IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 30, 2022, accepted April 14, 2022, date of publication April 28, 2022, date of current version May 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3171033

AUPOD: End-to-End Automatic Poster Design by Self-Supervision

DONGJIN HUANG, JINYAO LI, CHUANMAN LIU, AND JINHUA LIU
Shanghai Film Academy, Shanghai University, Shanghai 200072, China
Corresponding author: Dongjin Huang (djhuang@shu.edu.cn)

This work was supported by the Shanghai Natural Science Foundation of China under Grant 19ZR1419100.

ABSTRACT The automatic design has become a popular topic in the application field of computer vision technologies. Previous methods for automatic design are mostly saliency-based, relying on an off-the-shelf model for saliency map detection and hand-crafted aesthetic rules for ranking on multiple proposals. We argue that the multi-stage generation and the excessive reliance on saliency map hindered the progress of pursuing better automatic design solutions. In this work, we explore the possibility of a saliency-free solution in a representative scenario, automatic poster design. We propose a novel end-to-end framework to solve the automatic poster design problem, which is divided into the layout prediction and attributes identification sub-tasks. We design a neural network based on multi-modality feature extraction to learn the two sub-tasks jointly. We train the deep neural network in our framework with automatically extracted supervision from semi-structured posters, bypassing a large amount of required manual labor. Both qualitative and quantitative results show the impressive performance of our end-to-end approach after discarding the explicit saliency detection module. Our system learned on self-supervision performs well on the automatic design by learning aesthetic constraints implicitly in the neural networks.

INDEX TERMS Design automation, design aesthetic, artificial intelligence, neural networks, machine learning.

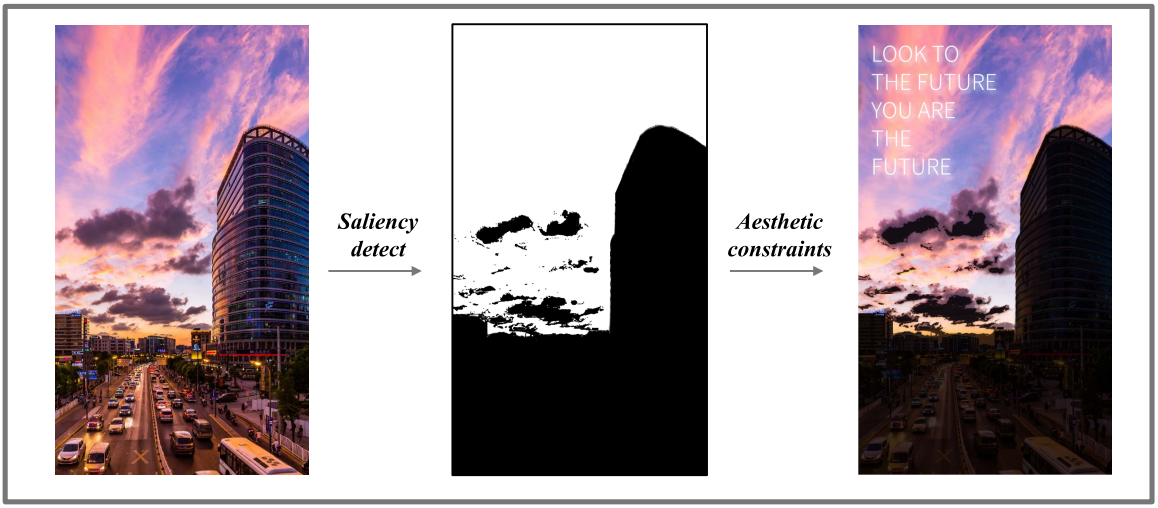
I. INTRODUCTION

Recent years have witnessed the rising interest in computer-aided automatic graphic design because of the explosive development of computer vision technologies. With the great success on various fundamental tasks in computer vision: image cropping [1], object detection [2], semantic segmentation [3] and others, some researchers turned to their application on automatic graphic design. The automatic graphic design aims to obtain graphic design works automatically by processing the basic graphic units and organizing them together based on reasonable aesthetic principles. Such applications can help significantly relieve the required human labor for relatively simple design works.

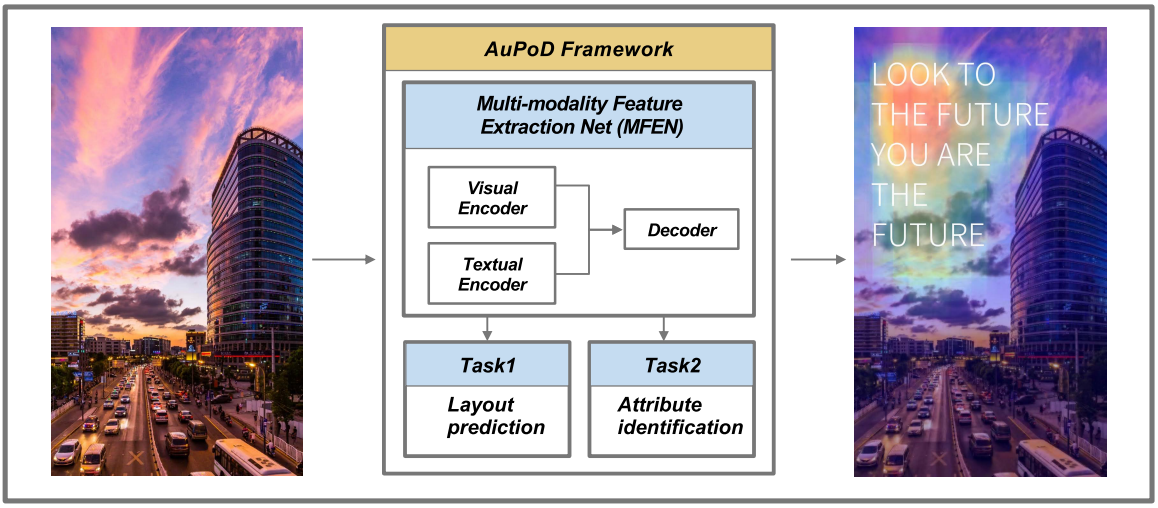
To our knowledge, most previous works on automatic design are saliency-based [4]–[6]. As shown in Fig. 1.(a), the saliency-based approaches for automatic design can be summarized as a two-step pipeline: They rely on off-the-shelf saliency detect algorithms to obtain the saliency map of original background images; They add aesthetic constraints to generate proper layout based on the saliency map. The aesthetic constraints can be explicitly added as templates [5] or implicitly modeled in an aesthetic evaluation module [6]. The pipeline-style framework suffers various drawbacks caused by the combination of independent components, including the problems of error propagation in the pipeline, the domain discrepancy of data used in each sub-stage, the difficulty of collecting and maintaining aesthetic constraints, etc. Those problems have hampered the progress of achieving better automatic design solutions.

In this paper, we focus on a specific and representative scenario for automatic design, automatic poster design based on photographs. To tackle those challenges in previous saliency-based solutions, we propose an end-to-end framework for **Au**tomatic **Po**ster **D**esign, named as **AuPoD**. As shown in Fig. 1.(b), we decompose the original challenging poster design task into two sub-tasks, including layout prediction and attributes identification. Our AuPoD framework is based on an end-to-end neural network, AuPoD Net, to generate harmonic posters from input background images and textual sequences (headlines). The AuPoD Net is centered on the multi-modality feature extraction net for generating a

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.



(a) Saliency-based Approaches



(b) End-to-end AuPoD Framework

FIGURE 1. The comparison between saliency-based approaches and our AuPoD framework. (a) Saliency-based approaches rely on off-the-shelf saliency detection algorithm and aesthetic constraints modeled in implicit/explicit way to obtain proper layout; (b) Our AuPoD framework tackles the two sub-tasks of poster design: layout prediction and attributes identification in an end-to-end manner based on multi-modality feature extraction.

density map of layout information. The layout prediction and attributes identification tasks are formulated as searching and classification problems based on the density map. The overall framework is trained on the self-supervision signals from semi-structured posters to capture the interdependencies between background images and headlines. Our AuPoD framework solves the problem of automatic poster design in a unified framework, instead of pipelines.

Specifically, our AuPoD framework surpasses the saliency-based approaches from the following perspectives:

(a) The unified framework of AuPoD prevents the data distribution discrepancy during training. The conventional saliency-based methods usually apply an off-the-shelf saliency map detection module that is trained on specific datasets. Some of them additionally train models for aesthetic evaluation. The data distribution of each component in the whole pipeline can be significantly different from each other, resulting in poor performance during inference. By training the unified framework on the automatically extracted consistent data, AuPoD refrains from such data discrepancy issue.

(b) AuPoD enjoys the benefits of end-to-end training by alleviating the error propagation in pipelines. Instead of training several components independently and using a cascaded pipeline for inference, our unified model is trained and used end-to-end, such that errors in previous components may be corrected later during inference.

(c) AuPoD does not require the hand-crafted aesthetic rules, but learns them implicitly from data. Many prior efforts on automatic design usually rely on hand-crafted aesthetic rules, which heavily depend on the expertise of aesthetic evaluation. The costly expertise and vagueness of aesthetic evaluation make it difficult to obtain and maintain

the explicit rules. Our model instead learns them implicitly from the data, thus enjoying better generalization and lower cost.

(d) AuPoD model the complex information for layout better by discarding the explicit saliency map detection. The layout prediction problem is closely related to the position of the salient object in the background image. However, it also depends on other information that cannot be described by the saliency map, e.g. color and position associations among multiple objects. Our AuPoD Net directly predicts the layout density map, considering that information jointly.

Our main contributions are listed below:

- We firstly propose the end-to-end framework for automatic poster design, avoiding the weakness of multi-step pipeline generation and the difficulty of maintaining aesthetic constraints.
- We design AuPoD Net that jointly learns the layout prediction and attributes identification by multi-modality feature extraction. We utilize the supervision signals extracted from semi-structured posters to train the network.
- The experimental results empirically show the effectiveness of our AuPoD framework. The extensive analysis of the results demonstrates the superiority of AuPoD by end-to-end data-driven learning on automatic design.

II. RELATED WORK

This work is related to the following research topics: **automatic design** (Sec. II-A) targeting at the automatic generation of various design works, **aesthetic evaluation** (Sec. II-B) focusing on the assessment of overall aesthetic quality, **image segmentation** (Sec. II-C) that divide images to multiple parts and **self-supervised learning** (Sec. II-D) aiming to utilize the self-supervision in data for training of neural networks.

A. AUTOMATIC DESIGN

Automatic graphic design has attracted lots of researchers. Jahanian *et al.* [4] started the automatic design problem and solve it by explicit aesthetic constraints by human priors. Yang *et al.* [5] focused on generating visual-textual layout targeting magazine cover design. They combined hand-crafted topic-dependent templates and pre-defined aesthetic principles to generate harmonious typography. The required expertise for template design, however, limited its practicality and diversity. Zhang *et al.* [6] and Li *et al.* [7] used a saliency-based framework and extra aesthetic evaluation for automatic design, considering only the layout prediction.

Automatic layout, identifying the geometric relations among multiple elements, is a classic problem for automatic design [8]. Previous work attempted to solve the problem by templates based on domain knowledge [4], saliency maps [9], attention mechanisms [10], etc. LayoutGAN [11] further applied generative models for layout generation of relational elements, but considering only objects with simple semantics. Lee *et al.* [12] explored the layout generation with given constraints from users, but not the natural constraints in data.

In our framework, we tackle the challenge of automatic design in the specific field of poster design. We consider both layout generation and the identification of corresponding attributes jointly.

B. AESTHETICS EVALUATION

An important component in prior methods for automatic design is aesthetic evaluation, which essentially defines the criteria for automatic poster design. The computational aesthetic evaluation plays a critical role in various visual generation tasks. Existing approaches for aesthetic evaluation can be divided into two categories: feature based and deep learning based. Feature based approaches rely on hand-crafted aesthetic features, including general global statistics [13], [14] (color distribution, brightness, etc.), generic graphical descriptors on local regions [15], [16], semantic-related features [17], [18] (scene topic, portrait attribute, etc.), and others. These features can be fed into regression models for quantitative aesthetic evaluation. More researchers have been dedicated to deep learning models for aesthetic scoring in recent years. The RAPID [19] model firstly attempted to use convolutional neural networks for aesthetic quality evaluation. The DMA Net [20] further improved the representation by using a multi-patch aggregation network instead of extracting features from a single patch. Zhang *et al.* [21] simulated the mechanism of human aesthetic evaluation by extracting fine-grained features on attended regions. Many other [22]–[24] approaches also contribute to the progress in this field. In our framework, instead of explicitly training an aesthetic evaluation model, the poster design results are expected to meet the implicit aesthetic constraints. The aesthetic constraints are learned in the model parameters by end-to-end training.

C. IMAGE SEGMENTATION

Image segmentation is a fundamental task in computer vision [25]–[27], targeting at segmenting the original image into two or more parts for visual understanding. As we mentioned, previous approaches for poster design are mostly based on salient object detection [28]–[30], a kind of segmentation for the most visually attractive object. Researchers have developed numerious image segmentation algorithms based on region growing [31], probabilistic graphical model [32], especially recent deep learning models [33], [34], boosting the performance significantly. We argue that a complicated pipeline system with a saliency detection component is not the best solution for the poster design problem. However, it is obvious that layout prediction is closely related to the segmentation problem. A straightforward aesthetic constraint is that usually, the headlines do not cover the salient object. The feature extraction network in our framework borrows the widely-used convolutional encoder-decoder architecture [35]–[37] in image segmentation, but learning from both visual and textual information jointly. We expect the architecture to draw benefits from segmentation-style models but consider richer information and constraints in poster design.

D. SELF-SUPERVISED LEARNING

Self-supervised learning has recently become a popular learning paradigm in various fields, such as natural language processing [38]–[41], computer vision [42]–[44], graph learning [45]–[48], and beyond. By utilizing the intrinsic associations in the data, self-supervised learning provides a solution for semi-automatically constructing supervision signals. The data-hungry training process of deep neural networks benefits from the cheap supervision built in this way. Generally, they obtain supervision from the data itself by predicting part of the input from transformed input [49], [50], corrupted input [51], [52], or other modalities of the original input [53]. Our AuPoD framework leverages the implicit supervision signals from two modalities in the semi-structured posters: the visual modality and the textual modality. We formulate the automatic poster design problem by predicting the missing layout and attribute information from the given two modalities, following the paradigm of self-supervised learning.

III. APPROACH

In this paper, we focus on tackling the problem of automatic poster design in an end-to-end manner. We formulate the problem as below.

Automatic Poster Design aims to generate a poster automatically from given background images and headlines. It provides a background image $I \in \mathbb{R}^{C\times H\times W}$, where C, H, W are the number of channels, the height, and the width of the image, respectively. Text T is also provided as the headline in the poster, which can be further formulated as a sequence of length n. The automatic poster design system solves two sub-problems jointly. They determine a specific region on which to put the text, and attributes of the headline, e.g. the color and the font family.

Unlike previous works based on pre-trained saliency map detection models, we adopt an end-to-end joint learning framework (Sec. III-A) to solve this problem. The main obstacle of end-to-end training is that the labeling of textual layout and attributes is expensive and time-consuming. In this paper, we construct labeled data by self-supervision (Sec. III-B) to bypass the required large amount of manual labor. The end-to-end AuPoD Net (Sec. III-C) is trained with the constructed supervision signals for producing harmonic posters automatically. The textual layout and attributes of the headline are learned jointly (Sec. III-D) to benefit each other. We propose to use a searching-based approximation for layout prediction during inference (Sec. III-E).

A. FRAMEWORK OVERVIEW

In this part, we describe the overview of our AuPoD system. The input of AuPoD is a background image $I = \{I_{c,i,j}\} \in \mathbb{R}^{C\times H\times W}$ and a textual sequence $T = \{T_1, T_2, \ldots, T_n\}$. We collect semi-structured posters to construct supervision signals for training. The end-to-end deep neural network AuPoD Net is trained on the collected data to learn the correlations between visual and textual objects, as well as corresponding attributes. The AuPoD Net extracts the image features with convolutional networks and text features with pre-trained context-aware token embeddings. The multi-modality features are then aggregated for predicting the position and size of the textual bounding box, and other corresponding attributes (font, color). The overall objective is decomposed into two sub-goals as layout prediction and attributes identification, learned jointly. Our AuPoD framework can automatically learn to generate harmonic posters by utilizing the multi-modality input features and capturing the association between features and underlying aesthetic constraints for poster design implicitly. We dive into the details of each part in the following sections.

FIGURE 2. The potential annotations in a semi-structured poster. Modern graphics editing systems for design store the relative position and corresponding attributes of each object in a semi-structured manner. The stored information may serve as intrinsic annotations for the finally rendered posters.

B. SELF-SUPERVISION

We adopt the widely-applied self-supervised learning paradigm for learning the poster design patterns from cheap supervision signals. The training of deep neural networks is usually label-intensive. Collecting a large number of annotations for design, however, requires the expensive expertise of professional designers. The costly labeling process for harmonic posters based on basic components has impeded the development of end-to-end approaches. Inspired by the recent success on self-supervised learning [41], [43], we utilize the self-supervision signals in semi-structured posters for training, thus bypassing the difficulty of fetching annotations.

The semi-structured posters, storing the intermediate products by human designers, naturally implies the required annotations representing the layout and attributes information (Fig. 2). Instead of collecting proper basic design elements (background images, headlines, etc.) and annotating them by human designers, we directly collect the semi-structured posters. We extract and reorganize the layout and attribute information in it as self-supervision signals. For example, as shown in Fig. 2, we can extract the textual sequence of the

headline, the relative coordinates and the size of the textual box, and other corresponding attributes. With the textual sequence and the background image regarded as input and the rendered poster as the target, we provide the required training signals for the neural network in our AuPoD framework.

The self-supervision signals obtained from semi-structured posters may slightly differ from those of professional annotations. More difficult control on quality and consistency of annotations brings challenges for training. However, we prove empirically that this is a tractable approach for constructing supervision for data-driven end-to-end poster design learning (Sec. IV).

C. AuPoD NET

In this part, we describe the details about our AuPoD Net, a deep neural network for end-to-end learning of the associations between background images and headlines. The main body of our AuPoD Net is based on an encoder-decoder architecture, namely the Multi-modality Feature Extraction Net, for feature extraction and density map decoding. It aggregates the multi-modality information from visual and textual objects, then decodes for the density map indicating layout information. We formulate the layout prediction and attributes identification sub-tasks as a constrained optimization problem and a classification problem, respectively. For layout prediction, we search for a proper region with a maximum score based on the density map. For attributes identification, we combine the generated density map and original visual features to predict the distribution on the predefined attributes set.

1) MULTI-MODALITY FEATURE EXTRACTION NET

The Multi-modality Feature Extraction Net (MFEN) aims to extract features from the input textual and visual objects, then model the score for each pixel to indicate the proper area for the textual object.

The overall structure follows the principled encoder-decoder architecture. Most previous segmentation neural networks can be decomposed in a similar way, with the encoder extracting high-level features from the input background image and the decoder performing up-sampling. As shown in Fig. 3, compared to segmentation networks, we do not only consider the feature from the image, but also the semantics of input text.

The encoder for the image features is a convolution-based deep neural network (details on the architecture described in Sec. IV-A). Briefly speaking, it encodes the graphic features in more channels with the feature map on broader receptive fields as the network gets deeper:

$$H_V = \texttt{Encoder}(I), \tag{1}$$

where $H_V \in \mathbb{R}^{C_V \times H \times W}$ denotes for the hidden image feature map extracted by the encoder. C_V, H, W are the number of channels, height, and width of the feature map.

$$U = \texttt{MLP}(\texttt{AVG}(\{E_1, E_2, \ldots, E_n\})). \tag{2}$$

The textual features, in the meantime, are encoded with the pre-trained contextual token embeddings as in (2). The token embedding for each token T_i is denoted as $E_i \in \mathbb{R}^d$. Those distributed vector representations are expected to carry the semantic information of the input text. We then use average pooling (denoted by AVG) to obtain the fixed-dimension distributed representation of the whole input sequence. A multi-layer perceptron (denoted by MLP) is followed to convert the textual representation to a similar vector space of the graphic representation. The final textual representation is a vector $U \in \mathbb{R}^{d'}$.

$$F = \texttt{Concat}(H_V, \texttt{REP}(U), \texttt{REP}(L)). \tag{3}$$

We aggregate the two main parts of critical information for poster design: the visual representation and the textual representation together as shown in (3). The textual representation is replicated across the height and width dimension to align with the visual representation (the repeat operation denoted by REP). We additionally add a scalar feature as the length of the input sequence $L \in \mathbb{R}$. The design is quite intuitive in that the number of tokens L is usually helpful for determining the size and height-width ratio of the textual bounding box. The same repetition operation is applied to the length feature. The concatenated representation is denoted by $F \in \mathbb{R}^{(C_V + d' + 1) \times H \times W}$.

$$M = \texttt{Decoder}(F). \tag{4}$$

Finally, we use the aggregated multi-modality feature F as the input for the decoder as in (4). The decoder outputs a density map $M \in \mathbb{R}^{H \times W}$ of the same size as the original image. Each element $M_{i,j}$ is a score corresponding to the pixel $I_{i,j}$, representing the weight of the pixel selected to be present in the textual area.

2) LAYOUT PREDICTION

The layout prediction module predicts the position and the size of the textual box by solving an optimization problem on the density map.

Based on the output of the decoder M, we assume that $\sigma(M_{i,j})$ represents the probability of $I_{i,j}$ located in the bounding box of given textual sequence. σ denotes the sigmoid function as shown below:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{5}$$

Given the density map M and the corresponding probability matrix $\sigma(M)$, we need to determine the corner coordinates (x_1, y_1) (left bottom) and (x_2, y_2) (right top) of the textual box. The corresponding prediction task can be formulated as the following constrained optimization problem as (6), shown at the bottom of the next page.

The optimization above essentially maximizes the joint probability of all pixels to be consistent with the assigned textual box, based on the assumption that the probabilities of pixels are independent. The exact inference algorithm is computationally expensive. We design an approximate algorithm for tractable computation in Sec. III-E.

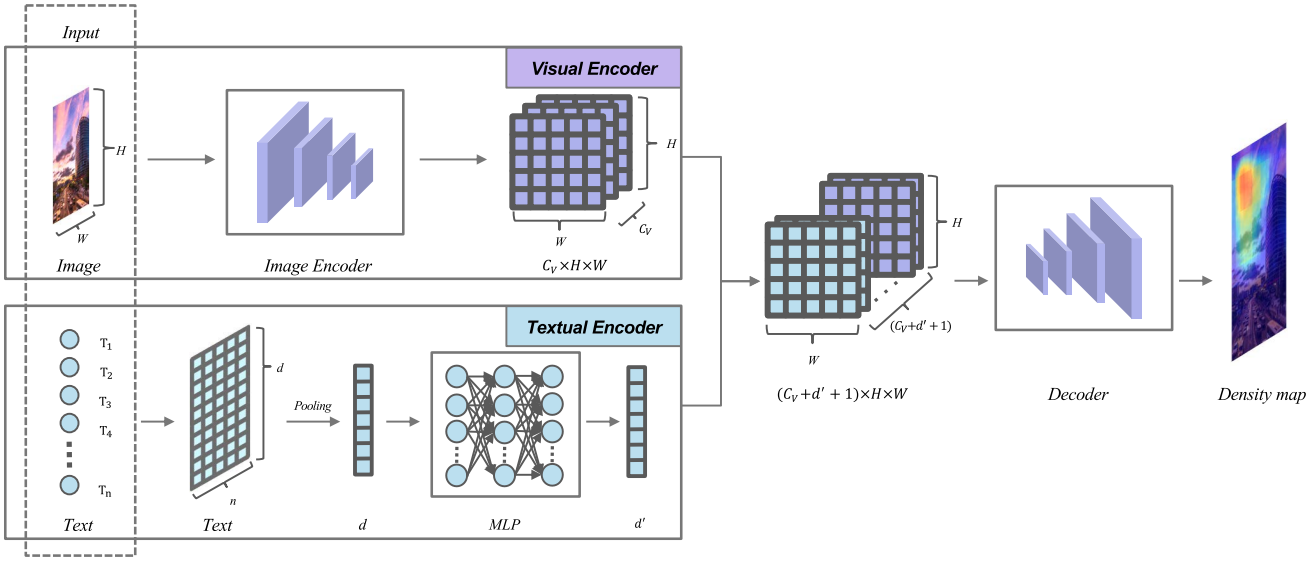


FIGURE 3. The structure of Multi-modality Feature Extraction Net (MFEN) in our AuPoD framework. The features from different modalities (visual & textual) are encoded with modality-specific encoders and aggregated for joint decoding of the final density map. The overall structure follows a classic encoder-decoder architecture.

3) ATTRIBUTES IDENTIFICATION

The identification of other attributes is regarded as classification problems based on extracted features. Continuous attributes are discretized to several classes for simplicity. Specifically, we collect features from two sources: On the one hand, we collect the hidden image features by the MFEN for the global summary of the overall input; On the other hand, we collect features from a weighted local view of the original raw image.

The features F from the Multi-modality Feature Extraction Net depict the global view of the input. By data-driven joint learning (Sec. III-D), The aggregated hidden features F will contain not only the critical information for layout prediction, but also for attributes identification.

We are also interested in the local view of the images because the pixels in the region of textual boxes may have a greater impact on attributes. For example, the color of the textual input is generally constrained by the color tone of the local textual area. The weighted raw image $\sigma(M) \circ I$ applies the probabilistic density map as attention weights on each pixel of the raw image. We use a similar convolutional encoder to extract the local view image features $F_l \in \mathbb{R}^{C_V \times H \times W}$ as below:

$$F_l = \texttt{Encoder}_l(\sigma(M) \cdot I). \tag{7}$$

We concatenate the global features and the local features, and use an MLP classifier for logits output. The logits are normalized to probabilistic distributions by the softmax function. p_i denotes for the probability of the attribute belonging to the i_{th} class as in (8).

$$\begin{aligned} \texttt{logit}_i &= \texttt{MLP}(F, F_l), \\ p_i &= \frac{\exp(\texttt{logit}_i)}{\sum_k \exp(\texttt{logit}_k)}. \end{aligned} \tag{8}$$

D. JOINT LAYOUT AND ATTRIBUTES LEARNING

Since we use a unified neural network for layout prediction and attributes identification, we can enjoy the benefits of joint training on the two sub-tasks. The objective for layout prediction $\mathcal{L}_{\text{layout}}$ and attributes identification $\mathcal{L}_{\text{attributes}}$ are listed below, respectively:

$$\begin{aligned} \mathcal{L}_{\text{layout}} &= -\sum_{i,j} (G_{i,j} \log \sigma(M_{i,j}) \\ &\quad + (1 - G_{i,j}) \log(1 - \sigma(M_{i,j}))), \\ \mathcal{L}_{\text{attributes}} &= -\sum_{a \in \text{Attributes}} \log p^a_{y_a}, \end{aligned} \tag{9}$$

where $G_{i,j}$ is a binary indicator for whether the pixel (i, j) is located in the textual area. $p^a_{y_a}$ is the probability of the attribute a belonging to the gold class y_a.

The overall objective is to minimize the sum of layout prediction loss and the attributes identification loss as in (10). It essentially maximize the log likelihood on our dataset.

$$\mathcal{L} = \mathcal{L}_{\text{layout}} + \mathcal{L}_{\text{attributes}}. \tag{10}$$

$$\begin{aligned} \underset{x_1, y_1, x_2, y_2}{\arg\max} \quad & \{ \sum_{\substack{i \in [x_1, x_2] \\ j \in [y_1, y_2]}} \log \sigma(M_{i,j}) + \sum_{\substack{i \notin [x_1, x_2] \\ j \notin [y_1, y_2]}} \log(1 - \sigma(M_{i,j})) \}, \\ s.t. \quad & x_1 < x_2, \ y_1 < y_2. \end{aligned} \tag{6}$$

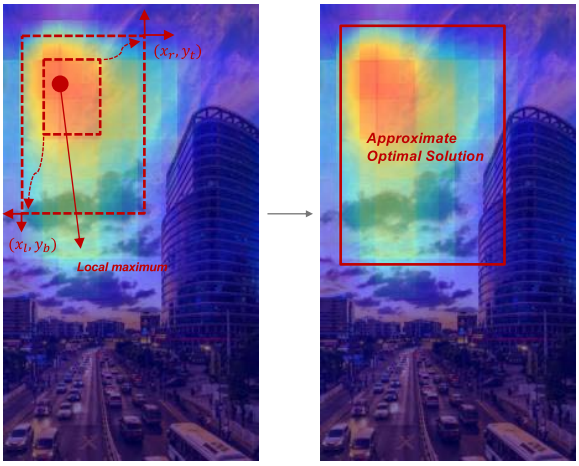


FIGURE 4. Approximate inference for layout prediction.

The joint learning helps our AuPoD Net to extract better features from the multi-modal (textual and visual) views of data. The final objective will benefit from the two sub-modules, promoting each other.

E. INFERENCE ALGORITHM

The inference of the framework for automatic poster design is non-trivial, especially the layout prediction part. The optimal solution for layout prediction is determined by the optimization problem defined in (6). Such optimization problem is intractable in practice. We design the approximate algorithm based on the locality property of neural network output. Specifically, we decide the region of the textual object by greedily enlarging the area of the rectangle from the local maximum of M (Fig. 4). The detailed algorithm is listed in Algorithm 1.

The approximate inference algorithm leverages the locality of the density map, assuming that the proper location of the textual input is approximately centered at the local maximum and the score for the candidate area is almost convex in terms of the distance from edges to the local maximum.

IV. EXPERIMENTS

To verify the effectiveness of the proposed AuPoD framework, we conduct extensive experiments from various perspectives to show the overall performance, the aesthetic interpretation, as well as the plausibility of our framework designs.

A. SET-UP

We first introduce the set-up including the used benchmark, metrics, and models as below.

1) BENCHMARK

Despite many visual benchmarks that have been explored on saliency detection or aesthetic evaluation, there exists no previous large-scale data for the whole process of automatic poster design. This also explains why end-to-end framework has not been explored in this field. As previously mentioned in Sec. III-B, we construct supervision signals from cheap semi-structured posters to tackle this challenge.

We collect over 6000 semi-structured posters. We discard the posters of which the annotations are not complete or too complicated to extract. After the filtering, we obtained around 4600 posters with topics varying from portraits, animals, food, scenes, to others. We use this dataset for both training and evaluation of the AuPoD Net in our framework. We split it according to the ratio of 7:1:2 for training, validation, and test, respectively.

Algorithm 1 Inference for Layout Prediction

function Score(M, (x_1, y_1), (x_2, y_2))
  $b \leftarrow \sum_{i \in [x_1, x_2]} \log \sigma(M_{i,j})$ ▷ bonus term
    $j \in [y_1, y_2]$
  $p \leftarrow \sum_{i \notin [x_1, x_2]} \log(1 - \sigma(M_{i,j}))$ ▷ penalty term
    $j \notin [y_1, y_2]$
  return $b + p$
end function
Find all local maximum positions in M, as $P = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}$
$R \leftarrow \phi$
for all $(x_i, y_i) \in P$ **do**
  $(x_l, y_b) \leftarrow (x_i, y_i)$
  $(x_r, y_t) \leftarrow (x_i, y_i)$
  $\text{score}_i \leftarrow \text{Score}(M, (x_l, y_b), (x_r, y_t))$
  repeat
    $C \leftarrow [\text{Score}(M, (x_l - 1, y_b), (x_r, y_t)), \text{Score}(M, (x_l, y_b - 1), (x_r, y_t)), \text{Score}(M, (x_l, y_b), (x_r + 1, y_t)), \text{Score}(M, (x_l, y_b), (x_r, y_t + 1))]$ ▷ scores of one-step expansions in four directions
    $\text{score}'_i = \max C$
    Update the corresponding coordinates accordingly
  until $\text{score}'_i < \text{score}_i$
  $R \leftarrow R \cup \{(x_l, y_b, x_r, y_t, \text{score}_i)\}$
  Exclude all positions in P covered by the newly determined region
end for
Sort all items in R according to the related scores in descending order
return R

2) METRICS

We evaluate the performance of automatic poster design systems according to the following quality measurements:

- **Test Loss** We use the test loss defined in Sec. III-D to automatically evaluate the overall performance of various models on our dataset, the lower, the better.
- **Jaccard Similarity** We use the principled Jaccard similarity to measure the consistency between two areas. Specifically, we use this measurement to evaluate the similarity between the generated layout and automatically extracted supervision, the higher, the better.

TABLE 1. Automatic quantitative evaluation of various models on our benchmark.

Model		Test Loss ↓	Jaccard Similarity ↑	ACC ↑
saliency-based	[9]	-	0.463	-
	Ours	-	0.637	-
	[6]	-	0.691	-
AuPoD-varieties	AuPoD-FCN32	0.431	0.479	0.134
	AuPoD-UNet	0.259	0.683	0.215
	AuPoD-BASNet	0.176	0.717	0.202
	AuPoD-U^{2}Net	**0.168**	**0.742**	**0.238**

- **ACC (macro)** We use macro-averaged accuracy to automatically evaluate the attributes of headlines in a poster, the higher, the better.
- **Manual Rating** Because of the subjectivity and natural vagueness of aesthetic evaluation, we additionally use human evaluation for more convincing evaluation results. Specifically, we set ratings from 1 to 5, and ask human designers to rate each generated poster. We demonstrate the distribution of human ratings as manual quantitative analysis results. The criteria for each rating is listed below:
  1) poor: The design is definitely unacceptable and hard to understand.
  2) inferior: The design is obviously worse than reasonable solutions.
  3) acceptable: The design is mostly harmonious and fits basic aesthetic constraints for location and coloring.
  4) good: The design is quite harmonious. There are also other reasonable choices.
  5) excellent: The design is excellent. No equal or better substitutions can be found.

3) MODELS

We implement and discuss the results of the following models to show the strengths of the proposed framework and the necessity of the design for each part.

- **saliency-based**: We introduce three saliency-based approaches for comparison on layout prediction. The first one is implemented by ourself, with an off-the-shelf saliency detection model(BAS-based) for saliency map and a set of rules for regularizing the position of the textual object not to conflict with the salient object. Besides, we choose two representative saliency-based state-of-the-art approaches in [6], [9] as baselines for our framework. Specifically, the saliency map is generated by BASNet and FCN-32, respectively. Then the diffusion-based method in [6] for proposal generation. Note that the **Test Loss** and **ACC** metrics are not applicable for this method, because they are not trained in a consistent manner with us and they consider only layout information.
- **AuPoD-varieties**: We compare a series of varieties of our AuPoD framework based on various backbone models for the visual encoder-decoder architecture. The classic architectures for saliency detection are used, e.g. FCN-32, UNet [35], BASNet [54], etc. Specifically, FCN-32 and BASNet are saliency detection backbones used in [9] and [6] for reference, respectively.

4) IMPLEMENTATION DETAILS

We use the bert-base-uncased model [38] to obtain the contextual token embeddings in our textual encoder. Note that we use the model only for the embedding inference purpose, which means we do not update its parameters during the training of our AuPoD framework. We train the whole framework with the AdamW optimizer and the learning rate initially set to 1e-3. The batch size for training is 16 and the dropout ratio is 0.1. We train the model for a maximum of 20K steps and early stop training when observing a performance drop on the validation set.

B. MAIN RESULTS

We compare and analyze the end-to-end evaluation results of various models accordingly in this part. We analyze them by comparing the automatic and human evaluation metrics for quantitative evaluation. We include qualitative analysis by demonstrating the end-to-end poster design results of our framework.

1) AUTOMATIC QUANTITATIVE ANALYSIS

We use the three automatic metrics described in Sec. IV-A for comparison. **Test Loss**, **Jaccard Similarity**, and **ACC** evaluate the overall quality, the layout generation quality and the attributes identification quality, respectively.

Our AuPoD network performs better than the saliency-based method on layout prediction sub-task. As shown in Table. 1, saliency-based approaches generally perform worse than our AuPod-varieties with identical backbone models. The results show superiority of end-to-end self-supervised learning, compared to the combination of saliency map and manually designed layout generation rules or modules. The statistics indicate that our end-to-end learning framework has a stronger capability for capturing layout associations than the saliency-based pipeline.

As the backbone structure gets more powerful capability, our AuPoD framework can obtain better design works in terms of the overall quality. Since the saliency-based method can not be fully evaluated with all the quantitative metrics, e.g. **ACC** for attributes identification, we further investigate the performance of our AuPoD framework on various backbone models. As shown in Table. 1, the overall performance,

as well as the effectiveness of each component, becomes better as the backbone model becomes stronger. The results demonstrate that our AuPoD framework can benefit from the better inductive bias for saliency detection. It is evidence showing the associations between poster design and saliency detection.

We conclude that automatic poster design is closely related to saliency detection but relies not only on the saliency information by the automatic quantitative analysis above.

2) MANUAL QUANTITATIVE ANALYSIS

We conduct a human evaluation for a more comprehensive and more precise evaluation on the quality of our system. We simulate a simple Turing Test by randomly mixing the generated posters by our AuPod framework, the posters generated by the best saliency-based approach in previous automatic quantitative analysis, and original posters in the data. We prepare 50 groups and ask human designers to give a proper rating from 1 to 5 for each poster. We analyze the rating distributions of each part for manual quantitative analysis.

As shown in Fig. 5, the rating distribution of our AuPoD system outputs has higher variance, whereas the rating distribution of original posters is more centralized. The results show that compared to automatically generated posters, manually designed posters generally have more stable quality. This is expected behavior because the outputs of deep learning models are easier to be affected by the uncertainty of neural networks. According to the rating criteria in Sec. IV-A, posters that are rated with 3 or greater scores are acceptable. The ratio of acceptable posters in Original and AuPoD groups are the same as 94%, while it is 78% for the Saliency-based group. These observations indicate the great potential of our AuPoD framework to become an important aid to human designers on poster design. However, the results also show that automatic design systems obviously fall behind human designers on near-optimal solutions (with the ratings $>= 4$). Thus We will not claim that our AuPoD system can completely substitute human designers, but provide high-quality candidates for boosting their efficiency.

3) QUALITATIVE ANALYSIS

For qualitative analysis, we show a few poster generation examples of our AuPoD framework by inference on the test data. For the saliency-based approach, we directly use the textual sequence with the same attributes as the original posters. For our AuPoD system, we use predicted attributes to generate the textual object. The identification of these attributes is quite subjective so we leave the justification to each reader. We mainly focus on layout generation in this part.

We compare the results of the saliency-based method, our framework, and the annotations in the semi-structured posters (Fig. 6). As shown in (b) and (c) columns, the saliency-based method first detects the salient object and identifies the layout by avoiding the region of the salient object. As shown in (d) and (e) columns, the high-density region in the density maps of AuPoD outputs directly indicates the textual regions in the posters.

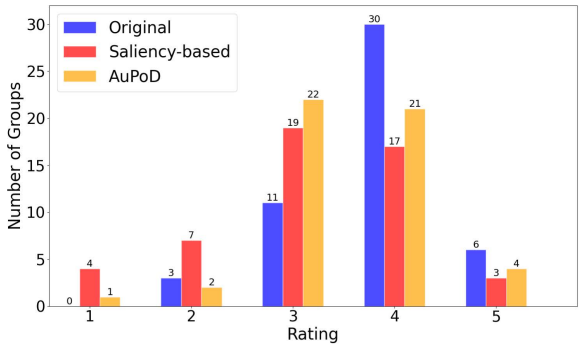


FIGURE 5. Human evaluation ratings.

TABLE 2. Inference speed of different poster design methods.

Method	Speed
saliency-based	3.4 examples/sec
AuPoD-U^{2}Net	2.7 examples/sec
human	0.063 examples/sec

For the background images with clear salient objects (ii and iv), the saliency-based methods can also generate harmonious posters. However, on background images without obvious and separated salient objects (i, iii, and v), our AuPoD framework shows superiority by generating posters based on implicit aesthetic constraints.

4) INFERENCE EFFICIENCY

The inference efficiency is an important factor for the deployment of poster design systems in production. We list the average inference speed of the saliency-based method and our AuPoD system for comparison in Table. 2.

The speed of automatic design systems, both the saliency-based and our AuPoD system, are tested and computed in a single-GPU (Tesla V100 16GB) machine with 32 2.3GHz CPU Cores and 128 GB memory. The human inference speed is tested as the average time used for thinking and dragging the textual elements on the background image with the help of professional design software.

As shown in the table, both automatic design systems, including the saliency-based and our AuPoD system, are significantly faster than humans. Our AuPoD system presents a slightly slower but similar speed when compared to the saliency-based approach. It's because of the more complex model architecture and the searching process during inference. Given the observation in Sec. IV-B2, we can see that our AuPoD system may help human designers to boost their efficiency a lot, with high-quality candidates generated.

C. ANALYSIS AND DISCUSSION

Apart from the overall generation quality of AuPoD framework, we care about the effectiveness and necessity of the system design on each part. We analyze the intermediate density map generation results for interpretability of potential

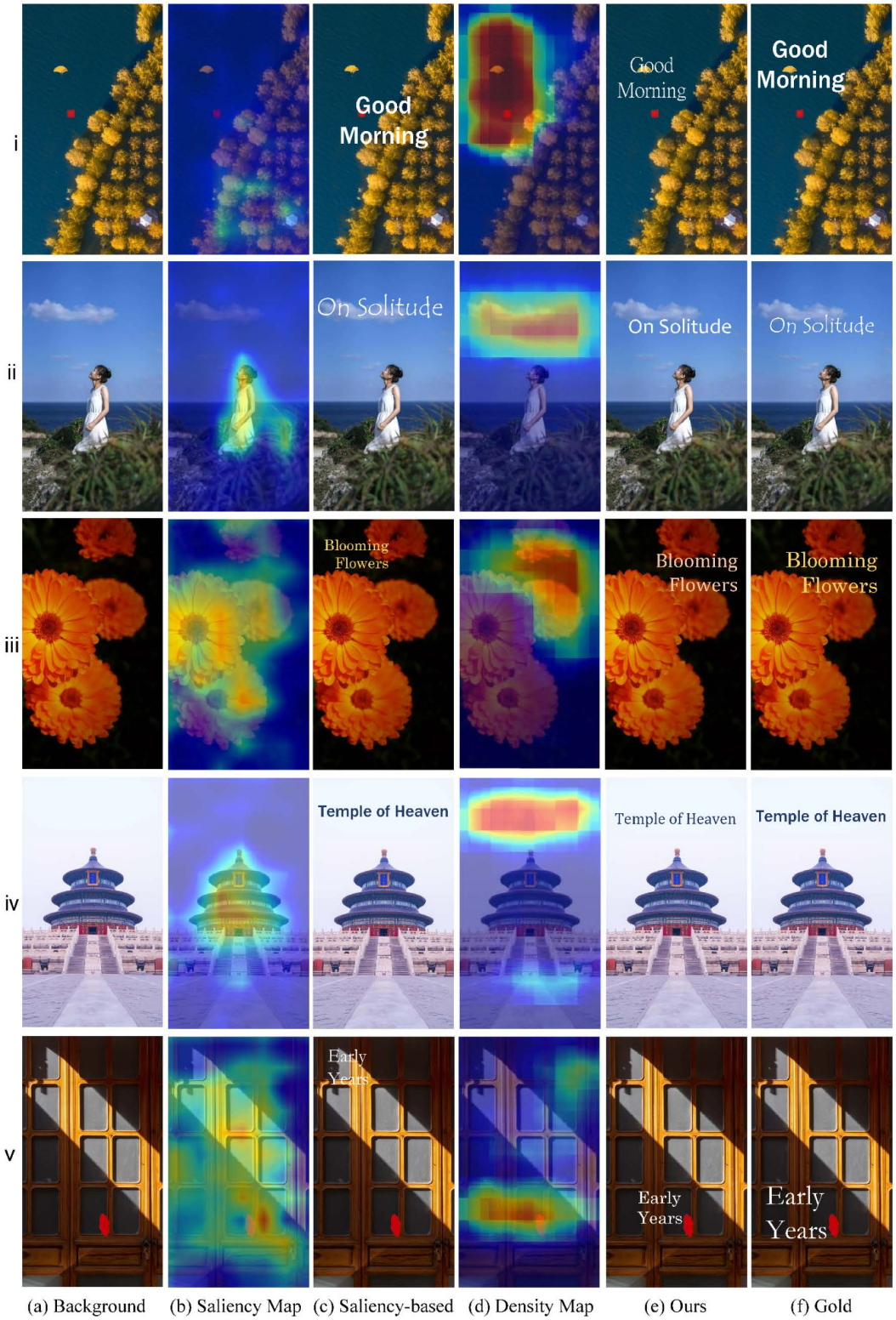


FIGURE 6. Qualitative analysis on the generated layout.

aesthetic patterns learned in the framework. We also discuss the benefits of joint learning and multi-modality feature extraction, respectively. These cases more directly exhibit the effectiveness of our AuPoD framework.

1) AESTHETICS INTERPRETABILITY

A thorough case study on the automatically generated posters of our AuPoD framework further reveals the intrinsic behavior of the framework. We manually traverse the intermediate

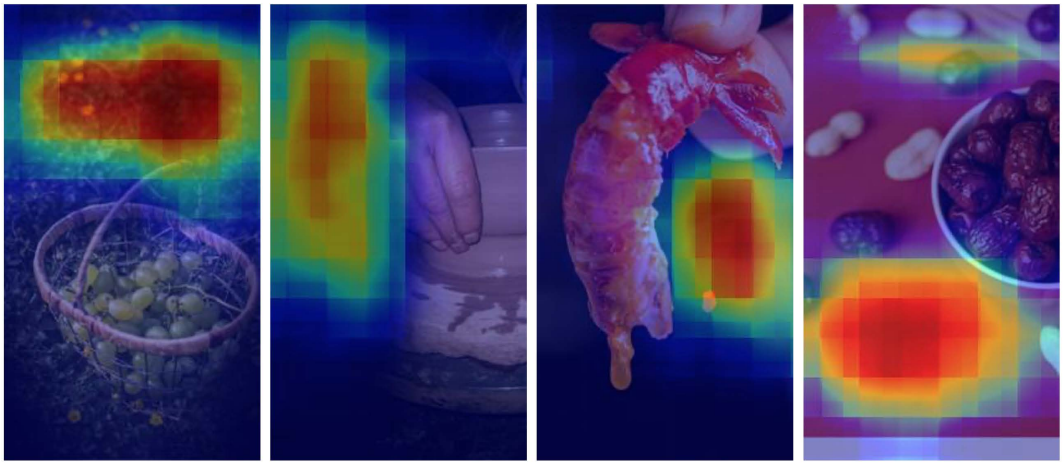


FIGURE 7. General symmetry demonstrated by density maps.

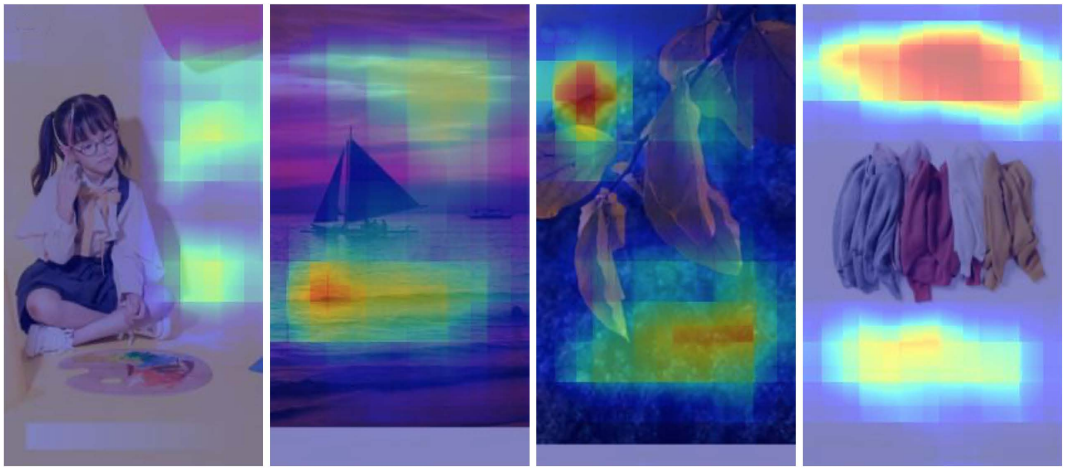


FIGURE 8. Multi-modes diversity demonstrated by density maps.

density map results of our system and find the following patterns that are encoded in our framework implicitly.

a: GENERAL SYMMETRY

Symmetry is one of the aesthetic properties for design works. As shown in Fig. 7, the predicted layout density map generally follow the symmetry property between the salient object in the image and the textual object. The symmetry property is ensured in various directions, showing the implicit aesthetic patterns learned in our AuPoD framework.

b: GENERATION DIVERSITY

The proper layout for a specific poster may not be exclusive but in different modes. Although the self-supervision in semi-structured posters always gives an exclusive layout, our AuPoD framework automatically learns to generate multi-modes density map based on implicit aesthetic constraints, as shown in Fig. 8.

The results and analysis above verify our claim that the AuPoD framework can learn implicit salient object concepts and construct aesthetic constraints on relative positions between the salient object and expected textual object. The data-driven end-to-end training helps discard the explicit saliency detection module and aesthetic rules, but learn them jointly in the model parameters.

2) BENEFITS OF JOINT LEARNING

We investigate the potential benefits of joint learning by conducting an ablation study on the learning objectives of our AuPoD framework. Specifically, we compare the results of joint learning and the results by learning each objective independently. In the independent learning setting, we duplicate the Multi-modality Feature Extraction Net and learn the two sub-tasks independently. We merge the results of two neural networks for comparison to the results of joint learning. As demonstrated in Fig. 9, results in the independent learning setting are generally worse than those of joint learning. It indicates that the two sub-tasks, layout prediction and attributes identification can benefit from each other. It is obvious that the layout prediction and attributes identification both require better density map output for deciding the textual region and extracting the attentive feature. The associations between the two sub-tasks result in the benefits of joint learning.

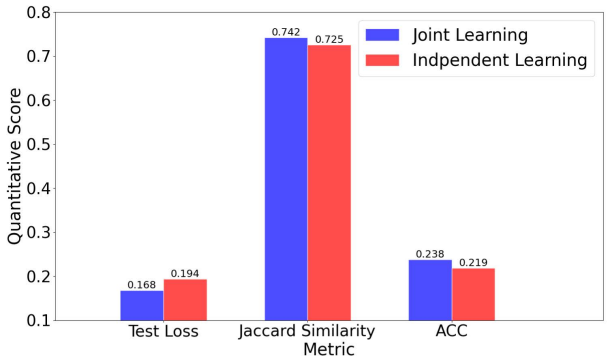


FIGURE 9. Benefits of joint learning.

TABLE 3. Multi-modality ablation study.

Model	Jaccard Similarity
AuPoD-U^{2}Net	0.742
AuPoD-U^{2}Net$_{\text{-text encoder}}$	0.649

3) MULTI-MODALITY ISSUE

A further question regarding the effectiveness of the proposed framework is that: Whether the multi-modality feature extraction and fusion boost the overall performance.

Visual representations are undoubtedly critical for both layout prediction and attributes identification. Besides, textual information is closely related to the attributes attached to the headlines. We mainly focus on the investigation of textual representations for layout prediction. We remove the textual encoder part in the MFEN for ablation study on the effect of textual representations. As shown in Table. 3, the Jaccard Similarity decreases after we remove the textual feature. The textual representation affects the layout prediction by input length and emotion of the input text.

V. CONCLUSION

In this paper, we propose an end-to-end poster design system, AuPoD. AuPoD is learned from the self-supervision mined from cheap semi-structured data. The Multi-modality Feature Extraction Net in our AuPoD framework effectively learns the patterns of organizing visual and textual objects together, along with assigning attributes to the textual object. Empirical results show that AuPoD provides a tractable solution for automatically learning aesthetic constraints from data and utilizing those constraints during poster generation. The superiority of our AuPoD framework on closing the gap for data discrepancy and fixing up the deficiencies of error propagation make it a better solution than the conventional saliency-based approaches.

In the current state, we focus on the combination of a single background graphical object and a single textual box. In real-world applications, more complicated posters may be composed of multiple graphical objects for decoration and more textual objects for multi-granularity description. Generating such complex posters requires more fine-grained modeling of the interdependencies among those objects. In the future, we will explore proper joint inference algorithms to extend our AuPoD system to scenarios of generating those complicated posters.

REFERENCES

[1] J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping: A computational complexity study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 507–515.
[2] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*.
[3] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.
[4] A. Jahanian, J. Liu, Q. Lin, D. Tretter, E. O'Brien-Strain, S. C. Lee, N. Lyons, and J. Allebach, "Recommendation system for automatic design of magazine covers," in *Proc. Int. Conf. Intell. User Interface (IUI)*, 2013, pp. 95–106.
[5] X. Yang, T. Mei, Y.-Q. Xu, Y. Rui, and S. Li, "Automatic generation of visual-textual presentation layout," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 2, pp. 1–22, Mar. 2016.
[6] P. Zhang, C. Li, and C. Wang, "Smarttext: Learning to generate harmonious textual layout over natural image," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[7] C. Li, P. Zhang, and C. Wang, "Harmonious textual layout generation over natural images via deep aesthetics learning," *IEEE Trans. Multimedia*, early access, Aug. 30, 2021, doi: 10.1109/TMM.2021.3097900.
[8] N. Hurst, W. Li, and K. Marriott, "Review of automatic document formatting," in *Proc. 9th ACM Symp. Document Eng. (DocEng)*, 2009, pp. 99–108.
[9] Z. Bylinskii, N. W. Kim, P. O'Donovan, S. Alsheikh, S. Madan, H. Pfister, F. Durand, B. Russell, and A. Hertzmann, "Learning visual importance for graphic designs and data visualizations," in *Proc. 30th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2017, pp. 57–69.
[10] X. Pang, Y. Cao, R. W. H. Lau, and A. B. Chan, "Directing user attention via visual flow on web designs," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–11, Nov. 2016.
[11] J. Li, J. Yang, A. Hertzmann, J. Zhang, and T. Xu, "LayoutGAN: Generating graphic layouts with wireframe discriminators," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[12] H.-Y. Lee, L. Jiang, I. Essa, P. B. Le, H. Gong, M.-H. Yang, and W. Yang, "Neural design network: Graphic layout generation with constraints," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 491–506.
[13] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 419–426.
[14] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 541–544.
[15] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1784–1791.
[16] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proc. CVPR*, 2011, pp. 33–40.
[17] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2206–2213.
[18] A. Lienhard, P. Ladret, and A. Caplier, "Low level features for quality assessment of facial images," in *Proc. 10th Int. Conf. Comput. Vis. Theory Appl.*, 2015, pp. 545–552.
[19] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 457–466.
[20] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 990–998.
[21] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2815–2826, Nov. 2019.
[22] H. Zhang and D. Xu, "Ethnic painting analysis based on deep learning," *Scientia Sinica Informationis*, vol. 49, no. 2, pp. 204–215, Feb. 2019.
[23] X. Zhang, X. Gao, W. Lu, L. He, and J. Li, "Beyond vision: A multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks," *IEEE Trans. Multimedia*, vol. 23, pp. 611–623, 2021.

[24] K. Sheng, W. Dong, M. Chai, G. Wang, P. Zhou, F. Huang, B.-G. Hu, R. Ji, and C. Ma, "Revisiting image aesthetic assessment via self-supervised feature learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 5709–5716.
[25] A. Rosenfeld, *Digital Picture Processing*. New York, NY, USA: Academic, 1976.
[26] R. Szeliski, *Computer Vision: Algorithms and Applications*. London, U.K.: Springer, 2010.
[27] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, 2011.
[28] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1028–1035.
[29] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2083–2090.
[30] A. Borji, "What is a salient object? A dataset and a baseline model for salient object detection," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 742–756, Feb. 2015.
[31] R. Nock and F. Nielsen, "Statistical region merging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1452–1458, Nov. 2004.
[32] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[33] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*.
[34] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2359–2367.
[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[36] X. Xia and B. Kulis, "W-net: A deep model for fully unsupervised image segmentation," 2017, *arXiv:1711.08506*.
[37] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.
[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1, 2019, pp. 4171–4186.
[39] A. Radford, J. Wu, R. Child, D. Luan, and D. Amodei, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[40] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1441–1451.
[41] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[42] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.
[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[44] J.-B. Grill, F. Strub, F. Altché, and C. Tallec, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.
[45] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: Large-scale information network embedding," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 1067–1077.
[46] H. Wang, J. Wang, J. Wang, M. Zhao, W. Zhang, F. Zhang, X. Xie, and M. Guo, "GraphGAN: Graph representation learning with generative adversarial nets," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018.
[47] F.-Y. Sun, J. Hoffman, V. Verma, and J. Tang, "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," in *Proc. Int. Conf. Learn. Represent.*, 2019.
[48] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[49] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[50] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Learning image representations by completing damaged jigsaw puzzles," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 793–802.
[51] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
[52] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.
[53] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 1192–1200.
[54] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

DONGJIN HUANG received the Ph.D. degree in computer application technology from Shanghai University, in 2011. He is currently an Associate Professor with the Shanghai Film Academy, Shanghai University. His research interests include computer graphics, virtual reality, and physical animation.

JINYAO LI received the B.S. degree in digital media technology from Shanghai University, Shanghai, China, in 2019, where she is currently pursuing the M.S. degree with the Shanghai Film Academy. Her research interests include machine learning and virtual reality.

CHUANMAN LIU received the B.S. degree in digital media technology from Shanghai University, Shanghai, China, in 2021, where she is currently pursuing the M.S. degree with the Shanghai Film Academy. Her research interest includes virtual reality.

JINHUA LIU received the B.S. degree in digital media technology from the Henan University of Urban Construction, Pingdingshan, China, in 2017, and the M.S. degree from the School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China, in 2020. She is currently pursuing the Ph.D. degree with the Shanghai Film Academy, Shanghai University, China. Her research interests include image processing, virtual reality, and physical animation.