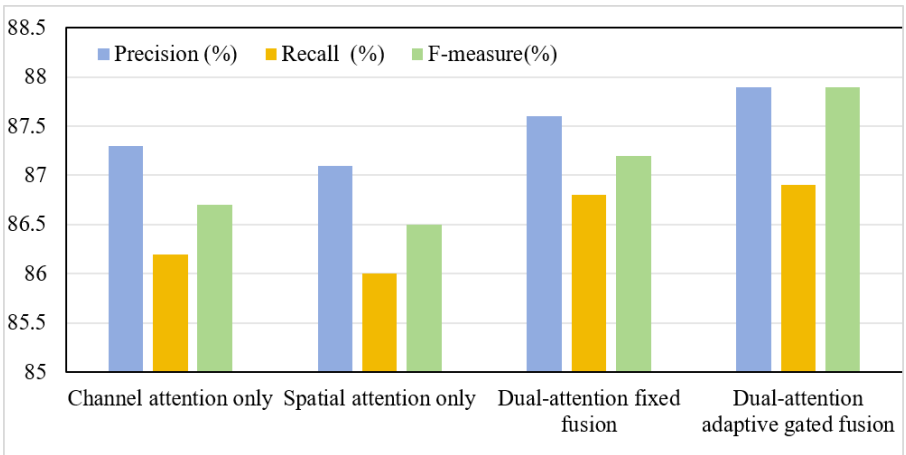

Figure 6. Component ablation experiment results for the dual-path attention feature enhancement and adaptive fusion

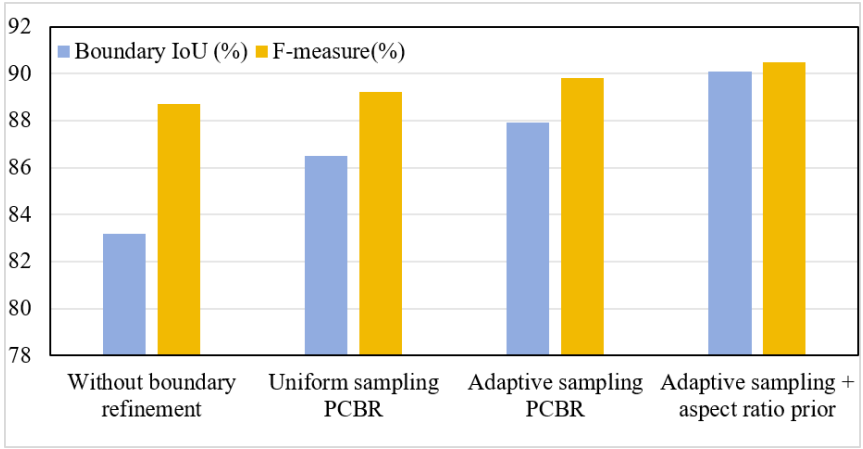

Figure 7. Component ablation experiment results for the character-level progressive boundary refinement

Table 5. Comparative experiments on atrous convolution dilation rate combinations

Dilation Rate Combination	F-measure (%)	Phenomenon Analysis
{1,2,4}	89.1	Receptive field coverage is relatively small, resulting in insufficient feature extraction for large-scale characters.
{2,4,8}	89.7	Multi-scale adaptability is improved, though extreme-scale adaptation deficiencies still exist.
{3,6,12}	90.5	Receptive field gradient distribution is reasonable, effectively accommodating both small-scale and large-scale character features.
{4,8,16}	90.1	Dilation rates are excessively large, introducing excessive background noise and causing a modest decline in precision.

Table 6. Comparative experiments on channel grouping numbers of the dual-path attention feature enhancement and adaptive fusion

Number of Channel Groups (G)	F-Measure (%)	Phenomenon Analysis
4	89.6	The number of groups is small, resulting in insufficient refinement of channel interaction modeling.
8	90.5	Channel grouping is reasonable, with intra-group correlation mining being fully achieved.
16	90.2	Excessive groups lead to fragmentation of correlated features across groups.
32	89.8	Severe channel fragmentation occurs, resulting in a decline in semantic representation integrity.

From the component ablation experiments of the dual-path attention feature enhancement and adaptive fusion, it is demonstrated that the collaborative optimization of the dual-path attention mechanism significantly outperforms the single-branch attention mechanism. The adaptive gated dynamic fusion strategy adaptively adjusts the fusion weights according to the scene characteristics, yielding an additional 0.4% improvement in F-measure compared with the fixed-weight fusion approach, thereby effectively addressing the insufficient scene adaptability of fixed fusion strategies.

From the component ablation results of the character-level progressive boundary refinement, it is demonstrated that the curvature-adaptive sampling mechanism, compared with conventional uniform sampling, precisely focuses on the critical regions of character contours, thereby enhancing contour fitting accuracy. The English character aspect ratio prior constraint effectively suppresses boundary regression distortions, further improving both the morphological plausibility and precision of the detection results.

3.4 Key hyperparameter sensitivity analysis experiments

To determine the optimal hyperparameter configuration for the model, gradient-controlled experiments were conducted in this work on the four core hyperparameters, with the influence of hyperparameter variations on model performance being analyzed.

From the dilation rate comparative experiments shown in Table 5, it can be observed that the {3,6,12} gradient combination forms a reasonable multi-scale receptive field coverage range, which is well adapted to the scale distribution characteristics of English characters in natural scenes. Dilation rates that are too small fail to cover large-scale character features, while excessively large dilation rates introduce invalid background features that interfere with model discrimination.

From Table 6, it can be observed that optimal model performance is achieved when the number of channel groups is set to 8. When the grouping number is too low, the channel dependency modeling becomes coarse; conversely, when the grouping number is too high, global semantic correlations across channels are disrupted, leading to a degradation in feature representation capacity.

From the sampling point count experiment results shown in Figure 8, it can be observed that optimal model performance is achieved when the number of sampling points is increased to 16, reaching a balanced state. Further increases in the sampling point count yield only marginal precision improvements, while computational cost is substantially increased and model inference efficiency is significantly degraded.

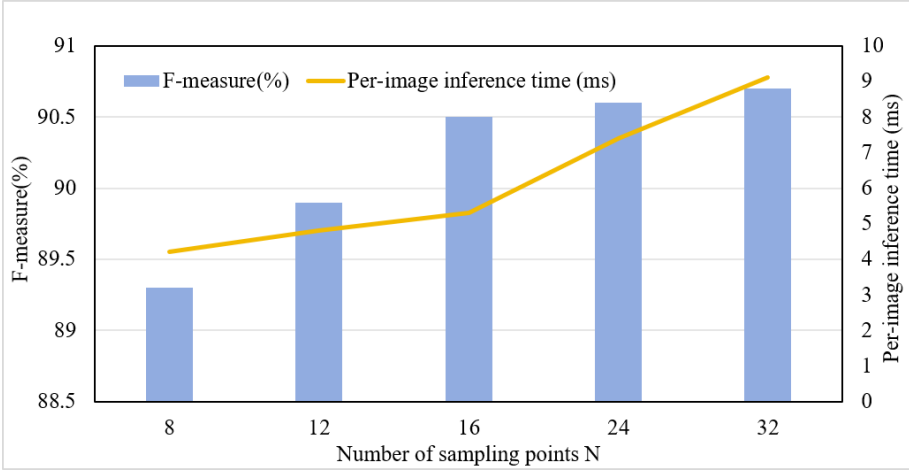

Figure 8. Comparative experiments on boundary sampling point count

Table 7. Comparative experiments on the prior loss weight

Weight λ_4	F-Measure (%)	Phenomenon Analysis
0	90.0	No morphological constraint is applied; a small number of character distortion false detections exist.
0.1	90.3	The prior constraint is relatively weak, with limited optimization effectiveness.
0.2	90.5	Constraint strength is moderate, achieving optimal balance between precision and generalization.
0.3	90.2	Constraint is overly strong, resulting in reduced adaptability for irregular character detection.
0.5	89.7	Excessive constraint restricts the model's adaptive optimization capacity.

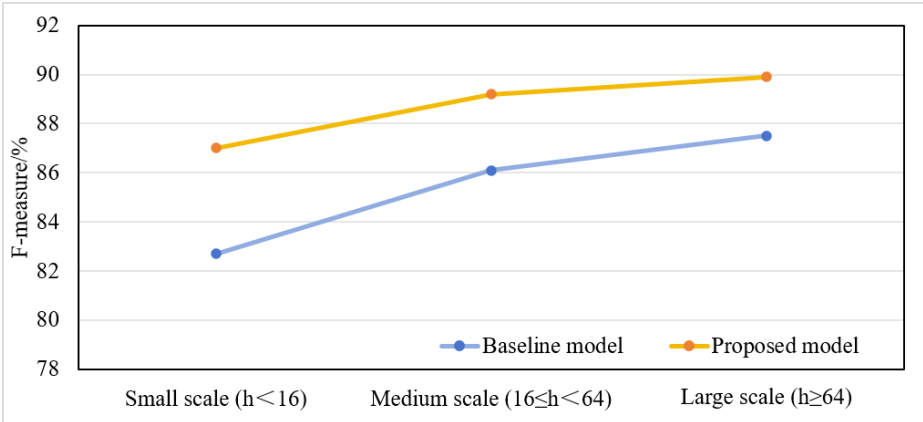

Figure 9. Performance comparison of multi-scale character detection

From Table 7, it can be observed that optimal overall model performance is achieved when the prior loss weight is set to 0.2. When the weight is too small, morphological distortions cannot be effectively constrained; conversely, when the weight is too large, the model's adaptive fitting capacity for special artistic fonts and deformed characters is excessively restricted, leading to a degradation in model generalization.

3.5 Specialized performance and robustness validation experiments

To specifically validate the adaptability of the proposed algorithm to challenging scenarios, three specialized experiments were conducted, focusing on multi-scale character detection, narrow character detection, and robustness against complex interference. The Total-Text test set was partitioned into three scale subsets according to character pixel height, with the comparison results between the baseline model and the proposed method being presented in Figure 9.

The experimental results demonstrate that the most significant performance improvement of the proposed algorithm is achieved for small-scale characters, with a gain of 4.3%, while simultaneous accuracy improvements are observed for medium-scale and large-scale characters. This confirms that the multi-scale contextual perception feature extraction module effectively adapts to scenarios with drastic character scale variations, addressing the issues of small-target miss-detection and insufficient feature representation for large targets that are inherent in conventional models.

A specialized test subset was constructed for English narrow characters, with the detection results being presented in Table 8.

Table 8. Performance comparison of narrow character detection

Model	Narrow Character F-Measure (%)	Boundary Intersection Over Union (%)
Baseline model	83.5	84.9
Proposed model	87.2	90.1

Compared with the baseline model, the proposed method achieves a 3.7% improvement in narrow-character F-measure and a 5.2% improvement in boundary intersection over union. These results fully validate that the dynamic upsampling module effectively suppresses contour distortions introduced by conventional sampling approaches, precisely preserves the edge details of narrow characters, and substantially enhances the boundary fitting accuracy for irregular narrow characters.

Robustness testing was conducted on three interference subsets of the ICDAR 2015 dataset, with the performance degradation rate of the model being adopted as the evaluation metric. The results are presented in Table 9.

Table 9. Performance degradation comparison under complex interference scenarios (F-measure degradation rate/%)

Model	Occlusion Interference	Low-Light Interference	Gaussian Noise Interference
Baseline model	4.8	5.1	4.5
Proposed model	2.7	2.9	2.4

Across the three types of complex interference scenarios—occlusion, low-light, and noise—the performance degradation rates of the proposed model are more than 2% lower than those of the baseline model in all cases. Benefiting from the feature selection capability of the attention mechanism and the long-range contextual modeling characteristics, the model effectively resists background interference, compensates for local feature deficiencies, and demonstrates superior scene robustness.

3.6 Visualization validation experiments

To intuitively validate the actual detection capability of the proposed method under conditions of complex backgrounds, scale variations, and narrow characters, two representative English scenes—street-view storefront signage and product packaging—were selected in this work for comparative visualization experiments. As can be observed from Figure 10, although the baseline model achieves preliminary localization of the main text regions, issues such as boundary offsets, local miss-detections, and insufficient contour adherence persist at small-scale telephone numbers, slanted handwriting, vertically arranged characters, and narrow-bodied text on packaging surfaces. Following the introduction of multi-scale contextual modeling and dual-path attention enhancement, the model's discriminative capacity for long-range characters, semantic associations between adjacent characters, and complex background interference is markedly improved, with text response regions becoming more complete and non-text textures being effectively suppressed. After the further adoption of dynamic upsampling, the continuity and detail fidelity of character edges are significantly enhanced, with particularly stable structural recovery being exhibited in character regions characterized by thin strokes, curved contours, and substantial aspect ratio variations. Following progressive boundary refinement, the final detection contours conform more closely to the true character boundaries, maintaining high consistency in typical challenging cases such as street-view numerals, vertical signage, and product brand and flavor labels. The above results demonstrate that the multi-scale contextual perception feature extraction, attention-based adaptive enhancement, dynamic contour reconstruction, and iterative boundary optimization mechanisms constructed in this work exhibit favorable synergistic effects, simultaneously improving the completeness, boundary precision, and robustness of character detection in complex English scenes, thereby providing a stable front-end detection foundation for subsequent high-reliability text recognition and industrial vision applications.

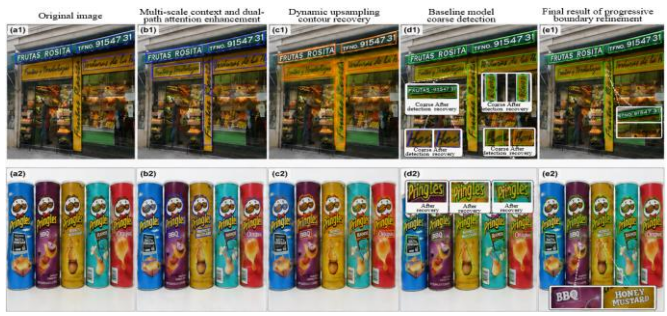

Figure 10. Implementation effect visualization of English scene character detection with multi-scale contextual feature fusion and attention mechanisms

4. DISCUSSION

The performance advantages of the proposed detection framework are derived from the progressive collaborative design and scene-adaptive optimization of each core module, with different functional modules addressing the inherent challenges of English scene character detection from the perspectives of feature extraction, feature fusion, image reconstruction, and boundary optimization. Through the construction of full-coverage receptive fields via gradient-dilated convolutions, coupled with lightweight long-range dependency modeling through criss-cross attention, the multi-scale contextual perception feature extraction module effectively resolves the insufficiency of conventional networks in capturing extremely large- and small-scale character features, thereby enhancing multi-scale scene adaptability at the foundational feature level. The dual-path attention adaptive fusion mechanism performs channel-wise semantic screening and spatial position enhancement separately, with adaptive feature fusion across different scenes being achieved through dynamic gating weights, thereby overcoming the limitations of fixed fusion strategies while balancing noise suppression in complex backgrounds with fine character detail preservation. By abandoning the inherent deficiencies of fixed-kernel sampling and adaptively correcting sampling coordinates based on neighborhood context, the dynamic upsampling mechanism thoroughly alleviates the destructive effects of checkerboard artifacts on narrow character contours. Through curvature-adaptive sampling for enhanced contour representation accuracy and aspect ratio prior constraints for reduced boundary regression solution space, the character-level boundary refinement strategy effectively avoids morphological distortions caused by unconstrained optimization, achieving simultaneous improvements in both detection accuracy and contour plausibility.

Although excellent detection performance is exhibited by the proposed method in conventional complex natural scenes,

significant limitations remain under extreme operating conditions and engineering deployment scenarios. In highly occluded scenes, when the effective character area occlusion ratio exceeds 50%, substantial loss of local character texture and neighborhood contextual information occurs, and the boundary iterative optimization module fails to obtain sufficient feature support for coordinate correction, resulting in notable detection accuracy degradation. For non-standardized characters such as artistically deformed fonts and irregular handwriting, the morphological scales deviate from the aspect ratio distribution patterns of conventional printed fonts, rendering the preset prior constraints inapplicable to such irregular character structures and substantially reducing their effectiveness. At the model deployment level, although steady accuracy improvements are achieved through the multi-module collaborative architecture, the parameter count remains relatively large compared with lightweight detection models, making it difficult to directly adapt to real-time detection requirements of low-computational-power terminal devices such as embedded and mobile platforms, thereby limiting the practical application scenarios of the algorithm.

To address the limitations of the current method, three future research directions are identified in this work: robustness optimization, lightweight deployment, and generalization expansion. In subsequent research, high-level textual semantic priors may be introduced, leveraging character sequence distribution regularities and English lexical knowledge to assist feature inference, thereby compensating for the deficiency of low-level image features in extreme occlusion and blurring scenarios and further enhancing the environmental robustness of the model. Concurrently, structured pruning and knowledge distillation techniques can be incorporated to streamline redundant network parameters and feature channels, with feature knowledge transfer being accomplished through high-precision teacher models, thereby reducing model size and computational cost while maintaining detection accuracy, and adapting to edge-device deployment requirements. Furthermore, generalization-oriented modifications can be performed on the existing network architecture to adapt to the morphological characteristics and structural regularities of multilingual characters, transcending the single-language English detection constraint and constructing a universal scene character detection framework applicable to multiple languages and fonts, thereby further expanding the academic value and engineering application scope of the algorithm.

5. CONCLUSION

The core challenges of English character detection in complex natural scenes—including severe scale variations, strong random background interference, vulnerability of narrow character contours to distortion, and insufficient boundary localization precision—were systematically investigated in this work, with an end-to-end character detection framework being constructed through the synergistic optimization of multi-scale contextual and dual-path attention mechanisms. A multi-scale contextual perception feature extraction module, a dual-path attention adaptive fusion module, a context-enhanced dynamic upsampling module, and a character-level progressive boundary refinement strategy were successively designed. Efficient long-range contextual modeling was achieved through lightweight atrous convolutions and criss-cross attention; fine-grained feature enhancement was accomplished via channel-wise and spatial dual-path adaptive attention mechanisms; contour distortion issues of conventional upsampling were resolved through dynamic coordinate-offset interpolation; and iterative boundary optimization was realized by combining curvature-adaptive sampling with character morphological prior constraints. Through this comprehensive pipeline spanning feature extraction, feature fusion, image reconstruction, and contour fitting, the technical bottlenecks of existing scene character detection algorithms were effectively overcome.

The effectiveness and superiority of the proposed method were thoroughly validated through multi-dimensional comparative experiments, module ablation studies, hyperparameter analyses, and specialized robustness tests. The overall detection performance of the model on the Total-Text and ICDAR 2015 standard datasets significantly outperformed recent state-of-the-art algorithms, with outstanding detection stability and adaptability being demonstrated in challenging scenarios involving small-scale characters, narrow characters, curved texts, and severe interference. This research establishes a high-accuracy technical framework for English scene character detection, effectively addressing the limitations of conventional approaches in detecting characters with extreme scale variations and irregular shapes. A novel technical paradigm is provided for multi-scale modeling and fine-grained boundary optimization in natural-scene text detection, while reliable technical support is offered for practical engineering optical character recognition tasks such as intelligent inspection of industrial equipment, scene perception for autonomous driving, and mobile intelligent translation, demonstrating substantial academic research value and significant application potential.

REFERENCES

[1] Long, S.B., He, X., Yao, C. (2021). Scene text detection and recognition: The deep learning era. International Journal of Computer Vision, 129(1): 161-184. https://doi.org/10.1007/s11263-020-01369-0

[2] Naiemi, F., Ghods, V., Khalesi, H. (2022). Scene text detection and recognition: A survey. Multimedia Tools and Applications, 81: 20255-20290. https://doi.org/10.1007/s11042-022-12693-7

[3] Rainarli, E., Suprapto, Wahyono. (2021). A decade: Review of scene text detection methods. Computer Science Review, 42: 100434. https://doi.org/10.1016/j.cosrev.2021.100434

[4] Guan, T.K., Gu, C.C., Lu, C.S., Tu, J.Z., Feng, Q., Wu, K.J. (2022). Industrial scene text detection with refined feature-attentive network. IEEE Transactions on Circuits and Systems for Video Technology, 32(9): 6073-6085. https://doi.org/10.1109/TCSVT.2022.3156390

[5] Cai, Y.Q., Liu, C., Cheng, P.R., Du, D.W., Zhang, L.B., Wang, W.Q. (2021). Scale-residual learning network for scene text detection. IEEE Transactions on Circuits and Systems for Video Technology, 31(7): 2725-2738. https://doi.org/10.1109/TCSVT.2020.3029167

[6] Cai, Y., Liu, Y.L., Shen, C.H., Jin, L.W., Li, Y.D., Ergu, D. (2022). Arbitrarily shaped scene text detection with dynamic convolution. Pattern Recognition, 127: 108608. https://doi.org/10.1016/j.patcog.2022.108608

[7] Liao, M.H., Zou, Z.S., Wan, Z.Y., Yao, C., Bai, X.

(2023). Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1): 919-931. https://doi.org/10.1109/TPAMI.2022.3155612

[8] Liu, Z.D., Zhou, W.G., Li, H.Q. (2021). MFECN: Multi-level feature enhanced cumulative network for scene text detection. ACM Transactions on Multimedia Computing, Communications, and Applications, 17(3): 1-22. https://doi.org/10.1145/3440087

[9] Zhu, Y.X., Du, J. (2021). TextMountain: Accurate scene text detection via instance segmentation. Pattern Recognition, 110: 107336. https://doi.org/10.1016/j.patcog.2020.107336

[10] Cheng, P.R., Cai, Y.Q., Wang, W.Q. (2020). A direct regression scene text detector with position-sensitive segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 30(11): 4171-4181. https://doi.org/10.1109/TCSVT.2019.2947475

[11] Du, B., Ye, J., Zhang, J., Liu, J.H., Tao, D. (2022). I3CL: Intra- and inter-instance collaborative learning for arbitrary-shaped scene text detection. International Journal of Computer Vision, 130: 1961-1977. https://doi.org/10.1007/s11263-022-01616-6

[12] Wu, Y.R., Zhang, W., Wan, S.H. (2022). CE-text: A context-aware and embedded text detector in natural scene images. Pattern Recognition Letters, 159: 77-83. https://doi.org/10.1016/j.patrec.2022.05.004

[13] Zhong, D.J., Lyu, S., Shivakumara, P., Pal, U., Lu, Y. (2022). Text proposals with location-awareness-attention network for arbitrarily shaped scene text detection and recognition. Expert Systems with Applications, 205: 117564. https://doi.org/10.1016/j.eswa.2022.117564

[14] Qin, S.X., Chen, L. (2022). Arbitrary-shaped scene text detection with keypoint-based shape representation. International Journal on Document Analysis and Recognition, 25: 115-127. https://doi.org/10.1007/s10032-022-00396-6

[15] Cao, M., Zhang, C., Yang, D.M., Zou, Y.X. (2022). All you need is a second look: Towards arbitrary-shaped text detection. IEEE Transactions on Circuits and Systems for Video Technology, 32(2): 758-767. https://doi.org/10.1109/TCSVT.2021.3068133

[16] Ma, H.Y., Lu, N.N., Mei, J.J., Guan, T., Zhang, Y., Geng, X. (2023). Label distribution learning for scene text detection. Frontiers of Computer Science, 17: 176339. https://doi.org/10.1007/s11704-022-1446-5

[17] Li, X.G., Yao, X.C., Liu, Y. (2024). Combining Swin Transformer and attention-weighted fusion for scene text detection. Neural Processing Letters, 56: 52. https://doi.org/10.1007/s11063-024-11501-7

[18] Xu, J.Y., Lin, A.L., Li, J.X., Lu, G.M. (2024). Text position-aware pixel aggregation network with adaptive Gaussian threshold: Detecting text in the wild. IEEE Transactions on Circuits and Systems for Video Technology, 34(1): 286-298. https://doi.org/10.1109/TCSVT.2023.3285096

[19] Fu, Z.L., Xie, H.T., Fang, S.C., Wang, Y.X., Xing, M.T., Zhang, Y.D. (2023). Learning pixel affinity pyramid for arbitrary-shaped text detection. ACM Transactions on Multimedia Computing, Communications, and Applications, 19(1s): 1-24. https://doi.org/10.1145/3524617

[20] Cheng, Q., Wang, G.D. (2021). Shape awareness and structure-preserving network for arbitrary shape text detection. Multimedia Tools and Applications, 80: 10761-10775. https://doi.org/10.1007/s11042-020-10039-9

[21] Fu, H.T., Liu, W.Z., Liu, Y.L., Cao, Z.G., Lu, H. (2023). SIERRA: A robust bilateral feature upsampler for dense prediction. Computer Vision and Image Understanding, 235: 103762. https://doi.org/10.1016/j.cviu.2023.103762

[22] Lu, H., Liu, W.Z., Fu, H.T., Cao, Z.G. (2025). FADE: A task-agnostic upsampling operator for encoder–decoder architectures. International Journal of Computer Vision, 133: 151-172. https://doi.org/10.1007/s11263-024-02191-8

[23] Cheng, P.R., Zhao, Y.Z., Wang, W.Q. (2023). Detect arbitrary-shaped text via adaptive thresholding and localization quality estimation. IEEE Transactions on Circuits and Systems for Video Technology, 33(12): 7480-7490. https://doi.org/10.1109/TCSVT.2023.3274673

[24] Zhang, S.X., Yang, C., Zhu, X.B., Yin, X.C. (2024). Arbitrary shape text detection via boundary transformer. IEEE Transactions on Multimedia, 26: 1747-1760. https://doi.org/10.1109/TMM.2023.3286657