https://doi.org/10.2352/EI.2025.37.9.IQSP-243


# 3D CG Image Quality Assessment in Vision and Language Based on Stable Diffusion

***Norifumi Kawabata, Research Division of Computational Imaging, Computational Imaging Lab, Tagami Honmachi, Kanazawa-city, Ishikawa, Japan***

## Abstract

*GPT-4, which is a multimodal large-scale language model, was released on March 14, 2023. GPT-4 is equipped with Transformer, a machine learning model for natural language processing, which trains a large neural network through unsupervised learning, followed by reinforcement learning from human feedback (RLHF) based on human feedback. Although GPT-4 is one of the research achievements in the field of natural language processing (NLP), it is a technology that can be applied not only to natural language generation but also to image generation. However, specifications for GPT-4 have not been made public, therefore it is difficult to use for research purposes. In this study, we first generated an image database by adjusting parameters using Stable Diffusion, which is a deep learning model that is also used for image generation based on text input and images. And then, we carried out experiments to evaluate the 3D CG image quality from the generated database, and discussed the quality assessment of the image generation model.*

## 1. Introduction

The progress of image generation technology by Artificial Intelligence (AI) in the 2020s and beyond is remarkable. GPT-4, a multimodal large-scale language model, will be released in March 2023. GPT-4 is a technology that can be applied not only to natural language generation but also to image generation, but its detailed specifications have not been released, and it is difficult to use it for research purposes. Therefore, this study deals with Stable Diffusion, an open-source image generation AI model.

Stable Diffusion [1] was released in August 2022. What makes Stable Diffusion completely different from other image-generating AI services is that anyone can freely use all of the AI's programs and data. If there is a personal computer in the home, the AI can generate as many images as desired. A program that serves as a front-end to Stable Diffusion appeared on the scene immediately. The AUTOMATIC1111 version of Stable Diffusion WebUI.

Stable Diffusion is a type of artificial intelligence that generates images that match text or images inputted into it. AI that generates images has existed for some time, and some even output images that are indistinguishable from the real thing. For example, image generation using adversarial generative networks [2], super-resolution technology [3], image generation using neural networks [4, 5] is considered to be one of them. However, most of them were AIs that could be used only for a single purpose, such as generating images of human faces or performing super-resolution. Image generation AIs that have emerged and become popular after 2022 use a new technology called the diffusion model. When text is input, it generates images that match the text. Most of the images are in English only, but they are close to the words we use in our daily life. The new technology can generate an unimaginably wide variety of images from text. In this trend, Stability AI's Stable Diffusion has emerged as an open source image generation AI that employs a diffusion model. Stable Diffusion includes text-to-image (txt2img) and image-to-image (img2img). In this study, we consider img2img from an image processing perspective.

So far, we have studied the optimal design of multi-view super-resolution images based on the structure of CNNs [4], Contrast enhancement for color laparoscopic imaging and optimal conditions of SRCNN super-resolution processing [3], Region segmentation of color laparoscopic contrast-enhanced images considering SRCNN super-resolution processing by image regions. We have separately conducted region segmentation of color laparoscopic contrast-enhanced images considering SRCNN super-resolution processing [5], and optimal design of color laparoscopic super-resolution images by hostile generation network [2]. Although the objective of generating a single image has been achieved, we thought that the adoption of a Stable Diffusion model that takes into account the fusion of vision and language would enable us to generate images with a wider field of view and to evaluate their image quality.

In this study, we used Stable Diffusion, a deep learning model that is also used for image generation based on text prompt input and image-based image generation. By varying the CFG Scale parameter, which adjusts the fidelity of the generated image to the prompt content, the prompt, and the input image, we generated 3DCG images with visual and verbal fusion. Finally, the quality of the image generation model was evaluated and discussed.

## 2. Related work

In this section, we describe related studies in terms of (1) image generation models and (2) image processing.

### *2.1. In terms of image generation models*

In terms of image generation models, there are studies on detecting images generated by Stable Diffusion using frequency artifacts as a case study in Disney-Style Art [6], Proposed framework for arbitrary style transformations using a diffusion model called diffusion-enhanced patch matching [7], Accurate diffusion inversion using a coupled transform [8], RGBT object tracking using a Bayesian dumbbell diffusion model [9], Improved visibility in bad weather using a patch-based denoising diffusion model citeFan2023, Patch-based denoising diffusion model for improved visibility in bad weather conditions.

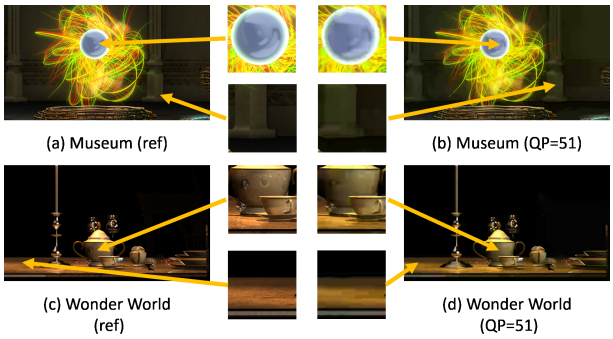


***Figure 1.*** *3D CG image content used in this study*

### 2.2. In terms of image processing

In terms of image processing, there are studies on Identification of distributed nonlinear systems in $\alpha$-stable noise [11], A hybrid model for image denoising combining a modified isotropic diffusion model and a modified Perona-Malik model [12], A new anisotropic 4th order diffusion model for low Dose CT image processing [13], Multiple B-value model-based residual network for accelerated high-resolution diffusion-weighted images [14].

### 2.3. Conclusion

In general, there are many studies on image generation from the viewpoint of image generation models and image processing, but the relationship between image generation AI and image quality evaluation is not clear, and this study examines whether image generation AI is effective in the field of image quality evaluation through experiments.

## 3. Experimental set

### 3.1. Images used in this study

The images used in this study are 3DCG contents (Museum (M), Wonder World (W)) such as Fig. 1 provided free of charge by NICT. For image generation, a CG camera for 8 viewpoints is first constructed, and then the camera work and rendering are performed. Camera work and rendering were performed to generate still images for the 8 viewpoints. Although this content is originally a multi-view 3D image content, we used a single viewpoint image out of the eight still images in this study.

### 3.2. Experimental procedure

Next, the experimental procedure is as follows:

1. Generate 3D CG images from 3D CG contents by setting up camera work and rendering.
2. Configure the AUTOMATIC1111 version of the Stable Diffusion WebUI. (See next chapter)
3. Load the generated 3DCG images into Stable Diffusion WebUI, set the prompts, CFG Scale, and start the image generation AI to generate images. In this process, not only image generation is performed, but also GPU processing time is measured.
4. Image quality evaluation (Structural SIMilarity (SSIM)) is performed on the images generated by the image generation AI. Comparisons with CFG Scale parameters and GPU processing time are also made and discussed.

### 3.3. Experimental and assessment method

As an experimental method, in this study, the AUTOMATIC1111 version of the Stable Diffusion WebUI is used as shown in Fig.2. The image generation AI was run on a desktop PC with an Intel Core i7-8700 processor running at 3.20 GHz (Intel Turbo Boost Technology 2.0 supported: maximum 4.60 GHz), a NVIDIA GeForce GTX 1050 Ti GPU, and 32 GB of memory. The experiment was conducted using a desktop PC. Stable Diffusion is basically recommended to be run on a GPU (Graphics Processing Unit). Although it is possible to run on a CPU (Central Processing Unit), preliminary experiments have shown that it takes more than five hours to generate one image on a CPU, compared to eight minutes on a GPU, so there is a difference in both time and processing efficiency. Therefore, it is better to work on a desktop PC or workstation than on a notebook PC (excluding GPUs). As described below, using Google Colaboratory is also an option. However, although the free version can use GPUs, it has time and performance limitations, and often initializes in a certain state. Therefore, when using Google Colaboratory, it is recommended to use the paid version, which basically has no limitations (but not completely).

As an evaluation method, this study uses the Structural SIMilarity (SSIM), an index that evaluates the visual impact of luminance, contrast, and structure. The overall index is obtained by combining and multiplying three terms as shown in the Eq. (1).

$$\mathrm{SSIM}(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \tag{1}$$

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{3}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{4}$$

Here, $\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy}$ in Eq. (2), (3) and (4) are the local mean, standard deviation and cross-covariance of image $x, y$. and $\alpha = \beta = \gamma = 1$ and $C_3 = \frac{C_2}{2}$, they can be expressed as follows Eq. (5).

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

This study evaluates the degree of similarity with the original image.

## 4. Image generation by Stable Diffusion

This section describes the procedure for performing Stable Diffusion and its overview.

1. Install Python. (Python 3.11.5)
2. Install git. (git 2.43.2)
3. Install AUTOMATIC1111/stable-diffusion-webui using git. Open a command prompt in the installation folder, and type and execute the following commands.

```
git clone https://github.com/AUTOMATIC1111/
stable-diffusion-webui.git
```

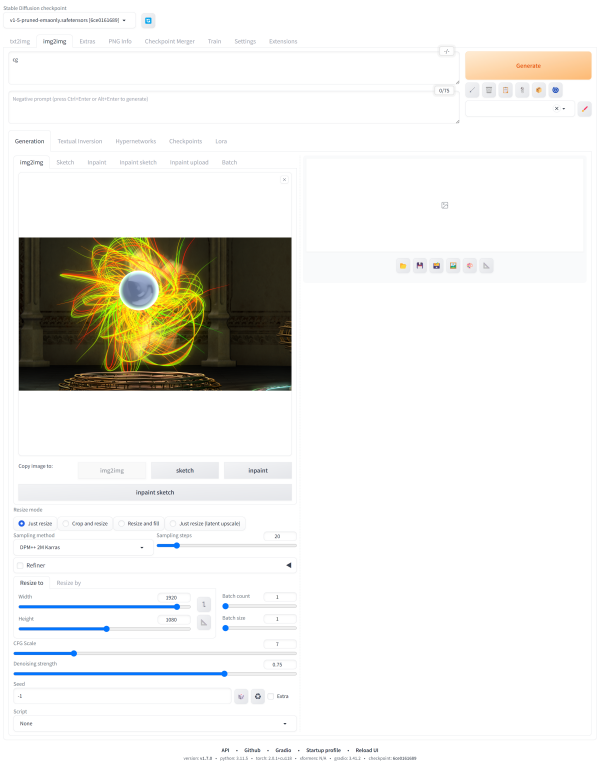

***Figure 2.*** *Stable Diffusion WebUI (AUTOMATIC1111)*

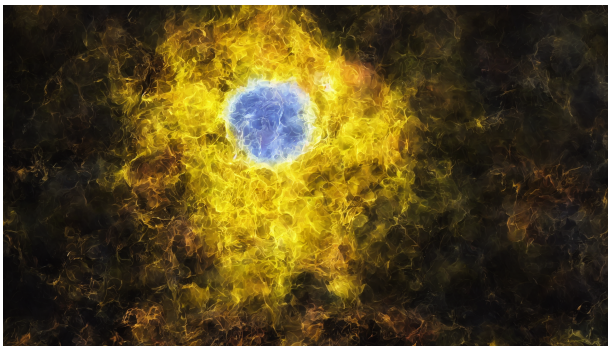

***Figure 3.*** *Experimental result (CFG scale=3)(“M”, prompt= “cg”)*

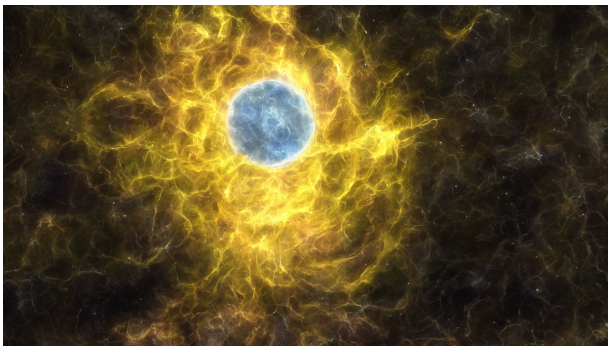

***Figure 4.*** *Experimental result (CFG scale=7)(“M”, prompt= “cg”)*

4. Run webui-user.bat to start the installation. A command prompt will open and the installation will begin. The installation is complete when the message “Running on local URL: http://127.0.0.1:7860” appears.
5. Create and execute webui-user-my.bat. The current WebUI can generate images without using the settings for low VRAM, but errors occur as soon as a slightly large image is created, so a batch file for running in low VRAM is created and used. Copy webui-user.bat and create “webui-user-my.bat” in the same folder. Change the line “set COMMANDLINE_ARGS=” in the created webui-user-my.bat as follows.

```
set COMMANDLINE_ARGS=--medvram --xformers
```

6. (Optional) Obtain and deploy additional training models. In this study, the default model was used for the runs.
7. Launch WebUI and select the img2img tab. The img2img tab allows the user to generate a new image based on an image and a spell. In this study, a 3D CG image was loaded and the prompts (Museum: cg, WonderWorld: cg,cup) were entered to generate the image. The CFG Scale was set to 3, 7 (default), 15, and 30.

## 5. Experimental results and discussion

The experimental results are shown in Figs. 3 to 10. Here, Figs. 3 to6, Figs. 7 to 10 show each results of image generation in “Museum”, “WonderWorld”.

The experimental results show that the “Museum” and “WonderWorld” produced smoky images at CFG Scale=3 and 7, respectively. Stable Diffusion generates images from the viewpoint of noise diffusion, and therefore, when the CFG Scale is low, the images may contain noise when the so-called accuracy is lowered.

In the case of “Museum”, the central pattern is an arc of thin lines, which may have generated fine noise. Furthermore, in the “Museum”, the noise becomes less smoke-like, more line-like, and more arc-like as CFG Scale=15 and 30. This is partly because the object area is arc-shaped, but it may also be due to the fact that “cg” is entered at the prompt, and the font of “cg” is arc-shaped. It can be inferred from the experimental results that as the CFG Scale is increased, the prompt is affected in addition to the image.

In the case of “WonderWorld”, we can see that the smoke of noise gradually disappears and objects are constructed as CFG Scale=7, 15, and 30. This can be attributed to the fact that the prompt was set to “cg,cup” and not only “cg” but also “cup” was added to the prompt. We tried to generate 3D CG content for “WonderWorld” using only the “cg” prompt, but it generated something that had nothing to do with cups, and we could not generate it successfully, so we added the prompt. Thus, depending on the 3D CG content, there are cases where only “cg” can be used, and other cases where more detailed keywords, such as “cg,cup”, must be entered as prompts, so it is necessary to enter prompts while carefully checking the image content.

On the other hand, Fig. 11 shows the structural similarity of images (SSIM) and Fig. 12 shows the generation time of images. The SSIM results show that all of the images in the “Museum” are above 0.7, and that the images in the “WonderWorld” are above 0.7 for CFG Scale=7, but below 0.7, there is little similarity between the images. The generation time of the images showed

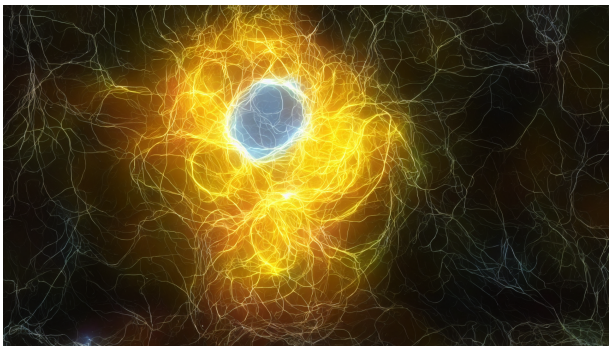

***Figure 5.*** *Experimental result (CFG scale=15)(“M”, prompt= “cg”)*

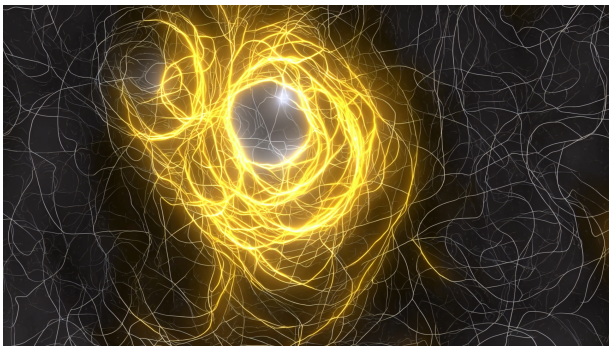

***Figure 6.*** *Experimental result (CFG scale=30)(“M”, prompt= “cg”)*

***Figure 7.*** *Experimental result (CFG scale=3)(“W”, prompt= “cg, cup”)*

***Figure 8.*** *Experimental result (CFG scale=7)(“W”, prompt= “cg, cup”)*

that for the “Museum” the generation time decreased as the CFG Scale increased, but this was not the case for the “WonderWorld”. The image generation process was completed in 8 mins 10 secs to 8 mins 50 secs in all cases, although there were differences depending on the content.

## 6. Conclusion

The results of this study suggest that Stable Diffusion is effective in evaluating the quality of 3D CG generated images. As future works, more detailed parameter settings and image processing patterns will be added. This paper is a developmental improvement of the content of the March 2024 Forum on Visual Expression and Arts & Sciences 2024 (Expressive Japan 2024).

## References


[1] “Stable Diffusion Demo,” `https://huggingface.co/spaces/stabilityai/stable-diffusion`, accessed October 15, 2024.

[2] N. Kawabata and T. Nakaguchi: “Optimal Design of Color Laparoscopic Super-Resolution Image Quality Based on Generative Adversarial Networks,” *Proc. of the 2023 International Conference on Computer Graphics and Image Processing (CGIP 2023)*, S2-6, pp.1–7, Tokyo, Japan, January 2023. DOI: `10.1109/CGIP58526.2023.00009`

[3] N. Kawabata and T. Nakaguchi: “Color Laparoscopic High-Definition Video Quality Assessment for Super-Resolution,” *Proc. of SPIE* (*The 25th International Workshop on Advanced Image Technology* (IWAIT2022)), 7A5, pp.1–6, The Hong Kong Polytechnic Univ., Hong Kong (Hybrid), January 4-6, 2022.
DOI: `10.1117/12.2626140`

[4] N. Kawabata: “HEVC Image Quality Assessment of the Multi-view and Super-resolution Images Based on CNN,” *Proc. of the 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE2018)*, vol.7, POS1A-3, pp.38–39, Nara, Japan, October 2018.
DOI: `10.1109/GCCE.2018.8574768`

[5] N. Kawabata and T. Nakaguchi: “Color Laparoscopic Image Region Segmentation after Contrast Enhancement Including SRCNN by Image Regions,” *Proc. of SPIE (The International Forum on Medical Imaging in Asia (IFMIA2021))*, vol.11792, no.1179209, 6 pages, April 2021.
DOI: `10.1117/12.2590852`

[6] J. Zhang, Y. Wang, H. R. Tohidypour, and P. Nasiopoulos: “DETECTING STABLE DIFFUSION GENERATED IMAGES USING FREQUENCY ARTIFACTS: A CASE STUDY ON DISNEY-STYLE ART,” *Proc. of ICIP 2023*, pp.1845–1849, 2023.
DOI: `10.1109/ICIP49359.2023.10221905`

[7] M. Hamazaspyan and S. Navasardyan: “Diffusion-Enhanced PatchMatch: A Framework for Arbitrary Style Transfer with Diffusion Models,” *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.797–805, 2023.
DOI: `10.1109/CVPRW59228.2023.00087`

[8] B. Wallace, A. Gokul, and N. Naik: “EDICT: Exact Diffusion Inversion via Coupled Transformations,” *Proc. of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2023)*, pp.22532–22541, 2023.
DOI: `10.1109/CVPR52729.2023.02158`

***Figure 9.*** *Experimental result (CFG scale=15)(“W”, prompt= “cg, cup”)*

***Figure 10.*** *Experimental result (CFG scale=30)(“W”, prompt= “cg, cup”)*

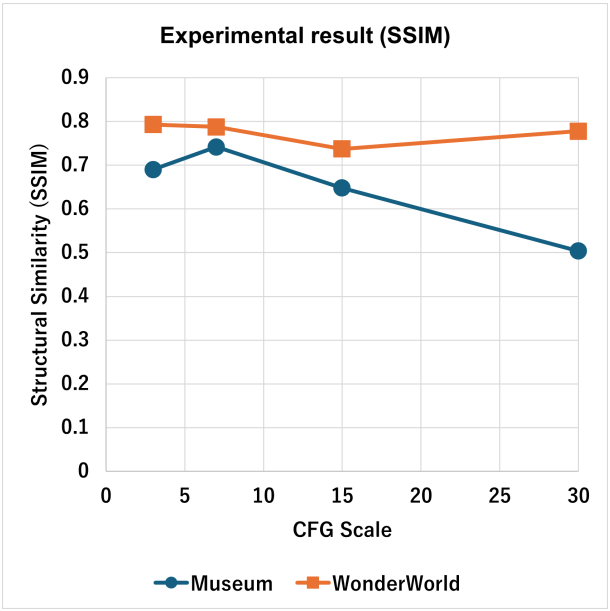


***Figure 11.*** *Structural similarity of images*

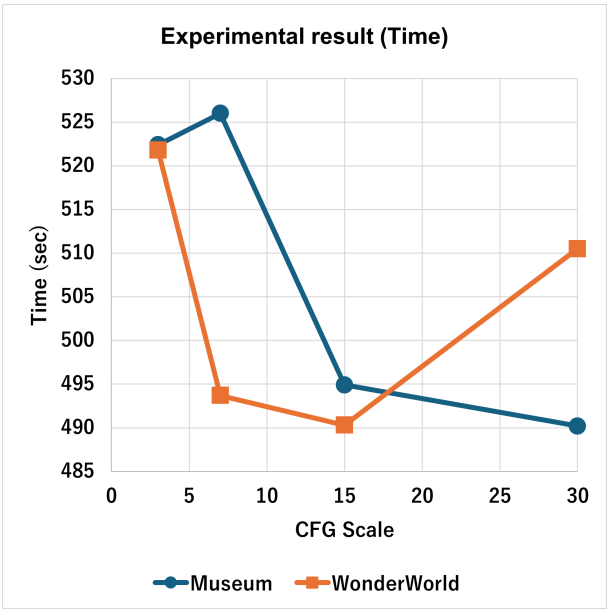


***Figure 12.*** *Image generation time*


[9] S. Fan, C. He, C. Wei, Y. Zheng, and X. Chen: “Bayesian Dumbbell Diffusion Model for RGBT Object Tracking With Enriched Priors,” *IEEE Signal Processing Letters*, Vol.30, pp.873–877, 2023.
DOI: 10.1109/LSP.2023.3295758

[10] O. Özdenizci and R. Legenstein: “Restoring Vision in Adverse Weather Conditions With Patch-Based Denoising Diffusion Models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.45, No.8, pp.10346–10357, August 2023.
DOI: 10.1109/TPAMI.2023.3238179

[11] L. Lu, H. Zhao, and B. Champagne: “Distributed Nonlinear System Identification in $\alpha$-Stable Noise,” *IEEE Signal Process Lett*, Vol.25, No.7, pp.979–983, July 2018.
DOI: 10.1109/LSP.2018.2835763

[12] N. Wang, Y. Shang, Y. Chen, M. Yang, Q. Zhang, Y. Liu, and Z. Gui: “A Hybrid Model for Image Denoising Combining Modified Isotropic Diffusion Model and Modified Perona-Malik Model,” *IEEE Access*, Vol.6, pp.33568–33582, 2018.
DOI: 10.1109/ACCESS.2018.2844163

[13] L. Wang, Y. Liu, R. Wu, Y. Liu, R. Yan, S. Ren, and Z. Gui: “Image Processing for Low-Dose CT via Novel Anisotropic Fourth-Order Diffusion Model,” *IEEE Access*, vol.10, pp.50114–50124, 2022.
DOI: 10.1109/ACCESS.2022.3172975

[14] F. Wang, H. Zhang, F. Dai, W. Chen, S. Xu, Z. Yang, D. Shen, C. Wang, and H. Wang: “Multiple B-Value Model-Based Residual Network (MORN) for Accelerated High-Resolution Diffusion-Weighted Imaging,” *IEEE J Biomed Health Inform*, Vol.26, No.9, pp.4575–4586, September 2022.
DOI: 10.1109/JBHI.2022.3193299