ED-YOLO: An Algorithm for Detecting Questions and Visual Elements in Educational Documents

1st Tao Lin
Faculty of Intelligent Technology
Shanghai Institute of Technology
Shanghai, China
lintao@sit.edu.cn
ORCID: 0000-0002-3440-3874

2nd Tao Wu
Faculty of Intelligent Technology
Shanghai Institute of Technology
Shanghai, China
246142161@mail.sit.edu.cn
ORCID: 0009-0008-4510-3528

***Abstract*—With the rapid development of digital education, the digital transformation of paper-based teaching resources has become a core task in the construction of educational informatization. Accurate identification and segmentation of questions, images, and tables in educational documents are essential for realizing the structured and intelligent utilization of resources. Existing methods for detecting questions in complex layouts, such as multi-column layouts, mixed text-image formats, and handwritten annotations, suffer from insufficient positioning accuracy and limited ability to suppress background interference. Additionally, the detection of diverse question types, such as multiple-choice, fill-in-the-blank, and solution questions, presents further challenges. To address these issues, this paper proposes an enhanced segmentation algorithm, ED-YOLO (Educational Document-YOLO), which improves upon YOLOv11 for detecting both questions and visual elements in educational documents. The algorithm introduces a Task-Aligned Dynamic Detection Head (TADDH) to enhance positioning and classification performance, optimizes feature fusion using a Context Guide Fusion Module (CGFM), and integrates DCNv4 with C3K2 blocks (DCNv4_C3k2) in both backbone and neck for adaptive feature extraction. Experimental results demonstrate that ED-YOLO achieves 98.32% mAP50 and 95.25% precision on a multi-subject dataset, showing significant robustness in complex document layouts. The proposed method provides valuable technical support for the efficient digital transformation of educational resources, enhancing both question detection and the identification of associated visual content such as images and tables.**

***Index Terms*—Educational Documents, Question and Visual Element Detection, YOLOv11, Feature Fusion, Task Alignment**

I. Introduction

The rapid digitization of educational resources has generated a large volume of document images, including examination papers, textbooks, and handwritten assignments [1]. Automatically detecting questions and their associated visual elements, such as images and tables, is a fundamental step toward structured document understanding and intelligent educational applications [2], [3]. Accurate localization of these elements enables downstream tasks such as question parsing, answer analysis, automatic grading, and educational resource indexing [4], [5].

Compared with natural scene images, educational documents exhibit distinctive characteristics that make detection particularly challenging, including complex multi-column layouts, dense text regions, diverse question formats, and handwritten annotations [6], [7]. Recent advances in deep learning, especially the YOLO family [8], [9], [11]–[13], have significantly improved document analysis performance. However, directly applying existing YOLO-based models presents notable limitations: inaccurate bounding box regression in dense layouts, misalignment between classification confidence and localization quality [16], [17], and limited flexibility in modeling irregular geometric structures [23], [24].

To overcome these limitations, this paper proposes ED-YOLO (Educational Document-YOLO), an enhanced detection framework for questions and visual elements in educational documents. Built upon YOLOv11 [13], ED-YOLO introduces several key improvements. First, a hybrid module combining Deformable Convolution v4 with C3K2 (DCNv4_C3k2) is deployed in both the backbone and neck networks to enhance adaptive feature extraction for irregular and complex document structures [26]. Second, a Context Guide Fusion Module (CGFM) is proposed to optimize multi-scale feature fusion by incorporating contextual guidance, enabling better discrimination between foreground elements and background noise [19], [20]. Third, a Task-Aligned Dynamic Detection Head (TADDH) is designed to strengthen the consistency between classification and localization, improving detection accuracy in dense layouts [18].

Extensive experiments conducted on a multi-subject educational document dataset demonstrate the effectiveness of the proposed method. ED-YOLO achieves superior performance in both precision and robustness under complex layouts, outperforming baseline YOLOv11 and other competitive approaches. The results indicate that ED-YOLO is well suited for practical educational document understanding tasks, providing reliable support for the digital transformation of educational resources.

To facilitate reproducibility and future research, upon acceptance of this paper, we will publicly release the core code snippets of the proposed model as well as the trained weights of all comparative models. The data and code are available on Zenodo at https://doi.org/10.5281/zenodo.18252348.

II. Related Work

A. Educational Document Analysis

Educational document analysis aims to extract structured information from teaching materials such as examination papers, textbooks, and homework sheets [1]. Early studies mainly relied on rule-based layout analysis and traditional image processing techniques, including connected component analysis, projection profiles, and heuristic segmentation. Although these approaches perform well under constrained layouts, they lack robustness when facing complex document structures, diverse question formats, and handwritten content commonly found in real-world educational documents.

With the development of deep learning, convolutional neural networks have been widely applied to document understanding tasks, including document layout analysis, text block segmentation, and question detection [29], [30]. Some methods formulate question detection as a layout segmentation problem, while others treat it as an object detection task. The emergence of transformer-based models has further advanced document understanding, with approaches such as LayoutLMv3 [5] and DiT [4] achieving state-of-the-art results on various benchmarks. DocLayout-YOLO [28] recently proposed a specialized YOLO-based approach for document layout analysis with significant improvements. VSR [31] introduced a unified framework combining vision, semantics and relations for layout analysis. DocFormer [32] presented an end-to-end transformer architecture for document understanding. Despite notable progress, educational documents remain challenging due to dense text distributions, subtle inter-class differences between question types, and strong background interference. These challenges motivate the adoption of more powerful detection frameworks with improved feature representation and task alignment.

B. Object Detection in Document Images

Object detection techniques have been extensively studied in both natural images and document images [34]. Two-stage detectors, such as Faster R-CNN [35], provide high detection accuracy but suffer from high computational cost, limiting their applicability in large-scale document processing systems. In contrast, one-stage detectors offer a favorable trade-off between speed and accuracy and have been increasingly adopted in document analysis tasks.

The YOLO series has evolved rapidly in recent years, with improvements in backbone design, feature pyramids, and detection heads. YOLOv5 introduced a series of engineering optimizations that significantly improved training stability and inference speed [9]. YOLOv8 further enhanced the architecture with a decoupled head design and anchor-free detection paradigm [10]. More recently, YOLOv9 proposed Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN) to address information bottleneck issues [11]. YOLOv10 introduced NMS-free training with consistent dual assignments, achieving state-of-the-art performance with reduced computational overhead [12]. YOLOv11 continued this evolution with improved spatial pyramid pooling and enhanced feature aggregation. YOLO26, the latest iteration, introduced key architectural enhancements for real-time object detection [14]. YOLO-World extended YOLO to open-vocabulary detection scenarios through vision-language modeling [15]. These models achieve strong performance in general object detection scenarios. However, directly applying generic YOLO models to document images is suboptimal. Document images contain densely packed objects with fine-grained boundaries, where inaccurate localization can significantly impact downstream tasks [6], [28]. In addition, conventional detection heads typically optimize classification and localization independently, which may lead to misaligned confidence scores in dense document layouts.

Several studies have attempted to adapt object detectors for document-specific scenarios by incorporating layout-aware features or multi-scale representations [2], [3]. Nevertheless, accurately detecting questions and associated visual elements in educational documents remains an open problem, particularly under complex layouts and noisy backgrounds [7].

C. Task Alignment in Detection Heads

Recent research has highlighted the importance of aligning classification confidence with localization quality in object detection [33]. Traditional detection heads often produce high classification scores for poorly localized bounding boxes, which degrades detection reliability. To address this issue, task-aligned learning strategies have been proposed to jointly optimize classification and regression tasks by sharing features or dynamically weighting predictions.

TOOD [16] introduced a task-aligned one-stage object detector that explicitly aligns the two tasks through a task alignment learning mechanism. YOLOX [17] proposed a decoupled head structure that separates classification and localization branches while maintaining efficient inference. DyHead [18] presented a dynamic head module that unifies scale-awareness, spatial-awareness, and task-awareness through attention mechanisms. Dynamic detection heads further enhance flexibility by adapting parameters or feature responses based on input characteristics. Such designs have shown effectiveness in dense object detection scenarios, where object sizes and spatial distributions vary significantly. However, most existing task-aligned or dynamic heads are designed for natural images and do not fully consider the unique challenges of document images, such as dense text blocks and subtle structural variations. This gap motivates the design of a task-aligned detection head specifically tailored for educational document analysis.

D. Feature Fusion and Deformable Convolution

Multi-scale feature fusion is a key component of modern object detection frameworks, enabling detectors to capture both global context and fine-grained details [19]. Feature pyramid networks (FPN) and their variants, such as PANet [20] and BiFPN [21], are widely used to aggregate features from different stages of the backbone. ASFF [22] proposed

adaptively spatial feature fusion to learn the optimal combination weights for different feature levels. However, simple feature aggregation may be insufficient for document images, where contextual relationships between text blocks, questions, and visual elements are crucial for accurate detection.

Deformable convolution has been proposed to enhance the modeling capacity of CNNs by introducing learnable sampling offsets, allowing convolution kernels to adapt to irregular object shapes [23]. DCNv2 [24] improved upon the original design by adding modulation mechanisms to better control the spatial sampling. Recent versions of deformable convolution, including DCNv3 [25] and DCNv4 [26], further improve stability and efficiency through optimized memory access patterns and aggregation strategies. In document images, deformable convolution is particularly beneficial for modeling non-uniform text layouts and distorted visual elements such as skewed tables or diagrams [27]. Integrating deformable convolution with effective feature fusion strategies can significantly improve feature representation in complex document scenarios.

III. Method

A. Overall Framework

This paper proposes ED-YOLO, an enhanced neural network architecture for detecting questions and visual elements in educational documents. The proposed method is built upon YOLOv11 and is specifically designed to address three key challenges in educational document analysis: inaccurate localization of question boundaries in dense layouts, insufficient suppression of background interference caused by mixed text–image content and handwritten annotations, and limited feature representation capability for irregular geometric structures.

As illustrated in Fig. 1, ED-YOLO introduces three progressive improvements over the baseline YOLOv11 framework. First, a DCNv4_C3k2 module, which combines Deformable Convolution v4 with C3K2 blocks, is integrated into both the backbone and neck networks to enhance adaptive feature extraction for non-rigid and distorted document structures. Second, a Context-Guided Fusion Module (CGFM) is embedded into the neck network to strengthen semantic interaction across multi-scale features. Third, a Task-Aligned Dynamic Detection Head (TADDH) is designed to replace the conventional detection head, explicitly aligning classification confidence with localization quality.

These components form a unified pipeline of feature enhancement, contextual fusion, and detection refinement, enabling robust detection performance in complex educational document scenarios.

B. Task-Aligned Dynamic Detection Head (TADDH)

In complex educational documents, dense layouts and background noise often lead to misalignment between classification confidence and localization accuracy. To address this issue, we propose the Task-Aligned Dynamic Detection Head (TADDH), whose structure is shown in Fig. 2. TADDH explicitly decouples and dynamically aligns classification and localization tasks to improve detection robustness.

Let the multi-scale feature maps extracted from the neck network be denoted as

$$\mathcal{F} = \{f_1, f_2, \ldots, f_N\}, \quad f_i \in \mathbb{R}^{C\times H\times W}. \tag{1}$$

Each feature map is first processed by a convolution–group normalization layer to unify channel dimensions and spatial resolution. The aligned features are concatenated to incorporate contextual information:

$$G_{\text{concat}} = \text{Concat}\left(\text{Conv}(f_1), \text{Conv}(f_2), \ldots, \text{Conv}(f_N)\right). \tag{2}$$

The concatenated feature representation is encoded by a shared convolutional layer and decomposed into task-specific feature spaces:

$$F_{\text{cls}} = \mathcal{H}_{\text{cls}}(G), \quad F_{\text{reg}} = \mathcal{H}_{\text{reg}}(G), \tag{3}$$

where $\mathcal{H}_{\text{cls}}(\cdot)$ and $\mathcal{H}_{\text{reg}}(\cdot)$ denote task-aware transformation functions implemented by lightweight convolutional blocks.

To enhance geometric modeling capability, dynamic deformable convolution is applied to the localization branch:

$$F'_{\text{reg}} = \text{DyDCNv2}(F_{\text{reg}}; \Theta_{\text{offset}}), \tag{4}$$

where the offset Θ_{offset} is dynamically generated from F_{reg} via a 3×3 convolution layer.

To explicitly align localization quality with classification confidence, a task alignment loss is introduced:

$$\mathcal{L}_{\text{align}} = \sum_{i=1}^{N} \text{IoU}(\hat{b}_i, b_i) \cdot \text{CE}(p_i, c_i), \tag{5}$$

where $\hat{b}_i$ and b_i denote the predicted and ground-truth bounding boxes, p_i is the predicted classification probability, and c_i is the target category.

C. Context-Guided Fusion Module (CGFM)

Educational documents often exhibit spatial fragmentation caused by multi-column layouts and semantic interference from mixed text–image content. To preserve semantic integrity and enhance contextual consistency, we introduce the Context-Guided Fusion Module (CGFM), as illustrated in Fig. 3.

CGFM takes shallow spatial features $x_0 \in \mathbb{R}^{C\times H\times W}$ and deep semantic features $x_1 \in \mathbb{R}^{C\times H\times W}$ as input. After channel alignment, the features are concatenated:

$$x_{\text{concat}} = [x_0, x_1] \in \mathbb{R}^{2C\times H\times W}. \tag{6}$$

A squeeze-and-excitation (SE) attention mechanism is employed to generate channel-wise importance weights:

$$w = \sigma\left(W_2 \cdot \delta\left(W_1 \cdot \text{GAP}(x_{\text{concat}})\right)\right), \tag{7}$$

where $\text{GAP}(\cdot)$ denotes global average pooling, W_1 and W_2 are learnable weight matrices, $\delta(\cdot)$ is the ReLU activation function, and $\sigma(\cdot)$ is the sigmoid function.

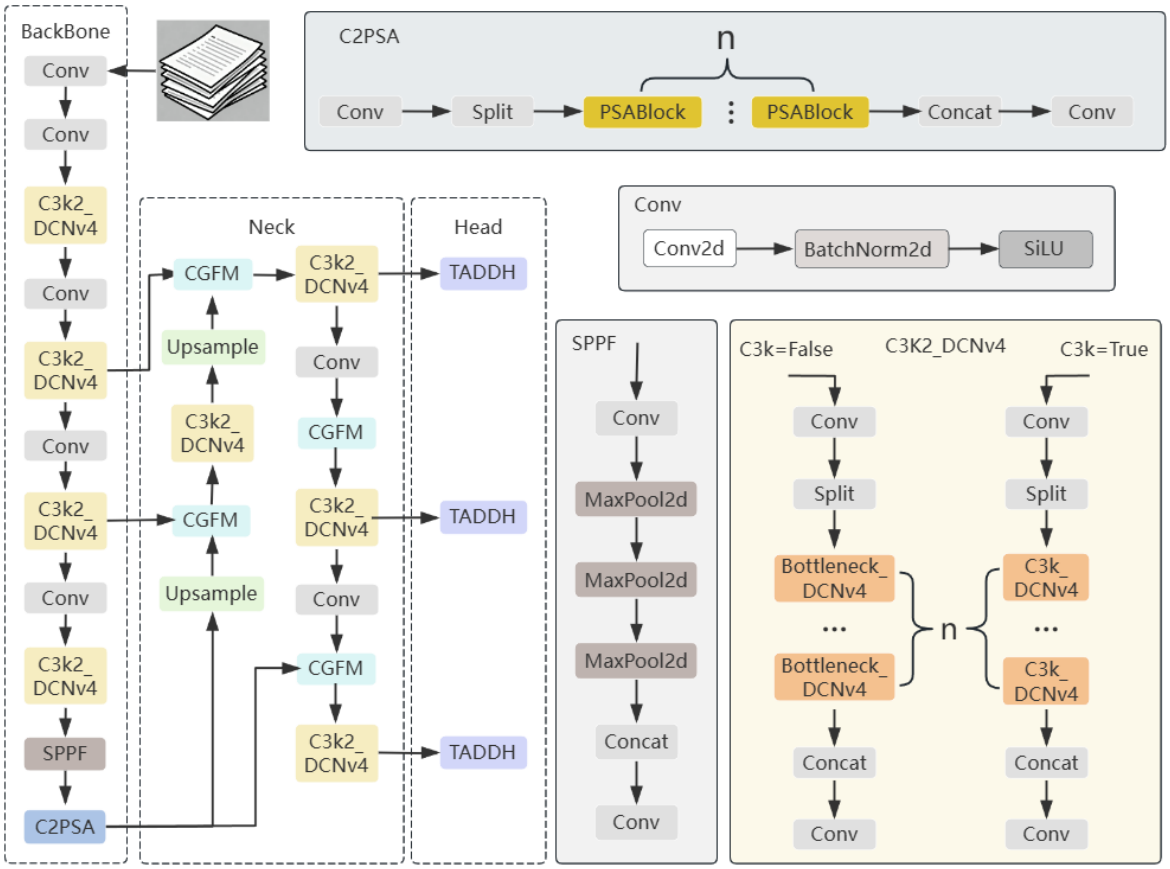


Fig. 1: Overall architecture of ED-YOLO. The framework integrates DCNv4_C3k2 in backbone and neck, CGFM for feature fusion, and TADDH for task-aligned detection.

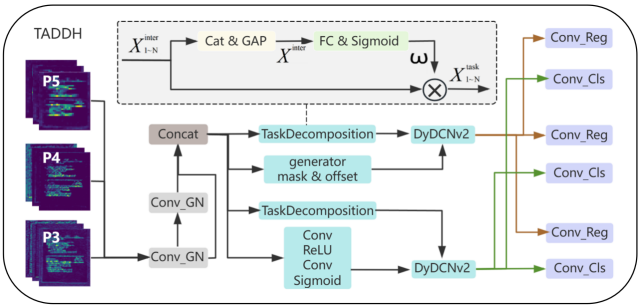


Fig. 2: Architecture of the Task-Aligned Dynamic Detection Head (TADDH). The design decouples classification and localization tasks while enforcing alignment through dynamic feature learning.

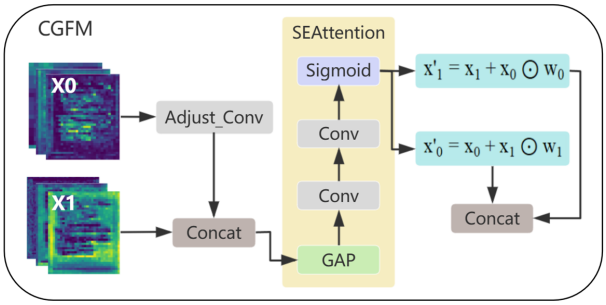


Fig. 3: Structure of the Context-Guided Fusion Module (CGFM), which enables bidirectional interaction between shallow spatial features and deep semantic features.

The weights are split into w_0 and w_1 and applied to the corresponding feature maps. Bidirectional residual fusion is then performed:

$$x_0' = x_0 + x_1 \odot w_1, \quad x_1' = x_1 + x_0 \odot w_0. \tag{8}$$

The final fused representation integrates fine-grained spatial details with high-level semantic context, improving boundary sensitivity and contextual coherence in question detection.

D. DCNv4_C3k2: Deformable Convolution Enhanced Feature Extraction

To enhance feature extraction under handwritten deformations and layout distortions, ED-YOLO incorporates a hybrid module that combines Deformable Convolution v4 (DCNv4) with C3K2 blocks, referred to as DCNv4_C3k2. Unlike conventional integration where DCNv4 is applied solely in the backbone, the proposed DCNv4_C3k2 module is strategically deployed in both the backbone and neck networks to enable

TABLE I: Environment Configuration

Environment	Configuration 1 (Local Application)	Configuration 2 (Server Training)
Operating System	Win10 64-bit	Ubuntu 22.04
Graphics Card	Nvidia RTX 4060 Ti (8GB)	Nvidia RTX 3090 (24GB)
Python	3.8.20	3.8.16
PyTorch	2.4.1	2.3.1

multi-level adaptive feature learning.

The C3K2 module is an efficient feature extraction block inherited from YOLOv11, which employs a cross-stage partial (CSP) structure with two convolutional branches to balance feature extraction capability and computational efficiency. By integrating DCNv4 into the C3K2 structure, the resulting DCNv4_C3k2 module gains the ability to dynamically adjust convolutional sampling locations through learnable offsets and modulation coefficients, while maintaining the efficient gradient flow characteristics of the CSP architecture.

The DCNv4 operation at position p can be formulated as:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \tag{9}$$

where K is the number of sampling points, w_k is the learnable weight for the k-th sampling location, p_k denotes the regular grid sampling offset, Δp_k is the learned adaptive spatial offset, and Δm_k is the modulation coefficient that controls the contribution of each sampling point.

In the backbone network, DCNv4_C3k2 enhances low-level feature extraction by adapting to irregular geometric structures such as distorted tables and non-uniform text blocks. In the neck network, it further refines multi-scale features by modeling complex spatial relationships between questions and visual elements.

By deploying DCNv4_C3k2 across both backbone and neck networks, ED-YOLO achieves comprehensive geometric adaptability throughout the entire feature extraction and fusion pipeline, complementing CGFM and TADDH to jointly enhance detection accuracy and robustness in complex educational document scenarios.

IV. Experiments and Results

A. Experimental Environment and Configuration

The experiments are conducted in two sets of experimental environments for model training and actual detection applications, with the configurations shown in Table I. All models are implemented using the PyTorch framework to ensure fair performance evaluation across different algorithms.

B. Dataset Description

The experiments utilize a self-constructed educational question dataset, which includes images of paper-based teaching resources across ten subjects: Chinese, Mathematics, English, Physics, Chemistry, Biology, History, Politics, Geography, and Science. The dataset contains 7,622 images, covering multiple educational stages (primary school, junior high school, senior high school) and various question types such as multiple-choice questions, fill-in-the-blank questions, solution questions, and discussion questions.

The dataset is divided into training (5,335 images), validation (1,524 images), and test (763 images) sets with a 7:2:1 split. All images are annotated by experienced teachers to ensure high-quality data reliability.

Notably, the dataset includes the following challenging scenarios that are commonly encountered in real-world educational documents:

- 2,156 samples with multi-column layouts (e.g., textbook columns, double-column test papers)
- 1,873 samples with handwritten annotations (e.g., teacher corrections and student notes)
- 1,642 samples with mixed text–image layouts (e.g., formulas, charts, and illustrations interspersed with text)
- 1,951 samples of dense question types (e.g., closely arranged multiple-choice options and fill-in-the-blank questions)

These categories are not mutually exclusive; many documents exhibit multiple characteristics simultaneously, further increasing task complexity and reflecting real-world scenarios.

C. Evaluation Metrics

To comprehensively evaluate the detection performance of ED-YOLO, we adopt the following evaluation metrics:

- **Precision (P)**: The proportion of actual positive samples among those predicted as positive.
- **Recall (R)**: The proportion of actual positive samples that are correctly predicted.
- **F1 Score**: The harmonic mean of precision and recall.
- **Mean Average Precision (mAP)**: The average of the average precision (AP) values across all question types. mAP50 represents the average precision at an Intersection over Union (IoU) threshold of 0.5, and mAP50:95 represents the average precision across IoU thresholds from 0.5 to 0.95.
- **Model Parameters**: The total number of parameters, indicating model complexity.
- **Model Size**: The storage size required by the model, important for deployment on educational devices.

The formulas for each metric are as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{10}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{11}$$

$$F1\text{-Score} = \frac{2 \times P \times R}{P + R} \tag{12}$$

$$AP = \int_0^1 P(t)\, dt \tag{13}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{14}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

D. Object Detection Comparison Experiment

To evaluate the effectiveness of ED-YOLO, a comparison experiment is conducted with current state-of-the-art object detection models. The experiment involves training models on the self-constructed dataset for 200 epochs and evaluating their performance on the test set. The results are shown in Table II. ED-YOLO outperforms the other models across all evaluation metrics, with the highest mAP50 and significant improvements in recall and precision, particularly for detecting dense questions and scenarios with handwritten annotations.

E. Ablation Study of Improved Modules

To assess the contribution of each module to the performance of ED-YOLO, we conduct an ablation study by sequentially adding the DCNv4_C3k2, CGFM, and TADDH modules and evaluating their impact on performance. The results, shown in Table III, demonstrate that each improvement contributes positively, with the combination of all three modules achieving the highest performance.

F. Parameter Analysis and Layer-Wise Visualization

To investigate the influence of regularization parameter λ and analyze module contributions, we conducted parameter tuning experiments and layer-wise visualization. As shown in Fig. 4, appropriate λ values improve generalization, while excessive values cause over-regularization. Fig. 5 visualizes feature sampling points using Grad-CAM, highlighting how DCNv4_C3k2, CGFM, and TADDH contribute to accurate localization and feature extraction.

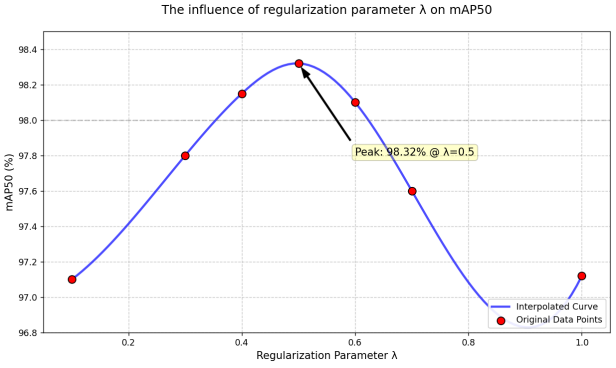


Fig. 4: Influence of regularization parameter λ on mAP50.

G. Subject-Specific Experiments

To evaluate ED-YOLO's adaptability to documents from different subjects, specific tests were conducted on subsets of the test set for each subject. Table IV shows the mAP50 results, with ED-YOLO outperforming YOLOv11-n across all subjects, especially in more complex subjects such as Chemistry and Biology.

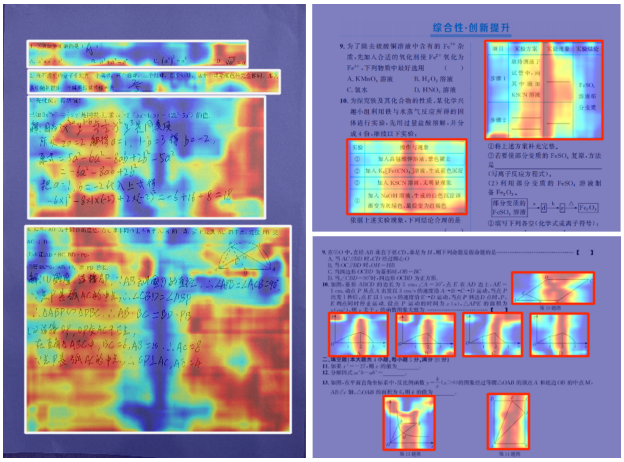


Fig. 5: Visualization of feature sampling points using Grad-CAM.

H. Experiments on Complex Scenarios

ED-YOLO's robustness in handling complex scenarios such as handwritten annotations, multi-column layouts, and mixed text-image layouts is demonstrated in Table V, showing significant improvements across all challenging cases.

I. Qualitative Results

To qualitatively evaluate the detection capability of ED-YOLO, representative visualization results on different subjects and document forms are presented in Fig. 6. The results cover both scanned and photographed educational documents, as well as single-column and double-column layouts, demonstrating the robustness and generalization ability of the proposed method.

As shown in Fig. 6(a), ED-YOLO accurately detects multiple question types, including multiple-choice, fill-in-the-blank, and long-answer questions, on a printed mathematics examination paper. The page contains dense function plots and a grid diagram, and the proposed method successfully distinguishes question regions from complex graphical elements. Fig. 6(b) presents detection results on a photographed mathematics problem sheet. Despite perspective distortion and illumination variation introduced during image capture, ED-YOLO maintains stable detection performance, indicating its adaptability to real-world photographed documents.

Fig. 6(c) illustrates detection results on a printed physics examination paper containing multiple-choice questions, function curves, experimental diagrams, and schematic illustrations. ED-YOLO effectively localizes both questions and diverse visual elements, such as physical phenomenon images and experimental setups. Fig. 6(d) shows results on a chemistry textbook page with a double-column layout, including tables and chemical reaction diagrams. The proposed method accurately handles cross-column content and complex visual

TABLE II: Performance Comparison of ED-YOLO with Other Object Detection Models

Algorithm	Precision (P)	Recall (R)	F1-Score	mAP50	mAP50:95	Parameters	Model size
YOLOv12-n	93.11%	86.79%	0.8983	95.31%	79.60%	2.50M	5.2MB
YOLOv11-n	93.40%	92.50%	0.9295	96.97%	81.47%	2.58M	5.2MB
YOLOv10-n	89.74%	90.24%	0.8999	95.22%	78.96%	2.26M	5.5MB
YOLOv9-t	91.67%	93.39%	0.9253	97.0%	82.14%	1.73M	4.0MB
YOLOv8-n	91.18%	94.28%	0.9270	96.55%	82.82%	2.68M	5.4MB
YOLOv6-n	94.76%	93.0%	0.9387	97.23%	82.72%	4.15M	8.2MB
YOLOv5-n	94.02%	92.70%	0.9336	97.10%	81.59%	2.18M	4.4MB
Ours (ED-YOLO)	95.25%	94.58%	0.9491	98.32%	86.75%	2.27M	4.6MB

TABLE III: Ablation Experiment Results

Modules	Precision (P)	Recall (R)	F1-Score	mAP50	mAP50:95	Parameters	Model size
Baseline (YOLOv11)	93.40%	92.50%	0.9295	96.97%	81.47%	2,582,347	5.2MB
+ DCNv4_C3k2	94.67%	93.59%	0.9313	97.19%	82.24%	2,517,515	5.1MB
+ DCNv4_C3k2 + CGFM	94.83%	93.72%	0.9367	97.82%	86.29%	2,673,675	5.4MB
+ DCNv4_C3k2 + CGFM + TADDH	95.25%	94.58%	0.9491	98.32%	86.75%	2,275,148	4.6MB

TABLE IV: Detection Results of Documents in Different Subjects

Subject	YOLOv11-n	ED-YOLO	Improvement
Chinese	96.25%	98.73%	+2.48%
Mathematics	93.17%	96.85%	+3.68%
English	95.82%	98.41%	+2.59%
Physics	92.64%	96.52%	+3.88%
Chemistry	91.85%	96.27%	+4.42%
Biology	93.56%	97.13%	+3.57%
History	96.43%	98.65%	+2.22%
Politics	95.97%	98.32%	+2.35%
Geography	92.98%	96.74%	+3.76%
Science	92.31%	96.48%	+4.17%

TABLE V: Detection Results in Complex Scenarios

Scenario	YOLOv11-n	ED-YOLO	Improvement
Handwritten Annotations	94.12%	97.85%	+3.73%
Multi-Column Layout	93.56%	97.23%	+3.67%
Mixed Text–Image Layout	90.83%	96.47%	+5.64%

structures, highlighting its robustness in dense and heterogeneous document layouts.

Overall, the visualization results demonstrate that ED-YOLO can reliably detect questions and associated visual elements across different subjects, document layouts, and acquisition conditions, significantly reducing missed detections and false detections in complex educational document scenarios.

V. Conclusion

This paper proposes ED-YOLO, a task-aligned and context-aware neural network for detecting questions and visual elements in educational documents. Aiming at the challenges posed by complex document layouts, dense question distributions, mixed text–image content, and handwritten annotations, ED-YOLO enhances the YOLOv11 framework through three complementary improvements: (1) DCNv4_C3k2 modules strategically deployed in both backbone and neck networks to enable multi-level geometric adaptability for irregular document structures; (2) a Context-Guided Fusion Module (CGFM) for enhanced multi-scale semantic interaction; and (3) a Task-Aligned Dynamic Detection Head (TADDH) for improving the consistency between classification confidence and localization accuracy.

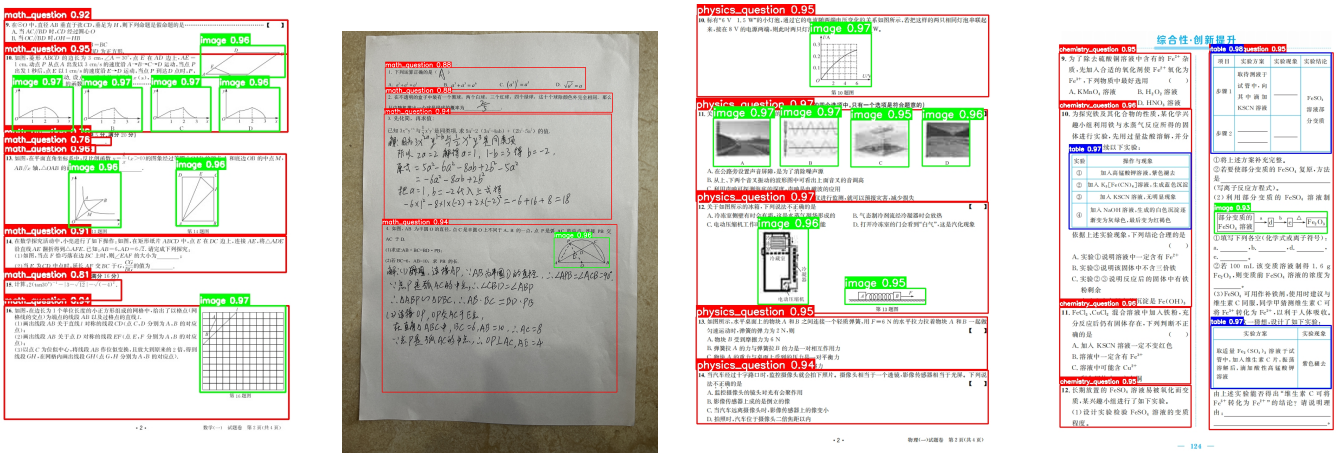
Extensive experiments conducted on a self-constructed multi-subject educational document dataset demonstrate that ED-YOLO achieves superior performance compared with state-of-the-art object detection methods, reaching 98.32% mAP50 and 95.25% precision. The proposed method consistently improves detection accuracy for both questions and associated visual elements, such as images and tables, while maintaining a lightweight model size suitable for practical deployment. Further analysis on subject-specific subsets and complex scenarios confirms the robustness and generalization capability of ED-YOLO across different educational disciplines and document layouts.

In future work, we plan to extend ED-YOLO toward finer-grained structural understanding of educational documents, such as hierarchical question parsing and semantic relationship modeling between questions and visual elements. In addition, integrating the proposed method with downstream educational applications, including automatic question bank construction and intelligent assessment systems, will be explored to further promote the intelligent utilization of educational resources.

References

[1] L. Cui et al., "Document AI: Benchmarks, Models and Applications," arXiv:2111.08609, 2021.

[2] X. Zhong, J. Tang, and A. J. Yepes, "PubLayNet: Largest Dataset Ever for Document Layout Analysis," in *Proc. ICDAR*, 2019, pp. 1015–1022.



(a) Printed mathematics examination paper with dense function plots and grid diagrams.

(b) Photographed mathematics problems with mixed question types and natural imaging distortions.

(c) Printed physics examination paper containing function curves, experimental diagrams, and schematic illustrations.

(d) Double-column chemistry textbook page with tables and chemical reaction diagrams.

Fig. 6: Visualization results of ED-YOLO on different educational document scenarios. The proposed method accurately detects questions and visual elements across subjects, layouts, and acquisition conditions.

[3] B. Pfitzmann et al., "DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis," in *Proc. ACM SIGKDD*, 2022, pp. 3743–3751.

[4] J. Li et al., "DiT: Self-supervised Pre-training for Document Image Transformer," in *Proc. ACM MM*, 2022, pp. 3530–3539.

[5] Y. Huang et al., "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking," in *Proc. ACM MM*, 2022, pp. 4083–4091.

[6] Y. Zhang et al., "Reading Order Matters: Information Extraction from Visually-rich Documents by Token Path Prediction," in *Proc. EMNLP*, 2023, pp. 12336–12348.

[7] M. Mathew et al., "DocVQA: A Dataset for VQA on Document Images," in *Proc. WACV*, 2021, pp. 2200–2209.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. CVPR*, 2016, pp. 779–788.

[9] G. Jocher, "YOLOv5: A State-of-the-Art Real-Time Object Detection System," 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[10] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8: Next-Generation Object Detection," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[11] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," arXiv:2402.13616, 2024.

[12] A. Wang et al., "YOLOv10: Real-Time End-to-End Object Detection," arXiv:2405.14458, 2024.

[13] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," arXiv:2410.17725, 2024.

[14] R. Sapkota and M. Karkee, *Ultralytics YOLO Evolution: An Overview of YOLO26, YOLO11, YOLOv8 and YOLOv5 Object Detectors for Computer Vision and Pattern Recognition*, arXiv preprint arXiv:2510.09653, 2025.

[15] T. Cheng et al., "YOLO-World: Real-Time Open-Vocabulary Object Detection," in *Proc. CVPR*, 2024, pp. 16901–16911.

[16] C. Feng et al., "TOOD: Task-aligned One-stage Object Detection," in *Proc. ICCV*, 2021, pp. 3490–3499.

[17] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," arXiv:2107.08430, 2021.

[18] X. Dai et al., "Dynamic Head: Unifying Object Detection Heads with Attentions," in *Proc. CVPR*, 2021, pp. 7373–7382.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proc. CVPR*, 2017, pp. 936–944.

[20] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," in *Proc. CVPR*, 2018, pp. 8759–8768.

[21] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in *Proc. CVPR*, 2020, pp. 10781–10790.

[22] S. Liu, D. Huang, and Y. Wang, "Learning Spatial Fusion for Single-Shot Object Detection," arXiv:1911.09516, 2019.

[23] J. Dai et al., "Deformable Convolutional Networks," in *Proc. ICCV*, 2017, pp. 764–773.

[24] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More Deformable, Better Results," in *Proc. CVPR*, 2019, pp. 9300–9308.

[25] W. Wang et al., "InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions," in *Proc. CVPR*, 2023, pp. 14408–14419.

[26] X. Xiong et al., "Efficient Deformable ConvNets: Rethinking Dynamic and Sparse Operator for Vision Applications," in *Proc. CVPR*, 2024, pp. 5572–5581.

[27] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time Scene Text Detection with Differentiable Binarization," in *Proc. AAAI*, 2020, vol. 34, no. 7, pp. 11474–11481.

[28] Z. Zhao et al., "DocLayout-YOLO: Enhancing Document Layout Analysis through Diverse Synthetic Data and Global-to-Local Adaptive Perception," arXiv:2410.12628, 2024.

[29] M. Li et al., "DocBank: A Benchmark Dataset for Document Layout Analysis," in *Proc. COLING*, 2020, pp. 949–960.

[30] A. Gu et al., "XYLayoutLM: Towards Layout-Aware Multimodal Networks for Visually-Rich Document Understanding," in *Proc. CVPR*, 2022, pp. 4573–4582.

[31] P. Zhang et al., "VSR: A Unified Framework for Document Layout Analysis Combining Vision, Semantics and Relations," in *Proc. ICDAR*, 2021, pp. 115–130.

[32] S. Appalaraju et al., "DocFormer: End-to-End Transformer for Document Understanding," in *Proc. ICCV*, 2021, pp. 993–1003.

[33] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection," in *Proc. NeurIPS*, 2020, pp. 21002–21012.

[34] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Proc. NeurIPS*, 2015, pp. 91–99.