Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Poetry4painting: diversified poetry generation for large-size ancient paintings based on data augmentation

Jiazhou Chen[a,*], Keyu Huang[a], Xinding Zhu[a], Xianlong Qiu[a], Haidan Wang[b], Xujia Qin[a]

[a]*College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China*
[b]*Research Centre for Digital Spatial Technology, Southeast Digital Economic Institutes, Quzhou, China*

ARTICLE INFO

Article history:
Received July 18, 2023

Keywords: Poetry generation, Painting poetry, Data augmentation

ABSTRACT

Chinese painting poetry is an extraordinary art form, which not only describes the painting contexts but also grasps the sentiment of the painters. In this paper, we propose an automatic poetry generation method *Poetry4painting*, which enhances the poetry diversity for large-size ancient paintings. The basic framework is based on multiple modern sentences, that are first captioned from the ancient painting and then used to generate a poem using CNN and LSTM. To solve the repeatability issue of this framework, four kinds of data augmentation are employed during online processing, including quantity, shape, surrounding, and object augmentation. In offline training, data augmentation is also used to create an image caption dataset with over 1500 painting images and 7500 captions. Through ablation studies, evaluations of poetry qualities and diversities, and comparisons with other methods, we demonstrate the validity of the proposed method.

1. Introduction

Poetry, painting, and calligraphy are known as the three perfections in ancient China. Painting poetry is a poem written in the empty space of a painting, regarded as a comprehensive integration of these three perfections. It significantly enhances the visual ambiance of the painting and expresses artists' personal feelings as well, thus guiding reviewers to better appreciate the painting. Taking the painting in Figure 1 as an example, poems on it express the context and emotions of both the painter and the other famous poet. Painting poetry has three common characteristics: 1) creation according to the theme, 2) transition from a static state to a dynamic state, and 3) imagination from nothing to something[1], thus heavily relying on artistic experiences to create. Therefore, only a small number of ancient Chinese paintings have poems on them.

Fig. 1. A real example "Ink brush plum blossom" of painting poetry created by the artist Wang Mian in the Yuan Dynasty.

To provide poem guidance for more ancient paintings, automatically generating painting poetry using artificial intelligence techniques has become a hot topic in both computer graphics and multimedia domains. Early methods for painting poetry generation are based on keywords[2]. Since keywords are all

*Corresponding author:
e-mail: `cjz@zjut.edu.cn` (Jiazhou Chen)

Fig. 2. Limitations of existing poetry generation methods for large-size ancient paintings. Two used-specified images are cropped and used to generate respective poems. Due to the content similarity of these two images, Jiuge[2] and iPoet[4] have a serious repeating issue. Note the repeated characters underlined by red dotted lines. ChatGPT[5] has sufficient diversity but does not conform to the painting content well. Therefore, it is very challenging to achieve the best balance between diversity and conformity.

extracted from the painting and used for poetry creation, the theme of the created poem is consistent with the input painting. But, keywords only describe static scenes, and can not reflect the dynamics of the scenery and people, thus lack of the transition from a static state to a dynamic state. To solve this issue, image caption techniques are employed to better describe the painting content[3], and a user-friendly interface is provided to achieve personalized poetry creation[4].

These learning-based methods work well for small-size ancient Chinese paintings, but not large-size ones, like the painting of "a panorama of rivers and mountains". For such a long-scroll painting, reviewers usually spend more time zooming in to watch its details, rather than have a glimpse of its global picture. Based on this observation, a user-friendly digital painting exhibition usually allows users to zoom in/out of the painting and drag a box of any size with their interests, like famous digital paintings in the online Google Culture & Art Project. However, the dragged patch from the same large-size painting often has quite repeated contents, which tends to result in unexpected repeatability for the painting poetry generation.

Figure 2 shows two different patches from the painting of "a panorama of rivers and mountains" (in the left column) and generated poems by three different existing methods on the right[4, 2, 5]. As shown in the middle two columns, poems generated by iPoet and Jiuge have many repeated words. The major reason for this repetition is twofold: 1) their algorithms are limited to the image content of the input image without sufficient reasonable imagination; 2) there is no large-size dataset dedicated to this task yet. In contrast, ChatGPT (Chat Generative Pre-trained Transformer)[5] shows a big power to answer questions, it can generate poems by integrating image parsing modules as well. The generated poems by ChatGPT have sufficient diversity but do not conform to the content of the input painting, as shown in the right column of Figure 2.

In this paper, we present a new painting poetry generation method, *Poetry4painting*, to achieve a balance between diversity and conformity. It adopts the modern-sentence-based poem generation framework from the iPoet[4], but it improves the diversity of the poetry generation by integrating data augmentation. To enlarge the dataset, we first build an iterative data annotation process that saves annotators a lot of labor and time. During the online poetry generation, four different data augmentation steps dedicated to our poetry generation task are proposed, including quantity, shape, surrounding, and object augmentation. In summary, the main contributions of our work are:

- A poetry generation method for large-size ancient paintings that reaches high diversity while preserving conformity of the painting.
- An enlarged dataset for painting poetry generation that is annotated in an efficient semi-automatic manner.
- A mass of experiments, including quantitive diversity evaluation, poetry quality evaluation, and ablation studies and comparisons, demonstrate the validity of the proposed method.

2. Related Work

2.1. Poetry generation

Automatic methods for generating poems have always been an important research direction in the computer graphics domain. Given several keywords, Zhou et al. employed genetic algorithms to generate poems[6], He et al. combined with a machine translation model to generate poetry sentences gradually[7]. When texts are provided, Wang et al. first extracted keywords from input texts and then generated corresponding poetry[8], Wang et al. converted modern literature to ancient poetry using neural networks[9], Yi et al. guided gradient updates through reinforcement learning to generate higher quality poetry[10].

To make the generating process more controllable, Hu et al. proposed different types of poetry generation under a unified framework[11], Chen et al. proposed the use of emotional control in poetry[12], Yi et al. showed the impact of different background factors on poetry and proposed a poetry generating method that can be controlled in a mixed latent space[13].

Given images, Guo et al.[2], Liu et al.[14] and Wu et al.[15] extracted keywords from images and then generate poetry using these keywords. Instead of generation, Xu et al.[16] and Liu et al.[17] directly match a written poem from the database according to keywords extracted from a painting. Chen et al. adopted image caption techniques to describe the input painting to modern sentences, and then generate poems[3].

Wang et al. proposed an unsupervised method to achieve an image-poetry conversion[18]. This method used latent codes to alleviate the collapse of the generative adversarial network and thus increased the poetry diversity, but it tended to generate painting-unrelated poetry words. Though automatically matching image-poetry pairs using MS COCO and CCPC[2] datasets saves a lot of annotation labor, it is prone to inappropriate correspondence between poetry results and images. Finally, it is still unknown whether this method can be directly applied to large-size paintings. And Feng et al. proposed a user interface to support personalized control of the content and the emotion of the resulting poems[4]. However, most of these methods focus on photos or small-size paintings, they can hardly generate high-quality poems with both high diversity and conformity for large-size paintings.

2.2. Painting semantic analysis

Object detection can capture and recognize objects in the image to obtain corresponding keywords. There are a mass of learning-based object detection algorithms in the literature. By extracting candidate boxes and classifying corresponding areas with deep learning methods, for example, Faster R-CNN[19], SSD[20], YOLO[21][22] and etc., Huang et al. opened the TensorFlow object detection API and made a detailed comparison of their performance[23].

The essence of image captioning is extracting image visual features and converting them into semantic information through a computer. In the early stage, Kulkarni et al. generated sentences in the form of template rules by extracting visual concepts[24], Vinyals et al. used the connection form of Encoder-Decoder[25], and combined CNN as an encoder with LSTM as decoder[26]. Based on this work, Xu et al. proposed attention mechanisms[27], and Lu et al. added adaptability[28]. Li et al. proposed a new feature extraction method to obtain a series of object detection boxes, which were taken as an image feature and delivered into an attribute detector[29]. Liu et al. employed an adversarial network to realize text and image mutual translation[30], and Chen et al. studied image adversarial samples[31]. Ashish et al.[32] described images with multiple languages rather than a single language. Recently, OpenAI proposed a multimodal model ChatGPT to answer questions according to provided texts and images[5], the answer can be a poem if required.

2.3. Data augmentation

Data augmentation is to produce more data whose content is close to the original data, it thus can not only increase the number but also improve the diversity of samples. The common data augmentation is addition and modification. In the computer vision domain, data augmentation methods such as flipping and rotation are used to increase image samples[33]. In the natural language processing domain, text data augmentation methods can be roughly divided into three categories: interpretation-based, noise-based, and sampling-based[34].

Interpretation can convey information consistent with the original text in natural language. Zhang et al. first used WordNet's synonym dictionary to classify and randomly replace data according to text similarity[35], and Zuo et al. used hypernym to randomly replace[36]. Then EDA was proposed, which includes data augmentation methods of synonym substitution, random insertion, random substitution, and random deletion[37]. Natural language processing techniques are used in a heuristic manner to augment data without changing the sentence semantics, including regular expressions[38], expanding the dictionary, and replacing the abbreviation[39]. The bidirectional translation, regarded as a kind of interpretation, can produce new sentences by backtracking [40][41].

Noise-based methods add noise that will not have a serious impact on semantics, including exchange, deletion, insertion, and replacement[42]. Spelling error lists are built to replace the original text[38, 39], and TF-IDF is used to select words to replace[43].

Sampling-based methods sample new data based on the distribution of the input data. It is usually designed according to specific tasks by artificial heuristic algorithms and training models, including pseudo-parallel sentence construction by non-training models[44], reinforcement sentence generation by pre-training models such as GPT-2 and DistilBERT[45, 46], and labeling unmarked sentence pairs using fine-tuned BERT in the input data[47].

The above methods can enrich data but lack additional information to expand the painting description. Therefore, this paper proposes an image-based data augmentation to solve the singularity problem in the work of large-size paintings, which can effectively add additional information to the limited text.

3. The overview of *Poetry4painting*

3.1. Framework

We propose a painting poetry generation framework, *Poetry4painting*. Taking modern Chinese as a medium, *Poetry4painting* automatically generates a poem according to part of the large-size painting image, which is cropped according to the box specified by the user. Our approach is mainly based on the framework of *iPoet*[4], which generates personalized painting poetry with the help of a visual interface.

However, as pointed out in the introduction section, cropped images at different places in the same painting are likely to have similar content, which leads to serious repetition for generated poems. We adopted it as a baseline without its visualization part. To solve the repetition issue, we integrated data augmentation into this framework to achieve a good balance between diversity and conformity.

Figure 3 shows the overall architecture of our *Poetry4painting* system, which mainly consists of three modules: analysis, augmentation, and generation. In offline training, a human-in-the-loop annotation tool is designed to significantly reduce the annotation time, and the annotated images and captions are both augmented by classical techniques. In the loop, a small number of datasets are trained to obtain results first, and then the low-quality model is used to generate a high-quality model by obtaining the corresponding annotations from large-size painting datasets with different slice sizes and correcting them manually.

Given the user-specified paintings, the online processing stage first extracts several keywords from paintings by object detection. The keywords' text and image are used as the input of the image2caption model to obtain modern Chinese through CNN and LSTM. The LSTM training is an iterative prediction

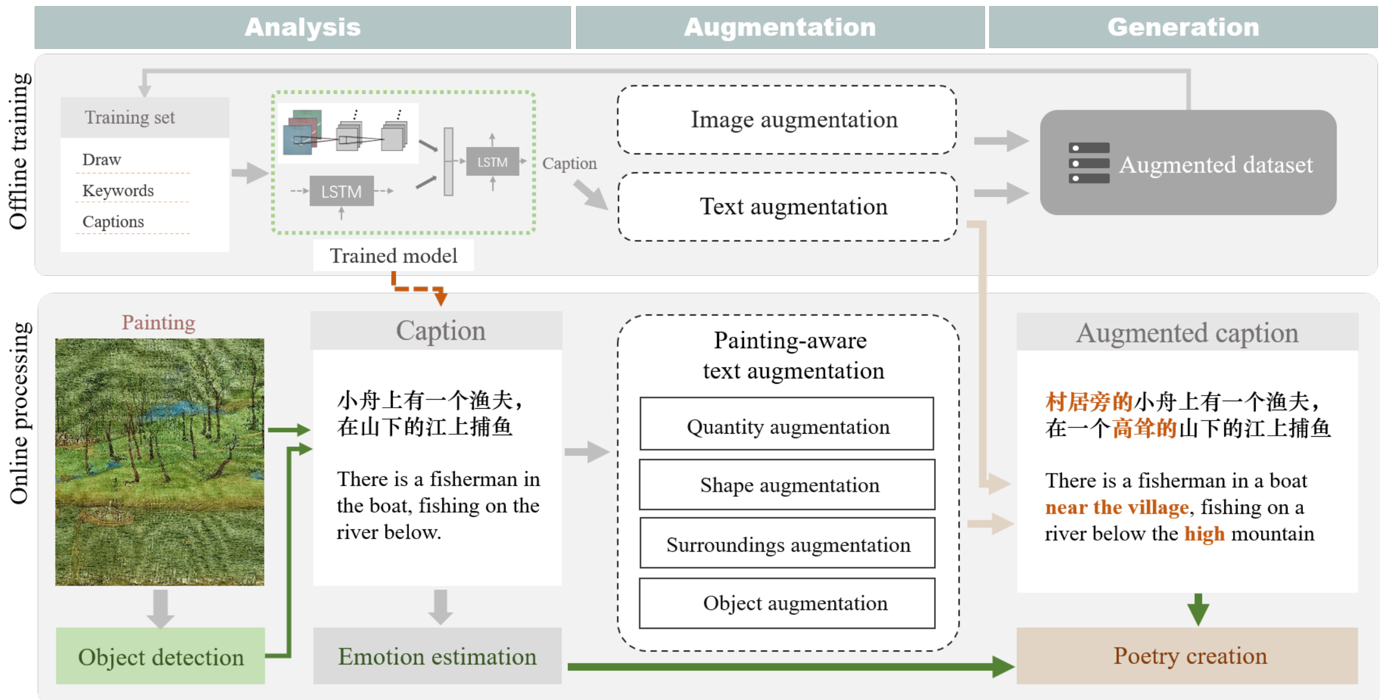


Fig. 3. An overview of our *Poetry4painting* framework. The first row shows the offline training stage, where both image and text augmentation are employed to facilitate the annotation in the dataset enlargement. The second row shows the online processing stage, where four different image-based text augmentation methods are proposed to improve the diversity of the painting poetry.

process of the next token, whose probability distribution is predicted according to embedding feature vectors of Chinese sentences, emotions, poetry tones, and contexts. The LSTM is implemented using TensorFlow and optimized using Adam.

To solve the repetition issue, we adopt 4 text augmentation methods that are aware of the input painting, including quantity augmentation, shape augmentation, surroundings augmentation, and object augmentation. The information from object detection can not only enrich sentences in the adjective way but also proofread and improve the content integrity between modern Chinese and input painting images. In terms of emotion, text information and image colors are used to extract emotions. Finally, emotion and the augmented modern Chinese are used as input to generate poetry.

3.2. Dataset

Small-size paintings. The dataset of this work consists of images from both small-size and large-size paintings. The set of small-size paintings is collected in [4]. It includes about 450 painting images and 2250 corresponding captions. Most of these small-size paintings are from Song Dynasty and are about life scenes.

Large-size paintings. Large-size paintings are characterized by long scrolls depicting large-scale scenes of a certain moment or a certain place. We thus collected 12 digital copies of large-size paintings, including “Thousand Miles of Rivers and Mountains”, “Dwelling in Fuchun Mountain”, “Clear Roaring in Yunshan”, etc. We implemented a program to evenly crop large-size painting images into small ones, whose width and height should be close to ρ pixels. The number of patches along the width and height is self-adaptive, as the side length of large-size paintings varies significantly(the longest/shortest side is 91406/1237 pixels). Each painting is segmented in multi-scale, thus ρ is set to 2000, 1500, 1000, and 500 pixels separately. After the automatic segmentation, we manually remove all unsuitable images, including ones that have only background content and a large area of calligraphy. We finally collected an image set with 1500 cropped images from these large-size 12 paintings.

Image augmentation. For all 1500 segmented images, we have to annotate 5 captions for each, that is a large caption set with 7500 captions in total, which costs too much labor and time. For this sake, we made use of the pre-trained image2caption model to first generate 5 captions, and then ask annotators to correct them manually. Such a semi-automatic annotation saves annotators a lot of time.

Text augmentation. After the annotation, we further augmented the dataset for both images and captions. For images, we employed horizontal flipping and noise addition to enlarge the image set. For captions, we used back-translation, synonym replacement, synonym insertion, and random deletion[37] to enlarge the caption set. Besides, keywords are automatically obtained from the annotation set by keyword extraction.

Thanks to these 462 (450 small-size + 12 large-size) paintings with various styles and 10 thousand corresponding captions, our dataset contains sufficient style diversities. Armed with the data augmentation below, our poetry generation method has the capability of dealing with different styles of Chinese ancient paintings, which is shown in case studies of Section 5.

4. Painting-aware text augmentation

As shown in Figure 3, the augmentation module is the core of our framework. Besides the image and text augmentation in the offline stage, online augmentation is crucial to achieving the diversity of painting poetry. To this end, we integrate four different data augmentation methods based on the input painting image and improve them for our diversified poetry generation task.

4.1. Quantity augmentation

The object quantity impacts the poetry result but is often ignored by existing methods. Figure 4(a-b) contains multiple mountains, while Figure 4(c-d) (in the dotted boxes) contains fewer mountains, where trees on these hills are paid more attention. We identify the number of objects and add modern Chinese with the corresponding word list. The quantitative addition can not only distinguish images with similar images but also provide assistance in obtaining emotion so that the information obtained can be incorporated into the poem as a reference.

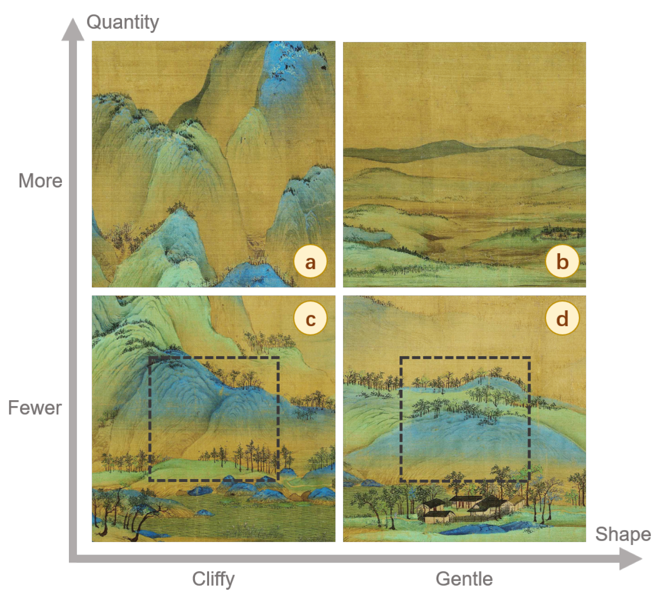


Fig. 4. Four cropped images from the same large-size painting, they have mountains with different quantities (vertical axis) and different shapes (horizontal axis). For this sake, we proposed quantity and shape augmentation. (a) Cliffy multiple mountains; (b) Gentle multiple mountains; (c) Fewer mountains nearby a river; (d) Fewer mountains nearby a village. Note that the cropping regions in (c-d) are drawn as dotted boxes, they are expanded for the surrounding augmentation.

4.2. Shape augmentation

Due to the fact that different shapes can express different emotions, it is important to modify captions based on object shapes. For instance, cliffy mountains in Figure 4(a) give a close look at a risky scene, while the gentle mountains in Figure 4(b) give a distant and peaceful landscape. They reveal opposite feelings for the object (mountain) with different shapes. The modifiers are distinguished according to the shape of the object and added to the modern Chinese sequence. Since the painting style of the author is fixed in the same large-size painting, each shape can be found in the corresponding description language by rules. A related adjective lexicon will be established for each shape. For example, the lexicon of (a) includes “towering”, “reaching into the sky”, etc., and the lexicon of (b) includes “continuous”, “multi-peaked”, etc. The related adjective will be randomly picked from the lexicon and added before the noun corresponding to the object “mountain”.

4.3. Surrounding augmentation

For cropped painting images that have contents with identical quantities and shapes, the generated poetry still lacks diversity. To meet the characteristic of “changing nothing into something”, we enlarge the user-specified region to detect more objects nearby, which is regarded as a reasonable imagination. For instance, although there are similar mountains in the dotted boxes of (c) and (d) of Figure 4, rivers outside the box in (c) and villages outside the box in (d) can be explored, thus their resulting poems will be distinguished. In contrast with a free imagination, our imagination is based on the painting content of the expanded region, which is obtained by enlarging twice the user-specified region with the center fixed.

The general idea is to add objects detected from the expanded region to augment the modern sentences describing the user-specified image, as shown in Figure 5. For this sake, target objects are detected for both regions, denoted as $P = \{p_1, p_2, \cdots, p_n\}$ for the user-specified region and $Q = \{q_1, q_2, \cdots, q_m\}$ for the expanded region respectively. To avoid word redundancy, we remove objects in Q that are already detected in P using label comparison. For each remaining object in Q, a correlated object in P is found:

$$\bar{p}_j = argmax_i \left\{ r(p_i, q_j) \right\} \tag{1}$$

The correlation of objects $r\left(p_i, q_j\right)$ is calculated according to the repetition coverage and distance:

$$r(p_i, q_j) = c(p_i, q_j) \cdot \beta + d(p_i, q_j) \cdot (1 - \beta) \tag{2}$$

where the distance $d(p_i, q_j)$ is the Euclidean distance of their centers and the repetition coverage is:

$$c(p_i, q_j) = \Omega(B_{p_i} \cap B_{q_j}) \,/\, \Omega(B_{q_j}) \tag{3}$$

where B_p is the bounding box of the detected object p and $\Omega()$ represents the area of this bounding box.

With the object correlation, we turn the remaining objects in Q into adjectives to insert into the modern Chinese of this image. More precisely, we establish an adjective word list for all objects and obtain the corresponding word list based on the correlation object set $\bar{p}_j$. The adjectives are selected from the word list and added in front of the objects that are detected from the original image, as shown in Figure 5. The way of adding adjectives will not change the original subjects, verbs, and objects, but can be added to the modern Chinese in a way of extra information to improve the effect of subsequent poems. Finally, we use EDA (Easy Data Augmentation)[37] to enrich modern Chinese by means of interpretation and noises.

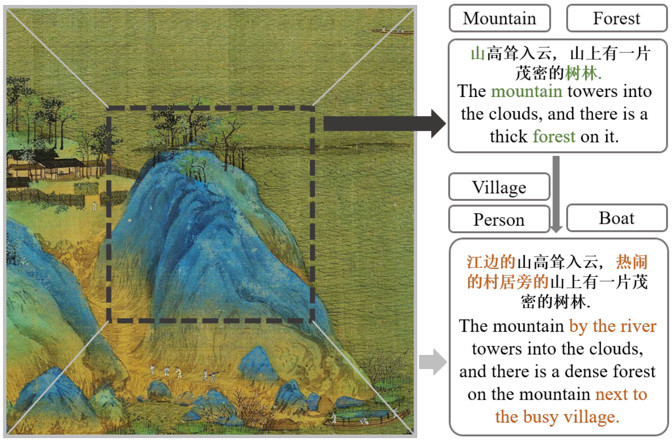


Fig. 5. An illustration of surrounding augmentation. The surrounding expansion is shown on the left. Two keywords and two sentences in the top right corner are extracted from the original image in the dotted box; while three keywords and two sentences in the bottom right corner are extracted from the expanded image in the solid box.

4.4. *Object augmentation*

4.4.1. *More objects*

To better generate modern Chinese that fits the input painting image, this paper first performs object detection on the painting, and then, the object is used as input together with the painting. Finally, modern Chinese is obtained by using a model trained with an extended dataset. To improve the completeness of the content and the matching between poems and paintings, missing keywords are added as reference texts for subsequent poems. Similar to previous works[27], we combine the image feature vector extracted from the painting using InceptionV3[48] with the object text converted into the word embedding vector as the input of the image description encoder, and the mixed representation is obtained through splicing and linear projection:

$$e(w_{-1}) = f(CNN(I), e(R(I))) \tag{4}$$

where f means the merging of vectors, $CNN(I)$ is the feature vector of image I , and $e(\cdot)$ represents the embedding, $R(I)$ is the object detected from I. The image feature vector uses the recursive neural network to take the variable length input encoding as a fixed length feature vector and input it to the decoding terminal to obtain the output result sequence. We use LSTM as a decoder to train our modal and generate Chinese sentences $v_i = \{w_t\}_{t=1}^{L_p}$, where w_t is the word, L_p means the sentence length. The hidden state of LSTM is:

$$s_t = LSTM(s_{t-1}, [e(w_t); e(w_{t-1}); c]), t \in \{0, \cdots, N-1\} \tag{5}$$

where $e(w_t)$ represents the embedding vectors that generate words, $e(w_{-1})$ is the initial image input from Equation 4. Generated texts are delivered to LSTM to generate the next word in sequence, the probability of the generation of each word w_t is:

$$p\left(w_t \middle| w_{0:N-1}, I, W\right) = Softmax(\gamma \times [s_t; c]) \tag{6}$$

where γ is a projective parameter, and the attention c is calculated as :

$$c = \sum_{k=1}^{L} r_k \times Softmax(s(r_k, q)) \tag{7}$$

where L represents features extracted at different image locations, $s(r_k, q)$ is a function of attention scoring, q is a query vector, k is the position sequence number, r_k represents the sequence of solving attention at location k.

Finally, we add the missing objects of results to the modern Chinese, and the final set of sequences v is:

$$v = \{w_t \mid t = 1, \ldots, N-1\} \cup R(I) \tag{8}$$

The caption can be filled with a specific length embedding according to the unmentioned image content. Similar to previous works[4], the extended four-sentence long sequence is encoded by GRU[49].

4.4.2. *Emotion*

Poetry can express thoughts and emotions, and painting poetry has no exception. To achieve "creation by theme", we obtain the painting's emotion through a combination of color features and object features:

$$E(I) = E_{color}(I) \cdot \omega + E_{object}(I) \cdot (1 - \omega) \tag{9}$$

where E_{color} is a color-based emotion estimation model[50] introduced by iPoet[4], and ω is a parameter to balance the two emotion factors, which is set to 0.5 for all our experiments. However, emotion can be expressed by not only colors but also objects. Thus, we propose an object-based emotion estimation E_{object} to improve:

$$E_{object}(I) = \sum_{i=1}^{n} E(p_i) \tag{10}$$

where $E(p_i)$ is the emotional vector of the object p_i[51].

5. Experiments

5.1. *Ablation study*

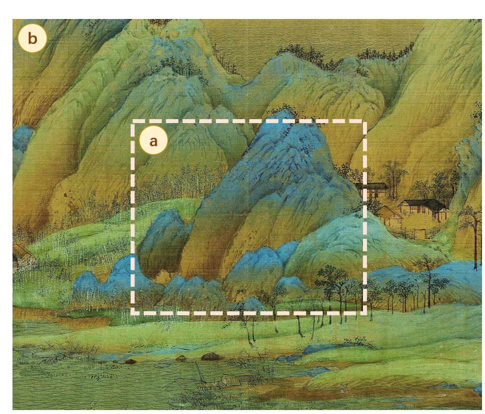


Fig. 6. A painting selected by the user. (a) user selection; (b) surroundings of user selection.

(a)Original	(b)Shape&Quantity	(c)Surroundings&Object	(d)Final result
山峰高耸入云，一片片树林长在山坡上 The peaks towered into the clouds, and patches of trees grew on the hillside.	碧绿的山峰高耸入，一片片树林长在连绵的山坡上 The green mountains rose high into the sky, and trees grew on the continuous slopes.	山峰在庭院旁高耸入云，一片片小溪边的树林长在山坡上，松树 The peaks towered into the clouds beside the courtyard, and patches of woods by the stream grew on the hillside, with pines.	碧绿的山峰在庭院旁高耸入云，一片片小溪边的树林长在连绵的山坡上，松树 The green mountains rose high above the courtyard, and the woods by the stream grew on the continuous hills, with pines.
峰山有合望云开， The mountains are in harmony with the clouds. 世树无生一半来。 Half the trees won't be born. 岂是秋关何处事， Is it anything to do in autumn? 时知不动故人哀。 Time cannot change the sorrow of friends.	山高树落碧崖来， The verdant hills are high and the trees are falling. 远路连山有此开。 There are winding roads on the continuous hills. 令疏风生何所恨， What do you hate when the wind blows? 林间隔草暗寒苔。 The grass between the trees is dark with moss.	山庐野寺上云台， There are simple huts, temples, and towering platforms on the hill. 一树浮仙万里来。 There was a fairy in the tree who had come from afar. 不觉天门先有石， Did not realize that there was a stone on the gate. 旅人相思过江开。 After crossing the river, the traveler aroused a lovesickness.	青山野色出庭凉， I feel cool in the courtyard on the green hills. 一树溪头对华堂。 By the stream there is a tree and a big, wide building. 洞壑连松高下里， There are many caves and pines on the continuous hillside. 时来只是数新塘。 Time passed and only many ponds remained.

Fig. 7. The generated poems with/without different augmentations. (a) Poems generated using the original painting without augmentations; (b) Poems generated using the original painting with only shape and quantity augmentations; (c) Poems generated using the original painting with only surrounding and object augmentations; (d) Poems generated using all augmentations.

Inspired by [10][13], We use similarity to automatically evaluate the impact of different factors. As shown in Figure 6, (a) represents the painting selected by the user, and (b) represents the painting of the surrounding objects. Figure 7 shows the effect of each step on the generation of the poem. Figure 7(a) is the result of painting poetry, which contains the main content of the selected ancient painting such as mountains and woods, but there is no more diverse description. In Figure 7(b), the extended modifier of the mountain is reflected in the verse by the word "verdant". The second line of the poem describes the continuous mountain roads and expresses the wonder of the magnificent mountains. The pine is derived from object detection, but it is not reflected in the original modern Chinese. The surrounding objects are used to add modifiers in the original modern Chinese, as shown in the blue part of Figure 7(c). These modifiers and complementary objects will be reflected in the generation of the poem. Finally, all the steps are combined to obtain the expanded poem, and the generated poems do not abandon any step of the supplement.

We obtained 60 randomly selected images from different large-size paintings to generate poems and used similarity calculation to conduct ablation experiments for each step. As suggested by Yi et al.[13], Jaccard is used to evaluate similarity automatically, TF-IDF and Levenshtein distance methods[52] are added to ensure fairness. The results of similarity calculations using different steps are compared in Table 1 to demonstrate that the augmentation approach helps diversify modern Chinese.

From the table, it shows that all the steps can diversify the modern Chinese, which is conducive to reducing similarity, and ultimately affecting the diversified generation of poetry.

Modern Chinese	Jaccard	TF-IDF	Levenshtein
Original sequence	36.0%	41.1%	36.2%
Shape & Quantity	32.5%	36.3%	31.8%
Surroundings & Object	23.1%	26.6%	28.2%
Final sequence	19.9%	21.1%	24.9%

Table 1. Three different similarity evaluations of modern Chinese texts using different steps of our method.

5.2. Diversity evaluation

The diversity of painting poetry requires evaluating not only the similarity of the results but also whether the content of the poems fits the paintings. Based on the similarity evaluation[13], we suggest using a mixture of thematic conformity and textual similarity calculations to obtain an overall diversity evaluation of painting poetry. The higher the evaluation of the diversity of the poems, the better and more diverse the results of the painting poetry.

The thematic keyword W comes from object detection, and the thematic conformity $C(A_k, W)$ of the poem and the keywords is calculated by Jaccard. If the keywords do not appear, it means that the poem does not match the topic. We first compute the similarity of two sets:

$$S(A_k) = \frac{\sum_{i \neq k} similar(A_k, A_i)}{N-1} \tag{11}$$

where N is the total number of poems in a collection, A is the vectorization result after word segmentation and eigenvalue calculation, $similar(A_k, A_i)$ calculates the similarity. And then the overall diversity evaluation D_k of the poems is:

$$D_k = C(A_k, W) \cdot \alpha + \frac{1}{S(A_k)} \cdot (1-\alpha) \tag{12}$$

where α is a hyper-parameter to balance the weights of $S(A_k)$ and $1/C(A_k, W)$, is set to 0.3 for all our evaluation.

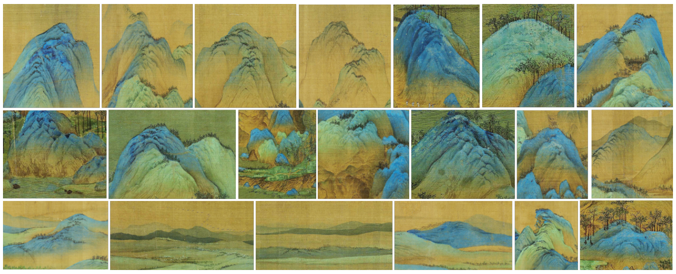


Fig. 8. Painting images that are randomly selected from the large-size ancient painting "A thousand Miles of Rivers and Mountains".

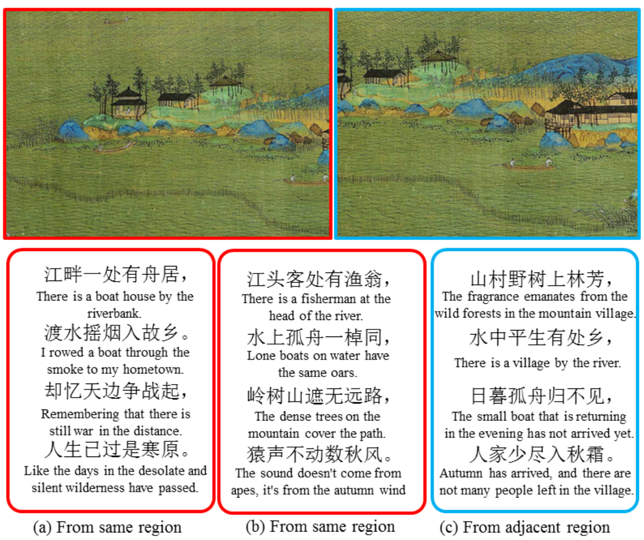


Fig. 9. Comparison with identical and adjacent regions of the large-size painting. The left and the middle poems are generated using the same left painting, while the right poem is generated using the right painting, which has half overlapping with the left painting.

To assess the overall diversity more precisely, we chose to capture images in the same painting, so that the input images are painted in the same style and the content is more repetitive. Twenty pictures with similar contents were randomly selected for experimental detection, as shown in Figure 8.

Figure 9 compares three poems generated by our method using an identical region and an adjacent region of a large-size painting. Though both of left two poems are generated by the same left painting above, they still show sufficient differences. And the right poem has more discrepancies, as it uses an adjacent region with half overlapping with the left one. This evaluation demonstrates the capability of our method for increasing poetry diversities, even if the input paintings are very similar.

The similarity is calculated using the average value of different methods, including Jaccard, Levenshtein distance, and TF-IDF. Our results were compared with other algorithms for painting poetry. We compare the following STOA (the-state-of-art) methods:

Jiuge[2]: a human-machine collaborative Chinese classical poetry generation based on keywords.

iPoet[4]: an automatic painting poetry generation based on modern Chinese with visual multimodal analysis.

ChatGPT[5]: a large-scale and multimodal model which can accept image and text inputs and produce text outputs, but not specifically for painting poetry.

As shown in Table 2, *Poetry4painting* gets the highest score compared to other models: *iPoet* can achieve results with high thematic conformity, however, due to the consistent content of the images, the generated modern Chinese is very similar resulting in a lack of diversity in the obtained poems. Since *ChatGPT* is not a dedicated model for painting poetry, the generated poems are more in lack of thematic conformity. And *Jiuge* lacks thematic conformity due to the method of keywords, it is easy for some important information in the images to be missing, and the keyword extensions to deviate from the paintings.

5.3. Poetry quality evaluation

Based on previous work[10], manual evaluation is used to evaluate the quality of poetry, because automatic methods such as BLEU deviate from the human evaluation manner. Following[53][54], we consider: consistency (is this poem related to painting in terms of content and emotion?), diversity (can poetry augment imagination in a less repetitive way?), and relevance (is the diversity of this poem augmented within the reasonable imagination of the painting?).

We invited 5 experts in the field of humanistic poetry and ancient painting, as well as 15 general users with no background in literature or painting, to conduct a user study. We conducted separate experiments and evaluations on the same and various large-size paintings. Each person independently selected 10 images of the same large-size painting and 10 images of various large-size paintings in order to obtain the corresponding painting poetry. The results of the poems generated by the different methods were put together and scored by the users based on the three scoring criteria above, the score out of 10, and the average score results are shown in Figure 10.

For both ordinary users and professional reviewers, the scores of our method are higher than other methods, indicating that our method can generate higher quality and more diverse large-size painting poetry. The results obtained by *iPoet* and *Jiuge* are highly repetitive, which make it difficult to achieve the diversity of poetry. And *Jiuge* is easy to generate poems that do not conform to the content, or generate poems with unreasonable expansion. The poetry results of *ChatGPT* are diverse but often inconsistent with the painting in terms of content.

In terms of consistency in our method, participants stated that "the poem accurately describes the objects in the painting, showing the magnificent", "emotionally indicating the author's attitude towards life". In terms of diversity, "the imaginary part of the method is innovative and more diverse results are generated than other methods". In terms of relevance, "combining people's experience, a mountain stream can be imagined from the layout of objects such as rocks and trees in the painting", "Other generative methods refer to 'fisherman' and 'sand', which is not consistent with the painting".

Models	Input	Similarity↓	Conformity↑	Diversity↑
Poetry4painting	Painting&Extended Painting	**5.3%**	**100%**	**28.8**
Jiuge[2]	Painting	5.4%	30%	21.4
iPoet[4]	Painting	11.4%	100%	19.1
ChatGPT[5]	Painting	5.8%	65%	23.8
Jiuge[2]	Extended Painting	6.2%	35%	19.6
iPoet[4]	Extended Painting	11.6%	100%	18.6
ChatGPT[5]	Extended Painting	5.7%	55%	23.0

Table 2. Automatic diversity evaluation of our results. *Poetry4painting* can automatically obtain information about extended painting based on input painting. The average similarity is calculated by Jaccard similarity, TF-IDF, and Levenshtein distance. ↑ indicates higher is better, ↓ indicates lower is better.

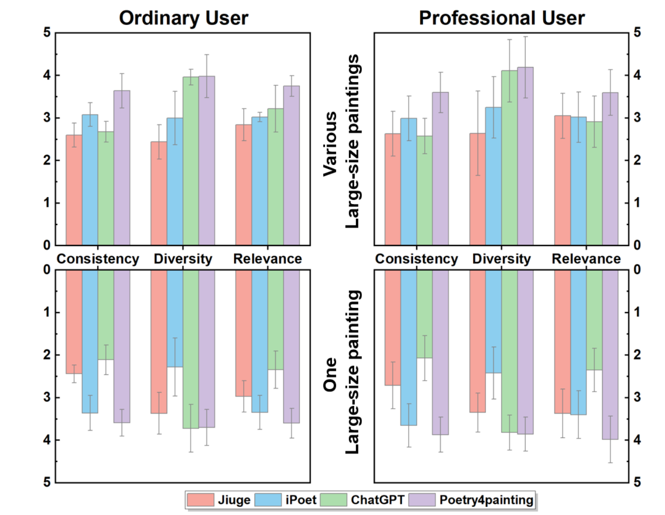


Fig. 10. Statistics of human evaluation. We evaluated both ordinary users (left) and professional users (right). The diversity is measured among cropped images from multiple large-size paintings (top) or from a single large-size painting (bottom).

5.4. Case study

First of all, *Poetry4painting* is compatible with both small-size paintings and large-size paintings. The surrounding augmentation will not work for small-size paintings, as the whole image is considered as the input. Since the focus of this paper is not small-size paintings, we take two case studies of large-size paintings here with very different painting styles, as shown in Figure 11. It shows the comparison of generated poems with the existing three SOTA methods. The source images come from two different large-size paintings. The dotted boxes show the user-specified regions, and our *Poetry4painting* always uses both these regions and the expanded regions. To reach fair comparison, other SOTA methods use the expanded images in Figure 11(a), while they use the user-specified images in Figure 11(b).

Figure 11 shows that *iPoet* and *Jiuge* tend to produce repetitive words and phrases, underlined by the red dotted lines. We also find that important information about the painting, such as "river" and "village", is lost in the poems generated by *Jiuge*, resulting in a shift in the focus of the poem. Thus the poems do not match the theme of the painting. Although *ChatGPT* can generate a variety of poems, the quality of painting poetry is unstable. *ChatGPT* may generate poems with inconsistent content or even completely unrelated poems. For instance, the left result in (b) is about ponds and animals, but the painting is about mountains.

Poetry4painting can create high-quality and diverse painting poetry. The *Poetry4painting* poem in (a) accurately describe "mountain", "village" and "river", and contain reasonable augmentation. On the left, the poem is augmented with additional objects such as "the bridge" and "the shadow of the willow", which combine with the word "quiet" to create a scene of a quiet village with willow trees growing around. The word "ferry" in the poem on the right reflects the dynamics of the canoe in the painting. This poem displays emotions from the sadness of parting to relief in the second and fourth sentences. According to the expanded painting in (b), "river" was added to the left poem, and "temple" was added to the right poem. Because the information is taken from the objects around the painting, we can generate more reasonable augmentations.

6. Conclusion and Future Work

In this work, we propose a poetry generation method for large-size ancient paintings, *Poetry4painting*, which achieves the best balance between diversity and conformity. To address the issue of insufficient training data, iterative expansion of offline training is carried out using both image and text augmentation methods. In the online processing stage, 4 different painting-aware text augmentation methods are proposed to enrich the modern Chinese, including quantity, shape, surrounding, and object augmentations. Then the data augmentations are integrated into the framework of painting poetry generation based on multi-sentence modern Chinese, and combined with emotional analysis. Through ablation study, quantitative diversity evaluation, poetry quality evaluation, and comparisons with SOTA methods (i.e. *iPoet*, *Jiuge*, *ChatGPT*), we demonstrate the effectiveness of our model.

Our framework is limited to large-size landscape paintings, not working well for other types of ancient paintings, such as character scenes, and bustling markets. In such scenarios, it may be important to identify specific events in ancient paintings and use them to generate poetry. Another potential direction is to enrich the categories of emotions, considering other factors to achieve more multi-dimensional acquisition of emotions, such as social background and biographies of historical characters.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62172367), Natural Science Foundation of Zhejiang Province (LGF22F020022), Zhejiang Science and Technology Project of Cultural Relic Protection (2020014) and the Science and Technology Project of Quzhou (2021K02).

References

[1] Q. Zhang, J. Meng, Painting poem appreciation strategy, Middle school Chinese: Greater Language Forum 1 (2011) 2.
[2] Z. Guo, X. Yi, M. Sun, W. Li, C. Yang, J. Liang, H. Chen, Y. Zhang, R. Li, Jiuge: A human-machine collaborative chinese classical poetry generation system, in: Proc. of ACL., 2019, pp. 25–30.
[3] J. Chen, K. Huang, Y. Feng, W. Zhang, S. Tan, W. Chen, Automatic poetry generation based on ancient chinese paintings, Computer-Aided Design and Computer Graphics 33 (7) (2021) 1038–1044.
[4] Y. Feng, J. Chen, K. Huang, J. K. Wong, H. Ye, W. Zhang, R. Zhu, X. Luo, W. Chen, ipoet: interactive painting poetry creation with visual multimodal analysis, Journal of visualization 25 (3) (2022) 671–685.
[5] OpenAI, Openai:gpt-4 technical report, arXiv.2303.08774 (2023).
[6] C. Zhou, W. You, D. Xiaojun, Genetic algorithm and its implementation of automatic generation of chinese songci, Journal of Software 21 (3) (2010) 427–437.
[7] J. He, M. Zhou, L. Jiang, Generating chinese classical poems with statistical machine translation models, in: Proc. of AAAI Conference on Artificial Intelligence., 2012, pp. 1650–1656.
[8] Z. Wang, W. He, H. Wu, H. Wu, W. Li, H. Wang, E. Chen, Chinese poetry generation with planning based neural network, in: Proc. of COLING 2016., 2016, pp. 1051–1060.
[9] D. Wang, Z. Bai, Y. Feng, The conversion method and device of modern prose to ancient poetry based on lstm model: China.
[10] X. Yi, M. Sun, R. Li, W. Li, Automatic poetry generation with mutual reinforcement learning, in: Proc. of EMNLP., 2018, pp. 3143–3153.
[11] J. Hu, M. Sun, Generating major types of chinese classical poetry in a uniformed framework, arXiv.1803.02994 (2018).
[12] H. Chen, X. Yi, M. Sun, W. Li, C. Yang, Z. Guo, Sentiment-controllable chinese poetry generation, in: Proc. of IJCAI., 2019, pp. 4925–4931.
[13] X. Yi, R. Li, C. Yang, W. Li, M. Sun, Mixpoet: Diverse poetry generation via learning controllable mixed latent space, in: Proc. of the 2020 Conference on Artificial Intelligence., 2020, pp. 9450–9457.
[14] D. Liu, Q. Guo, W. Li, J. Lv, A multi-modal chinese poetry generation model, in: 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1–8.
[15] L. Wu, M. Xu, S. Qian, J. Cui, Image to modern chinese poetry creation via a constrained topic-aware model, ACM Trans. Multimedia Comput. Commun. Appl. 16 (2) (2020) 1–21.
[16] L. Xu, L. Jiang, C. Qin, Z. Wang, D. Du, How images inspire poems: Generating classical chinese poetry from images with memory networks, arXiv.1803.02994 (2018).
[17] L. Liu, X. Wan, Z. Guo, Images2poem: Generating chinese poetry from image streams, in: Proc. of ACM Multimedia Conference., 2018, pp. 1967–1975.
[18] J. Wang, H. Li, C. Wu, F. Gong, L. Wang, Generating diverse chinese poetry from images via unsupervised method, Neurocomputing 492 (C) (2022) 188–200.
[19] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis Machine Intelligence 39 (6) (2017) 1137–1149.
[20] C. B. Alexander, C. Fu, S. Christian, A. Dragomir, E. Dumitru, R. Scott, W. Liu, Ssd: Single shot multibox detector, in: Proc. of ECCV., 2016, pp. 21–37.
[21] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proc. of CVPR., 2016, pp. 779–788.
[22] J. Redmon, A. Farhadi, Yolo9000: Better, faster, stronger, in: Proc. of CVPR., 2017, pp. 6517–6525.
[23] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, K. Murphy, Speed/accuracy trade-offs for modern convolutional object detectors, in: Proc. of CVPR., 2017, pp. 3296–3297.
[24] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, T. L. Berg, Baby talk: Understanding and generating simple image descriptions, in: Proc. of CVPR., 2011, pp. 1601–1608.
[25] K. Cho, D. Bahdanau, F. B. Holger Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv.1406.1078v3 (2020).
[26] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proc. of CVPR., 2015, pp. 3156–3164.
[27] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proc. of ICML., 2015, pp. 2048–2057.
[28] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: Proc. of CVPR., 2017, pp. 3242–3250.
[29] N. Li, Z. Chen, Image captioning with visual-semantic lstm, in: Proc. of IJCAI., 2018, pp. 793–799.
[30] B. Liu, K. Song, Y. Zhu, M. G. de, A. Elgammal, Time: text and image mutual-translation adversarial networks, arXiv.2005.13192 (2020).
[31] H. Chen, H. Zhang, P. Chen, J. Yi, C. Hsieh, Attacking visual language grounding with adversarial examples: A case study on neural image captioning, in: Proc. of ACL., 2018, pp. 2587–2597.
[32] V. T. Ashish, P.-T. Jordi, X. Chen, R. Soricut, Crossmodal-3600: A massively multilingual multimodal evaluation dataset, arXiv.2205.12522 (2022).
[33] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, F. Shen, A survey on image data augmentation for deep learning, Journal of Big Data 6 (1) (2019) 1–48.
[34] B. Li, Y. Hou, W. Che, Data augmentation approaches in natural language processing: A survey, arXiv.2110.01852 (2022).
[35] X. Zhang, J. Zhao, L. Yann, Character-level convolutional networks for text classification, in: Proc. of NIPS., 2015, pp. 649–657.
[36] X. Zuo, Y. Chen, K. Liu, J. Zhao, Knowdis: Knowledge enhanced data augmentation for event causality detection via distant supervision, in: Proc. of International Conference on Computational Linguistics., 2020, pp. 1544–1550.
[37] J. Wei, K. Zou, Eda: easy data augmentation techniques for boosting performance on text classification tasks, in: Proc. of EMNLP-IJCNLP., 2019, pp. 6381–6387.
[38] C. Coulombe, Text data augmentation made simple by leveraging nlp cloud apis, arXiv.1812.04718 (2018).
[39] M. Regina, M. Meyer, S. Goutal, Text data augmentation: Towards better detection of spear-phishing emails, arXiv.2007.02033 (2021).
[40] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, Q. V. Le, Qanet: Combining local convolution with global self-attention for reading comprehension, in: Proc. of ICLA., 2018, pp. 1–16.
[41] A. R. Fabbri, S. Han, H. Li, H. Li, M. Ghazvininejad, S. Joty, D. Radev, Y. Mehdad, Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation, in: Proc. of NAACL-HLT., 2021, pp. 704–717.
[42] D. Zhang, T. Li, H. Zhang, B. Yin, On data augmentation for extreme multi-label classification, arXiv.2009.10778 (2020).
[43] Z. Xie, S. I. Wang, J. Li, D. Lévy, A. Nie, D. Jurafsky, A. Y. Ng, Data noising as smoothing in neural network language models, in: Proc. of ICLR., 2017, pp. 1–12.
[44] Y. Zhang, T. Ge, X. Sun, Parallel data augmentation for formality style transfer, in: Proc. of ACL., 2020, pp. 3221–3228.
[45] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Zwerdling, Do not have enough data? deep learning to the rescue!, in: Proc. of AAAI., 2020, pp. 7383–7390.
[46] H. Abonizio, S. B. Junior, Pre-trained data augmentation for text classification, in: Proc. of BRACIS., 2020, pp. 551–565.
[47] N. Thakur, N. Reimers, J. Daxenberger, I. Gurevych, Augmented sbert: data augmentation method for improving bi-encoders for pairwise sentence scoring tasks, in: Proc. of NAACL-HLT., 2021, pp. 296–310.
[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proc. of CVPR., 2016, pp. 2818–2826.
[49] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv.1412.3555 (2014).
[50] F. Takahashi, Y. Kawabata, The association between colors and emotions for emotional words and facial expressions, Color Research Application

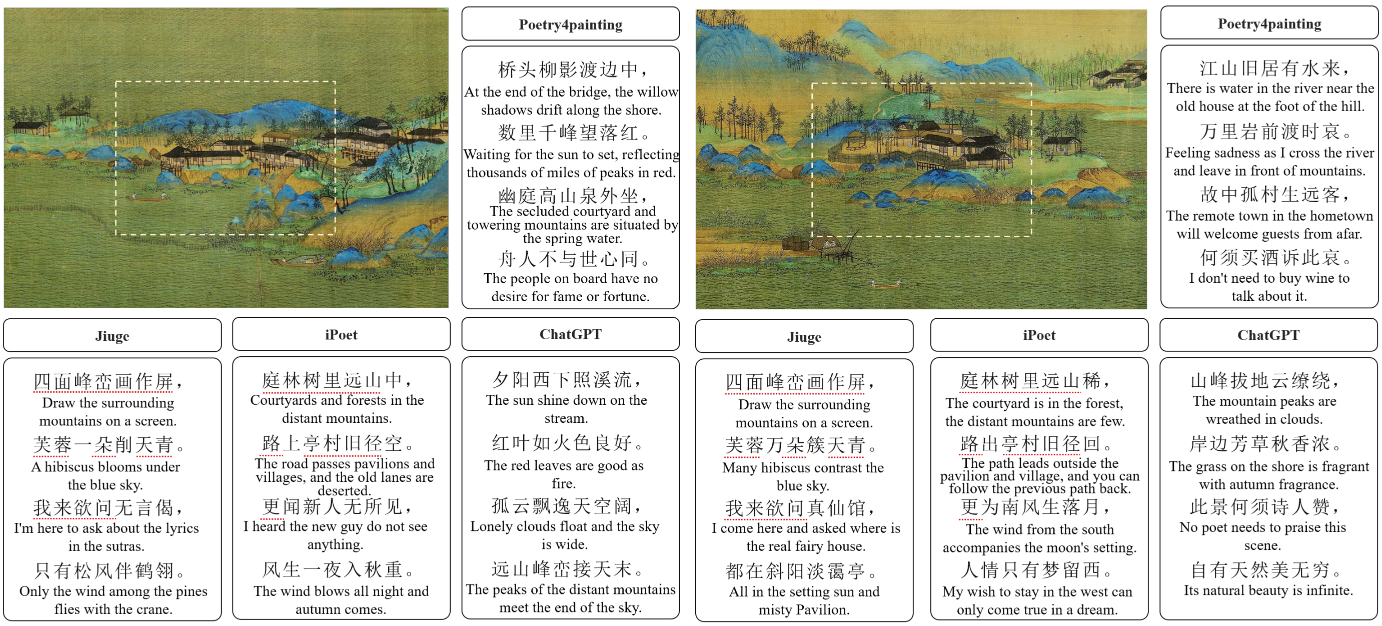
43 (2) (2018) 247–257.

[51] Y. Jiang, Language, object, image, and artistic conception, Literary Review 3 (2002) 69–75.

[52] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, K. Kochut, A brief survey of text mining: Classification, clustering and ex-traction techniques., arXiv.1707.02919 (2017).

[53] R. Yan, i, poet: automatic poetry composition through recurrent neural networks with iterative polishing schema, in: Proc. of IJCAI., 2016, pp. 2238–2244.

[54] J. Zhang, Y. Feng, D. Wang, Y. Wang, A. Abel, S. Zhang, A. Zhang, Flexible and creative chinese poetry generation using neural memory, in: Proc. of IJCAI., 2017, pp. 1364–1373.



(a) Comparisons of the painting "Thousand miles of rivers and mountains" with expansion

(b) Comparisons of the painting "Range upon range of mountains" without expansion

Fig. 11. Comparisons of generated poems with the-state-of-the-art methods. Dotted boxes shows user-specified regions, while the overall images are the expanded ones. Our method always use both of them, while other methods use either the expanded ones (a) or the original ones (b). Repeated characters in poems are underlined by red dots, which can be found in large amounts for iPoet and Jiuge. Though ChatGPT is free of repetition, it usually creates content without any correlation.