

Developing an Automated Synthetic Korean Painting Dataset Generation System for AI Training

Sungwon Moon*, Jiwon Lee*, Dowon Nam*

*Electronics and Telecommunications Research Institute (ETRI), Korea

moonstarry@etri.re.kr, ez1005@etri.re.kr, dwnam@etri.re.kr

Abstract— This paper describes research on automatically generating training datasets to develop AI specialized for Korean paintings. The goal of our research is to use generative AI to generate synthetic Korean paintings while automatically labeling object areas within the paintings. In addition, we quantitatively evaluated the training dataset generated by the proposed system, and qualitatively evaluated the dataset by Korean traditional culture experts. Based on these experimental results, we suggest directions for future research and improvement.

Keywords— Data Generation, Auto Labeling, AI Art, Object Detection, Data Augmentation

I. INTRODUCTION

With the rapid development of generative AI, text-to-image AI can generate high-quality images in various styles. However, it is difficult to generate high-quality images of domains that are a small part of the total data, such as Korean paintings. When fine-tuning an AI model to generate images for a domain with little data, such as Korean painting, the model loses generality. Techniques such as textual inversion and hypernetwork have been proposed as a way to generate images of a specific domain without losing the generality of the model. Although these techniques have helped to generate Korean painting style images, there is a lack of analysis on how they affect cross-attention in image generation [2],[3].

Cross-attention can be used as a good indicator to visualize the text-image attribution between text prompt input and generated results. Existing research has shown that it is useful to visualize the correlation between the textual prompt input and the pixel area of the generated image output using cross-attention. However, it is not yet clear whether it is possible to use cross-attention when using techniques such as text inversion or hypernetwork. If cross-attention between text prompt input and output results is possible even after using textual inversion or hypernetwork, it would be possible to use this to create Korean-style images and automatically generate labels for them.

In this research project, we applied additional techniques such as textual inversion and hypernetwork to generate Korean paintings using Stable Diffusion, and constructed a pipeline that maintains the text-to-image attribution based on cross-attention. We aim to show that objects in Korean paintings can be automatically labeled using the above method. In this paper, the training dataset generated in the above

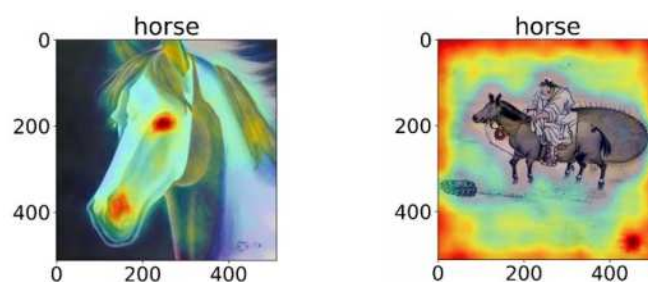


Figure 1. Example of text-to-image attribution before(left) and after(right) using text inversion

manner was used for AI training to quantitatively evaluate the impact on object detection performance, and the results of qualitative evaluation of the generated images and labels by Korean traditional culture experts were presented.

II. OUR APPROACH

A. Generate Synthetic Korean Paintings with Diffusion Model

Text-to-image AI with a large diffusion model such as Stable Diffusion can generate images for various text prompt inputs [1]. However, it is difficult to generate images in domains with little data, such as Korean paintings, and in particular, it is nearly impossible to generate Korean paintings in the style of a specific Korean artist using only text prompts. To overcome these limitations, various techniques have been explored to generate images of the desired style without training on a large diffusion model. Among them, we tested whether it is possible to generate a large number of Korean paintings from a small amount of data using textual inversion and hypernetwork with Stable Diffusion [1],[2]. Both techniques can generate Korean paintings, but we chose to use the hypernetwork technique and will explain why in Section B.

B. Visualize Text-Image Attribution

Tang proposed DAAM, a technique that uses cross-attention to visualize the pixel area of the image affected by words in the form of a heat map in text-to-image AI [4]. By applying this, a bounding box and semantic segmentation mask can be created for the object in the image by using the word corresponding to the object in the text prompt input as a class. However, in order to apply this approach, it is necessary

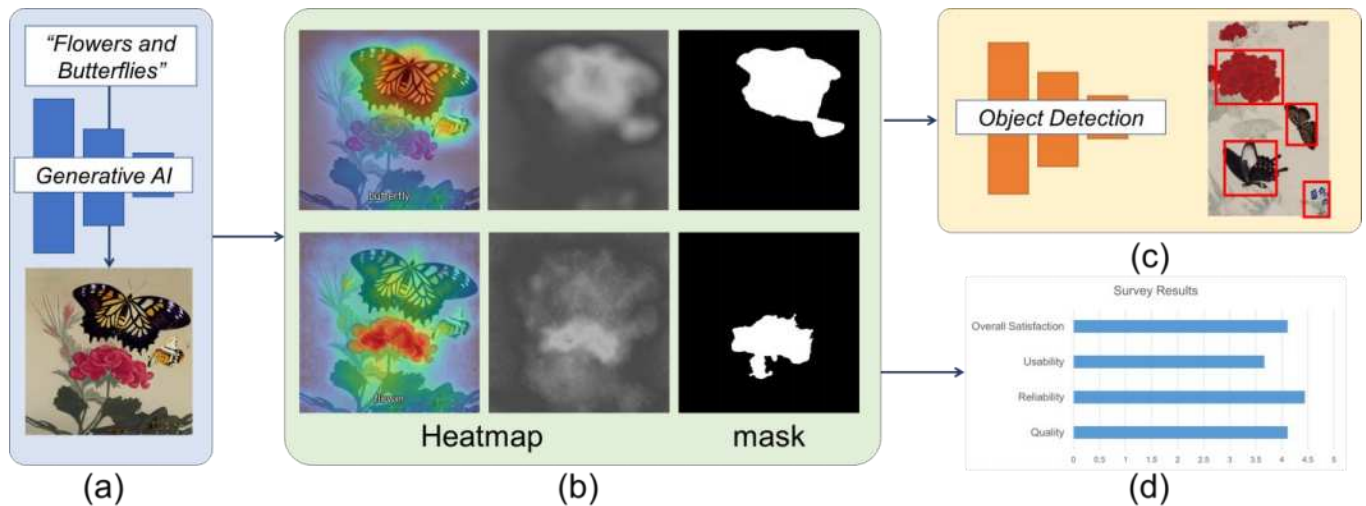


Figure 2. Automated synthetic Korean painting dataset generation system: (a) synthetic image generation with text-to-image model, (b) automatic labeling with text-to-image attribution, (c) object detection AI training, (d) dataset evaluation survey results from Korean cultural experts

to confirm that the cross-attention between text embedding and latent representation is maintained even after applying the technique to generate Korean Painting. We visualized the DAAM for Korean Painting generated by textual inversion and hypernetwork, respectively. As shown in Figure 1, when textual inversion was used, the cross-attention between the word corresponding to the object and the object in the image was weakened during the process of applying Korean painting style, so it could not be used for labeling. On the other hand, when using the hypernetwork, it was confirmed that the text-image attribution was well maintained while creating images that reproduced the Korean painting style. As shown in Figure 3, we used the text-image attribution visualized above to automatically label the generated images, and also used it as data to evaluate the AI-generated training dataset by Korean traditional culture experts.

C. Visualize Text-Image Attribution Automatic Labeling of Objects in Synthetic Korean Paintings

The method of automatically labeling objects in Korean paintings generated by text-to-image AI using text-image attribution in the form of a heat map in the proposed system is as follows. First, the DAAM for the word corresponding to the object converts the color space from RGB to grayscale. Next,

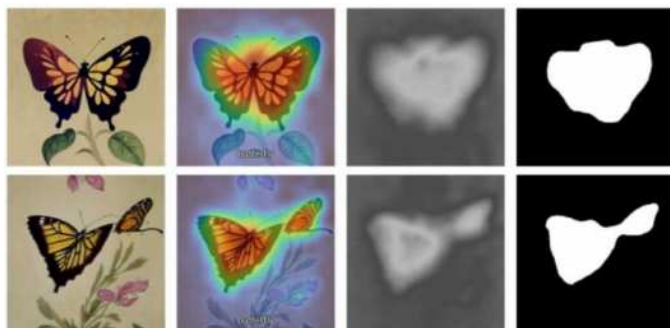


Figure 3. Example of visualized text-image Attribution

the DAAM is refined using an empirical threshold, converted into a binary mask form, and noise is removed by applying erosion and dilation according to morphology image processing. Finally, the largest contour in the noise-removed DAAM was used as a semantic segmentation mask for each object, and the rectangular area surrounding it was used as a bounding box.

III. EXPERIMENTAL RESULTS AND SURVEY

We verified the practicality of the proposed system as follows. First, for quantitative evaluation, we used a training dataset created by automatically labeling 256 Korean pictures automatically generated by the proposed object detection AI training and validating its performance. The validation data used were 14 pieces of "Flowers and Butterflies" by "Nam Gyeu", a painter of the Joseon Dynasty, and the detection target class was limited to butterflies. As a result of using DAMO-YOLO as an object detection model, it was confirmed that butterflies could be detected even without training actual Korean paintings [5]. The quantitative evaluation result was AP50 0.322, which is not a very high performance, but it was sufficient to confirm the usability of the proposed system. In addition, to qualitatively evaluate the generated dataset, we provided sample data to nine Korean cultural experts and conducted a survey on aspects of data quality, labeling reliability, dataset usability, and overall satisfaction. According to the survey results, the overall satisfaction with the dataset was 4.1 out of 5 as shown in Figure 4.

IV. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a system that generates a dataset for AI training by using generative AI to generate synthetic Korean paintings and automatically label objects in the generated paintings. In addition, the experimental results showed that the dataset generated by the proposed system can be used for AI training, and the survey showed that the generated data and labels were satisfactory even for experts. We hope that the proposed system will be used as an AI-based

	#1	#2	#3	#4	#5	#6	#7	#8	#9	Avg
Quality		4	4	5	3	5	4	5	3	4.111111
Reliability		4	5	5	5	5	5	3	4	4.444444
Usability		4	4	5	3	4	3	4	2	3.666667
Overall Satisfaction		4	4	5	4	5	3	4	4	4.111111

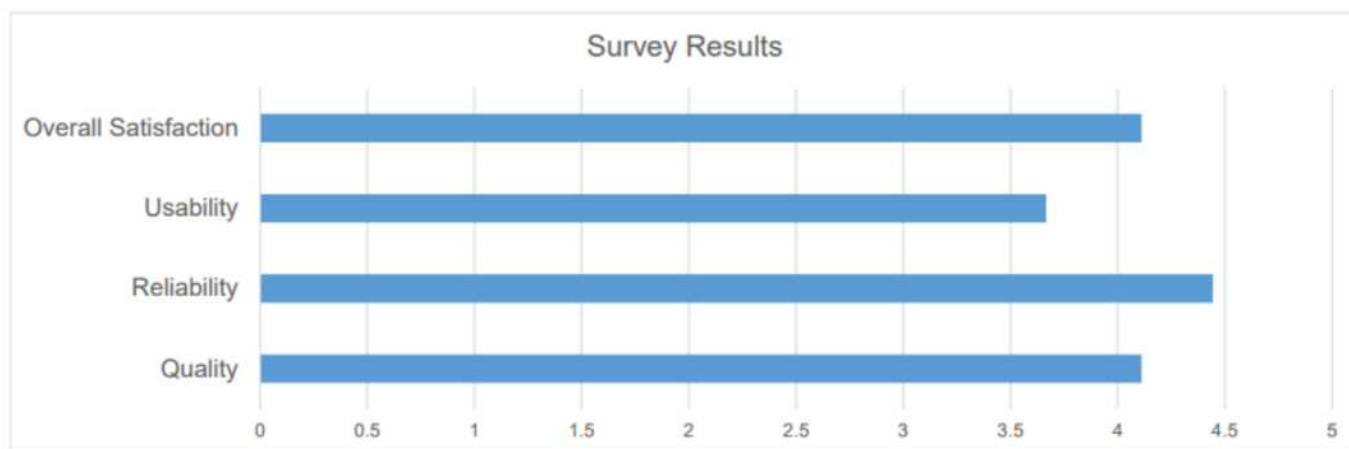


Figure 4. Dataset evaluation survey results from nine Korean cultural experts

technology for classification and analysis of Korean cultural heritage. Based on suggestions from cultural heritage experts, we plan to conduct the following follow-up research. First, in order to prevent objects created by generative AI from being mixed together, we plan to automatically create a layout with an appropriate composition and create a Korean painting based on it. In addition, we plan to improve the proposed system so that it can classify and label multiple objects at the instance level.

ACKNOWLEDGMENT

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2023(Project Name: Development of storytelling AI technology for cultural heritage tailored to the various interests of users, Project Number: RS-2023-00220195, Contribution Rate: 100%)

REFERENCES

- [1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models." *arXiv*, 2021.
- [2] Rinon Gal et al., "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion," *arXiv*, 2022.
- [3] A. Vaswani et al., "Attention Is All You Need". *arXiv*, 2023.
- [4] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to Follow Image Editing Instructions". *arXiv*, 2023.
- [5] R. Tang et al., "What the DAAM: Interpreting Stable Diffusion using cross attention", *arXiv*, 2022.
- [6] Xianzhe Xu et al., "DAMO-YOLO : A Report on Real-Time Object Detection Design," *arXiv*, 2023.



Sungwon Moon received his B.S and M.S degrees of computer science from KAIST, Korea in 2010 and 2012. Since 2012 he has been a researcher in the Sportainment Section, SW.Content Reserch Laboratory, Electronics and Telecommunications Research Institute (ETRI), Republic of Korea. His research interests are Digital Watermarking, Video Forensic, and Video Processing.



Jiwon Lee received the B. S. degree in Computer Engineering from Kyungpook National University, Republic of Korea, in 2008, and the Ph. D. degree in Computer Science from Korea Advanced institute of Science and Technology (KAIST), Republic of Korea, in 2013. Since 2013 he has been a senior researcher in the Sportainment Section, SW.Content Reserch Laboratory, Electronics and Telecommunications Research Institute (ETRI), Republic of Korea. His research interests include multimedia security, image/video watermarking, and image/video processing.



sports video analysis.

Do-Won Nam received the B.S. degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in 1996 and M.S. degree in Information Technology from Pohang University of Science of Technology (POSTECH), Korea in 1998. He is working as a principal researcher in the Electronics and Telecommunications Research Institute (ETRI) since 2001. His research interests include data mining, digital rights management, digital cinema system and