Using SAM to Extract Semantic Boundaries in Complex Artistic Images to Assist in the Optimization of Graphic Layout Design

Shuo Liu
School of Visual Arts Design , Hubei Institute of Fine Arts, Wuhan, Hubei, 430205, China
Corresponding author: Shuo Liu (e-mail: lius1117@126.com)

ABSTRACT Although the Segment Anything Model (SAM) exhibits strong capability in visual boundary extraction, its limited high-level semantic understanding restricts its application to complex artistic images and intelligent design-oriented scenarios. This limitation is particularly evident in tasks requiring semantic interpretation beyond low-level edge detection. To address this issue, this study proposes a semantically enhanced segmentation framework that integrates textual prompts with design priors to improve image decomposition and structural understanding. The proposed framework provides end-to-end support from semantic image analysis to layout generation through interactive optimization and hierarchical boundary construction. Experimental results demonstrate that the method achieves an mIoU of 0.78 and a pixel accuracy of 0.95 on the ADE20K validation set. For visual center recognition and text avoidance, which are critical indicators of layout quality, the proposed approach attains a visual overlap of 0.82 and a text avoidance accuracy of 0.91, with an expert evaluation score of 4.2. Compared with baseline methods, the framework delivers superior segmentation performance, layout rationality, and operational efficiency. Beyond graphic design applications, the proposed semantic enhancement strategy provides an effective computational paradigm for multimodal visual information processing and electromagnetic imaging interpretation, offering valuable methodological insights for intelligent sensing systems and advanced image analysis in engineering applications.

INDEX TERMS segment anything model, semantic boundary extraction, multimodal visual information processing, intelligent sensing, graphic layout optimization

I. INTRODUCTION

WITH the rapid development of digital design, complex artistic images are increasingly widely used in various typography scenes, ranging from advertising design and bookbinding to web interfaces. This ubiquitous role of the artistic image as a core carrier of conveying visual information and emotional value is equally evident [1,2]. However, most of the traditional image processing methods fail to capture the semantic structure of the abstract and irregular artistic images accurately and invariably necessitate substantial manual intervention during the typing process, which leads to low efficiency and difficulties in guaranteeing the aesthetic consistency of the design results [3,4]. Although early AI-related methods achieved certain advancement on image segmentation tasks, due to the limitations of fixed semantic categories and their inability to understand abstract artistic expressions, they could not meet the requirements of extracting flexible semantic boundaries in typography design [5,6].

With the revolutionary zero-sample segmentation capability, SAM brought new possibilities to the semantic understanding of complex images. Therefore, this model can segment any image at a pixel level according to the user's prompts, represented as points, boxes, and text descriptions, without training data, and theoretically extract boundary information about all visible objects in the image [7]. Bridging the visual-semantic gap has become a key leap from computational analysis to design generation. On the one hand, it is necessary to construct a semantic enhancement model to map the underlying contour probability map of SAM into a "semantic boundary map" [8], integrate composition rules, color emotions and visual movement priors [9,10], and realize the dimensional upgrade from pixel-level edges to design-level semantics; on the other hand, it is necessary to develop an interactive text prompt interface at the practical level, allowing designers to redistribute the weights of the segmentation results through sparse dots or semantic short sentences [11,12]. The model dynamically

learns "design context" rather than "object outline" within the zero-sample framework, and then outputs "typography-ready" semantic units with hierarchical hierarchy, rhythmic breathing and negative space topology. The core objective of this study is to bridge the gap between the underlying visual boundary generated by SAM and the high-level semantic structure required for typography design. Design priors are applied on top of the underlying segmentation of SAM, and the fragments are aggregated into structured semantic units that conform to the typography logic, and further transformed into parameterizable design elements. This paper proposes a "semantic enhancement-structural transformation" framework. This study enhances the SAM's ability to perceive design-related semantics in artistic images through multimodal cue fusion and interactive optimization strategies.

II. SEMANTIC BOUNDARY OPTIMIZATION AND HIERARCHICAL CONSTRUCTION FOR TYPOGRAPHY DESIGN

ANALYSIS OF THE DIRECT APPLICATION AND LIMITATIONS OF SAM IN ARTISTIC IMAGES

SAM, as a general segmentation model, has an architecture mainly consisting of three parts.

(1) Image encoder based on Vision Transformer[13]:

Responsible for converting images into high-dimensional semantic features;

(2) Cue encoder[14]:

Supports the unified encoding of points, boxes, masks, and CLIP (Contrastive Language-Image Pre-training) text prompts into vector representations;

(3) Lightweight mask decoder[15]:

Dynamically outputs segmentation masks by fusing image and cue features;

The SAM model provides all-mode segmentation[16], automatically generates full-image masks, which is prone to redundancy; the interactive prediction mode[17] allows users to specify targets by clicking or text, which is more in line with the extraction needs of specific semantic units in design scenarios.

FRAGMENTED SEGMENTATION RESULTS, DISCONNECTED FROM OVERALL AESTHETIC STRUCTURE

SAM, in its all-encompassing segmentation mode, tends to decompose images into a large number of small semantic units. While this over-segmentation characteristic can capture local details, it destroys the overall aesthetic structure of the artistic image, as shown in Figure 1.

Figure 1 shows two oil paintings containing animals and figures. SAM segments the figures' hair, face, and clothing, and the animals' forelimbs and hindlimbs into separate masks, even identifying patterns on the clothing as individual objects. This fragmented segmentation ignores the holistic nature of the "figure"/"animal" as a complete subject, making it difficult for subsequent layout design to grasp the core visual focus of the image. Many abstract elements in artistic images do not correspond to specific objects in the physical world. SAM's forced segmentation of these elements is meaningless, applying numerous noisy boundaries and interfering with design decisions.

LACK OF HIERARCHICAL RELATIONSHIP REFLECTING DESIGN PRIORITIES AND VISUAL LOGIC

Layout design requires a clear hierarchical relationship of elements that can guide the organization of visual information, such as distinguishing between main elements, secondary elements, and background elements. However, the SAM segmentation result is flat; all extracted masks are semantically equal, failing to reflect the required hierarchy and visual logic of the design. In a landscape illustration, SAM segments the foreground flowers, midground trees, and background mountains into separate masks but cannot recognize that "mountains" as background should be at the bottom in the visual hierarchy, while "flowers" as foreground subjects should be at the top. The lack of hierarchical information in SAM's outputs hinders their direct applicability to layout design, compelling designers to make numerous manual adjustments in segmentation; there are so many manual adjustments designers need to make in order to achieve a satisfying design, which hugely reduces design efficiency.

SEMANTIC ENHANCEMENT SEGMENTATION STRATEGY

This system combines SAM with CLIP text prompts to construct a semantic boundary extraction framework, the working of which is shown in Figure 2.

Figure 2 illustrates the complete workflow of SAM-assisted artistic image typography. First, CLIP text prompts and human visual cues jointly guide SAM to segment the main subject; The mask is refined through lightweight interactive iterations; Boundaries are simplified and smoothed, and SVG (Scalable Vector Graphics) vectors are exported; A hierarchical tree of "subject-secondary-background" is automatically clustered based on features such as area/contrast/position; Finally, the vector paths and hierarchical JSON (JavaScript Object Notation) are embedded into the design software, automatically adapting the layout for visual text wrapping, stacking order, and negative space balance.

MULTIMODAL CUE FUSION

Combining the CLIP model, textual semantic guidance (such as "subject", "background texture", "decorative element") is applied. The multimodal cue fusion strategy aims to compensate for the lack of SAM's understanding of abstract art semantics by applying textual semantic information. The specific implementation method is to combine the CLIP model with SAM, and use CLIP's cross-modal understanding ability to transform design-related textual

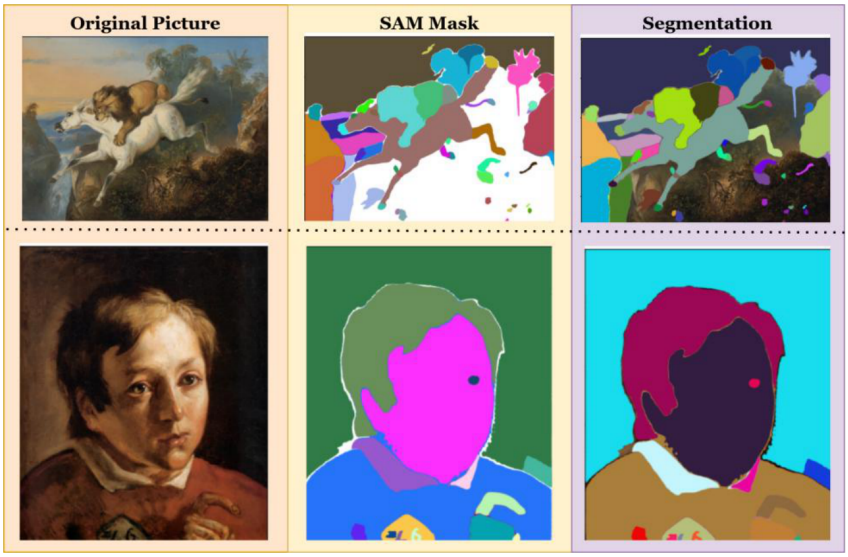

FIGURE 1. Limitations of SAM segmentation

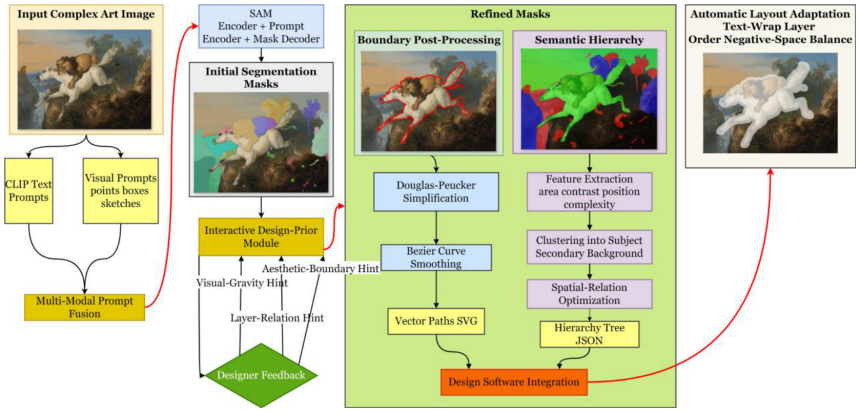

FIGURE 2. The entire process of semantic boundary extraction framework

concepts (such as "subject", "background texture", "decorative element", "visual center") into visual feature cues, guiding SAM to perform segmentation that better meets design requirements. When the user inputs textual cues to extract the main elements in the picture, the system uses the CLIP model to transform the text into a feature vector, and uses this vector as an additional input to SAM, which is then fused with image features to jointly guide the mask generation process. This activates SAM's perception of semantic units with design significance in art images and reduces over-segmentation of irrelevant details. To adapt to the ambiguity and polysemy of art image semantics, multimodal cue fusion adopts a dynamic weight adjustment mechanism [18].

$$w_{\text{text}} = \sigma(aA + b) \tag{1}$$

In formula (1), σ is the Sigmoid function; A represents the image abstraction score (0–1); a and b are empirical coefficients.

Generation of Textual Prompts:

To ensure practicality and efficiency, our system adopts a semi-automatic prompting strategy rather than requiring expert-written prompts per image. Upon image input, the system first analyzes global features and suggests a set of default textual prompts from a predefined template library (e.g., "main subject", "background", "decorative element"). The user can then select, combine, or minimally edit these suggestions via a simple interface to match their specific design intent. For fully automatic operation, the system can directly employ the most relevant default prompt based on a preliminary scene analysis.

INTERACTIVE OPTIMIZATION BASED ON DESIGN PRIORS

In this study, design priors refer to a set of pre-defined compositional rules and data-driven visual features that guide the semantic segmentation and layout optimization process. These priors are integrated into the framework through a hybrid approach: static rules ensure layout stability and adherence to design principles, while dynamic parameters allow adaptation to image-specific semantics via

multimodal fusion and user interaction. The key design priors employed in our system are categorized as follows:

(1) Composition Rules:

Rule of thirds for positioning key elements. Visual center gravity to prioritize the main subject. Symmetry/asymmetry balance for layout harmony.

(2) Visual Hierarchy Principles:

Perceptual importance is estimated through a linear combination of area, contrast with background, and proximity to the image center.

(3) Color Emotion Priors:

Warm/cool tone associations derived from color theory. Saturation and brightness guidelines to maintain emotional consistency with the original artwork.

(4) Text Avoidance Constraints:

No-text zones are automatically defined as the union of masks corresponding to “main” and “secondary” objects in the hierarchical semantic map.

(5) Negative Space Balance:

Minimum white-space thresholds are applied to ensure aesthetic breathing room in the layout, adjusted based on image abstraction score.

These priors are implemented as configurable parameters within the system. Default values are derived from design theory and empirical studies, but users can adjust their weights through an interactive interface or via CLIP-guided semantic reweighting in response to textual prompts. These priors are encoded as a configurable but self-contained rule set. Once configured, they are automatically applied to any input image based on the extracted visual and semantic features, requiring no manual re-specification per image. The interactive optimization strategy based on design priors simulates the thinking logic of designers when analyzing images, providing interactive prompts that conform to design habits and guiding SAM to gradually approach the ideal segmentation result. Unlike traditional point-and-click interaction, the interactive prompt design of this strategy is based on the basic principles of typography design, mainly including:

(1) Visual center of gravity prompts:

Allowing users to click or select the visual center of gravity area in the layout, the system automatically regards the area as a high-priority semantic unit, guiding SAM to prioritize the segmentation of the main elements in the area, which conforms to the designer’s working habit of first determining the focus of the picture.

$$L_v = \|c_m - c_v\|^2 \tag{2}$$

In formula (2), cm represents the center of the predicted mask; cv is the visual centroid coordinate of the user’s click.

(2) Hierarchical Relationship Hints:

Users can define hierarchical relationships between different elements through drag-and-drop operations, transforming these relationships into constraints to guide SAM in adjusting the mask’s containment relationships and occlusion order.

(3) Aesthetic Boundary Hints:

Since most semantic boundaries in artistic images are "perceptual" rather than physical boundaries caused by color gradients, the system lets users directly input aesthetic boundary hints through hand-drawn curves. Based on the shape and position of the hand-drawn curves, a smooth segmentation boundary that matches them is generated.

DESIGN-ORIENTED BOUNDARY POST-PROCESSING AND HIERARCHICAL CONSTRUCTION

BOUNDARY SIMPLIFICATION AND SMOOTHING

Smooth paths were generated using the Douglas-Puk algorithm [19] by iteratively deleting vertices that did not have much effect on the overall shape, which significantly reduced the number of vertices while maintaining the basic shape of the boundary.

$$\epsilon = \epsilon_{\text{base}} + k(1 - A) \tag{3}$$

In formula (3), ϵ_{base} represents the baseline threshold; A is the image abstraction score; and k is the scaling factor. Key parameters of the algorithm can be automatically adjusted according to the degree of abstraction of an image: for figurative images, a smaller threshold can be used in order to keep the features more detailed, while a larger threshold for abstract images can provide a simpler geometric form.

SEMANTIC HIERARCHICAL REASONING

Based on the spatial inclusion relationship, area, and position of the mask, a hierarchical tree of "subject-secondary-background" is automatically constructed to segment the visual attributes and spatial relationships of the mask, automatically building a hierarchical structure that conforms to the layout design logic. This mainly includes the following steps:

Feature extraction:

Multi-dimensional features are extracted for each segmentation mask, including area size, color contrast (brightness difference with the background), position information (distance from the image center), contour complexity (number of boundary vertices), etc. These features comprehensively reflect the visual importance of the semantic unit in the layout.

$$S_i = \alpha\,\text{Area}_i + \beta\,\text{Contrast}_i + \gamma(1 - \text{Dist}_{\text{center},i}) \tag{4}$$

In formula (4), Area_i is the normalized area; Contrast_i represents the difference in brightness with the background; $\text{Dist}_{\text{center},i}$ represents the normalized distance to the center of the image; $\alpha + \beta + \gamma = 1$ is the weighting coefficient.

III. OPTIMIZATION OF SEMANTIC BOUNDARIES IN GRAPHIC TYPOGRAPHY DESIGN

The system further develops and expands the semantic boundary and hierarchical relationship, transforming structured semantic boundary data into practical design value and

achieving a closed loop from image understanding to layout generation. It applies two important strategies, namely, visual center-driven text layout and hierarchical constraint-based negative space balance, to automatically generate layouts that meet the requirements of design aesthetics, which can dramatically improve the efficiency and quality of design in complex artistic image scenarios. The overall graphic layout design framework is shown in Figure 3.

Figure 3 shows the complete optimization path of typography design combined with semantic boundaries. Through three major engines, namely intelligent avoidance and wrapping, composition-aware layout, and stylistic derivation, intermediate results such as dynamic text wrapping, visual flow, decorative graphics and data color matching are generated respectively. All intermediate assets are uniformly imported into the design software interface, and finally the automatically adapted layout is output with one click, realizing the overall optimization of text wrapping, hierarchical order and negative space balance.

INTELLIGENT CONTENT AVOIDANCE AND TEXT WRAPPING

The transition from accurate semantic segmentation to effective layout optimization is core to our framework. The hierarchical semantic boundaries and labels ("subject-secondary-background") directly inform and drive the automated layout process in three key ways:

(1) The semantic labels automatically define functional zones for text placement (e.g., avoidance of "subject" areas) and element positioning;

(2) The precise contour geometry of dominant elements is converted into Bézier curves that guide dynamic text wrapping; and

(3) The computed visual features (area, contrast, position) of each semantic region are used to calculate visual flow and to generate adaptive, non-uniform grids.

Consequently, the accuracy and design-awareness of the initial segmentation are not merely preliminary steps but are deterministic factors for the feasibility and aesthetic quality of the final layout.

DYNAMIC TEXT WRAPPING

By directly leveraging semantic boundaries for the path of text arrangement, text can flow along natural contours of image elements and create an irregular, fluid typography effect. Semantic boundaries are first converted into Bézier curve paths, while distribution of the text on the curve, length, font size, character spacing, are automatically calculated by the system based on text content.

$$k_s = k_0 \cdot (1 - \kappa(s) \cdot w) \tag{5}$$

In equation (5), k_s is the instantaneous character spacing at curve s; k_0 is the baseline character spacing; $\kappa(s)$ represents curvature; and w is the compensation coefficient. In parts of the curve with larger curvature, for example, a protruding part of the object, make the character spacing smaller in such a way that these characters are set closely but not overlapping; in parts of the curve with lesser curvature, for instance, a flat part of an object, increase the character spacing so that the characters are not too sparse. Dynamic text wrapping supports several wrapping modes to adapt to different design needs:

The "Tight Wrap" mode fits the text snugly against the semantic boundary for maximum space utilization;

"Loose Wrap" mode retains some white space between the text and the boundary;

"Outline Path Follow" mode makes the text oriented according to the tangent direction of a Bézier curve, creating a strong sense of movement;

"Inner Fill" mode lays out the text inside the area bounded by the semantic enclosure, which fits the treatment of main elements with closed outlines.

Designers can seamlessly toggle between modes via parameter adjustment, enabling instant preview of diverse typographic effects.

SMART ELEMENT LAYOUT

The automatic planning of the other graphic elements' placement incorporates icons, decorative graphics, and auxiliary text blocks, based on the importance and spatial characteristics of the semantic regions so that balance is achieved in the overall layout. According to the semantic hierarchy tree, it autonomously divides layout areas into a forbidden placement area (the main element is located there), a restricted placement area, and a free placement area–the background area. Based on "visual weight balance" principles, it can position the element in an optimal position in the free placement area.

LAYOUT DESIGN BASED ON COMPOSITIONAL PERCEPTION

VISUAL FLOW ANALYSIS

Identify the main elements and key secondary elements in the semantic hierarchy tree, and extract the center coordinates and contour feature points of these elements; based on the visual organization principle in Gestalt psychology [20], calculate the attraction weights between these feature points [21]:

$$W_{ij} = \exp(-\lambda \cdot d_{ij}/D_{\max}) \cdot L_i \cdot L_j \tag{6}$$

In formula (6), W_{ij} represents the attraction between feature points i and j; d_{ij} represents the Euclidean distance between feature points i and j; $D_{\max}$ represents the diagonal of the page; and $L_i \cdot L_j \in \{0, 1\}$ represents whether it belongs to the main body/key secondary. A directed graph path search algorithm [22] is used to find the optimal visual path from the entrance (usually the upper left corner or visual center) to the exit (usually the lower right corner). This path is the suggested visual flow.

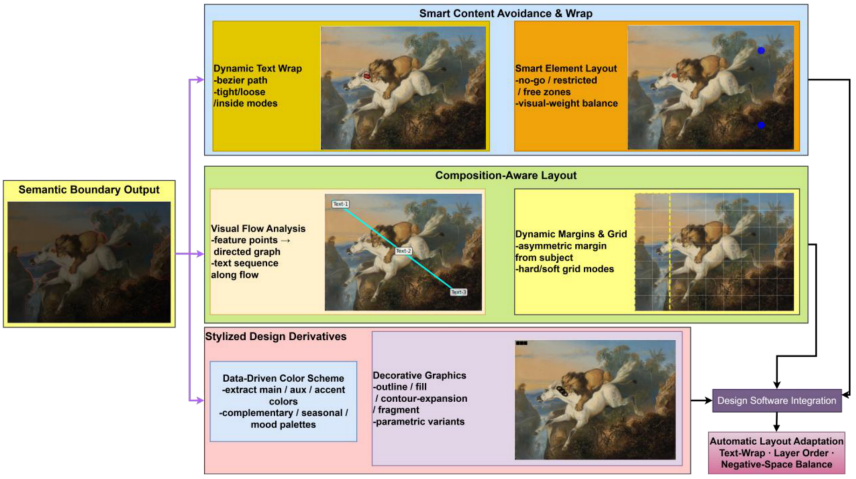

FIGURE 3. Layout Support Framework

DYNAMIC VERSION CORE AND MESH DEFINITION

Traditional typography typically uses fixed rectangular areas for the text area and grid, which are difficult to adapt to the compositional needs of irregular artistic images. Dynamic text area and grid definition technology automatically generates an asymmetrical text area and grid system that matches the inherent rhythm of the image, based on the semantic boundaries of the main elements. The system analyzes the outline shape and position of the main elements, calculates the smallest enclosing area that can contain the main elements while retaining appropriate white space, and uses this area as the dynamic text area. Based on the direction of the outline feature lines of the main elements (e.g., horizontal, vertical, slanted), adaptive grid lines are generated. The density of the grid lines is automatically adjusted according to the complexity of the elements within the area; the grid lines are denser in densely populated areas and sparser in sparsely populated areas.

DERIVATION OF STYLIZED DESIGN ELEMENTS

GENERATION OF DECORATIVE GRAPHICS

The system vectorizes and creatively transforms the core semantic boundaries into reusable decorative graphics. The system provides various algorithms for transformation: The algorithm "Contour Extraction" uses semantic boundaries directly as decorative graphics; The "Internal Fill" algorithm creates closed graphics with gradient or pattern fills depending on the boundary shape. The algorithm "Contour Expansion" provides layered contour graphics by offsetting, thickening and dotting the boundaries; Another approach, the "Fragment Recombination" algorithm, decomposes complex boundaries into a number of simple sub-shapes which get recombined through rotation, scaling, and mirroring to generate abstract decorative patterns. The generated decorative graphics can be used in multiple positions on the page, including headers, footers, title borders, text backgrounds, and corner decorations that form visual elements echoing the main image. The system allows for parametric adjustments of the decorative graphics and thus enables designers to change parameters such as color, transparency, rotation angle, and repetition density to swiftly generate various variations. What's more, the system can automatically analyze the stylistic characteristics of decorative graphics, like whether they are geometric, organic, or have a hand-drawn feel, and can recommend matching fonts and layout styles that correspond to the overall style being used.

DATA-DRIVEN COLOR SCHEME

The system extracts primary and secondary colors from segmented semantic regions to generate harmonious design color schemes. For each semantic region, color analysis is performed to retrieve the primary color of the highest percentage, the secondary color of the second highest percentage, and the accent color with the highest saturation. According to color theory, the colors extracted can go through a selection to create an aesthetically harmonious combination of colors. The system offers several color scheme modes to fit diverse design requirements: "Subject Priority" mode uses the color of the main element as the core, complemented by other colors; "Emotion-Oriented" mode automatically adjusts color saturation and brightness in line with the design theme; "Seasonally Adaptive" mode suggests color combinations that best match the current season.

$$C_{\text{final}} = \frac{\sum_k A_k \cdot S_k \cdot c_k}{\sum_k A_k \cdot S_k} \tag{7}$$

In formula (7), C_{final} represents the primary output color; A_k represents the area of the k-th semantic region; S_k represents saturation; and c_k represents the average RGB (Red, Green, Blue) of the region. Each color scheme includes a primary color for titles and key content, secondary color for secondary content and background, accent color for interactive elements like buttons and icons, and detailed color value parameters to make it easy for designers to apply directly in various design software. Data-driven color schemes ensure that, in terms of color, the connection

between the typographic design and the original artistic image can be inherent, avoiding stylistic fragmentation caused by manual color matching and offering rich creative inspirations to designers.

IV. EXPERIMENTAL VERIFICATION AND RESULT ANALYSIS

EXPERIMENTAL DESIGN AND DATASET CONSTRUCTION

DATASET INTRODUCTION

This study uses the ADE20K dataset (https://ade20k.csail.mit.edu/) as the experimental benchmark. ADE20K contains 150 semantic categories, covering indoor, outdoor, natural and urban scenes, as well as artistic images. These images contain complex occlusion relationships, diverse object scales and layouts, effectively simulating the challenges of complex artistic images in terms of composition and semantic level. Figure 4 shows a dataset that provides pixel-level semantic segmentation, instance segmentation, and background segmentation, which can be used to verify the framework's semantic boundary extraction and hierarchical construction capabilities. For example, it can verify whether the framework can correctly group the various parts of a "chair" (legs, seat) into a complete semantic unit. Furthermore, the deep integration of the dataset with the mainstream deep learning framework PyTorch lowers the barrier to experimental reproduction.

DATA PREPROCESSING

While ADE20K is not a purely artistic image dataset, its complex scene understanding requirements align closely with the core objectives of this study. To serve graphic layout design evaluation, the following preprocessing operations were performed:

Image Selection:

Images from the dataset that demonstrate greater aesthetic appeal and challenge in composition, color, and element layout were prioritized, including scenes with rich foreground-background interaction and clear visual focus and movement.

Annotation Transformation:

The original annotations from ADE20K were transformed into hierarchical design structure annotations suitable for this study.

Design Hierarchy Definition:

Based on object category, spatial relationships, and area size, 150 categories were automatically mapped to a three-level design hierarchy: "Subject-Secondary-Background." For example, large-scale foreground objects such as "people" and "cars" are often defined as subjects; "chairs" and "tables" are defined as secondary elements; and "sky," "walls," and "ground" are defined as background elements.

Generating Design Usability Ground Values:

Based on the above hierarchy, two key design-relevant ground truths were generated:

(1) Visual Center Region:

Five professional designers (with $\geq$3 years of design experience) independently annotated the "visual center" for each image, defined as the primary area that naturally draws viewer attention. The final ground truth was obtained by taking the pixel-wise union of these five annotations. This consensus region serves as the reference for evaluating visual center recognition.

(2) Text Avoidance Region:

This was automatically derived as the union of masks for all instances labeled as "Subject" and "Secondary" in the design hierarchy, representing areas where text placement should be avoided to maintain readability and visual balance.

Training Set:

20,210 images. Used for training the model fine-tuning or cue word learning modules that may be involved in this framework.

Validation Set:

2,000 images. Used for hyperparameter tuning.

Test Set:

2,000 images from the validation set are used for ablation studies and results analysis.

INTRODUCTION TO THE COMPARISON MODEL

The following baseline model was set up for comparison in the experiment.

Table 1 shows five comparison methods. Traditional methods include GrabCut[23] and SLIC (Simple Linear Iterative Clustering)[24] superpixel segmentation. General semantic segmentation models include DeepLabV3+[25] (a semantic segmentation model based on convolutional neural networks) and Mask2Former[26] (an instance segmentation model based on Transformer). SAM basic mode: This is the SAM fully automatic segmentation mode without adding any semantic enhancement or post-processing.

TABLE 1. Introduction to the Comparison Model

Baseline	Algorithm	Description
GrabCut	Graph-cut energy minimization with user box; Gaussian color models	Classic interactive segmentation; tests if deep methods outperform manual-init algorithms.
SLIC	k-means on 5-D color-position space; linear complexity	Representative low-level grouping; evaluates boundary adherence vs. artistic complexity.
DeepLabV3+	ASPP(Atrous Spatial Pyramid Pooling) + encoder-decoder; atrous separable conv; multi-scale	CNN(Convolutional Neural Networks)-based semantic baseline; measures convolutional limit on fine-art boundaries.
Mask2Former	Masked-attention Transformer; query-based instance/semantic	State-of-the-art Transformer instance segmentation; compares latest Transformer design to SAM.
SAM-Baseline	ViT (Vision Transformer)image encoder; promptable mask decoder; zero-shot	SAM without any semantic enhancement; quantifies gain produced by the proposed proposed framework.

EVALUATION INDICATORS

The proposed framework is evaluated using both quantitative segmentation metrics and human-centric design rationality assessments. The following indicators and protocols are employed.

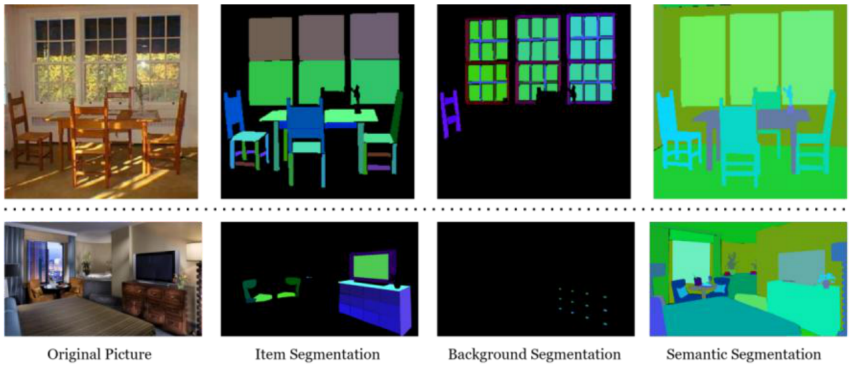

FIGURE 4. Introduction to the image segmentation criteria of the ADE20K dataset

SEGMENTATION ACCURACY METRICS

Pixel precision and mIoU (Mean Intersection over Union) are used as quantitative metrics for segmentation accuracy. These are calculated by comparing the semantic boundaries generated by the system with manually annotated "structured deconstruction annotations." Pixel precision is measured as the proportion of correctly classified pixels out of the total number of pixels.

$$\text{Pixel Accuracy} = \frac{\sum_{i=1}^{N} I(p_i = \hat{p}_i)}{N} \tag{8}$$

In formula (8), p_i represents the artificial label of pixel i; $\hat{p}_i$ is the predicted label of pixel i; and N is the total number of pixels. mIoU calculates the average intersection-union ratio (the ratio of the area of the intersection of the predicted mask and the ground truth mask to the area of the union) for each category. These metrics mainly reflect the extraction accuracy of the underlying visual boundaries.

$$\text{mIoU} = \frac{1}{K} \sum_{k=1}^{K} \frac{|G_k \cap P_k|}{|G_k \cup P_k|} \tag{9}$$

In formula (9), K represents the total number of categories (including background); G_k represents the set of true mask pixels for category K; and P_k represents the set of predicted mask pixels for category K.

DESIGN RATIONALITY INDICATORS

Design rationality indicators evaluate the quality of semantic boundaries from a typography application perspective, including four indicators: expert scoring, consistency of design annotations (measured by Visual-Weight IoU), time efficiency and accuracy of text avoidance (Text-Avoid Acc).

Expert Scoring:

A formal expert evaluation was conducted to holistically assess the layout quality. 50 professional designers (25 graphic designers and 25 designers, each with $\geq$3 years of experience) participated in a double-blind review. Each expert evaluated 20 pairs of layout schemes in a randomized order. Scoring was based on a 5-point Likert scale (1: Very Poor, 5: Excellent) across three predefined dimensions: layout harmony, information clarity, and visual appeal.

$$\text{ExpertScore} = \frac{1}{M} \sum_{m=1}^{M} s_m \tag{10}$$

In formula (10), M represents the number of participating experts; s_m represents the score given by the mth expert.

Design annotation consistency:

Visual-Weight IoU:

This metric quantifies the system's ability to recognize the visual center of an image, which is crucial for determining layout focus. It is calculated as the intersection-over-union between the system-extracted "main subject" region and the human-annotated visual center region. The formula is defined as:

$$\text{Visual-Weight IoU} = \frac{|B_{\text{sys}} \cap B_{\text{human}}|}{|B_{\text{sys}} \cup B_{\text{human}}|} \tag{11}$$

In formula (11), B_{sys} represents the "main" region extracted by the system; B_{human} represents the manually labeled "visual center of gravity" region.

Text-Avoid Acc:

This evaluates whether the text wrapping path automatically generated by the system successfully avoids the manually labeled "text no-go zones," and is measured by calculating the proportion of unavoided pixels to the total no-go zone pixels.

$$\text{Text-Avoid Acc} = 1 - \frac{|T \cap R|}{|T|} \tag{12}$$

In formula (12), T is the set of manually marked "text restricted area" pixels; R is the set of pixels actually occupied by the system text; $T \cap R$ is the number of pixels not avoided.

Time Efficiency Metric:

To quantify the practical utility of the system in accelerating the design workflow, we measure the task completion time. This metric records the total time a designer takes from receiving the original artistic image to producing a satisfactory layout, using a specified system. The reported time reduction percentage is calculated by comparing the average task time when using our full system against the baseline SAM setup. This end-to-end measurement encompasses

all stages: image loading, system interaction, segmentation processing, and final layout adjustment/generation.

Table 2 includes scoring items such as layout harmony (whether the arrangement of elements conforms to the principle of visual balance), clarity of information delivery (whether key information is easy to identify), and visual appeal (the aesthetics and uniqueness of the overall design). After each designer scores independently, the average score is calculated as the final expert score.

TABLE 2. Rating Dimensions

Score	Layout Harmony	Information Clarity	Visual Appeal
5	Excellent: elements are perfectly balanced; no sense of tension or emptiness.	Excellent: key message instantly readable; hierarchy is unmistakable.	Excellent: highly original and beautiful; evokes strong positive reaction.
4	Good: minor imbalance that does not disturb the viewer.	Good: key message clear within 2 s; minor hierarchy issues.	Good: pleasing and somewhat distinctive; above average.
3	Fair: noticeable imbalance but still acceptable.	Fair: key message needs conscious search; average hierarchy.	Fair: neutral look; neither attractive nor unattractive.
2	Poor: clear imbalance; layout feels uncomfortable.	Poor: key message hard to find; weak hierarchy.	Poor: uninteresting or slightly unattractive.
1	Very poor: chaotic layout; impossible to find visual anchor.	Very poor: information buried; no discernible hierarchy.	Very poor: strongly unattractive or clichéd.

TIME EFFICIENCY EVALUATION PROTOCOL

The task completion time is measured to quantify workflow acceleration.

Participants:

100 designers participated, evenly split into experienced ($\geq$5 years) and novice (<2 years) groups.

Task & Design:

Each participant completed 5 distinct layout design tasks for both (a) our Full System and (b) the Baseline SAM setup, in a counterbalanced order to mitigate learning effects.

Measurement:

Time was recorded from the moment the raw image was loaded until the final layout was exported and deemed satisfactory by the participant, encompassing all interaction.

EXPERIMENTAL RESULTS ANALYSIS

The experimental results are shown in Figure 5.

Figure 5 shows the comparison results. The proposed system achieves a mIoU of 0.78 and a pixel accuracy of 0.95, ranking first. Compared to the second-best method, Mask2Former (mIoU 0.75), the IoU is improved by 0.03 and the pixel accuracy by 0.02, indicating that the semantic enhancement module can still significantly optimize boundary localization accuracy in complex artistic scenes. The proposed system simultaneously achieves the best in all three design rationality indicators: expert score of 4.2, an improvement of 1.1 points compared to the SAM baseline; Visual-Weight IoU of 0.82, an improvement of approximately 19% compared to SAM (0.69); and text avoidance accuracy of 0.91, an improvement of 0.05 compared to the SAM baseline (0.86). This simultaneous leap in indicators indicates that the more accurate segmentation mask can be directly converted into a typographically friendly semantic region, receiving unanimous praise from professional designers for its layout balance, information readability, and visual appeal.

ABLATION EXPERIMENT ANALYSIS

To rigorously evaluate the individual contributions of the two core enhancement components (textual prompts and design priors) and to validate the necessity of their integration, we conducted a comprehensive ablation study on the ADE20K validation set. The study systematically isolates the effects of each component by constructing four distinct system configurations. This design allows us to quantify how much performance gain originates from semantic guidance via text, how much from the application of design-specific rules and constraints, and how their combination yields synergistic benefits. The four configurations are defined as follows:

Configuration A (Full System):

Incorporates all proposed components: SAM + Textual Prompts (via CLIP) + Design Priors (Interactive Optimization & Post-processing). This represents our complete framework.

Configuration B (Textual Prompts Only):

Retains SAM and the Textual Prompts module (multimodal cue fusion via CLIP), but removes all Design Priors. This includes disabling the interactive optimization based on visual center/hierarchy hints and the design-oriented boundary post-processing & hierarchical construction. Segmentation results are the raw masks from SAM guided only by text, without any design-specific refinement or structuring.

Configuration C (Design Priors Only):

Retains SAM and the Design Priors modules (interactive optimization and post-processing), but removes the Textual Prompts. The system uses only default point/box cues (e.g., a single central point prompt) to initiate SAM segmentation, after which the full suite of design priors (interactive refinement and structural transformation) is applied. This configuration tests the value of design knowledge without high-level semantic guidance from language.

Configuration D (Baseline SAM):

Uses only the original SAM model in its fully automatic segmentation mode, without any text prompts, design priors, or post-processing. This serves as the baseline to quantify the net gain provided by our enhancements. The comparison of key evaluation metrics for each configuration is shown in Figure 6:

The comparative performance of these four configurations across all five evaluation metrics is presented in Figure 6, which visualizes the results for segmentation accuracy (mIoU and Pixel Accuracy) and design quality (Expert Score, Visual-Weight IoU, and Text-Avoid Acc) in a unified layout. A clear and consistent performance hierarchy emerges from the figure, with each configuration demonstrating distinct strengths aligned with its core components.

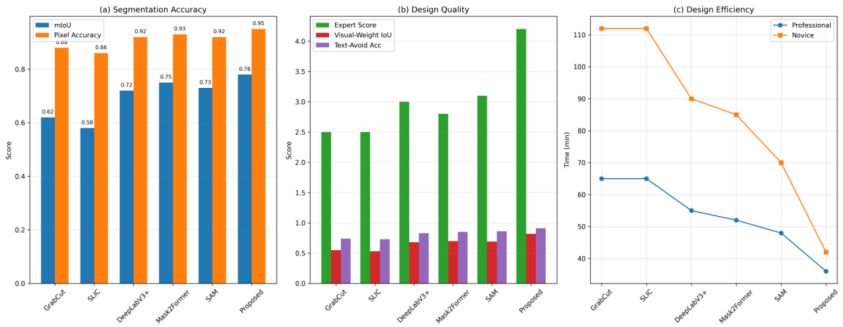

FIGURE 5. Analysis of Model Comparison Results

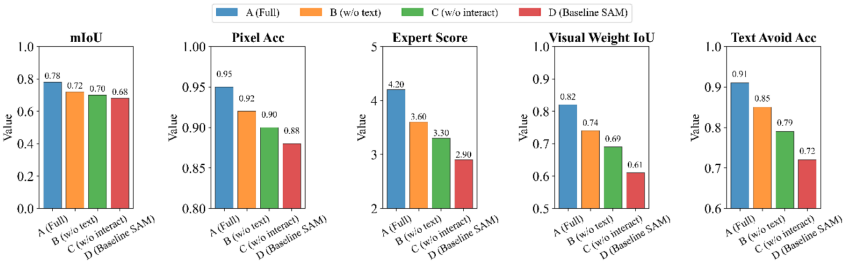

FIGURE 6. Analysis of ablation test results

Configuration D (Baseline SAM) serves as the foundational benchmark across all metrics. It achieves an mIoU of 0.68 and a Pixel Accuracy of 0.88 for segmentation tasks, along with an Expert Score of 2.9, a Visual-Weight IoU of 0.61, and a Text-Avoid Acc of 0.72 for design quality. Building upon this baseline, Configuration C (w/o interact) incorporates design priors but lacks the interactive optimization module. This configuration yields notable improvements in design-related metrics: its Visual-Weight IoU increases to 0.69 (+0.08 over Configuration D), its Expert Score rises to 3.3 (+0.4), and its Text-Avoid Acc reaches 0.79 (+0.07). These gains confirm that predefined design rules and constraints play a crucial role in aligning segmentation outputs with layout-oriented semantics. Meanwhile, Configuration B (w/o text) introduces textual prompts via CLIP while omitting the interactive design priors. This configuration exhibits consistent improvements across segmentation accuracy, achieving an mIoU of 0.72 (+0.04 over Configuration D) and a Pixel Accuracy of 0.92 (+0.04), demonstrating that high-level semantic guidance effectively mitigates oversegmentation and provides a more robust foundation for subsequent layout optimization. Crucially, Configuration A (Full) outperforms all other configurations across every metric, achieving state-of-the-art results. It attains the highest segmentation accuracy with an mIoU of 0.78 and a Pixel Accuracy of 0.95. For design quality, it also leads with an Expert Score of 4.2, a Visual-Weight IoU of 0.82, and a Text-Avoid Acc of 0.91. What makes this performance particularly striking is that the gains achieved by Configuration A are not merely additive but often surpass the sum of the individual improvements from Configurations B and C. For instance, in Visual-Weight IoU—a key indicator of design alignment—Configuration A achieves a +0.21 gain over the baseline Configuration D, exceeding the +0.08 gain contributed by Configuration C alone, which is the primary driver for visual layout optimization. This synergistic effect demonstrates that textual prompts and design priors are not only essential components but also highly complementary to each other. Their seamless integration constitutes the core innovation of the proposed framework, as validated by the ablation study results shown in Figure 6.

CASE ANALYSIS

Through two complete case studies, this paper demonstrates the entire process from image input, semantic boundary extraction, hierarchical construction, to final layout optimization, comparing the intermediate results and final design drafts of SAM and this framework.

Case Study 1:

Impressionist painting, using "A Sunday Afternoon on the Ile de la Grande Jatte" as an example. The foreground of the painting features brightly colored figures and pets; the middle ground shows the core area of crowds of citizens and trees; and the background features a river, sailboats, and distant buildings with blurred details. The three-layered layout is clear and has a sense of depth.

Case Study 2:

Abstract painting, using "Revolving Doors IV (La Rencontre)" as an example. The image is constructed from the collision of irregular color blocks and flowing lines. The color blocks in the picture are irregular in shape and have sharp edges, with strong color contrast. The spiral lines and thin straight lines running through the picture convey a sense of flow. The semantic boundaries are blurred, yet full of

tension.

Figure 7 demonstrates the whole process of layout optimization for two real-world cases.

Case 1:

Compared with the SAM basic model, while it can segment some elements like people and trees, it always faces over-segmentation. This results in segmenting clothing and accessories on people into independent masks and failing to distinguish the hierarchical relationship between trees and rivers. Through multimodal prompts to "extract trees and rivers," the proposed system has obtained complete forest boundaries and river outlines and constructed a hierarchical structure of "people-trees-rivers." Based on this boundary, the layout scheme generated can make the title text avoid the central focus, putting the text in the background area, highlighting the main subject while maintaining visual balance.

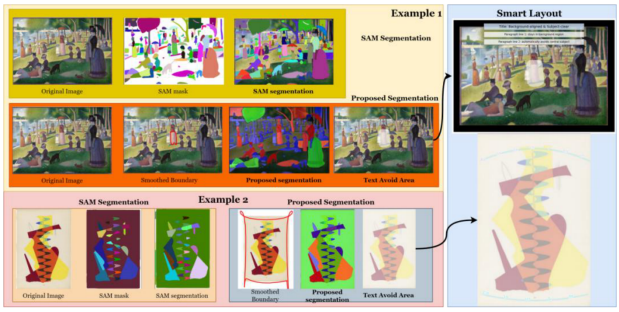

FIGURE 7. Case Analysis

In Case 2, the SAM basic model performs poorly on abstract art images, generating extremely fragmented results of mask segmentation. The proposed system, based on the text prompts, extracts major color blocks and flowing lines; working together with three key control points clicked by the user, it successfully extracts two main color block areas and one core flowing line. Hierarchical reasoning identifies the biggest color block as the main subject, while the rest are taken as subordinate elements. A layout scheme defines flowing lines to determine the visual flow, along which text is laid out. Primary and secondary colors are extracted from color blocks to serve as the color for text, creating a uniform design style.

V. CONCLUSION

This work addresses the core contradiction of the SAM model regarding complex artistic images: the disconnection between low-level visual boundaries and high-level design semantics. Resolving this contradiction is crucial for applications in the textile industry, providing a vital technical framework for the automated analysis of fabric patterns and embroidery designs where the visual segmentation must accurately reflect the complex semantic intent of the textile artist. It constructs the technical framework from semantically enhanced segmentation to layout optimization applications. This work reaches an mIoU of 0.78 and a pixel accuracy of 0.95 on the ADE20K validation set, surpassing the baseline SAM segmentation. In terms of design rationality, it reached an expert comprehensive score of 4.2, a visual center of gravity recognition overlap of 0.82, and a text avoidance accuracy of 0.91, which significantly improved the layout harmony and information clarity of the typography scheme. For user testing, the system reduced task time for experienced designers by 25% and novice designers by 40%. The ablation experiments validated the effectiveness of the three modules: multimodal cue fusion, interactive optimization, and boundary post-processing, which further demonstrated that post-processing is the most critical contribution to design rationality.

AUTHOR CONTRIBUTIONS

Shuo Liu designed, collected and analyzed the data, and drafted the manuscript. Shuo Liu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Shuo Liu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] C. Soreanu, “From media to mediums of expression: Visual art communication and meaning from fine arts to advertising,” Anastasis Research in Medieval Culture and Art, vol. 7, no. 2, pp. 261-276, 2020, doi: 10.35221/armca.2020.2.08.

[2] N. Kolesnyk, O. Piddubna, O. Polishchuk, et al., “Digital art in designing an artistic image,” Ad Alta, vol. 31, no. 12, pp. 128-133, 2022, [Online]. Available: https://eprints.zu.edu.ua/id/eprint/36188.

[3] L. Wang, “Machine learning-based environmental art automated design method,” Journal of Computational Methods in Sciences and Engineering, vol. 25, no. 2, pp. 1866-1879, 2025, doi: 10.1177/14727978241306041.

[4] Y. Wu and K. Kim, “Automatic generation of traditional patterns and aesthetic quality evaluation technology,” Information Technology and Management, vol. 25, no. 2, pp. 125-143, 2024, doi: 10.1007/s10799-022-00356-w.

[5] M. Vijendran, J. Deng, S. Chen, E. S. L. Ho, and H. P. H. Shum, “Artificial intelligence for geometry-based feature extraction, analysis and synthesis in artistic images: A survey,” Artificial Intelligence Review, vol. 58, no. 2, pp. 64, 2024, doi: 10.1007/s10462-024-11051-3.

[6] Sakshi and V. Kukreja, “Image segmentation techniques: Statistical, comprehensive, semi-automated analysis and an application perspective analysis of mathematical expressions,” Archives of Computational Methods in Engineering, vol. 30, no. 1, pp. 457-495, 2023, doi: 10.1007/s11831-022-09805-9.

[7] E. Song, D. Oh, and B. S. Oh, “Visual prompt selection framework for real-time object detection and interactive segmentation in augmented reality applications,” Applied Sciences, vol. 14, no. 22, Art. no. 10502, 2024, doi: 10.3390/app142210502.

[8] P. Ni, X. Li, D. Kong, and X. Yin, "Scene-adaptive 3D semantic segmentation based on multi-level boundarysemantic-enhancement for intelligent vehicles," IEEE Transactions on Intelligent Vehicles, vol. 9, no. 1, pp. 1722-1732, 2023, doi: 10.1109/TIV.2023.3274949.
[9] Y. Zhang, M. Fjeld, M. Fratarcangeli, A. Said, and S. Zhao, "Affective colormap design for accurate visual comprehension in industrial tomography," Sensors, vol. 21, no. 14, pp. 4766, 2021, doi: 10.3390/s21144766.
[10] L. P. Yuan, Z. Zhou, J. Zhao, Y. Guo, F. Du, and H. Qu, "Infocolorizer: Interactive recommendation of color palettes for infographics," IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 12, pp. 4252-4266, 2021, doi: 10.1109/TVCG.2021.3085327.
[11] S. Chan, W. Zhou, Y. Lei, C. Li, J. Hu, and F. Hong, "Sparse point annotations for remote sensing image segmentation," Scientific Reports, vol. 15, no. 1, Art. no. 27347, 2025, doi: 10.1038/s41598-025-12969-6.
[12] T. Zhang, M. Zhu, Y. Ren, C. Wang, L. Zhang, W. Zhang, et al., "BPA-SAM: Box prompt augmented SAM for traditional Chinese realistic painting," Journal of Graphics, vol. 46, no. 2, pp. 322, 2025, doi: 10.11996/JG.j.2095-302X.2025020322.
[13] S. Liu, F. Wang, H. You, N. Jiao, G. Zhou, and T. Zhang, "Context-aggregated and SAM-guided network for ViTbased instance segmentation in remote sensing images," Remote Sensing, vol. 16, no. 13, pp. 2472, 2024, doi: 10.3390/rs16132472.
[14] Z. Wang, Y. Zhang, Z. Zhang, Z. Jiang, Y. Yu, L. Li, et al., "Exploring semantic prompts in the segment anything model for domain adaptation," Remote Sensing, vol. 16, no. 5, pp. 758, 2024, doi: 10.3390/rs16050758.
[15] C. Li, L. Qi, and X. Geng, "A SAM-guided two-stream lightweight model for anomaly detection," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 21, no. 2, pp. 1-23, 2025, doi: 10.1145/3706574.
[16] K. Fan, L. Liang, H. Li, W. Situ, W. Zhao, and G. Li, "Research on medical image segmentation based on SAM and its future prospects," Bioengineering, vol. 12, no. 6, pp. 608, 2025, doi: 10.3390/bioengineering12060608.
[17] Z. Chen and Q. Sun, "Weakly-supervised semantic segmentation with image-level labels: From traditional models to foundation models," ACM Computing Surveys, vol. 57, no. 5, pp. 1-29, 2025, doi: 10.1145/3707447.
[18] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," ACM Computing Surveys, vol. 56, no. 9, pp. 1-36, 2024, doi: 10.1145/3649447.
[19] Y. E. Leichsenring and F. Baldo, "An evaluation of compression algorithms applied to moving object trajectories," International Journal of Geographical Information Science, vol. 34, no. 3, pp. 539-558, 2020, doi: 10.1080/13658816.2019.1676430.
[20] K. Mohamed and F. Adiloglu, "Analyzing the role of gestalt elements and design principles in logo and branding," International Journal of Communication and Media Science, vol. 10, no. 2, pp. 33-43, 2023, doi: 10.14445/2349641X/IJCMS-V10I2P104.
[21] R. Feng, H. Shen, J. Bai, and X. Li, "Advances and opportunities in remote sensing image geometric registration: A systematic review of state-of-the-art approaches and future research directions," IEEE Geoscience and Remote Sensing Magazine, vol. 9, no. 4, pp. 120-142, 2021, doi: 10.1109/MGRS.2021.3081763.
[22] D. Di Caprio, A. Ebrahimnejad, H. Alrezaamiri, et al., "A novel ant colony algorithm for solving shortest path problems with fuzzy arc weights," Alexandria Engineering Journal, vol. 61, no. 5, pp. 3403-3415, 2022, doi: 10.1016/j.aej.2021.08.058.
[23] Z. Wang, Y. Lv, R. Wu, and Y. Zhang, "Review of GrabCut in image processing," Mathematics, vol. 11, no. 8, pp. 1965, 2023, doi: 10.3390/math11081965.
[24] C. Wu, J. Zheng, Z. Feng, H. Zhang, L. Zhang, J. Cao, et al., "Fuzzy SLIC: Fuzzy simple linear iterative clustering," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 6, pp. 2114-2124, 2020, doi: 10.1109/TCSVT.2020.3019109.
[25] Y. Wang, L. Yang, X. Liu, and P. Yan, "An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+," Scientific Reports, vol. 14, no. 1, pp. 9716, 2024, doi: 10.1038/s41598-024-60375-1.
[26] S. Guo, Q. Yang, S. Xiang, S. Wang, and X. Wang, "Mask2Former with improved query for semantic segmentation in remote-sensing images," Mathematics, vol. 12, no. 5, pp. 765, 2024, doi: 10.3390/math12050765.