

# Image Captioning on Fine Art Paintings via Virtual Paintings

Yue Lu

*The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences*  
Beijing, China  
luyue2016@ia.ac.cn

Xingyuan Dai

*The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences*  
Beijing, China  
daixingyuan2015@ia.ac.cn

Chao Guo

*The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences*  
Beijing, China  
guochao2014@ia.ac.cn

Fei-Yue Wang\*

*The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences*  
Beijing, China  
feiyue.wang@ia.ac.cn

**Abstract**—Machine learning in fine art paintings is attracting increasing attention recently. Image captioning of paintings is of great importance for painting analysis, but it is rarely studied. The paintings have abstract expressions and lack annotated datasets, leading to the data-hungry problem in painting captioning. Thus, painting captioning has more significant challenges than photographic image captioning. This paper makes a novel attempt at generating content descriptions of paintings. We generate virtual paintings using the style transfer technique to deal with the data-hungry problem, then train the painting captioning model via a two-step manner. We evaluate our method on an annotated small-scale painting captioning dataset and demonstrate our improvements.

**Index Terms**—image captioning, fine art paintings, style transfer

## I. INTRODUCTION

Machine learning in fine art paintings is attracting increasing attention recently, and many tasks have been explored [1]. The image captioning of paintings is a valuable task due to its high information output and wide applications, but it is rarely studied. This paper investigates how to generate the content descriptions of paintings. Specifically, we investigate the image captioning task of impressionism painting as it is a representative painting style and contains many daily life scenes. This task has potential value in automatic painting description generation, text-based painting searching, visually impaired assistance in art appreciation, and early childhood education in art lessons.

Although the paintings have their digitalized version in online museums, there are no painting datasets annotated with proper content descriptions. Besides, the paintings usually have abstract expressions, and the painting captioning task needs a large quantity of annotated data. Thus, the painting captioning task meets the problem of data-hungry, which is a big challenge.

This paper uses the photographic image captioning dataset to help in settling the problem of data-hungry. To use the

data from the photographic domain, we use virtual data as a bridge to reduce the gap between paintings and photographic domains. Specifically, we generate annotated virtual paintings from the annotated photographic images by the style transfer technique [2]. The virtual paintings are a stylized version of photographic images, and they share the same annotations. We train an image captioning model on the photographic dataset and then fine-tune it on our virtual paintings, and we do not train the model on any real paintings. We annotate a real painting dataset to test our model's performances, which presents meaningful improvements. We hope this work to be a potential basis for captioning more complex paintings and a starting point for further study in painting analysis.

The main contributions of our work are as follows:

- We make a novel attempt at generating content descriptions of fine art paintings.
- We train the model on generated virtual paintings to solve the data-hungry problem in image captioning of paintings, presenting meaningful improvements.
- We annotate a small-scale image captioning dataset of fine art paintings for evaluation, hoping to provide an evaluation reference for future works.

## II. RELATED WORKS

### A. Machine Learning in Art

Previous works have been explored many machine learning tasks in paintings, including classification [3] [4], object detection [5], visual question answering [6], emotion extraction [7], and robot painting [8]–[10]. Some preliminary works explored text image cross-searching of paintings [11] rather than generating contextual sentences directly. This paper investigates an image captioning task on paintings, which can generate informative sentences, useful for future painting analysis.

### B. Image Captioning

Most of the recent image captioning methods are based on a neural image captioning method [12], which use Convolutional

\*Corresponding author.

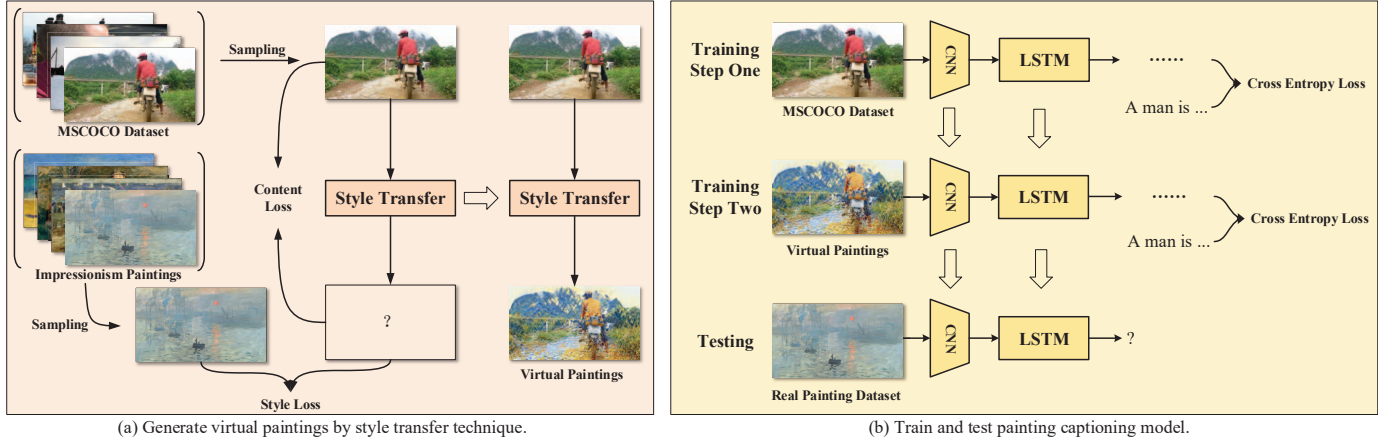


Fig. 1. Our overall framework.

Neural Networks (CNN) to extract image features and Long Short Term Memory Networks (LSTM) to generate descriptions. Training an image captioning model requires a large quantity of annotated images, but there are little annotated data for the paintings. This paper generates virtual paintings to settle this data-hungry problem.

### III. METHODOLOGY

#### A. Overview

Fig. 1 shows our overall framework. To deal with the problem of data-hungry in training, we train a style transfer model using an unpaired photographic image set and an impressionism painting set. Then we use the trained style transfer model to transfer images in MSCOCO dataset [13] to virtual paintings. After that, we train the image captioning model via a two-step manner. First, we train an image captioning model on MSCOCO dataset, and we use this trained model as our baseline model. Second, we use the baseline model as the starting point and fine-tune it on virtual paintings to get our final model. We annotate a real painting dataset to test the performance of the final model.

#### B. Virtual Paintings

Virtual paintings are generated by the style transfer technique, which transfers an image to another style while maintaining its content. We choose a real-time variant [14] of the style transfer method as it shows subtle nature in its generated impressionism style images. The style transfer model is trained using an unpaired photographic image set and an impressionism painting set together with content loss and style loss, and then it can be used as an end-to-end style transfer tool. We use the implementation and pre-trained parameters of the style transfer model from the original paper [14]. We create the virtual paintings by transferring images in MSCOCO dataset to impressionism style and copying the original captions as the annotations for virtual paintings.

#### C. Caption Generation

We use the classical neural image captioning method proposed by paper [12], where a CNN is used to extract image

features, and an LSTM is used to generating sentences base on the image features. As the CNN and LSTM are two main parts of the image captioning model, we compare the model performances under different fine-tuning strategies in Section IV. Although we use the classical neural image captioning method, our framework can also extend to other image captioning methods.

#### D. Relations to Parallel Learning

Our method is based on the parallel learning framework [15]. Fig. 2 shows the relations between our mechanism and parallel learning framework by adapting our main steps on the original parallel learning diagram. Our methods are corresponding to two main parts of parallel learning, including big data generating and computational experiments, and the unused parts are in gray. We use MSCOCO dataset and unlabeled paintings as original data and generate virtual paintings using the style transfer technique. Our big data contains MSCOCO dataset and virtual paintings. We use a two-step training procedure on the big data to get the final model.

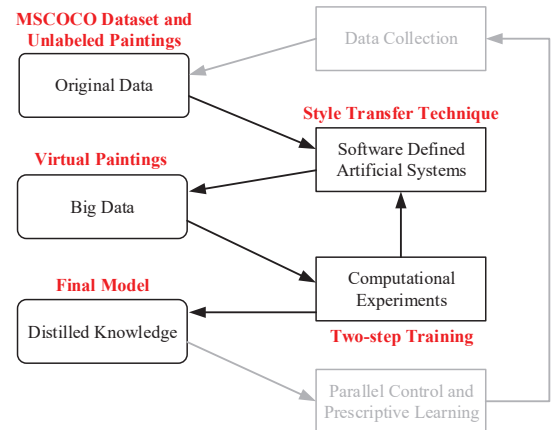


Fig. 2. The relations between our painting image captioning mechanism and parallel learning framework. This diagram is adapted from the framework in the parallel learning paper [15]. The unused parts are in gray.

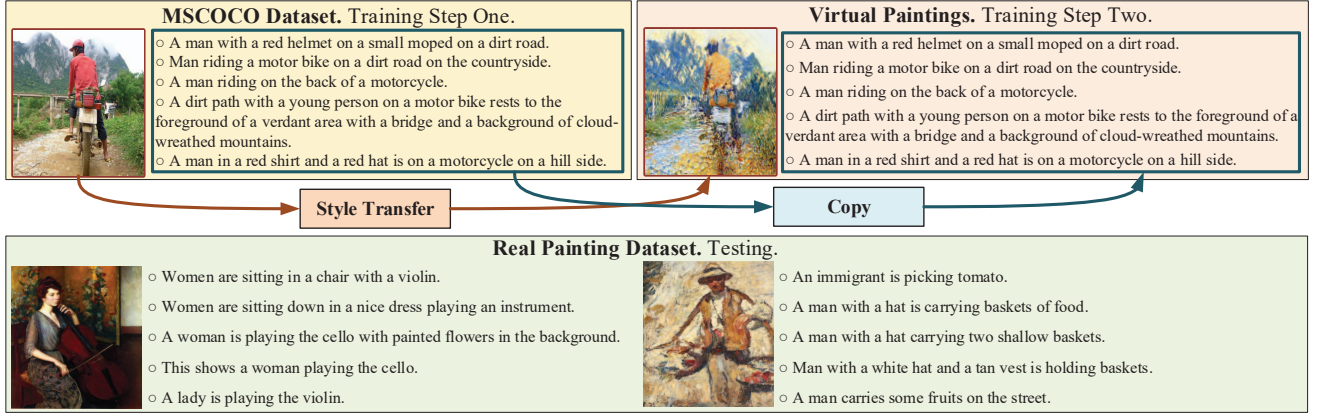


Fig. 3. Samples of data used in training and testing.

TABLE I  
PERFORMANCES OF OUR MODELS ON REAL PAINTING DATASET

	BLEU1	BLEU2	BLEU3	BLEU4	CIDEr	METEOR	ROUGE_L	SPICE
Baseline	47.44	25.75	13.03	6.81	13.13	11.69	31.01	5.74
Ours (CNN)	<b>49.61</b>	26.69	13.88	7.42	11.83	12.06	32.88	5.59
Ours (LSTM)	48.10	26.58	13.57	7.59	12.36	11.76	<b>33.25</b>	5.14
Ours (CNN+LSTM)	48.78	<b>26.92</b>	<b>14.76</b>	<b>8.47</b>	<b>14.76</b>	<b>12.58</b>	32.92	<b>6.22</b>

CNN: Training only CNN in training step two; LSTM: Training only LSTM in training step two; CNN+LSTM: Training both CNN and LSTM in training step two. The best scores are in bold.

#### IV. EXPERIMENTS

##### A. Datasets

a) *MSCOCO Dataset*: MSCOCO dataset is a commonly used image captioning dataset. Like previous works, we use the Karpathy splits [16] to separate images into 82,783 for training and 5,000 for testing.

b) *Real Painting Dataset*: To evaluate our proposed method, we annotate a real painting caption dataset by Amazon Mechanical Turk. The annotated dataset contains 100 images of genre paintings in impressionism style, with each image annotated with five sentences by five workers, like MSCOCO dataset. Fig. 3 shows the samples of data we used.

##### B. Training Strategy

Our image captioning networks consist of CNN and LSTM, responsible for extracting features and generating sentences, respectively. To find the best fine-tuning way and study the influence of each part of the model, we tried three fine-tuning settings in training step two, including fine-tuning CNN only, LSTM only, and both CNN and LSTM.

##### C. Implementation Details

a) *Vocabulary*: We use the vocabulary built from MSCOCO dataset by common practice [12], including 9,487 words. This vocabulary contains 183 (99.5%) of the 184 words in painting vocabulary. Although we do not use any information from real paintings, the vocabulary we use presents its good generalization capability.

b) *Network Structure*: We use 224\*224 as input image size and ResNet-152 [17] to get a 2048 dimension feature. For the LSTM networks, we use 512 as the input feature size, hidden feature size, and cell state size.

##### D. Results

We evaluate our model on the common metrics of image captioning task including BLEU [18] (BLEU<sub>n</sub> means calculating BLEU score using *n*-gram tokenizing), CIDEr [19], METEOR [20], ROUGE\_L [21], and SPICE [22]. Table. I shows the performances of our models on real painting dataset compared with the baseline model that trained only using MSCOCO dataset. The best scores all come from our models, with an average improvement of 4.79%.

Fig. 4 shows the selected samples of our painting captioning results, where we compare our CNN+LSTM model with the baseline model and the ground truth. The good and bad expressions are in bold. We can see our model presents the improvements in correcting the inaccurate descriptions of the baseline model, including recognition of objects, scenes, and relations between objects.

##### E. Discussion

This paper generates virtual paintings by the style transfer technique, so the painting captioning quality depends on the style transfer quality. The results show that our model can understand the objects and their relations in the paintings under the challenge of abstract expressions. Thus the distribution of our generated virtual paintings is successfully getting close to the distribution of real paintings.





	<p><b>Ground Truth:</b> People are standing on small boats along a body of water.</p> <p><b>Baseline:</b> A group of people standing <b>on a beach</b>.</p> <p><b>Ours (CNN+LSTM):</b> A group of people on a boat <b>in the water</b>.</p>
	<p><b>Ground Truth:</b> A man and a woman are drinking in a bar.</p> <p><b>Baseline:</b> A man sitting <b>on top of</b> a wooden table.</p> <p><b>Ours (CNN+LSTM):</b> A couple of people sitting <b>at a table</b> in a restaurant.</p>
	<p><b>Ground Truth:</b> A woman is sitting alone with a covered white dress.</p> <p><b>Baseline:</b> A woman sitting in front of a <b>teddy bear</b>.</p> <p><b>Ours (CNN+LSTM):</b> A close up of a person <b>wearing a red tie</b>.</p>
	<p><b>Ground Truth:</b> A person in a field of grass in front of a house.</p> <p><b>Baseline:</b> A black and white photo of a man <b>on a bench</b>.</p> <p><b>Ours (CNN+LSTM):</b> A man <b>on a field</b> with a frisbee.</p>

Fig. 4. Improvements of our model compared with the baseline model. The improvements are shown in bold.

We generate training data for painting captioning using the style transfer technique on a large quantity of annotated photographic images. Our method can be generalized to solve the problem of data-hungry in other machine learning tasks when there are enough labeled data in a related domain, serving as a potential data generation technique. Since relying on the style transfer, our method has limited influence by domain gap. Thus there may be larger improvements when dealing with more abstract paintings such as cubism, fauvism, and abstract expressionism.

Most of the best scores come from our CNN+LSTM model, and for the CIDEr score, only our CNN+LSTM model performs better than the baseline model. These results are consistent with our intuition that adaptation of the whole model brings the most considerable improvement. Two of the best scores come from the other two models, possibly due to the noise from our small-scale testing dataset.

## V. CONCLUSION

This paper makes a novel attempt at the image captioning task of fine art paintings. We generate virtual paintings to train the painting captioning model. The experiment results on our annotated real painting dataset show the improvements of our proposed method.

We hope that our work can be a starting point for more advanced painting understanding tasks, and our annotated dataset can establish an evaluation reference for future works. More advanced image captioning methods and more annotated paintings can be used to evaluate the performances of our methods in future work.

## ACKNOWLEDGMENT

This work is supported in part by Skywork Intelligence Culture & Technology LTD.

## REFERENCES

- [1] Y. Lu, C. Guo, Y. Lin, F. Zhuo, and F.-Y. Wang, "Computational aesthetics of fine art paintings: The state of the art and outlook," *Acta Automatica Sinica*, vol. 46, no. 11, pp. 2239–2259, 2020.
- [2] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.
- [3] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, "Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification," in *Proceedings of the IEEE International Conference on Image Processing*, 2016, pp. 3703–3707.
- [4] S.-H. Zhong, X. Huang, and Z. Xiao, "Fine-art painting classification via two-channel dual path networks," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 1, pp. 137–152, 2020.
- [5] D. Kadish, S. Risi, and A. S. Løvlie, "Improving object detection in art images using only style transfer," *arXiv preprint arXiv:2102.06529*, 2021.
- [6] N. Garcia, C. Ye, Z. Liu, Q. Hu, M. Otani, C. Chu, Y. Nakashima, and T. Mitamura, "A dataset and baselines for visual question answering on art," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 92–108.
- [7] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. Guibas, "Artemis: Affective language for visual art," *arXiv preprint arXiv:2101.07396*, 2021.
- [8] F.-Y. Wang, "Parallel art: from intelligent art to artistic intelligence," The Alfred North Whitehead College, Tech. Rep., 2017, the Alfred North Whitehead Academy.
- [9] C. Guo, Y. Lu, Y. Lin, F. Zhuo, and F.-Y. Wang, "Parallel art: artistic creation under human-machine collaboration," *Chinese Journal of Intelligent Science and Technology*, vol. 1, no. 4, pp. 335–341, 2019.
- [10] C. Guo, T. Bai, Y. Lu, Y. Lin, G. Xiong, X. Wang, and F.-Y. Wang, "Skywork-davinci: A novel cpss-based painting support system," in *Proceedings of the IEEE 16th International Conference on Automation Science and Engineering*, 2020, pp. 673–678.
- [11] N. Garcia and G. Vogiatzis, "How to read paintings: semantic art understanding with multi-modal retrieval," in *Proceedings of the European Conference on Computer Vision Workshops*, 2018, pp. 676–691.
- [12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
- [13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.
- [14] A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer, "A style-aware content loss for real-time hd style transfer," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 698–714.
- [15] L. Li, Y. Lin, D.-P. Cao, N.-N. Zheng, and F.-Y. Wang, "Parallel learning: a new framework for machine learning," *Acta Automatica Sinica*, vol. 43, no. 1, pp. 1–8, 2017.
- [16] K. Andrej and F.-F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [19] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.
- [20] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization*, 2005, pp. 65–72.
- [21] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Text Summarization Branches Out*, 2004, pp. 74–81.
- [22] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 382–398.