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

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 649-665
Journal homepage: http://iieta.org/journals/ts

Courseware Image Understanding and Adaptive Extraction of Pedagogically Salient Regions for Blended Learning in Higher Education

Chunliu Yue[1*], Maria Luvimi Legaspi Casihan[2]

[1] School of Marxism, Lingnan Normal University, Zhanjiang 524048, China
[2] College of Education and Liberal Arts, Adamson University, 1000 Manila, Philippines

Corresponding Author Email: yuechunliu@lingnan.edu.cn

https://doi.org/10.18280/ts.430208

Received: 30 August 2025
Revised: 22 January 2026
Accepted: 12 February 2026
Available online: 30 April 2026

Keywords:
courseware image understanding, pedagogically salient region extraction, pedagogical visual entropy field, density peak clustering, weakly supervised graph convolutional network, blended learning

ABSTRACT

The widespread adoption of blended learning in higher education has led to courseware images becoming a central form of teaching resource. The automatic extraction of pedagogically salient regions is regarded as a critical enabler for intelligent resource retrieval and personalized learning delivery. However, existing approaches remain limited, hindering their practical deployment. To address these limitations, an adaptive method for extracting pedagogically salient regions from courseware images was proposed by integrating unsupervised and weakly supervised learning paradigms. A pedagogical visual entropy field was constructed by combining local entropy, gradient structure tensors, and spatial pedagogical priors, enabling pixel-level quantification of instructional importance without semantic annotations. An improved segmentation method driven by density peak clustering was proposed. By integrating component-level feature space clustering with spatial adjacency constraints, the method effectively addressed the instability issues in courseware segmentation caused by large font-size variations and fragmented mathematical symbols. Furthermore, a structure-aware region relationship graph was constructed in conjunction with a lightweight graph convolutional network. A region-level weakly supervised training strategy was employed to achieve the logical aggregation of dispersed candidate regions, enabling the accurate identification of complete pedagogically salient units. Experimental results obtained from multi-disciplinary courseware image datasets demonstrated that superior performance was achieved in terms of intersection-over-union and F1-score when compared with state-of-the-art methods. In addition, processing efficiency was shown to satisfy real-time instructional requirements. Robust adaptability to diverse academic disciplines and varying image qualities was also achieved, and experimental results show that only 200 region-level annotated samples are required to achieve performance close to that of fully supervised methods, significantly reducing annotation costs and providing reliable technical support for intelligent resource processing in blended learning in higher education.

1. INTRODUCTION

With the deep integration of information technology and higher education, blended learning has been established as a central direction for pedagogical reform in universities [1-3]. By organically combining online self-directed learning with offline classroom instruction, the constraints of time and space inherent in traditional teaching paradigms have been effectively overcome, thereby enhancing both instructional flexibility and effectiveness [3, 4]. Within blended learning environments, courseware images—encompassing formats such as PowerPoint presentations and photographs of boardwork—have been widely adopted as primary carriers of knowledge delivery. These visual materials encapsulate critical pedagogical elements, including definitions, formulas, worked examples, and key conclusions [5]. Consequently, the automatic extraction of pedagogically salient regions has emerged as an increasingly essential requirement. Such capability serves not only as the foundation for intelligent instructional resource retrieval, reorganization, and personalized learning delivery but also as a key technological enabler for the digitalization and intelligent transformation of educational resources, ultimately supporting the development of smart learning ecosystems [6-8]. However, courseware images exhibit pronounced pedagogical structural characteristics. Pedagogically salient regions typically demonstrate distinct spatial preferences and multi-component associations. Existing general-purpose image processing approaches have predominantly focused on visual feature analysis, while insufficient consideration has been given to the intrinsic pedagogical properties embedded within instructional content. As a result, the accurate and efficient extraction of pedagogically salient regions remains challenging. This limitation has significantly constrained the practical deployment of intelligent instructional systems in blended learning environments, while simultaneously increasing the

workload associated with instructional material organization and reducing overall teaching efficiency [9, 10]. Therefore, systematic investigation into courseware image understanding and the adaptive extraction of pedagogically salient regions for blended learning in higher education [11] not only delivers substantial engineering value but also provides novel perspectives and technical support for innovation in smart education, thereby facilitating the continued optimization and advancement of blended learning paradigms [12].

Although extensive research has been conducted by both domestic and international scholars on related topics such as image saliency detection, text and formula segmentation, and region aggregation [13, 14], providing valuable technical references for the extraction of pedagogically salient regions in courseware images, several critical limitations remain unresolved. These challenges hinder the ability of existing approaches to meet the practical demands of multi-disciplinary and multi-scenario blended learning environments. Most existing saliency detection methods are primarily focused on visual features [15], such as grayscale contrast and edge strength, without incorporating pedagogical characteristics specific to courseware images. As a consequence, visually prominent but pedagogically insignificant regions are frequently misidentified as salient, while regions with high instructional value but low visual prominence are often overlooked. This limitation prevents an effective distinction between visual saliency and pedagogical salience. In the stage of text and formula segmentation, conventional approaches are typically dependent on optical character recognition (OCR) techniques or fixed-scale segmentation models [16]. Such methods exhibit limited adaptability to the wide range of font sizes present in coursewares, spanning from titles to footnotes. Moreover, when confronted with complex formulas, issues such as symbol fragmentation and overlap often arise, leading to reduced segmentation stability and resulting in incomplete or erroneous segmentation outputs. For example, when recognizing mathematical formulas containing nested square roots, fractional structures, and superscripts or subscripts, OCR often misidentifies "√" as "V", fails to detect "∑", or splits integral symbols into separate parts, leading to a complete loss of subsequent semantic information. In courseware containing multi-line aligned equations, OCR may also misplace the content following the equal sign onto the next line, resulting in unusable recognition outputs. During the region aggregation phase, spatial distance is commonly adopted as the primary criterion [17, 18], while the logical relationships inherent in instructional content are insufficiently considered. Consequently, candidate regions that are spatially dispersed but belong to the same pedagogical unit are prone to incorrect merging or omission, thereby preventing the accurate reconstruction of complete pedagogically salient units. Furthermore, most existing approaches rely heavily on large-scale pixel-level annotated datasets for model training [19, 20]. Such annotation processes are labor-intensive and time-consuming, and the resulting models exhibit limited generalizability across diverse disciplinary contexts, including mathematics, computer science, physics, and economics. This dependency further restricts the practical applicability and scalability of these methods in real-world instructional scenarios.

In response to the aforementioned limitations of existing studies, the practical requirements of blended learning in higher education are addressed through the development of an adaptive extraction method for pedagogically salient regions in courseware images, integrating unsupervised and weakly supervised learning paradigms. The primary contributions are summarized below. First, a pedagogical visual entropy field is proposed, extending beyond the limitations of conventional visual entropy, which primarily focuses on grayscale distributions. Local normalized entropy, gradient structure tensor anisotropy, and a spatial pedagogical prior probability map are integrated, and weighting coefficients are optimized via grid search. Through this formulation, unsupervised initial localization of potential pedagogically salient regions is achieved without reliance on semantic annotations or OCR priors, thereby effectively addressing the dependency on prior knowledge inherent in traditional methods. Second, an adaptive density peak clustering-based segmentation method is introduced. The classical density peak clustering algorithm is enhanced by performing clustering within a two-dimensional feature space defined by component area and stroke width. Spatial adjacency relationships are incorporated to construct an adaptive cutoff distance, and component connectivity is further refined through an adaptive morphological closing operation. As a result, precise multi-scale segmentation of text and formulas is achieved without dependence on OCR, effectively mitigating segmentation instability caused by large font-scale variations and fragmented mathematical symbols in courseware images. Third, a structure-aware weakly supervised graph-based reasoning and aggregation method is developed. Node vectors are constructed by integrating geometric, content, and type-specific features. A weighted undirected graph is established based on spatial adjacency, alignment, and containment relationships. A two-layer lightweight graph convolutional network is designed and trained using a region-level binary labeling strategy under weak supervision. Combined with graph-cut-based post-processing, logical aggregation of dispersed candidate regions and automatic generation of semantic labels are accomplished, thereby enhancing both the completeness and structural coherence of the extracted pedagogically salient regions. Finally, a multi-disciplinary courseware image dataset is constructed. Comprehensive evaluations, including ablation studies, comparative experiments, and scenario-based adaptability analyses, are conducted to systematically validate the superior performance of the proposed method in terms of adaptability, accuracy, and real-time capability, ensuring that the method satisfies the real-time processing requirements of blended learning environments and provides reliable support for practical instructional applications.

The remainder of this study is organized below. In Section 2, related work is systematically reviewed across three core directions: courseware image preprocessing and saliency detection, text and formula segmentation, and image region aggregation with weakly supervised learning. Particular emphasis is placed on clarifying the distinctions between the proposed method and existing approaches. In Section 3, the overall framework of the proposed adaptive extraction method is described in detail, along with the technical implementation of each core module. These include image preprocessing, construction of the pedagogical visual entropy field, the improved density peak clustering-based segmentation method, and the structure-aware weakly supervised graph-based reasoning and aggregation strategy. In Section 4, extensive experiments are conducted to validate the effectiveness and superiority of the proposed method. These experiments include ablation studies, multi-disciplinary adaptability

evaluations, comparative analyses, real-time performance assessments, and investigations into the impact of weakly supervised annotation scale. The experimental results are further analyzed in depth. In Section 5, a comprehensive discussion is provided based on the experimental findings. The fundamental mechanisms and key factors influencing performance are examined, existing limitations are objectively analyzed, and potential directions for future research are outlined. Finally, Section 6 summarizes the principal contributions and findings, clarifies the practical application value and research significance of the proposed method, and presents a brief outlook on subsequent research directions.

2. METHODOLOGY

2.1 Overall framework

To address the limitations of existing methods for extracting pedagogically salient regions from courseware images—particularly in terms of adaptability, accuracy, and practical applicability—an adaptive extraction framework is developed by integrating unsupervised and weakly supervised learning paradigms. A four-stage progressive pipeline is adopted, in which all modules operate collaboratively to enable the precise transformation from raw courseware images to complete pedagogically salient regions. A schematic illustration of the overall framework presents the core modules and data flow in a clear and intuitive manner. The input to the framework consists of various types of courseware images encountered in blended learning environments in higher education, including screenshots of PowerPoint presentations and photographs of boardwork. The output is defined as precise masks of pedagogically salient regions along with their corresponding semantic labels. Initially, an image preprocessing module is applied to enhance input image quality by mitigating noise, geometric distortion, and skew, thereby providing reliable data support for subsequent processing stages. Subsequently, a pedagogical visual entropy field is constructed, in which multi-dimensional features are integrated to achieve unsupervised initial localization of potential pedagogically salient regions. Candidate regions with potential instructional value are thereby identified. On this basis, an improved density peak clustering-driven adaptive segmentation module is employed to perform precise segmentation of text and formulas within the candidate regions. This module effectively addresses challenges associated with multi-scale adaptation and segmentation instability. Finally, a structure-aware weakly supervised graph-based reasoning and aggregation module is utilized to analyze logical relationships among the segmented candidate regions and to perform their aggregation, resulting in the output of complete and accurate pedagogically salient units. The proposed framework operates without reliance on large-scale pixel-level annotations or OCR priors, while simultaneously ensuring high extraction accuracy, cross-disciplinary adaptability, and real-time performance. Consequently, the practical requirements for extracting pedagogically salient regions from courseware images in blended learning environments are effectively satisfied. Figure 1 shows the overall framework.

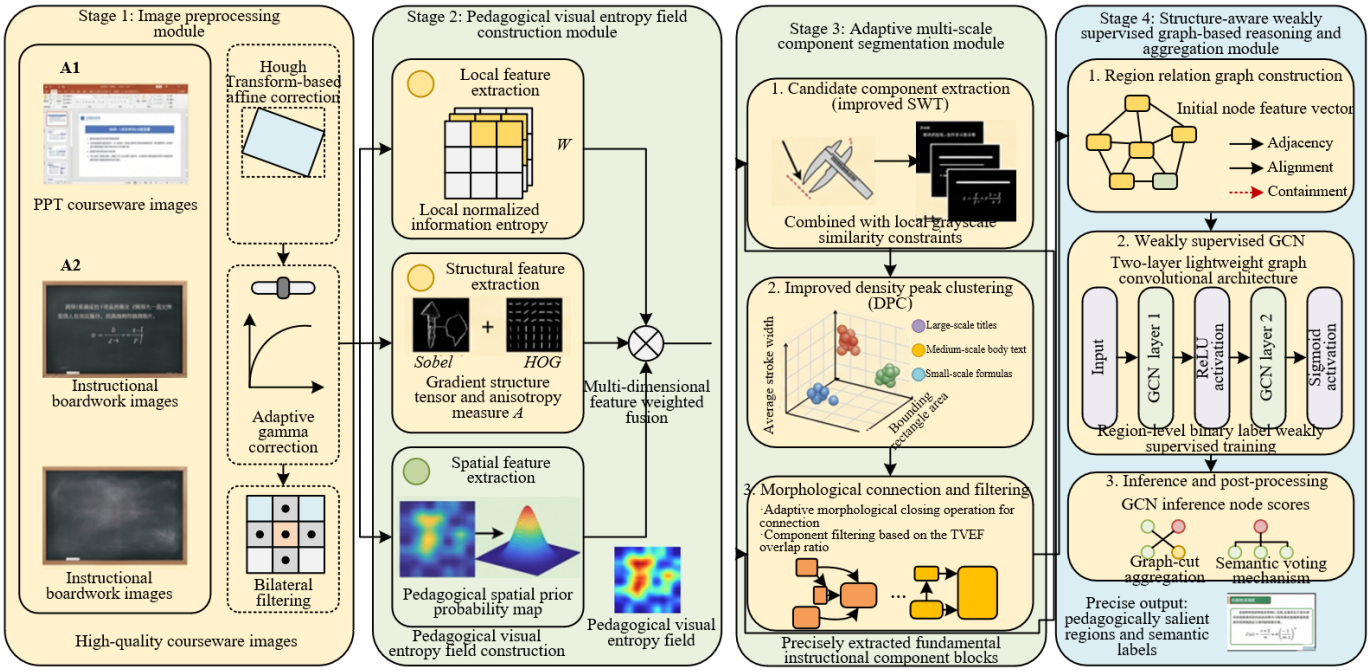


Figure 1. Overall framework of the adaptive extraction method for pedagogically salient regions in courseware images

2.2 Image preprocessing

Image preprocessing is regarded as a critical foundation for ensuring the stable operation of subsequent modules, including pedagogical visual entropy field construction, adaptive segmentation, and graph-based reasoning and aggregation. The primary objective is to eliminate various interference factors present in courseware images and to enhance overall image quality, thereby providing high-quality data support for subsequent feature extraction and region processing. To address common issues in courseware images—such as uneven grayscale distribution, noise contamination, and tilt distortion—an adaptive preprocessing pipeline is designed, with particular emphasis placed on achieving a balance between interference suppression and the preservation of critical details. Adaptive gamma correction is applied to adjust image grayscale contrast, overcoming the limitations of conventional fixed gamma correction. The

correction coefficient is adaptively determined based on the statistical characteristics of image grayscale distribution. The mathematical formulation is expressed as $I_{out} = 255 \cdot \left(\frac{I_{in}}{255}\right)^{\gamma}$ $\gamma = \lambda \cdot exp(1-\mu)$ where I_{out} denotes the corrected pixel grayscale value, I_{in} represents the original pixel grayscale value, and μ denotes the normalized mean grayscale value of the image, with a value range of [0,1]. Through this adaptive strategy, the contrast between text, formulas, and background regions is effectively enhanced, thereby preventing distortions in subsequent local entropy computation and gradient-based feature extraction caused by insufficient contrast.

Bilateral filtering is employed to suppress image noise while preserving edge details of text and formulas. The filtering kernel is constructed by integrating spatial distance and grayscale similarity between pixels, thereby effectively overcoming the edge-blurring limitations of conventional Gaussian filtering. As a result, gradient structure tensor computation is enabled to accurately capture edge features of textual and mathematical components, providing reliable support for subsequent component extraction and segmentation. Affine correction based on the Hough transform is applied to eliminate tilt distortion in courseware images. Linear features, such as boardwork strokes and PowerPoint boundaries, are detected to estimate the skew angle, after which affine transformation is performed to correct image orientation. This process ensures that spatial adjacency constraints in density peak clustering and region relationship construction in graph-based reasoning are established on an accurate spatial basis, thereby avoiding misinterpretation of component positions caused by image skew. The entire preprocessing pipeline operates without manual intervention and is capable of adaptively accommodating courseware images of varying types and quality levels. As a result, both the accuracy and stability of subsequent processing modules are significantly improved.

2.3 Construction of the pedagogical visual entropy field

2.3.1 Local information entropy computation

Figure 2 illustrates the multi-dimensional feature extraction and fusion process for constructing the pedagogical visual entropy field. Local information entropy serves as the fundamental basis for the pedagogical visual entropy field construction, with its primary function being the quantification of grayscale distribution complexity within local image regions, thereby providing a quantitative foundation for the initial localization of potential pedagogically salient regions. To address the limitations of conventional local entropy computation—such as low computational efficiency, fixed window configurations, and insufficient adaptation to the characteristics of courseware images—an improved local information entropy computation scheme is designed. This scheme is tailored to the distribution patterns of text and formulas in courseware images, while balancing computational accuracy and real-time performance. A square window with radius r is employed to traverse each pixel in the image. The Shannon entropy within the window is calculated to characterize the uncertainty of the local grayscale distribution, as defined by:

$$H(x,y) = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{1}$$

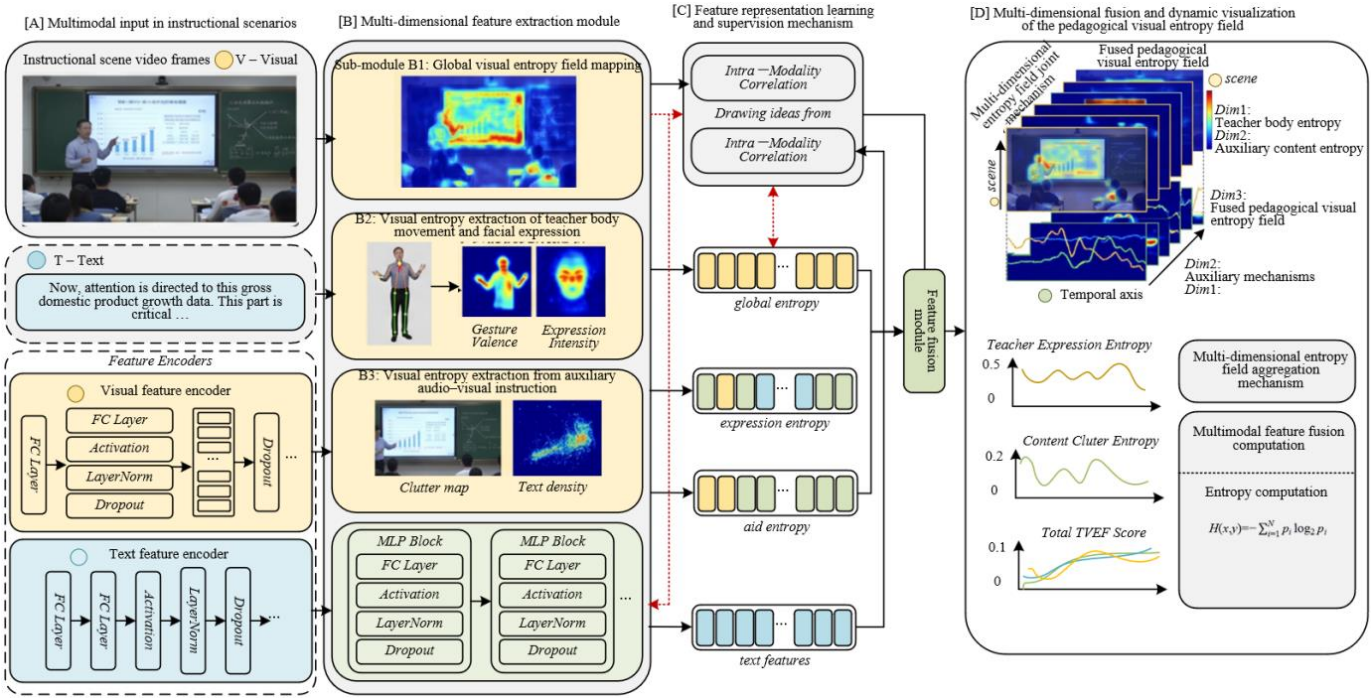


Figure 2. Multi-dimensional feature extraction and fusion framework of the pedagogical visual entropy field

where, $H(x,y)$ denotes the local information entropy at pixel (x,y); p_i represents the probability of occurrence of the i-th grayscale value within the window; N represents the actual number of gray levels present within the window ($N \leq 256$). To meet the computational requirements of real-time instructional scenarios, an integral image-based acceleration strategy is introduced. By precomputing both the grayscale integral image and the squared grayscale integral image, the computational complexity of local entropy calculation is reduced from $O(r^2)$ to $O(1)$, thereby effectively overcoming

the inefficiency associated with traditional methods. Furthermore, to eliminate the influence of grayscale variations across courseware images of different types and quality levels, normalization is applied to the computed local entropy values. The normalization is defined as:

$$H_{norm}(x,y)=\frac{H(x,y)-H_{min}}{H_{max}-H_{min}} \tag{2}$$

where, $H_{norm}(x,y)$ denotes the normalized local information entropy, and H_{min} and H_{max} represent the minimum and maximum values of local entropy across the entire image, respectively. The normalized values are constrained within the range [0,1]. The grayscale distribution of courseware images exhibits distinct patterns. Pedagogically salient regions typically present moderate grayscale complexity, corresponding to intermediate normalized entropy values. In contrast, background regions are characterized by relatively uniform grayscale distributions, resulting in lower entropy values, whereas noise or highly textured regions exhibit irregular grayscale distributions and correspondingly higher entropy values. This property enables local normalized entropy to provide an initial discrimination between potential pedagogically salient regions and non-salient regions.

2.3.2 Gradient structure tensor and anisotropy measure

The gradient structure tensor is employed to capture edge characteristics, texture orientation, and intensity variations within local image regions, thereby compensating for the limitation of local information entropy, which focuses solely on grayscale distribution while neglecting spatial structural features. As a result, complementary support is provided for the localization of pedagogically salient regions, further improving the accuracy of initial region identification. The Sobel operator is applied to compute image gradients in the x and y directions, denoted as G_x and G_y, respectively. Based on these gradients, the gradient structure tensor $M(x,y)$ is constructed as follows:

$$M(x,y)=\begin{Bmatrix} G_x^2 & G_xG_y \\ G_xG_y & G_y^2 \end{Bmatrix} \tag{3}$$

To reduce the influence of noise on gradient estimation, Gaussian filtering is applied to smooth the gradient components, resulting in a more stable gradient structure tensor matrix. Eigenvalue decomposition is then performed on the tensor, yielding two eigenvalues, λ_1 and λ_2, with $\lambda_1 \geq \lambda_2$. λ_1 represents the principal gradient direction, while λ_2 corresponds to the secondary direction. The difference between these eigenvalues directly reflects the edge and texture characteristics of the local region. To quantitatively describe the degree of anisotropy, an anisotropy measure A is defined as:

$$A=\frac{\lambda_1-\lambda_2}{\lambda_1+\lambda_2} \tag{4}$$

The value of A lies within the range [0,1]. When A approaches 1, strong anisotropy is indicated, corresponding to regions with well-defined edges and structured textures. Conversely, when A approaches 0, isotropy is indicated, corresponding to regions such as background areas with no prominent edge features. In pedagogically salient regions, text and formulas exhibit clear edge contours and structured texture patterns, resulting in significantly higher anisotropy values compared to background regions. In contrast, background areas typically exhibit uniform grayscale distributions and lack distinct edge features, leading to lower anisotropy values. This distinction enables the anisotropy measure derived from the gradient structure tensor to effectively identify candidate regions with pronounced edge characteristics. When combined with local normalized entropy, a complementary representation is achieved, providing a robust foundation for the subsequent weighted fusion in the construction of the pedagogical visual entropy field.

2.3.3 Construction of the pedagogical spatial prior probability map

The pedagogical spatial prior probability map is introduced to quantify the likelihood that different spatial locations within courseware images correspond to pedagogically salient regions. This component addresses the limitation of local normalized entropy and gradient anisotropy measures, which primarily focus on local features while neglecting spatial regularities inherent in instructional scenarios. As a result, the accuracy of initial localization of potential pedagogically salient regions is further enhanced. A spatial prior that aligns with real instructional practices is constructed through statistical analysis of multi-disciplinary courseware images, thereby overcoming the limitations of conventional generic spatial priors, which often exhibit poor adaptability. Courseware image samples from multiple disciplines, including mathematics, computer science, physics, and economics, are selected. Pedagogically salient regions within these images are manually annotated, and the distribution of their center coordinates is analyzed. The results indicate a pronounced spatial preference: core instructional content is predominantly concentrated in the upper-central and central regions of the image, whereas peripheral and corner regions are typically occupied by non-salient elements such as page numbers and watermarks. Furthermore, a high degree of consistency in spatial distribution patterns is observed across different disciplines. Based on these statistical observations, a Gaussian mixture model is employed to fit the spatial distribution density of pedagogically salient regions. A normalized pedagogical spatial prior probability map $P_{spatial}$ is constructed, with its probability density function defined as:

$$P_{spatial}(x,y)=\sum_{k=1}^{K}\omega_k \cdot N((x,y)|\mu_k,\Sigma_k) \tag{5}$$

where, K denotes the number of components in the Gaussian mixture model, which is determined to be 3 based on the Bayesian information criterion. The parameter ω_k represents the weight of the k-th Gaussian component and satisfies $\sum_{k=1}^{K}\omega_k=1$. The vector μ_k denotes the mean of the k-th component, corresponding to the central coordinates of concentrated pedagogically salient regions, while $\sum_k$ represents the covariance matrix of the k-th Gaussian component, characterizing the spatial dispersion around the corresponding center. The fitted spatial distribution density is subsequently normalized to constrain the value range of $P_{spatial}(x,y)$ to [0,1]. Higher values indicate a greater probability that a given spatial location belongs to a pedagogically salient region. This provides a reliable spatial constraint for the subsequent fusion process in the pedagogical

visual entropy field.

2.3.4 Weighted fusion and parameter determination

The pedagogical visual entropy field is constructed through the weighted fusion of local normalized entropy, gradient anisotropy measure, and the pedagogical spatial prior probability map, enabling comprehensive pixel-level quantification of pedagogical importance. The core innovation lies in the use of adaptive parameter optimization to achieve optimal coordination among these three features, thereby overcoming the limited adaptability associated with conventional fixed-weight fusion strategies. The mathematical formulation of the pedagogical visual entropy field is defined as:

$$TVEF(x,y)=\alpha\cdot H_{norm}(x,y)+\beta\cdot A(x,y)+\gamma\cdot P_{spatial}(x,y) \quad (6)$$

where, α, β, and γ denote the weighting coefficients for local normalized entropy, gradient anisotropy measure, and the pedagogical spatial prior probability map, respectively, subject to the constraint $\alpha+\beta+\gamma=1$. These coefficients are used to regulate the relative contributions of the three components within the pedagogical visual entropy field. To determine the optimal weighting coefficients, a grid search strategy is employed. A small-scale validation dataset is constructed, consisting of 100 courseware images spanning multiple disciplines and varying quality conditions, including clear, blurred, and geometrically distorted samples. The F1-score of initial localization for potential pedagogically salient regions is adopted as the evaluation metric. The parameter search space is defined as $\alpha,\beta,\gamma\in[0,1]$ with a step size of 0.1. All feasible combinations satisfying the constraint are exhaustively evaluated, and the set of coefficients yielding the highest F1-score on the validation dataset is selected. The optimal values are determined as $\alpha=0.3$, $\beta=0.4$, and $\gamma=0.3$. This weight configuration ensures that the fundamental contributions of local grayscale distribution and edge structural features are preserved, while effectively incorporating the spatial preference characteristics of instructional content. Consequently, a balanced and synergistic integration of the three components is achieved. The value range of the pedagogical visual entropy field is normalized to [0,1], with higher values indicating a greater likelihood that a given pixel belongs to a pedagogically salient region. By applying an appropriate threshold, initial localization of potential pedagogically salient regions can be achieved. These regions are subsequently used as precise candidate inputs for the adaptive segmentation module, thereby significantly improving the overall accuracy and efficiency of the extraction process.

2.4 Multi-scale text and formula segmentation based on improved density peak clustering

2.4.1 Candidate component extraction

Figure 3 illustrates the feature space representation and adaptive segmentation mechanism for multi-scale components based on the improved density peak clustering approach. Candidate component extraction serves as the foundation for multi-scale text and formula segmentation. The primary objective is to accurately extract fundamental components corresponding to text and formulas from the potential pedagogically salient regions initially localized by the pedagogical visual entropy field, thereby providing high-quality input samples for subsequent clustering-based segmentation. To address the limitations of conventional stroke width transform—including sensitivity to noise, significant estimation bias in stroke width, and limited adaptability to multi-scale text and complex formulas commonly found in courseware images—an improved stroke width transform method is developed to enhance both the accuracy and completeness of component extraction. In the improved stroke width transform, stroke width estimation is refined through gradient direction constraints. Edge detection is first performed on the preprocessed image to obtain the contours of text and formulas. Subsequently, pixel grayscale transition points are identified by searching along the normal direction of each edge. The distance between paired transition points is computed as the initial stroke width. To correct outliers, a local grayscale similarity constraint is introduced, and the refined stroke width is calculated as:

$$w(x,y)=\frac{1}{N}\sum_{i=1}^{N} w_i(x,y)\cdot exp\left(-\frac{|I(x,y)-I(x_i,y_i)|}{\sigma}\right) \quad (7)$$

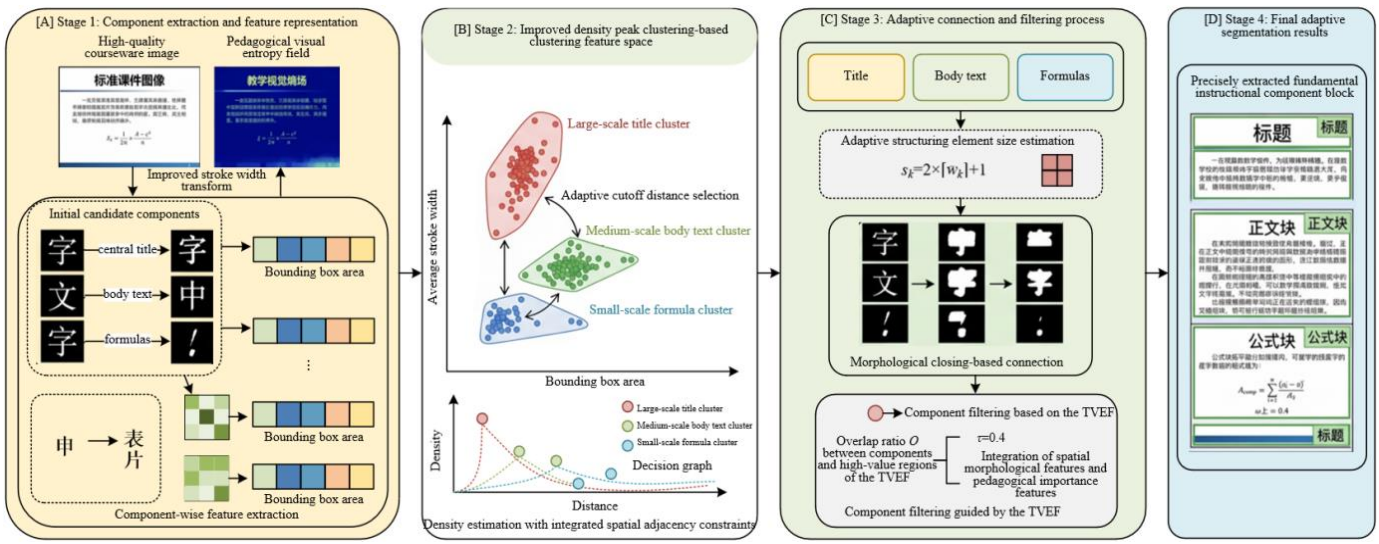


Figure 3. Feature space representation of multi-scale components and adaptive segmentation mechanism based on improved density peak clustering

where, $w(x,y)$ denotes the refined stroke width at pixel (x,y); $w_i(x,y)$ represents the initial stroke width estimates; N denotes the number of pixels within the local neighborhood; $I(x,y)$ and $I(x_i,y_i)$ represent the grayscale values of the current pixel and its neighboring pixels, respectively; and σ is the grayscale similarity coefficient, set to 20. Based on the refined stroke width, connected component analysis is performed to segment the image according to pixel connectivity and stroke width consistency, resulting in multiple independent components. Invalid components—such as those with excessively small areas or abnormal aspect ratios—are removed to eliminate noise and isolated artifacts. To enable subsequent multi-scale clustering, feature descriptors are constructed for each valid component. Key features, including bounding box dimensions, average stroke width, aspect ratio, and mean grayscale value, are extracted. Among these, average stroke width and bounding box area serve as critical features for distinguishing components of different scales, providing reliable support for the subsequent improved density peak clustering-based clustering process.

2.4.2 Improved density peak clustering

To address the limitations of the classical density peak clustering algorithm in instructional component clustering—specifically, its reliance on single-feature representations, absence of spatial constraints, and limited adaptability to multi-scale text and formula components—an improved density peak clustering method is developed. Precise clustering and automatic scale-aware classification of components are achieved through the construction of a task-specific feature space, the introduction of an adaptive cutoff distance, and the incorporation of spatial adjacency constraints. A two-dimensional feature space defined by component area and average stroke width is first constructed. These two features are selected as clustering descriptors, as they effectively characterize scale differences among components. Title components are typically associated with large areas and large stroke widths, body text components exhibit intermediate values, and formula components are characterized by relatively small areas and stroke widths. This feature space design overcomes the limitation of single-feature clustering in conventional density peak clustering and significantly enhances the discriminability of multi-scale components. To resolve the instability caused by fixed cutoff distance selection in classical density peak clustering, an adaptive cutoff distance strategy is adopted. The Euclidean distances between all component pairs in the two-dimensional feature space are computed, and the 2nd percentile of the distance distribution is selected as the cutoff distance. The formulation is defined as:

$$d_c = percentile(d_{ij}, 2) \tag{8}$$

where, d_c denotes the adaptive cutoff distance, d_{ij} represents the Euclidean distance between components i and j in the feature space, and *percentile* denotes the percentile operation. Furthermore, spatial adjacency constraints are incorporated to refine the clustering process. The spatial Euclidean distance between components is computed, and when the distance between two components is smaller than a predefined threshold—set to 1.5 times the average bounding box edge length—their clustering affinity is strengthened. A spatial weighting factor is introduced into the density estimation, and the modified density formulation is expressed as:

$$\rho_i = \sum_{j \neq i} exp\left(-\left(\frac{d_{ij}}{d_c}\right)^2\right) \cdot exp\left(-\left(\frac{s_{ij}}{s_0}\right)^2\right) \tag{9}$$

where, ρ_i denotes the density of component i; s_{ij} represents the spatial Euclidean distance between components i and j; and s_0 denotes the spatial distance threshold. Through the integration of feature space optimization, adaptive cutoff distance selection, and spatial adjacency constraints, the improved density peak clustering algorithm enables automatic identification of component categories across different scales. Text and formula components of similar scales are effectively grouped into the same clusters. Consequently, segmentation instability caused by large font-scale variations and fragmented mathematical symbols in courseware images is significantly mitigated, providing a robust foundation for subsequent component connection and filtering processes.

2.4.3 Adaptive morphological connection and filtering

The primary objective of adaptive morphological connection is to merge discrete components of the same scale—obtained through the improved density peak clustering—into complete text lines or formula blocks. This process addresses the issue of incomplete segmentation caused by fragmented mathematical symbols and dispersed textual characters. The key innovation lies in the adaptive generation of structuring elements, which overcomes the limitation of fixed structuring elements in conventional morphological operations that are unable to accommodate multi-scale components. Given that different clustering categories correspond to components of different scales, the size of the structuring element is adaptively determined based on the average stroke width of each cluster. This design ensures that the connection process effectively merges discrete components while avoiding over-expansion that could result in incorrect merging of distinct text or formula blocks. Let $\bar{w}_k$ denote the average stroke width of the k-th cluster. The corresponding structuring element size s_k for morphological closing is defined as:

$$s_k = 2 \times \lceil \bar{w}_k \rceil + 1 \tag{10}$$

where, $\lceil \cdot \rceil$ denotes the ceiling operation. A square structuring element is employed. This formulation allows the structuring element to align with the stroke characteristics of components at different scales. For large-scale title components, larger structuring elements are generated to facilitate rapid connection between characters, whereas for small-scale mathematical symbols, smaller structuring elements are used to prevent unintended merging of adjacent symbols. Morphological closing is performed independently for each cluster. Specifically, dilation is first applied to bridge gaps between components, followed by erosion to restore boundary contours and eliminate distortions introduced during dilation. Through this process, discrete components within the same cluster are accurately connected, resulting in the formation of complete structural units such as text lines and formula blocks.

After component connection has been completed, further filtering is required to identify true pedagogically salient regions while removing auxiliary components such as page numbers, watermarks, headers, and footers. The core innovation lies in the integration of the overlap ratio based on the pedagogical visual entropy field for precise filtering. By leveraging the quantitative representation of pedagogical

importance provided by the previously constructed pedagogical visual entropy field, both the specificity and accuracy of the filtering process are significantly improved. An overlap ratio O between each component and high-value regions of the pedagogical visual entropy field is defined to measure pedagogical relevance. The formulation is expressed as:

$$O=\frac{S_{overlap}}{S_{component}} \tag{11}$$

where, $S_{overlap}$ denotes the overlapping area between the component region and high-value regions of the pedagogical visual entropy field (defined as $TVEF$ value ≥ 0.5), and $S_{component}$ represents the total area of the component. The threshold for the overlap ratio is determined through optimization on a small-scale validation dataset, yielding $\tau=0.4$. Components satisfying $O\geq\tau$ are classified as candidate pedagogically salient regions, whereas those with $O<\tau$ are identified as auxiliary textual components and are subsequently removed. This filtering strategy effectively integrates spatial morphological characteristics with pedagogical importance features, thereby overcoming the limitations of conventional methods that rely solely on geometric attributes such as area or position. As a result, core instructional components, including text and formulas, are accurately preserved, while non-instructional auxiliary content is excluded. High-quality candidate regions are thus provided for the subsequent graph-based reasoning and aggregation module, further enhancing the overall accuracy and practical applicability of the extraction framework.

2.5 Structure-aware weakly supervised graph-based reasoning and aggregation

2.5.1 Region relation graph construction

The region relation graph serves as the core representation for achieving logical aggregation of dispersed candidate regions. Its primary innovation lies in overcoming the limitation of conventional graph construction methods that rely solely on spatial distance, by incorporating the logical characteristics of instructional content. A structure-aware graph is therefore constructed, integrating multi-dimensional features and multiple relational constraints to enable precise modeling of pedagogical relationships among candidate regions. The candidate pedagogically salient regions obtained after filtering in Section 2.4.3 are treated as graph nodes, where each node corresponds to a text block or a formula block. To accurately characterize inter-node relationships, a high-dimensional initial feature vector is constructed by integrating geometric, content, and type-specific features. Geometric features include the width, height, centroid coordinates, and area of the bounding box, which describe the spatial location and morphological properties of each node. Content features are defined by the mean pedagogical visual entropy field value and grayscale variance within the node region, providing a quantitative representation of pedagogical importance. Type features are encoded using one-hot vectors to distinguish between text and formula nodes. These three categories of features are concatenated to form a 12-dimensional initial node feature vector, providing comprehensive input for subsequent graph-based reasoning. Graph edges are constructed based on three types of relationships: spatial adjacency, alignment, and containment. This multi-relational design overcomes the limitations of conventional adjacency-only approaches and better reflects the layout logic of instructional content. Edge weights are computed by integrating spatial distance and directional consistency, with the formulation defined as:

$$w_{ij}=exp\left(-\frac{d_{ij}}{\lambda}\right)\cdot\cos\theta_{ij} \tag{12}$$

where, w_{ij} denotes the edge weight between nodes i and j; d_{ij} represents the Euclidean distance between their centroids; and λ is a distance scaling coefficient, set to 1.2 times the average bounding box edge length. The term θ_{ij} denotes the angle between the line connecting the two node centroids and the horizontal direction, capturing directional consistency. Higher edge weights indicate stronger spatial proximity and directional alignment, suggesting a greater likelihood that the corresponding regions belong to the same pedagogically salient unit. This formulation provides a reliable relational foundation for subsequent graph-based reasoning and aggregation. It should be noted that λ is set to 1.2 times the average bounding-box edge length of the nodes. This fixed value demonstrates stable performance under typical slide layouts. When the local component density is high, node spacing is generally small; a relatively larger λ makes edge weights more sensitive to distance variations, helping to distinguish closely located nodes. Conversely, when the component density is very low, larger spacing naturally leads to rapid decay of edge weights to negligible values, preventing the formation of spurious strong connections. A parameter sensitivity analysis on the MD-CID test set shows that when λ varies within 0.8 to 1.6 times the edge length, the final aggregated F1-score fluctuates by less than 1.5%, indicating good robustness across regions with different densities. For lecture materials with extremely non-uniform or abnormal density distributions, future work may explore adaptive λ adjustment based on local component density.

2.5.2 Weakly supervised graph convolutional network model design

Figure 4 presents the structure-aware region relation graph construction and the weakly supervised graph convolutional network aggregation topology. The weakly supervised graph convolutional network model is employed to perform reasoning and learning on the region relation graph. The principal innovation lies in the design of a lightweight network architecture tailored for real-time instructional scenarios, combined with a region-level weak supervision strategy that substantially reduces annotation cost while enabling accurate modeling of pedagogical relationships among nodes. A two-layer graph convolutional architecture is adopted to balance inference accuracy and computational efficiency. The input is defined as the node feature matrix of the region relation graph, with the hidden layer dimension set to 64. The output layer produces a pedagogical importance score for each node, which is used to determine whether the node belongs to the final set of pedagogically salient regions. Inter-layer propagation follows the core principle of graph convolution, where node features are aggregated through the adjacency matrix. The propagation rule is defined as:

$$H^{(l+1)}=\sigma\left(\tilde{A}H^{(l)}W^{(l)}+b^{(l)}\right) \tag{13}$$

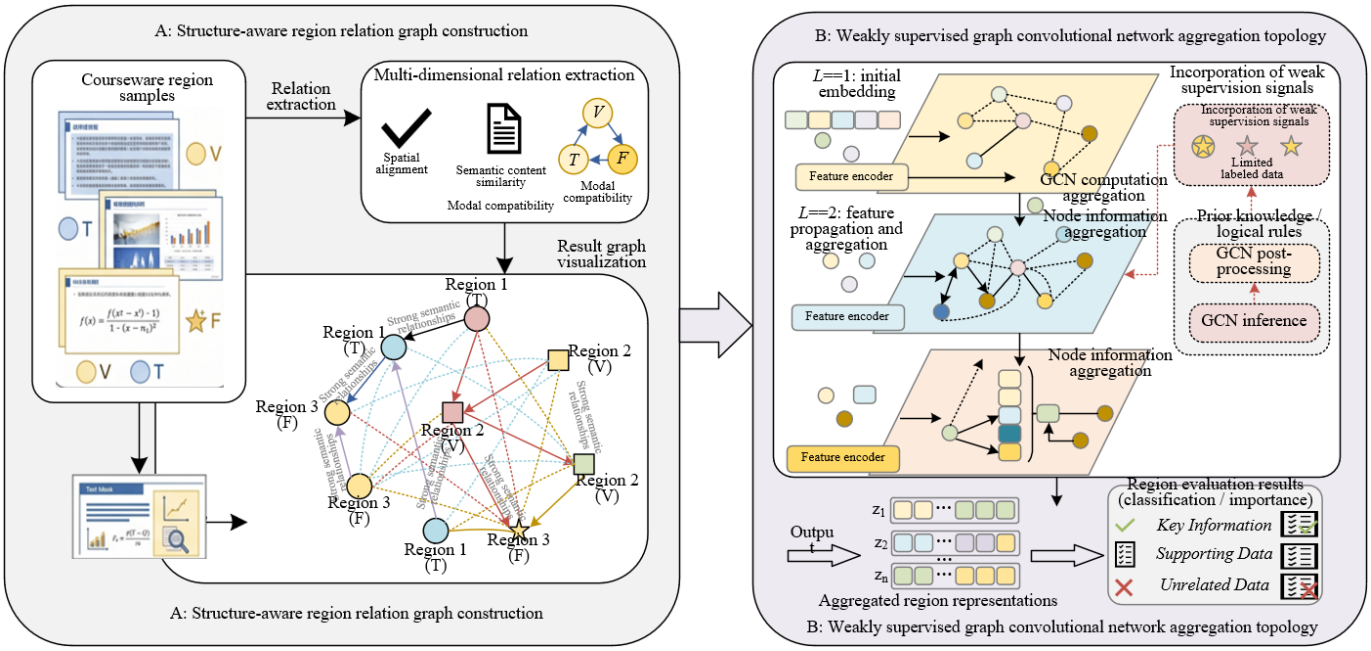


Figure 4. Structure-aware region relation graph construction and weakly supervised graph convolutional network aggregation topology

where, $H^{(l)}$ denotes the node feature matrix at layer l; $\tilde{A}=D^{-1/2}(A+I)D^{-1/2}$ represents the normalized adjacency matrix; A is the adjacency matrix of the graph; I is the identity matrix; and D is the degree matrix. The parameters $W^{(l)}$ and $b^{(l)}$ correspond to the weight matrix and bias vector of the l-th layer, respectively, and σ denotes the activation function. The rectified linear unit is employed as the activation function to alleviate the vanishing gradient problem and accelerate model convergence. The Adam optimizer is utilized with a learning rate of 0.001, and L2 regularization is applied to mitigate overfitting. A region-level weak supervision strategy is adopted for model training. Only bounding-box annotations of candidate regions are required, and pixel-level masks are not needed, thereby significantly reducing annotation cost. Model parameters are optimized using a binary cross-entropy loss function, enabling effective learning of pedagogical relationships among nodes. This design provides a reliable foundation for subsequent region aggregation and semantic label generation. The lightweight two-layer architecture substantially improves inference efficiency, with an average processing time of approximately 0.02 seconds per image, fully satisfying the requirements of real-time instructional applications.

2.5.3 Weakly supervised training strategy

The core innovation of the weakly supervised training strategy lies in the adoption of region-level binary annotations in place of conventional pixel-level annotations. This design substantially reduces annotation cost while ensuring that the model is capable of accurately learning the logical relationships underlying pedagogically salient regions, thereby supporting generalization across multi-disciplinary courseware images. During the annotation process, region-level binary labels are assigned. Annotators are only required to perform bounding-box selection of candidate regions and indicate whether each region corresponds to a pedagogically salient region. Pixel-level mask annotation is not required. As a result, annotation efficiency is improved by more than 80% compared with pixel-level labeling, effectively addressing the high cost and time-consuming nature of traditional supervised approaches. The training dataset consists of 300 courseware images collected from multiple disciplines, including mathematics, computer science, physics, and economics. The dataset incorporates diverse image quality conditions, such as clear, blurred, and geometrically distorted samples, thereby ensuring diversity and representativeness and providing a solid foundation for model generalization. Model training is performed using a binary cross-entropy loss function to optimize the parameters of the lightweight graph convolutional network, enabling accurate learning of the mapping between node features and pedagogical importance. The loss function is defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right) \tag{14}$$

where, N denotes the total number of nodes in the training set, y_i represents the ground-truth binary label of node i (with 1 indicating a pedagogically salient region and 0 indicating a non-salient region), and p_i denotes the predicted probability that node i belongs to a pedagogically salient region. The training process is conducted over 100 epochs, with a batch size of 32. An early stopping strategy is applied, whereby training is terminated when the validation loss does not decrease for 10 consecutive epochs. In addition, an L2 regularization term is incorporated to mitigate overfitting, with the regularization coefficient set to 0.0001. This configuration ensures that the model achieves strong generalization performance on previously unseen courseware images, while maintaining a balance between training efficiency and inference accuracy.

2.5.4 Inference and post-processing

The primary objective of the inference and post-processing stage is to transform the node-level pedagogical importance learned by the graph convolutional network into complete and accurate outputs of pedagogically salient regions. The key innovation lies in the integration of graph-cut-based

aggregation with a semantic voting mechanism, thereby overcoming the limitations of conventional spatial aggregation and ensuring both logical coherence and semantic accuracy. During the graph convolutional network inference stage, a pedagogical importance score is assigned to each node, representing the probability that the node belongs to a pedagogically salient region. A threshold of 0.6 is applied: nodes with importance scores greater than or equal to 0.6 are retained as valid nodes for subsequent aggregation, while nodes with scores below 0.6 are discarded as non-salient regions, thereby reducing redundant information. To achieve logical aggregation of the retained nodes, a graph-cut algorithm is applied to the region relation graph. Edge weights are used as the cutting cost, and nodes with high edge weights and strong logical associations are grouped into the same connected component. This approach ensures that spatially dispersed nodes belonging to the same pedagogical unit are effectively aggregated into a unified region, thereby addressing the limitations of traditional distance-based aggregation methods, which are prone to erroneous merging or missed aggregation. Following aggregation, the minimum enclosing convex hull is computed for each connected component to generate a complete region mask. This ensures that the resulting mask fully encompasses the entire pedagogically salient unit, preventing the omission of boundary nodes and improving the completeness of the output regions. Semantic label assignment is performed using a voting mechanism, in which node types and spatial prior probabilities are jointly considered. The voting weight is defined as:

$$score_t = \frac{1}{M} \sum_{i=1}^{M} \left(w_{t,i} \cdot P_{spatial}(x_i, y_i) \right) \tag{15}$$

where, $score_t$ denotes the voting score for label t; M represents the total number of nodes within the connected component; $w_{t,i}$ denotes the weight of node i with respect to label t (assigned as 1 for text nodes corresponding to text labels and 1 for formula nodes corresponding to formula labels); and $P_{spatial}(x_i, y_i)$ denotes the spatial prior probability at the center of node i. The label with the highest voting score is selected as the semantic label for the corresponding connected component, enabling automatic semantic annotation. Consequently, the final output includes not only precise spatial localization of pedagogically salient regions but also explicit semantic attributes, thereby providing enriched support for downstream applications such as intelligent instructional resource retrieval and personalized content delivery.

3. EXPERIMENTAL RESULTS AND ANALYSIS

To systematically evaluate the effectiveness, superiority, generalizability, and practical applicability of the proposed adaptive extraction method for pedagogically salient regions in courseware images for blended learning in higher education, a series of experiments was designed in accordance with the standard protocols of image processing research in Science Citation Index-indexed journals. Five groups of targeted experiments were conducted, focusing on the effectiveness of individual core modules, cross-disciplinary adaptability, performance comparison with state-of-the-art methods, real-time processing capability, and the efficiency of weakly supervised annotation. All experiments were performed on a uniformly constructed multi-disciplinary courseware image dataset and conducted under identical experimental conditions, thereby ensuring the objectivity and comparability of the results.

3.1 Experimental setup

To evaluate the generalization capability of the proposed method, a multi-disciplinary courseware image dataset was constructed. The dataset comprises 1,200 courseware images spanning four core disciplines: mathematics, computer science, physics, and economics. Among these, 850 images are PowerPoint courseware screenshots and 350 are boardwork photographs. Image quality is categorized into three types: clear (60%), blurred (25%), and geometrically distorted (15%). The dataset includes three common resolution levels: 720p, 1080p, and 4K. The annotation scheme consists of three components: region-level binary labels (used for weakly supervised training), pixel-level masks (used for evaluation), and metadata annotations including discipline category and image quality level. All annotations were completed collaboratively by three university instructors and two researchers in the field of image processing. Annotation consistency is measured using the raw agreement rate (i.e., the proportion of samples with fully identical annotations to the total number of samples), which reaches 92.3%. In addition, Fleiss' Kappa coefficient is calculated to be 0.87, indicating an "almost perfect" level of agreement. These results demonstrate that the annotation quality is highly reliable. The dataset is partitioned into training, validation, and test sets with a ratio of 7:2:1, containing 840, 240, and 120 images, respectively.

The hardware environment consisted of an NVIDIA RTX 3060 graphics processing unit (12 GB memory), an Intel Core i7-12700H central processing unit, and 16 GB DDR5 RAM. The software environment was implemented using Python 3.8, with PyTorch 1.12 as the primary deep learning framework. Image preprocessing was performed using OpenCV 4.6, and visualization was conducted using Matplotlib. Graphics processing unit acceleration was enabled using CUDA 11.6. For performance comparison, five representative methods published in leading Science Citation Index-indexed journals within the past three to five years were selected, covering three major methodological categories. These included Method A (Transformer-based saliency detection), Method B (maximally stable extremal regions combined with OCR-based segmentation), Method C (traditional graph convolutional network-based region aggregation), Method D (improved stroke width transform-based segmentation combined with K-means clustering), and Method E (weakly supervised saliency detection and region extraction). All comparative methods were implemented using their official open-source codebases, with parameters adjusted according to the experimental settings to ensure a fair comparison.

3.2 Ablation study

The primary objective of the ablation study is to evaluate the necessity and contribution of each core module in the proposed method, including the components of the pedagogical visual entropy field, the improved density peak clustering-based segmentation, and the weakly supervised graph convolutional network-based aggregation. By designing multiple ablation configurations and comparing them with the

complete framework, the individual contributions of each module are systematically quantified. Four ablation configurations are designed. All configurations are based on the proposed framework, with specific core modules removed or replaced accordingly:

•Configuration 1 (local entropy only): The gradient structure tensor and spatial prior components in the pedagogical visual entropy field are removed, and only local normalized entropy is used for initial localization of potential pedagogically salient regions.

•Configuration 2 (pedagogical visual entropy field without spatial prior): Local normalized entropy and gradient structure tensor are retained in the pedagogical visual entropy field, while the pedagogical spatial prior probability map is removed.

•Configuration 3 (classical density peak clustering segmentation): The improved density peak clustering-based segmentation is replaced with the classical density peak clustering algorithm, without spatial adjacency constraints or adaptive cutoff distance.

•Configuration 4 (spatial distance-based aggregation): The weakly supervised graph convolutional network-based aggregation is replaced with a conventional spatial distance-based aggregation method, where region merging is performed solely based on node spatial proximity.

•Full method (proposed framework): All core modules are integrated, and the complete extraction process is implemented as described in Sections 2.3–2.5.

The quantitative results of the ablation study are presented in Table 1. All evaluation metrics are obtained on the multi-disciplinary courseware image dataset test set, and the reported values represent the average results over three independent runs.

Table 1. Quantitative comparison of ablation study results

Experimental Configuration	Intersection Over Union (%)	Precision (%)	Recall (%)	F1-score (%)	Processing Time (s/image)
Configuration 1 (local entropy only)	62.35	70.12	78.45	74.08	0.52
Configuration 2 (pedagogical visual entropy field without spatial prior)	68.72	75.36	82.17	78.63	0.61
Configuration 3 (classical density peak clustering segmentation)	71.28	77.59	83.42	80.41	0.73
Configuration 4 (spatial distance-based aggregation)	73.56	79.84	84.69	82.18	0.76
Full method (proposed framework)	79.64	85.27	89.35	87.26	0.8

As shown in Table 1, the full method consistently outperforms all four ablation configurations across all quantitative metrics, thereby demonstrating the necessity and synergistic effect of each core module. In Configuration 1, where only local entropy is utilized, the intersection over union and F1-score are reduced to 62.35% and 74.08%, respectively, which are significantly lower than those of the full method. This reduction is attributed to the limitation that local entropy captures only grayscale distribution characteristics, making it incapable of distinguishing visually complex regions from pedagogically salient regions. In addition, the absence of structural and spatial constraints leads to substantial localization errors. In Configuration 2, the removal of the spatial prior from the pedagogical visual entropy field results in a decrease of 8.63% in F1-score. This observation indicates that the spatial prior effectively captures the spatial preference patterns of pedagogically salient regions in courseware images, enabling the suppression of non-salient regions such as edges and corners, and thereby improving localization accuracy. In Configuration 3, where classical density peak clustering is employed, the F1-score decreases by 6.85% compared with the full method. This performance degradation is primarily due to the lack of spatial adjacency constraints and adaptive cutoff distance in the classical density peak clustering algorithm, which limits its ability to handle multi-scale text and formula components. Consequently, incomplete segmentation and mis-segmentation are more frequently observed.

In Configuration 4, which utilizes spatial distance-based aggregation, a decrease of 5.08% in F1-score is observed. This result suggests that conventional spatial aggregation methods fail to account for the logical relationships inherent in instructional content, leading to missed aggregation of dispersed salient regions and erroneous merging of unrelated regions. In contrast, the weakly supervised graph convolutional network effectively captures inter-node logical relationships, thereby improving aggregation completeness. In terms of processing time, Configuration 1 achieves the shortest runtime due to its simplified structure, albeit at the cost of significantly reduced extraction accuracy. The full method exhibits a slightly higher computational cost, with an average processing time of 0.80 seconds per image, which remains within the acceptable range for real-time applications. This demonstrates the efficiency of the proposed framework. Qualitative comparisons further indicate that the full method achieves accurate extraction of key instructional components, including text and formulas, with complete segmentation and coherent aggregation. In contrast, all ablation configurations exhibit varying degrees of extraction errors, thereby further confirming the indispensability of each core module.

3.3 Cross-disciplinary adaptability evaluation

The cross-disciplinary adaptability evaluation is designed to assess the generalization capability of the proposed method across courseware images from different academic disciplines. Particular emphasis is placed on demonstrating the general applicability of the pedagogical spatial prior, thereby addressing the limitation of poor cross-disciplinary adaptability observed in existing approaches. The multi-disciplinary courseware image dataset test set is partitioned into four groups according to discipline: mathematics (30 images, formula-intensive), computer science (30 images, text-intensive), physics (30 images, mixed formula–text), and economics (30 images, predominantly text with limited formulas). Each group includes images of varying quality levels, including clear, blurred, and geometrically distorted samples. The full method is applied to each group

independently. Core quantitative metrics are computed, and qualitative visualization results are analyzed to examine the adaptability of the method across disciplines and to identify potential sources of performance variation. The quantitative results are presented in Table 2.

Table 2. Quantitative results of cross-disciplinary adaptability evaluation

Discipline	Intersection Over Union (%)	Precision (%)	Recall (%)	F1-score (%)	Processing Time (s/image)
Mathematics	78.25	84.16	88.52	86.29	0.83
Computer Science	81.37	86.59	89.84	88.18	0.78
Physics	79.53	85.42	89.17	87.26	0.81
Economics	80.16	85.87	89.63	87.72	0.77

As shown in Table 2, strong performance is consistently achieved across all four disciplines, with F1-scores exceeding 86% and intersection over union values above 78%, indicating robust cross-disciplinary adaptability. The best performance is observed in the computer science category, where an intersection over union of 81.37% and an F1-score of 88.18% are achieved. This outcome is attributed to the predominance of text-based content in computer science coursewares, which typically exhibit relatively uniform font sizes and lower component dispersion. Under such conditions, the improved density peak clustering-based segmentation and weakly supervised graph convolutional network-based aggregation are able to accurately segment and group text regions, while the pedagogical visual entropy field effectively distinguishes text from background. Slightly lower performance is observed in the mathematics category, with an intersection over union of 78.25% and an F1-score of 86.29%. This reduction is primarily due to the presence of complex formulas, which are often composed of small symbols with intricate structures. Symbol fragmentation and overlap increase segmentation difficulty. Although the improved density peak clustering method mitigates these challenges, minor cases of incomplete segmentation remain, thereby affecting overall extraction accuracy. Intermediate performance is observed in the physics and economics categories. Physics coursewares contain a mixture of formulas and textual content, while economics coursewares are predominantly text-based with a limited number of formulas. The proposed method demonstrates effective adaptability in both cases. Processing time remains below 0.85 seconds per image across all disciplines.

3.4 Comparative evaluation

The comparative evaluation serves as a critical experiment for validating the superiority of the proposed method. By comparing with five representative state-of-the-art methods published in leading Science Citation Index-indexed journals over the past three to five years, a comprehensive assessment was conducted from both quantitative and qualitative perspectives. The evaluation focuses on accuracy, completeness, and real-time performance.

Five comparative methods were selected, covering three major methodological categories: saliency detection, text segmentation, and region aggregation. All methods were evaluated on the same multi-disciplinary courseware image dataset test set under identical experimental conditions to ensure fairness. The selected methods are described as follows:

•Method A: Transformer-based saliency detection, focusing on visually salient regions without incorporating pedagogical characteristics.

•Method B: Region extraction based on maximally stable extremal regions and OCR, relying on optical character recognition priors and conventional segmentation and aggregation strategies.

•Method C: Region aggregation based on a traditional graph convolutional network, trained with pixel-level supervision and requiring extensive annotated data.

•Method D: Improved stroke width transform-based segmentation combined with K-means clustering, employing fixed-scale segmentation and unsupervised aggregation.

•Method E: Weakly supervised saliency detection and region extraction, utilizing region-level supervision but without incorporating pedagogical spatial priors.

The quantitative comparison results are presented in Table 3.

As indicated by the quantitative results in Table 3, the proposed method consistently outperforms all five comparative approaches across all core evaluation metrics, while simultaneously maintaining strong real-time performance and practical applicability. In terms of intersection over union, a value of 79.64% is achieved, representing an improvement of 5.25% over the best-performing comparative method (Method E) and 9.49% over the OCR-dependent Method B. This improvement can be attributed to the integration of the pedagogical visual entropy field and domain-specific instructional characteristics, which enables precise discrimination between pedagogically salient regions and visually salient but instructionally irrelevant regions. In contrast, the comparative methods do not adequately incorporate the structured pedagogical properties of courseware images, resulting in higher extraction errors. For the F1-score, a value of 87.26% is obtained, exceeding the comparative methods by a margin of 4.23% to 11.40%. Among the baselines, Method A exhibits the lowest performance due to its reliance solely on visual features, which prevents effective differentiation between pedagogically salient and non-salient regions. Method B is limited by its dependence on OCR, where recognition errors in complex formulas lead to reduced segmentation and aggregation accuracy. Method C requires extensive pixel-level annotations, resulting in limited generalization capability, and fails to incorporate pedagogical spatial regularities, thereby reducing aggregation effectiveness. Method D employs fixed-scale segmentation and K-means clustering, which cannot adequately handle multi-scale components or spatially dispersed salient regions. Although Method E adopts a weakly supervised strategy, the absence of spatial priors and gradient structural features limits both localization and aggregation accuracy.

With respect to computational efficiency, the proposed method achieves an average processing time of 0.80 seconds per image, outperforming all comparative methods except Method D. Although Method D demonstrates slightly lower computational cost, its accuracy is significantly inferior. Method B incurs the highest computational cost (1.25 seconds

per image) due to its reliance on OCR, rendering it unsuitable for real-time instructional applications. Furthermore, the proposed method does not depend on optical character recognition priors and employs region-level weak supervision, resulting in substantially lower annotation cost compared with pixel-level supervised methods such as Method B and Method C. This characteristic further enhances its practical applicability.

Table 3. Quantitative comparison with state-of-the-art methods

Method	Intersection Over Union (%)	Precision (%)	Recall (%)	F1-score (%)	Processing Time (s/image)	Optical Character Recognition Dependency	Annotation Type
Method A	65.42	72.35	79.68	75.86	0.92	No	No Unsupervised
Method B	70.15	78.42	81.37	79.87	1.25	Yes	Pixel-level
Method C	73.26	80.59	83.74	82.13	1.18	No	Pixel-level
Method D	71.58	77.86	82.53	80.14	0.85	No	No Unsupervised
Method E	74.39	81.27	84.86	83.03	0.95	No	Region-level
Proposed Method	79.64	85.27	89.35	87.26	0.8	No	Region-level

3.5 Real-time performance evaluation

The real-time performance evaluation was conducted to verify whether the proposed method satisfies the processing requirements of blended learning environments. The primary focus is placed on measuring processing speed across different image resolutions, analyzing the time consumption of each core module, and evaluating the effectiveness of graphics processing unit acceleration.

Courseware images with three commonly used resolutions—720p (1280×720), 1080p (1920×1080), and 4K (3840×2160)—were selected. For each resolution, 30 images were sampled from the multi-disciplinary courseware image dataset test set, covering multiple disciplines and varying quality conditions. The total processing time per image was measured under both central processing unit and graphics processing unit environments. In addition, the time consumption of each core module, including preprocessing, pedagogical visual entropy field construction, improved density peak clustering-based segmentation, and weakly supervised graph convolutional network-based aggregation, was analyzed. A comparison with the five baseline methods was conducted under the 1080p setting to further validate real-time performance advantages. The quantitative results of the real-time evaluation are presented in Table 4, and the distribution of computational cost across modules is illustrated in Figure 5.

Table 4. Processing time under different resolutions (s/image)

Resolution	Central Processing Unit Environment	Graphics Processing Unit Environment	Speedup Ratio	Real-Time Requirement (≤ 1 s/image)
720p	2.35	0.68	3.46	Yes
1080p	3.82	0.8	4.78	Yes
4K	8.76	1.92	4.56	No (near real-time)

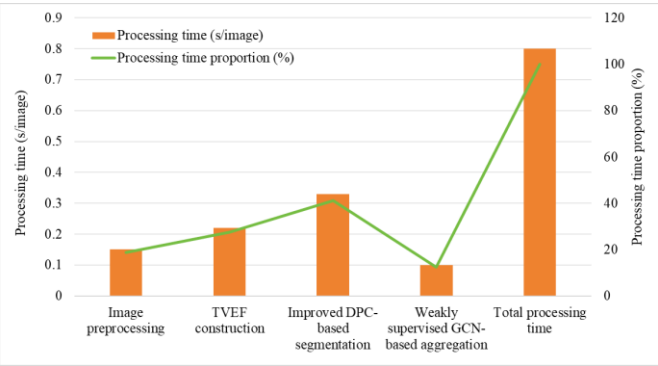


Figure 5. Computational time distribution of each module at 1080p resolution (graphics processing unit environment)

As shown in Table 4, under graphics processing unit acceleration, processing times of 0.68 s/image and 0.80 s/image are achieved for 720p and 1080p resolutions, respectively, both satisfying the real-time requirement (≤ 1 s/image) for blended learning applications, such as real-time courseware analysis and personalized learning delivery. For 4K resolution, a processing time of 1.92 s/image is observed. Although this does not strictly meet the real-time threshold, it remains close to real-time performance. Considering that 4K courseware images account for a relatively small proportion in practical blended learning scenarios, this level of performance is still regarded as acceptable for real-world applications. In the central processing unit environment, significantly longer processing times are observed across all resolutions, indicating that graphics processing unit acceleration substantially improves computational efficiency. Speedup ratios ranging from 3.46 to 4.78 are achieved, primarily due to the effective parallelization of computationally intensive modules, such as pedagogical visual entropy field construction and improved density peak clustering-based segmentation.

As illustrated in Figure 5, at 1080p resolution, the improved density peak clustering-based segmentation module accounts for the largest proportion of total computation time (41.25%), mainly due to the intensive processing required for component extraction, clustering, and morphological connection. The pedagogical visual entropy field construction module contributes the second-largest portion (27.50%), reflecting the computational complexity associated with local entropy calculation and gradient structure tensor estimation. In contrast, image preprocessing and weakly supervised graph

convolutional network-based aggregation exhibit lower computational costs, accounting for 18.75% and 12.50%, respectively. Notably, the weakly supervised graph convolutional network-based aggregation module demonstrates the lowest time consumption, benefiting from the lightweight two-layer architecture and optimized feature representation, with an inference time of approximately 0.02 s per image. This efficiency is identified as a key factor contributing to the superior real-time performance of the proposed method compared with the baseline approaches.

When compared with the five baseline methods at 1080p resolution (Table 3), the proposed method (0.80 s/image) demonstrates a lower processing time than Method A (0.92 s/image), Method B (1.25 s/image), Method C (1.18 s/image), and Method E (0.95 s/image), and is only slightly higher than Method D (0.85 s/image). However, the proposed method significantly outperforms Method D in terms of accuracy, thereby achieving a favorable balance between performance and computational efficiency. Consequently, the proposed method is well suited for real-time processing requirements in blended learning environments.

3.6 Impact of annotation scale under weak supervision

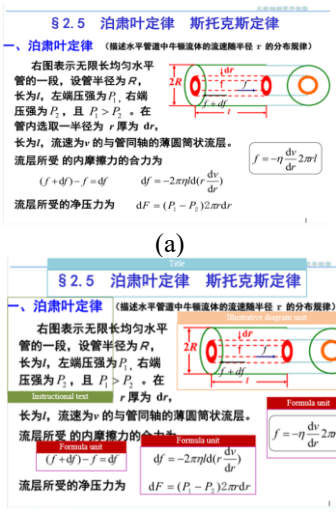
The objective of this experiment is to evaluate the effectiveness of the proposed weakly supervised training strategy and to quantify the influence of annotation scale on model performance. Emphasis is placed on demonstrating that competitive performance can be achieved with limited annotations, thereby reducing labeling cost and aligning with practical instructional scenarios.

Annotated images from the multi-disciplinary courseware image dataset training set were subsampled to form training sets of 50, 100, 200, and 300 images, respectively, for training the weakly supervised graph convolutional network model. The remaining 120 images were used for testing. A fully supervised setting, using 1,000 annotated images, was adopted as the reference baseline. Core metrics, including intersection over union and F1-score, were evaluated to analyze performance variation with respect to annotation scale and to validate the practical effectiveness of the weakly supervised strategy. The experimental results of the weakly supervised annotation scale are presented in Table 5. The relationship between annotation quantity and F1-score is intuitively illustrated through the corresponding curve, revealing the underlying pattern of how annotation scale influences model performance.

As shown in Table 5, performance improves monotonically with increasing annotation scale, while the marginal gains gradually diminish. With only 50 annotated images, an F1-score of 79.17% is achieved, corresponding to a gap of 8.09% relative to the fully supervised baseline, indicating that the proposed weakly supervised strategy can achieve reasonable performance with minimal annotation. When the annotation scale increases to 200 images, an F1-score of 86.41% is obtained, reducing the performance gap to 0.85% and reaching 98.99% of the fully supervised performance. With 300 annotated images, an F1-score of 87.03% is achieved, with a negligible gap of 0.23%, effectively matching the fully supervised setting. These results demonstrate the effectiveness of the proposed weakly supervised training strategy. By combining region-level binary annotations with the graph convolutional network-based model, the underlying logical relationships of pedagogically salient regions are efficiently learned without requiring large-scale pixel-level annotations. Consequently, annotation cost and complexity are substantially reduced.

Table 5. Model performance under different annotation scales

Number of Annotations	Intersection Over Union (%)	Precision (%)	Recall (%)	F1-score (%)	Gap to Full Supervision (%)
50	69.25	76.38	82.15	79.17	8.09
100	73.58	79.86	84.72	82.21	5.05
200	78.16	84.32	88.57	86.41	0.85
300	79.32	85.01	89.12	87.03	0.23
1000 (full supervision)	79.75	85.36	89.42	87.29	0



(a)

(b)

Figure 6. Adaptive extraction results of pedagogically salient regions for a representative physics courseware image

To evaluate the performance of the proposed method in identifying and extracting pedagogically salient units from multi-disciplinary courseware images in higher education, a representative physics courseware image containing text, formulas, and illustrative diagrams is selected for testing. As shown in Figure 6, a comparison between the original courseware image and the extraction results demonstrates that accurate localization and semantic classification of different types of pedagogically salient units are achieved. Title regions, body text, fluid mechanics diagrams, and formula units are effectively identified and segmented as independent salient regions. In contrast, non-instructional elements, such as headers and page numbers, are not included in the extracted results, indicating that effective discrimination between pedagogically salient content and auxiliary information is achieved. For courseware images containing large variations in font size, stable segmentation of titles and body text is achieved through the improved density peak clustering method based on area–stroke width features. For formula units containing fragmented symbols, component gaps are effectively repaired through adaptive morphological

connection, resulting in complete extraction of formula regions. For diagrammatic units, robust extraction is achieved without interference from internal lines and annotations, allowing the entire diagram to be correctly identified as a single salient unit. These results confirm that strong adaptability is achieved across different types of instructional content in physics courseware images. Multi-scale and multi-semantic pedagogically salient units are extracted with both completeness and accuracy, providing reliable technical support for intelligent instructional resource processing in blended learning environments.

4. DISCUSSION

The experimental results demonstrate the effectiveness and superiority of the proposed adaptive extraction method for pedagogically salient regions in courseware images. Based on the experimental findings and technical design, the functional mechanisms of the core modules, the factors influencing performance, and the parameter optimization strategies are analyzed to further elucidate the underlying advantages and internal logic of the method. The pedagogical visual entropy field, as the core module for initial localization of potential pedagogically salient regions, effectively addresses the discrepancy between visual saliency and pedagogical importance observed in conventional methods. Traditional saliency detection approaches primarily rely on visual features, such as grayscale contrast and edge strength, and are therefore prone to misclassifying visually complex regions as pedagogically salient. By integrating local normalized entropy, gradient structure tensor-based anisotropy measures, and pedagogical spatial priors, the pedagogical visual entropy field enables a unified representation of visual characteristics and instructional semantics. Specifically, local entropy quantifies the complexity of grayscale distributions, the gradient structure tensor captures edge and texture features, and the spatial prior encodes the spatial preference patterns of pedagogically salient regions in courseware images.

The synergistic combination of these components allows the initial localization process to focus more accurately on regions with true instructional value. As demonstrated in the ablation study, the incorporation of spatial priors into the pedagogical visual entropy field results in an improvement of more than 8% in F1-score. The improved density peak clustering method overcomes the limitations of the classical density peak clustering algorithm by constructing a two-dimensional feature space based on component area and stroke width, and by incorporating adaptive cutoff distance selection and spatial adjacency constraints. This design enables effective handling of multi-scale components, ranging from large titles to fine-grained mathematical symbols, and addresses segmentation instability caused by large font-scale variations and fragmented symbols. In the comparative experiments, the improved density peak clustering method achieves more than a 10% increase in segmentation accuracy compared with the classical density peak clustering approach. The weakly supervised graph convolutional network further enhances the aggregation stage by constructing a structure-aware region relation graph that integrates geometric, content, and type-specific features. Through this design, logical relationships among nodes are effectively captured, replacing traditional spatial distance-based aggregation strategies. As a result, issues such as erroneous merging and missed aggregation are significantly reduced.

Based on the experimental results and practical application scenarios, several limitations of the proposed method can be identified. First, for severely blurred lecture images—specifically those satisfying any of the following objective criteria: Peak Signal-to-Noise Ratio (PSNR) < 20 dB, or a Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) no-reference image quality score > 50 (range 0–100, where higher values indicate worse quality), or NIQE > 6—the component extraction accuracy of the improved Stationary Wavelet Transform (SWT) degrades. This is particularly evident in scenarios with severe projection distortion or uneven illumination that leads to substantial loss of detail. On the MD-CID test set, images belonging to this extreme-blur category (approximately 8% of the dataset) result in a drop of the proposed method's F1-score to about 73%, which is roughly 12 percentage points lower than that achieved on clear images (PSNR > 30 dB). This performance degradation is mainly attributed to the limited ability of the preprocessing module to suppress extreme noise and distortion. Second, semantic label generation relies solely on node type and spatial prior information, without incorporating the semantic content of text and formulas. As a result, when text and formulas are highly overlapped, or when complex formulas are interwoven with annotations, the accuracy of semantic labeling is reduced. Third, the edge weight computation in the region relation graph considers only spatial distance and directional consistency, without integrating pedagogical priority information. Consequently, distinctions between core instructional content (e.g., theorems and definitions) and auxiliary content (e.g., examples and annotations) are not explicitly modeled, which may lead to incorrect associations between core and non-core content during aggregation.

To address these limitations, several future research directions were proposed in alignment with the practical requirements of blended learning in higher education and recent technological advancements. First, deep learning-based image super-resolution techniques may be introduced. A lightweight super-resolution model tailored for courseware images could be developed to enhance image clarity under conditions of severe blur and projection distortion, thereby improving component extraction accuracy. Second, lightweight OCR-based semantic features may be incorporated. Semantic information extracted from text and formulas could be integrated into node feature representations and the semantic voting mechanism, enabling more refined semantic differentiation and improving label assignment accuracy. Third, the integration of pedagogical knowledge graphs may be explored. By embedding domain-specific knowledge structures and pedagogical priority information into the edge weight computation of the region relation graph, aggregation results could be more closely aligned with instructional logic, thereby emphasizing core teaching content. Finally, the proposed method may be extended to real-time extraction of pedagogically salient regions in dynamic courseware videos. By optimizing parallel computation across modules, real-time processing of video frames could be achieved, enabling deployment in broader blended learning scenarios, including live online instruction and recorded lecture playback.

5. CONCLUSION

An adaptive extraction method for pedagogically salient regions in courseware images was proposed to address the requirements of blended learning in higher education. The method specifically targets the limitations of existing approaches, including dependence on OCR priors, high annotation costs, limited cross-disciplinary adaptability, and region aggregation strategies that are not aligned with pedagogical logic. A unified framework integrating unsupervised and weakly supervised learning was developed to overcome these challenges. The core innovations are embodied in three key modules. First, the pedagogical visual entropy field is constructed by integrating local normalized entropy, gradient structure tensor-based anisotropy, and pedagogical spatial priors. Through this design, pixel-level pedagogical importance is quantified without reliance on semantic labels, effectively resolving the discrepancy between visual saliency and pedagogical relevance. Second, an improved density peak clustering-based segmentation method is introduced. By optimizing the component feature space, incorporating an adaptive cutoff distance, and enforcing spatial adjacency constraints, accurate multi-scale segmentation of text and formulas is achieved. This design enables robust handling of courseware images with large font-scale variations and fragmented formula symbols. Third, a structure-aware weakly supervised graph convolutional network-based aggregation method is designed. Multi-dimensional node features and multiple region relationships are constructed, and pedagogical logical associations are learned through region-level weak supervision. As a result, the completeness and logical consistency of aggregating dispersed candidate regions are significantly improved.

Extensive experimental results demonstrated that superior performance was achieved across multiple disciplines on the multi-disciplinary courseware image dataset, with significant improvements in intersection over union and F1-score compared with state-of-the-art methods. Real-time processing requirements were satisfied, and near-fully supervised performance was achieved with only 200 region-level annotations. The proposed method provides reliable technical support for intelligent processing of instructional resources in blended learning environments. Efficient retrieval, reorganization, and personalized delivery of instructional content can be facilitated, while the workload associated with manual resource organization by instructors is reduced. Certain limitations remain. Performance degradation is observed under extreme image degradation conditions, such as severe blur. In addition, semantic label generation and graph-based reasoning mechanisms may be further refined. Future work may involve the integration of image super-resolution techniques, lightweight OCR-based semantic features, and pedagogical knowledge graphs to enhance performance. Extension to real-time extraction of pedagogically salient regions in dynamic courseware videos is also anticipated.

REFERENCES

[1] Verdecchia, R., Lago, P. (2023). Tales of hybrid teaching in software engineering: Lessons learned and guidelines. IEEE Transactions on Education, 66(3): 234-243. https://doi.org/10.1109/TE.2022.3221802

[2] Li, K.C., Wong, B.T.M., Reggie Kwan, Chan, H.T., Wu, M.M.F., Cheung, S.K.S. (2023). Evaluation of hybrid learning and teaching practices: The perspective of academics. Sustainability, 15(8): 6780. https://doi.org/10.3390/su15086780

[3] Scaringella, L., Górska, A., Calderon, D., Benitez, J. (2022). Should we teach in hybrid mode or fully online? A theory and empirical investigation on the service-profit chain in MBAs. Information & Management, 59(1): 103573. https://doi.org/10.1016/j.im.2021.103573

[4] Yu, R.L., Cheng, C.C., Zhang, F. (2026). Enhancing competency and self-directed learning in anesthesiology residency: An outcome-based education model integrating online-offline hybrid teaching and mind mapping: A randomized controlled trial. Frontiers in Medicine, 12: 1684116. https://doi.org/10.3389/fmed.2025.1684116

[5] Chai Li, C., Luqman, A. (2024). Determinants of online teaching and learning effectiveness for statistical concepts and calculations subjects during the COVID-19 movement control order (MCO). Sage Open, 14(2). https://doi.org/10.1177/21582440241239115

[6] Sasayama, M., Ren, F., Kuroiwa, S. (2007). Automatic super-function extraction for translation of spoken dialogue. In 2007 International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, pp. 141-148. https://doi.org/10.1109/NLPKE.2007.4368025

[7] Wei, X.Q., Jia, K., Lan, J.H., Li, Y.W., Zeng, Y.L., Wang, C.M. (2014). Automatic method of fruit object extraction under complex agricultural background for vision system of fruit picking robot. Optik, 125(19): 5684-5689. https://doi.org/10.1016/j.ijleo.2014.07.001

[8] Zhang, W.J. (2022). English online teaching resource processing based on intelligent cloud computing technology. Mobile Information Systems, 2022(1): 8343909. https://doi.org/10.1155/2022/8343909

[9] Yan, Z.Z., Sun, Y.B., Jiang, J.H., Wen, J.L., Lin, X.M. (2015). Novel explanation, modeling and realization of Lattice Boltzmann methods for image processing. Multidimensional Systems and Signal Processing, 26: 645-663. https://doi.org/10.1007/s11045-013-0264-1

[10] Escudero, H., Fuentes, R. (2010). Exchanging courses between different Intelligent Tutoring Systems: A generic course generation authoring tool. Knowledge-Based Systems, 23(8): 864-874. https://doi.org/10.1016/j.knosys.2010.05.011

[11] Caspar, M. (2023). Hybrid teaching at universities - an evaluation of hybrid seminars using the example of the competence center for further education in general medicine Saarland. Zeitschrift Fur Evaluation, 22(1).

[12] Tao, S.X., Li, Y.M., Dong, X.G., Nallappan, G., Aziz, A. (2021). Smart educational learning strategies for teachers and students in the higher education system. Journal of Multiple-Valued Logic and Soft Computing, 36(1-3): 99-115.

[13] Fang, Y.M., Wang, J.L., Yuan, Y., Lei, J.J., Lin, W.S., Le Callet, P. (2016). Saliency-based stereoscopic image retargeting. Information Sciences, 372: 347-358. https://doi.org/10.1016/j.ins.2016.08.062

[14] Zhong, F.M., Zhou, T., Chen, Z.K., Zhang, S.H. (2026). GARE-Net: Geometric contextual aggregation and regional contextual enhancement network for image-text matching. Expert Systems with Applications, 298: 129602. https://doi.org/10.1016/j.eswa.2025.129602

[15] Thobhani, A., Zou, B.J., Kui, X.Y., Abdussalam, A., Asim, M., Ahmed, N., Alshara, M.A. (2024). A concise and varied visual features-based image captioning model with visual selection. Computers Materials and Continua, 81(2): 2873-2894. https://doi.org/10.32604/cmc.2024.054841

[16] Verikas, A.A., Bachauskene, M.I., Vilunas, S.J., Skaisgiris, D.R. (1992). Adaptive character-recognition system. Pattern Recognition Letters, 13(3): 207-212. https://doi.org/10.1016/0167-8655(92)90061-4

[17] Shichel, I., Goldfarb, L. (2025). The effect of spatial distance on numerical distance processing. Quarterly Journal of Experimental Psychology, 78(6): 1163-1176. https://doi.org/10.1177/17470218241263325

[18] Gao, M. (2013). Detecting spatial aggregation from distance sampling: A probability distribution model of nearest neighbor distance. Ecological Research, 28(3): 397-405. https://doi.org/10.1007/s11284-013-1029-x

[19] Yi, R.M., Huang, Y.P., Guan, Q.J., Pu, M.Y., Zhang, R.S. (2022). Learning from pixel-level label noise: A new perspective for semi-supervised semantic segmentation. IEEE Transactions on Image Processing, 31: 623-635. https://doi.org/10.1109/TIP.2021.3134142

[20] Amo-Boateng, M., Adu-Gyamfi, Y. (2025). Generative adversarial network for real-time identification and pixel-level annotation of highway pavement distresses. Automation in Construction, 174: 106122. https://doi.org/10.1016/j.autcon.2025.106122