

Research on Automatic Segmentation and Layout of Poster Theme Visual Elements Using Text-Prompt-Driven 3D Base Model (SAT) and Contrastive Learning

Ge Wang

Visual Communication Design Teaching and Research Office of the School of Fine Arts and Design, Xinyang University, Xinyang 464000, Henan, China

Corresponding author: Ge Wang (e-mail: 18503765775@163.com)

ABSTRACT Automatic segmentation and layout of poster visual elements require joint modeling of visual structure, theme semantics, and spatial configuration. To overcome weak theme understanding, unclear element boundaries, and limited semantic relevance in existing poster-design automation methods, this study proposes a text-prompt-driven segmentation and layout framework based on a SAT spatial-semantic foundation model and contrastive learning. A text-image-layout triplet dataset is constructed, containing poster theme descriptions, pixel-level masks, bounding boxes, and element category labels. The SAT encoder extracts multi-scale visual features, while a BERT-based prompt encoder projects theme semantics into a high-dimensional representation to guide visual attention toward theme-relevant regions. A contrastive-learning module constructs semantic association matrices among elements, and an attention fusion module outputs element masks and layout coordinates. Experiments on a self-constructed poster dataset show that the proposed method achieves an average IoU of 0.870, a layout rationality score of 8.6/10, and an inference time of 0.30 s per image. Ablation experiments confirm the contribution of text prompting, semantic guidance, and contrastive learning. The framework supports visual signal processing, multimodal feature alignment, and spatial layout optimization in intelligent design systems.

INDEX TERMS visual element segmentation, contrastive learning, text-driven prompting, visual signal processing, layout optimization

I. INTRODUCTION

BEING a significant conveyor of visual messages, the quality of the design of the poster has a direct impact on the performance of the information delivery process and the mental impression of the audience. The design process must fully take into account the correct division of visual objects and the rational structure of the space layout, yet the current approaches are usually confronted by the problem of the lack of thematic semantic knowledge. Despite the element segmentation technology in the computer vision field having progress, the semantic information of theme guidance is hard to capture accurately, thus there is deviation in the results of the segmentation and the design intention. The layout generation algorithms are largely based on rule constraints or statistical models and do not capture the deep semantic connection between elements and the layouts generated are not very visual in terms of balance or even thematic. The issues limit the possible use of automated poster design, and there is an immediate necessity to investigate intelligent approaches that can combine semantic

knowledge and visual processing [1,2].

The proposed paper builds an automatic poster visual element segmentation and layout framework, which is built on text-based SAT 3D basic model and contrastive learning. The topic description is encoded into a high dimensional semantic vector by text-based encoder, and guide SAT model to attend on the visual parts of topic during feature extration, which solves the problem of weak semantic perception power of traditional method. Pairs of contrastive learning forms semantic distance matrix between items and build hierarchical feature association, which provides a semantic constraint basis on which layout optimization is based. The attention fusion unit fuses the multimodal data and provides correct mask of elements and layout coordinates. The proposed method is much better in segmentation and layout quality than the current used technologies, which gives it a new technical direction in the development of intelligent design tools, and benefits the transfer of poster design paradigm from currently used human-led cooperation to human-machine cooperation [3,4].

II. RELATED WORK

The methods of visual element segmentation have developed the traditional image processing into deep learning. Early algorithms relied on color clustering and edge detection, extracting foreground objects using threshold segmentation and region growing algorithms, but were very sensitive to complicated backgrounds and changing light. Semantic segmentation methods have been developed as a result of the introduction of convolutional neural networks. Fully convolutional networks are end-to-end trained to perform pixel-level classification, dilated convolution increases the receptive field without compromising feature resolution, and encoder-decoder architectures combine multi-scale contextual information to enhance boundary accuracy. The instance segmentation approaches present object detection schemes based upon semantic segmentation in order to differentiate among different representatives of the same type, yet all of them are based on visual features alone and have no chances to comprehend the meaning of the task. Contrastive learning in multi modal tasks acquires the representations of discrimination by maximizing similarity in positive sample pairs and reducing similarity in negative sample pairs [5,6]. The concentration of feature distribution is controlled by temperature parameters, negative sampling strategies to improve the efficiency of training are intra-batch, and encoders of momentum are used to maintain feature consistency. Methods of poster layout generation are rule-based template matching, optimization-based constraint solving, and learning-based generative models. In this study, we specifically compare our framework against LayoutVAE (a generative model) and a classic constraint-based optimization solver to benchmark spatial rationality. Template methods are inflexible, optimization methods are computationally expensive and although generative models can produce a wide range of layouts, they do not explicitly model design principles, so that the generated results cannot be sufficiently controlled or rational.

III. METHODS

A. OVERALL FRAMEWORK DESIGN AND DATASET CONSTRUCTION

The general structure built in this paper utilizes the SAT architecture to extract deep spatial hierarchies. While the input consists of 2D RGB images, the '3D' aspect refers to the high-dimensional latent space representation where visual features, text semantics, and layout depth are jointly embedded to simulate visual layering in design. The dataset is built from 15000 samples of commercial posters, and each sample is described with title text description, element-level pixel mask, bounding box coordinates and label of the category. The types of visual elements were divided into four primary classes during annotation: title text, main image, decorative graphics, and background texture. Although these categories cover the core structure of commercial posters, we recognize that real-world designs often involve more complex and overlapping element types. This study serves

as a foundational step, and future work will expand the taxonomy to include finer-grained categories like logos and subtitles to enhance the model's generalizability. Labelme tool was used for pixel annotation with high precision, and the consistency of annotation obtained Cohens Kappa coefficient of 0.92, which was cross-validated by three professional designers.

The framework uses a three-layer architecture of encoder-fusionist-decoder. RGB images and topic text can be inputted into the system at the same time. The images are resized into 1024x1024 pixel resolution and augmented. The SAT encoder supports three different output feature map sizes of 32x32, 64x64, 128x128, and channel sizes of 256, 512 and 1024 respectively. The text prompt encoder extracts semantic vectors of 768 dimensions with BERT model, and then concatenate to the same dimensional space with visual features by two multilayer perceptrons [7,8]. The contrastive learning module forms intra-batch negative sample pairs, and a temperature coefficient values 0.07. The semantic distance matrix is calculated with the help of The InfoNCE form in the loss function. Attention fusion module combines cross-modal characteristics using 8-head self-attention. Finally, the decoder gives the element masks and layout parameters.

B. MULTI-SCALE VISUAL FEATURE EXTRACTION BASED ON SAT

The SAT model adopts a 2D Vision Transformer design optimized for spatial depth perception, segmenting the input image into a sequence of image patches of 16x16 to simulate hierarchical layer relationships in poster design. Each image patch is embedded into a 768-dimensional embedding vector by a linear projection layer and then fed into a 12-layer Transformer encoder with further positional encoding. The multi-head self-attention mechanism of each encoder has 12 heads, the expansion ratio of feedforward network is 4, and the activation function is GELU. To acquire the visual data of different granularities, we get the intermediate feature at layer 4, 8 and 12 respectively. This is achieved by resizing the number of channels to 256, 512 and 1024 with 1x1 convolutions and then upscale the result to the respective resolution to generate the feature pyramid.

In the Table 1 we present a summary of the receptive field range and semantic expressive power of features extracted by network layers. It can be concluded that shallow features contain rich texture data, their receptive field is 32x32 pixels to capture the edge of text and decorative patterns; mid-layer features 128x128 pixels to capture local structural relationship; deep features 512x512 pixels to capture global semantic layout [9,10]. In feature extraction, we use drop out ratio as 0.1 to avoid over-fitting, and adopt residual connection and layer normalization to make the gradient more stable.

Table 1 shows that although the number of parameters in deep features increases to 28.9M, the global semantic information captured by deep features plays a decisive role

TABLE 1. SAT Multi-Scale Feature Extraction Parameter Configuration

Network layer	Feature resolution	Channel Dimension	Receptive Field Size	Number of parameters (M)	Semantic hierarchy
Layer-4	128×128	256	32×32	7.2	Low-level textures
Layer-8	64×64	512	128×128	14.5	Intermediate structure
Layer-12	32×32	1024	512×512	28.9	High-level semantics
Fusion layer	64×64	768	256×256	5.3	Mixed features
Output layer	128×128	512	384×384	3.8	Segmentation mask

in understanding the overall layout of the poster, while the 7.2M parameters of shallow features focus on accurately locating element boundaries.

C. TEXT PROMPT ENCODING AND SEMANTIC GUIDANCE MECHANISM

Use of templates in the text leads to make the poster theme a template: Design a poster about [theme], [list of elements]. encoder (BERT) takes as input sequence of segments text, of up to 128 long, and outputs the contextual semantics learning a 12-layer bidirectional Transformer. The global semantic representation is the hidden state the position of start marker and the sequence of word positions 768-dimensional vectors are stored to be matched precisely.

Semantic guidance has cross attention algorithm to calibrate the textual and visual feature. Output textual query endpoint and visual features as key endpoint to form an attention matrix. For each word in the text and each spatial location in the spatial location of visual feature map, score of similarity is determined and then normalized to yield weight distribution. The attention weight matrix has number of rows equal to the number of words in the text and has number of columns equal to all spatial locations in the visual features. Then filter the words by threshold to get semantic activation masks, and use it to locate the topic visual content. The guidance process uses a feature modulation approach, which computes two sets of parameters, scaling coefficient and offset, depending on text semantics. The sigmoid function is used to map the scaling coefficient to a range of 0 to 1, which determines the strength of the activation of visual features; the offset is simply added to the feature vector to modify the response baseline. Both groups of parameters are produced by independent fully connected layers, with an input dimension of 768 representing text features and an output dimension of 1024 representing the number of channels in the visual system. The modulated features, which do not change the original spatial structure, dynamically increase the expressive force of topic-related regions, depending on the text description, and silencing the background information of interference.

D. CONTRASTIVE LEARNING-DRIVEN ELEMENT ASSOCIATION AND LAYOUT OPTIMIZATION

How do we create semantic relationships between training examples using contrastive learning? We use positive and negative sample pairs (e.g. elements in the same poster are positive sample pairs, while elements in other categories and from other posters are negative sample pairs). The batch

size will be 64, meaning that 256 instances of elements, and 32768 pairs of samples will be formed for contrastive training. Element features will be L2 normalized and embedded into a 128-dimension embedding space, and a temperature parameter of 0.07 is used. InfoNCE will pull similar samples more closely together, and push distant samples further apart. By epoch 50, element class discriminatory is 0.94.

The spatial configuration is done by the layout optimization module on the learned semantic distance matrix. The distance matrix is square of the number of elements and the element values are between 0 and 1, with the less value depicting a stronger semantic relationship between two elements. Layout constraints are defined, and three types are viewed, namely: non-overlapping constraints, boundary constraints, and alignment constraints. Non-overlapping constraints, used to guarantee between elements a spacing of more than 8 pixels, by calculating the intersection-union ratio of the bounding boxes, and boundary constraints, used to ensure that the distance between the center line of the title text and the main image does not exceed 5 pixels, by ensuring at least a spacing of 32 pixels exists between the center points of the elements and the canvas edge.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL SETUP AND EVALUATION INDICATORS

We conducted the experiments on server with NVIDIA A100 GPU (40GB VRAM), Intel Xeon Gold 6248R processor and 128GB RAM. The Ubuntu 20.04 system was used and the deep learning framework was PyTorch 1.12.0 and CUDA 11.6. We split the dataset into training set with 10,500 images, validation set with 3,000 images and test set with 1,500 images (7:2:1). We used AdamW optimizer and set the initial learning rate as 0.0001, and decayed it using cosine annealing. The batch size was 32 and the training time was 72 hours for 100 epochs. We used intersection-union ratio, pixel accuracy and segmentation task boundary F1 score as evaluation metrics of segmentation task; layout task alignment error, layout task overlap rate and layout task aesthetic score as evaluation metrics of layout task. User research recruited 45 professional designers to subjectively evaluate the generated results by rating 1 to 10 points. To establish a gold standard, a control group of 50 human-designed posters was included in the blind evaluation. Furthermore, we applied a one-way ANOVA followed by Tukey's HSD test to determine the statistical significance of the performance improvements between our method and the baselines. Finally, three people reviewed each poster

and scored it independently, and then calculated the average score. Element completeness: whether all elements appear in the poster; Layout rationality: whether the arrangement of elements in the poster is reasonable; Visual aesthetics: how aesthetically pleasing the poster looks.

B. ANALYSIS OF SEGMENTATION ACCURACY AND LAYOUT QUALITY

The testing was done using five experiments having 1500 poster samples. In every experiment, the model parameters were re-initialized, and the order of the data was shuffled. The segmentation task determined the Intersection-over-Union (IoU) of each of the four types of each of the four types of visual elements and statistically compared the variability of performance on each of the four categories of the test sets. Alignment error and user ratings were used as a measure of layout quality. The pixel deviation between the center point and ideal position of the element was used as a measure of the alignment error. The design team received user ratings by blind judging 200 randomly chosen posters. Three reviewers rated each poster independently and the average score of all three was taken as the final score.

TABLE 2. Comparison of segmentation accuracy and layout quality across different methods

Index	Mask R-CNN	DeepLab V3+	This article's method
Test 1 IoU	0.742	0.768	0.868
Test 2 IoU	0.738	0.772	0.872
Test 3 IoU	0.745	0.765	0.870
Test 4 IoU	0.741	0.770	0.869
Test 5 IoU	0.739	0.769	0.871
Average IoU	0.741	0.769	0.870
Standard deviation	0.0026	0.0025	0.0015
Layout rating	7.2	7.8	8.6

To provide a comprehensive evaluation, we compare our method against general segmentation backbones (Mask R-CNN, DeepLab V3+) for element extraction, as well as the specialized layout generation framework 'CrePoster' [9] to benchmark aesthetic arrangement.

The experimental results indicate that the mean IoU of the proposed algorithm is 0.870, which is 17.4% and 13.1% higher than Mask R-CNN and DeepLab V3+, respectively. The text based prompting mechanism to correctly locate topic related elements and contrastive learning to strengthen the semantic discriminability between elements is largely credited to the improvement in performance. The standard deviation of 0.0015 implies that the stability of the model is higher than that of the control method and the range of fluctuations is 0.004 over five tests. The score of a layout of 8.6 confirms that the semantic distance-based spatial configuration approach is effective in enhancing visual balance and the user satisfaction rating of 8.6 represents a 10.26% improvement compared to DeepLab V3+ (7.8) and a 19.44% improvement compared to Mask R-CNN (7.2). While general models provide a baseline for pixel accuracy, our framework significantly outperforms the

specialized CrePoster model in layout rationality (8.6 vs 8.1) due to the integration of contrastive semantic constraints. Notably, our method significantly outperformed LayoutVAE (7.5) and approached the quality of human designs (9.2). Statistical analysis ($p < 0.01$) confirms that the integration of contrastive learning provides a non-trivial advantage in layout rationality compared to existing generative baselines.

C. ABLATION EXPERIMENTS AND MODEL PERFORMANCE COMPARISON

Ablation Experiments: We gradually remove the corresponding modules to check the importance of every module to the final performance. The baseline is only the SAT feature extractor. We generate 4 control experiments by adding text prompt encoding module, semantic guidance mechanism and contrastive learning module one by one respectively. The experiment was repeated 3 times on the same test set and we calculated the mean of each experiment. We computed three different measures ratio of intersection over union, layout score and inference time. We fixed the batch size as 16 for keeping the same testing environment. We employed the same data augmentation strategies and hyper-parameter settings to make the experiment more reliable.

TABLE 3. Comparison of Ablation Experiment Results

Model Configuration	Average IoU	Layout rating	Inference time (seconds)
Baseline (SAT only)	0.732	6.8	0.25
+Text prompt encoding	0.804	7.4	0.27
+Semantic guidance mechanism	0.841	8.1	0.29
+Comparative Learning (Complete)	0.870	8.6	0.30

Ablation analysis results show that text-based prompting encoding improves the Intersection over Union (IoU) ratio from 0.732 to 0.804, an increase of 9.8%, demonstrating the significant impact of topic semantics on segmentation accuracy. The semantic guidance mechanism further improves the IoU to 0.841, achieving a layout score of 8.1, indicating that cross-modal feature alignment effectively enhances element localization capabilities. The contrastive learning module brings a final performance improvement of 2.9%, with a layout score increase of 0.5, validating the optimization effect of semantic association modeling between elements on spatial configuration. Inference time increases from 0.25 seconds to 0.30 seconds, an increase of 20%, but still meets the requirements of real-time applications. Compared to the baseline, the complete model improves the IoU ratio by 18.9% and the layout score by 26.5%, showcasing the synergistic gain effect of each module.

V. CONCLUSION

We present an automatic poster visual element segmentation and layout system based on the text-based SAT spatial-semantic foundation model and contrastive learning to

bridge 2D visual extraction with 3D-aware layout depth. We solve some problems existing in poster design, such as poor understanding of theme, unclear division of elements, poor semantic link, etc. Multi-scale visual features are extracted by constructing text-image-layout triplet dataset and SAT model. Text-hint encoder is designed to encode thematic semantics to encode into high dimensional space to guide the model to focus on the right place. Contrastive learning mechanism is presented to provide hierarchical feature association among elements that achieve correct element segmentation and sensible spatial layout arrangement. The future work will be extend 3D spatial layout modeling, interactive editing system, cross-cultural aesthetic adaptation strategy to improve the application range of the model in more fields of design.

AUTHOR CONTRIBUTIONS

Ge Wang designed, collected and analyzed the data, and drafted the manuscript. Ge Wang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Ge Wang 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] H. Liu, "The influence of the composition system of Chinese painting on poster design," *Design Art Research*, vol. (1), pp. 4, 2020, doi: CNKI:SUN:JCGS.0.2020-01-027.
- [2] A. Ma, "A brief analysis of the application of creative programming in poster design," *Footwear Technology and Design*, vol. 2, no. 20, pp. 64-66, 2022, doi: 10.3969/j.issn.2096-3793.2022-20-020.
- [3] Z. Deng W and Wu Y. Application Research of AR (Augmented Reality) Technology in Creative Interaction of Poster Design[J], "2021," doi: 10.2991/ASSEHR.K.210106.131.
- [4] L. Zhang, "A study on the impact of visual design of League of Legends global event posters on communication effect: focusing on composition, color and cultural symbols," *China Design*, pp. 8(10), 2025, doi: 10.12428/zgsj2025.10.100.
- [5] N. Wang and N. Sun, "Visual grammar analysis of the movie poster "The Water Gate Bridge of Changjin Lake"," *Modern Linguistics*, vol. 11, no. 7, pp. 3171-3176, 2023, doi: 10.12677/ML.2023.117430.
- [6] Y. Pyo, H. Cho, T. Lee, et al., "P-236: Late-News Poster: Layout Optimization of AMOLED Pixel Circuits based on Deep Reinforcement Learning," *SID International Symposium Digest of Technical Papers*, vol. 56, no. 1, pp. 1635-1638, 2025, doi: 10.1002/sdtp.18514.
- [7] Z. Long, G. Wu, Y. Zhang, et al., "Poster: Repair Cross Browser Layout Issues by Combining Learning and Search-based technique," *IEEE*, 2021, doi: 10.1109/ICST49551.2021.00062.
- [8] C. Xu, K. Han, and W. Xu, "Image-aware layout generation with user constraints for poster design," *The Visual Computer*, vol. 41, no. 6, pp. 4239-4252, 2025, doi: 10.1007/s00371-024-03657-z.
- [9] M. Zhang, F. Liu, and M. C. Ran, "CrePoster: Leveraging multi-level features for cultural relic poster generation via attention-based framework," *Expert Systems with Application*, vol. 245, no. Jul., pp. 123136.1-123136.14, 2024, doi: 10.1016/j.eswa.2024.123136.
- [10] E. Barker and V. Phillips, "Creating conference posters: Structure, form and content," *Journal of Perioperative Practice*, vol. 31, no. 7-8, pp. 296-299, 2021, doi: 10.1177/1750458921996254.