TESD: TILING-BASED EXPANSION MODEL IN STABLE DIFFUSION

Zhang WANG[1], Tiejun PAN*[2], Yaojie FEI[3], Junyi CHAI[4], Zhengqi PAN[5], Leina ZHENG[6]

Artificial Intelligence Generated Content (AIGC) refers to techniques such as Denoising Diffusion Probabilistic Models (DDPM) and large pre-trained models to automatically generate content. Stable Diffusion is a typical AIGC system for text-to-image generation, which has been widely applied in many fields. However, the AIGC technology requires a significant amount of GPU memory to generate high-resolution images. In addition, the generated images are random and need to be adjusted multiple times to meet the needs of the users. To address these problems, a Tiling-based Expansion model in Stable Diffusion (TESD) is proposed: (1) Tiling diffusion is used to generate relatively sharp images on low GPU memory devices, (2) The Image Feature Controller (IFC) is used to eliminate the randomness of the image and enhance the color level, (3) AIGC functions are implemented on embedded devices by deploying Stable Diffusion in the cloud. A straightforward patch based on the partitioning framework was integrated into the upscaling of the model, thereby achieving reduced GPU memory utilization and accelerated image processing speeds in contrast to conventional upscaling models. Through a comparison with seven similar enlargement models, our model outperforms all challenging solutions in terms of generation speed and effectiveness, with a very significant advantage and prospect.

Keywords: multi diffusion, tiled diffusion, amplification model, zero convolution, cloud computing

1. Introduction

Text-to-image generation has emerged as a highly dynamic and burgeoning field within Artificial Intelligence Generated Content (AIGC) in the modern technological landscape, pervading numerous facets of our daily lives. Stable Diffusion is a form of AIGC. Based on the diffusion model of deep learning, it generates images by gradually adding noise and then reverse denoising. Compared with the traditional text-to-image method, Stable Diffusion has a significant improvement in picture diversification, control, and economic benefits. Users can

[1] College of Science & Technology, Ningbo University, Ningbo, China, wangzhang@nbu.edu.cn
[2] * College of Science & Technology, Ningbo University, Ningbo, China, corresponding author, pantiejun@nbu.edu.cn
[3] College of Science & Technology, Ningbo University, Ningbo, China, 525507516@qq.com
[4] College of Science & Technology, Ningbo University, Ningbo, China, chaijunyiningbo@icloud.com
[5] School of Economics, Zhejiang Gongshang University, Hangzhou, China, 2364708494@qq.com
[6] Bushiness School, Zhejiang Wanli University, Ningbo, China, leina.zheng@zwu.edu.cn

customize the development process based on their own needs. For instance, they can adjust the model structure and training parameters to better adapt to various application scenarios and requirements. The functions of text - to - image can be applied to a wide range of industries, including game development, illustration design [1], healthcare [2], and e-commerce [3].

In Stable Diffusion, UNet and random seeds are two vital components. They respectively control the style and image segmentation of the image, both of which are significant aspects of the field of image generation. An important role in Stable Diffusion is played by UNet, an important neural network architecture known for its proficiency in image-to-image translation tasks. UNet acts as a noise predictor, progressively removing noise from the image during the back diffusion process. Through a series of convolutional layers, upsampling, and downsampling operations, the input noisy image is processed to predict and subtract the noise to produce a result closer to the original image. Fernando et al. [4] show in their research that UNet is characterized by its unique U-shaped structure and is good at capturing both low- and high-level features of images. It operates by conditionally processing the random latent image representation in an iterative denoising manner, leveraging the text embeddings as guiding cues. The Variational Autoencoder (VAE) is a generative model mainly used to learn the latent representation of data and generate new data samples through these representations. It consists of two parts: the encoder and the decoder, which are used to process the images generated afterward. Specific functionality is illustrated in Fig 1.

The random seed in Stable Diffusion critically influences the generated image's style and content: varying the seed produces distinct synthesis results. Xu et al. [5] showed that the basic operational framework of text-to-image generation is as follows: large-scale models first ingest potential and text prompts as their main inputs. The latent seed then acts as a catalyst for the generation of an initial, randomly configured latent image representation. Parallel to this, in a study by Luo et al. [6], it was shown that text prompts were converted into text embeddings by using a Contrast Language Image Pre-training (CLIP) text encoder.

However, to generate high-resolution images, computer computation needs to be greatly increased. Research by Li et al. [7] shows that the main text-to-image method, Stable Diffusion, uses the Latent Diffusion Models (LDMS) method and usually requires a large dataset to build the model. Although the model performs well in most cases, there are still some limitations and challenges. LDMS still cannot handle high-resolution images. Pan et al. [8] showed that LDMS may be limited by its reconstruction ability in tasks requiring high precision. At the same time, since the diffusion model needs a lot of function evaluation and gradient calculation during the training process, the algorithm needs to be further optimized in practical application to improve efficiency. Additionally, since LDMS is a probability-based model, more research is needed to explore how to effectively utilize prior knowledge

to improve model performance. The purpose of this paper is to reduce the memory required by the block algorithm.

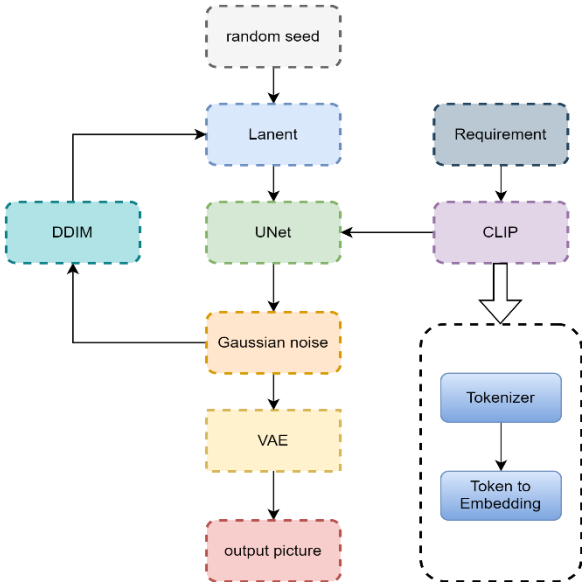

Fig. 1. Stable Diffusion framework diagram.

In general, the following limitations of traditional text-to-image generation methods can be identified:

When Stable Diffusion is used to generate high-precision pictures, the phenomenon of stashing or even memory explosion often occurs. Yariv et al. [9] found that Stable Diffusion occupied too much GPU memory. For Stable Diffusion, due to the use of large models, powerful encoders and decoders are used for data transmission, mainly using large models like CLIP. Research by Ting et al. [10] shows that it uses a Latent Diffusion Model (LDM). Subban R et al. found that in terms of image generation, although VAE can be used to compress images to appropriately reduce the GPU memory required for image generation, after decompression, the details of the image are rough and the image quality is reduced, as shown in Fig. 2. When the effect of the generated image is not obvious, it is often necessary to redraw the image. Chu et al. [11] found that AIGC redraws cause image distortion. In Hu et al. [12]. The study found that the less scope to redraw, the image is closer to the original image, the less the AI play space. Zhang et al. [13] found that this will greatly increase the possibility of image distortion.

Three more areas for innovation have been proposed based on these two points of shortcomings. We intend to add a tile block algorithm to the magnification model to reduce the memory pressure in the text-to-image process. In order to achieve the purpose of generating high-definition pictures with smaller memory.

Fig. 2. The difference when using traditional VAE.

In this work, we make the following contributions:

The tiling algorithm is combined with the magnification algorithm to reduce running GPU memory. The regional Block module, Residual Swin Transformer Block (RSTB) module and Swin Transformer Layer (STL) module are introduced, which emphasize more on the overall coordination and generation speed of the image.

We use ControlNet to limit the randomness of images. The minimal cell structure of the ControlNet model has two zero convolution modules whose weights and biases are initialized to zero. This allows ControlNet to fine-tune training on the capabilities of the original Stable Diffusion base model. The Recolor model in Convolutional Neural Network (CNN) and ControlNet combined with the tiling algorithm is used to achieve fine recoloring of black and white photos. We implement AIGC on an embedded device by deploying Stable Diffusion in the cloud. Implementing the magnification model based on AIGC in the cloud can enhance image generation speed and reduce the consumption of computer GPU memory.

2. Related Works

2.1 Selective Magnification Algorithm

Xu et al. [14] found that the main purpose of the image amplification algorithm is to recover high-resolution details from low-resolution images. These magnification algorithms are widely used in fields such as medical imaging, satellite remote sensing, video processing, and image processing. Liu et al. [15] found that common amplification algorithms can be divided into two categories: one is traditional image magnification algorithms, such as Lantent, Lanczos, Nearest, etc.; the other is AI-based image magnification algorithms, such as 4x-UltraSharp,

BSRGAN, ESGAN, etc. Rombach et al. [16] found that the current tiling work of general image amplification is usually achieved through a 4x-UltraSharp algorithm combined with a tiling algorithm, which has an obvious enhancement and amplification effect on images, but there are still shortcomings in details and computing speed. Therefore, this paper intends to compare Lantent, Lanczos, Nearest, ESGAN-4x, SCuNET, SCuNET PSNR, SwinIR 4x, 4X-ultrasharp and other amplification algorithms with TESD models combined with tiling algorithms.

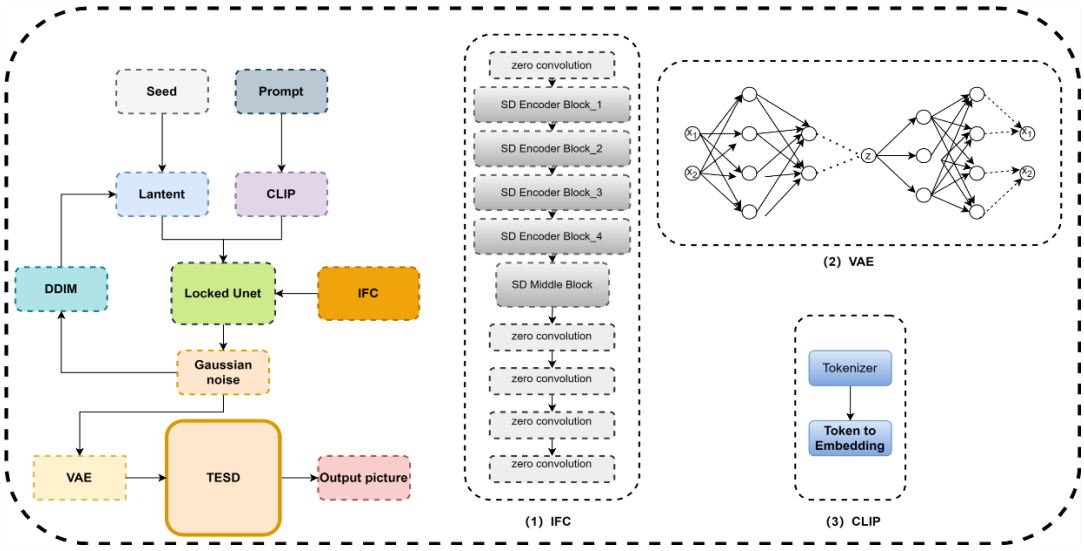

Fig. 3. Flowchart of TESD specific work on the cloud.

2.2 Zero Convolutional Layer

The research of Li and Wang et al. [17], indicates zero convolution layer is often used to eliminate the influence of random noise on the image generation process and prevent harmful noise from interfering with hidden states. Many control network models are mainly used to control human posture, facial expression, edge redrawing, etc. In the study of Wu et al. [18], this was found to be used in a control network model called recolor. The main principle of combining Recolor with ControlNet is to enter an additional condition into the neural network block. It then makes a copy of the same parameters as the original block for training. This trainable copy takes the external condition vector as input and uses the large pre-trained recolor model to build a powerful back to handle the various input conditions. In the study of Tong et al. [19], zero convolution has been shown to protect and eliminate noise, thus significantly improving the quality of the generated image.

2.3 Cloud Computing

In the research of Rao et al. [20], it is shown that the technical architecture of cloud computing is usually divided into the following layers: hardware layer (physical layer), including servers, storage devices, network hardware, and other infrastructure, providing computing and storage resources. In the study of Gui et al. [21], through virtualization technology, physical resources are abstracted into virtual resource pools, so that multiple users can share the same group of physical hardware and improve resource utilization. Resource management refers to the scheduling, allocation, and management of resources, including the dynamic adjustment of computing resources, storage resources, and network resources. The service layer provides IaaS, PaaS, SaaS, and other services to meet the needs of different users. Application layer: the application or service that the user uses directly, usually through a Web interface or API. The clouds we use in our daily lives include SaaS, PaaS, and LaaS. Due to the large number of files required for Stable Diffusion deployment and the large project, we choose SaaS for local deployment here. In the study of Gao et al. [22], it is shown that the ordinary Text-to-image graph project relies on strong computing power support, which has a huge demand for cloud computing resources. Especially for the training and inference of large-scale models, many GPU resources are required, resulting in computational problems.

3. Proposed Methods

The Tile-based extension model in the Stable Diffusion (TESD) network architecture proposes an innovative multi-level feature fusion architecture, and its core innovation point lies in the three collaborative designed modular components. It mainly includes three key elements: shallow feature extraction, deep feature extraction and high-quality image reconstruction. Compared with the traditional extended model, TESD introduces hierarchical Transformer in the diffusion super-resolution model for the first time to enhance the ability of cross-block global relationship modeling. Flowchart of TESD specific work on the cloud is depicted in Fig. 3. The key features are extracted from the prompt words through the CLIP model, combined with random seeds, and added to the UNet. The image is processed by dividing it into blocks, retaining the required features, and generating the image through the TESD amplification model and zero convolution.

3.1 TESD Amplification Module

Shallow feature extraction uses only one convolutional layer for feature extraction. A 3x3 convolutional high-frequency signal filter (HSF) was used to extract shallow features. The next step is deep feature extraction. The deep feature extraction module is composed of several residual Swin Transformer Blocks (RSTB) and convolutional blocks, and its specific structure is shown in Fig. 4., firstly, the

feature map of the shallow feature extraction module is divided into multiple non-overlapping patches embedded and then processed by several series residual Swin Transformer blocks. These blocks recombine multiple non-overlapping patch embeddings into one output with the same resolution as the input feature map. Secondly, a convolution layer outputs the result, with residual joins introduced inside each RSTB. In the residual RSTB, STL refers to the Swin Transformer layer, the structure of which is also shown in Fig. 4., it starts with a layer normalization layer, followed by a multi-head self-attention module. A residual connection is introduced at the end of the multi-head self-attention, followed by another layer of normalization. Finally, it passes through a Multilayer Perceptron (MLP), again introducing a residual connection at the end.

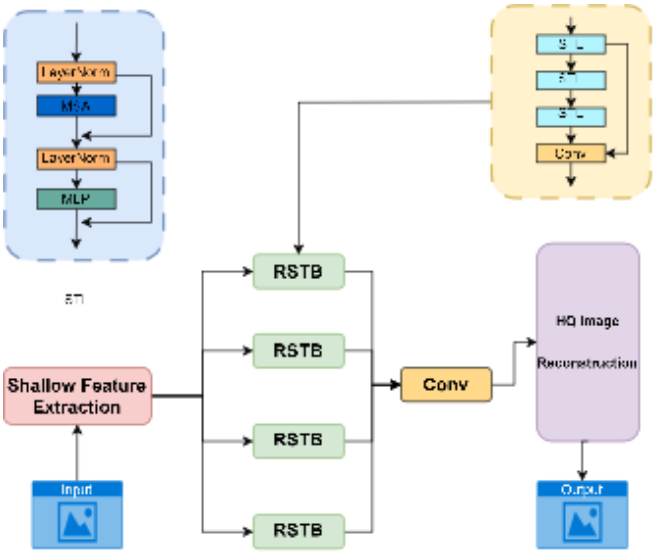

Fig. 4. TESD amplification module architecture diagram.

3.2 Zero Convolution Layer Combination

Combining the tiling magnification algorithm and zero convolution to restore black and white photos using AI. For this, we will need a ControlNet model called recolor, which works by first using a preprocessor to extract the grayscale image, then dividing the image into various regions through the tiling algorithm and recognizing each region to apply color. This model is the main one used in AIGC for coloring, and incorporating zero convolution and the tiling magnification algorithm can effectively reduce the interference of noise on image generation. The tiling magnification algorithm model can make the generated image clearer. We can also use the same quantitative evaluation method as mentioned above, first fixing the random seed and prompt words, then varying the zero convolution and tiling magnification algorithm, comparing the fineness of the images generated under

different conditions to find the best combination for restoring images. The specific structure is shown in Fig. 5.

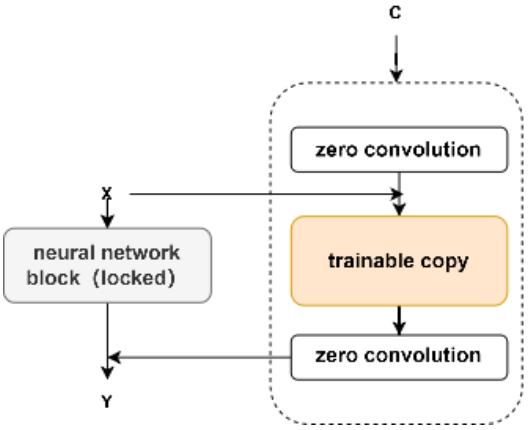

Fig. 5. Zero convolution structure diagram.

3.3 Cloud Deployment

In the context of the experimental design framework, the initial phase of establishing the environment emphasizes the selection of a Stable Diffusion environment configuration option, as offered by the Alibaba Cloud controller. This selection forms the foundation for constructing the fundamental scaffold of Stable Diffusion. However, it is important to recognize that the simplicity of this framework restricts its operational scope to the execution of basic text-to-image transformation tasks. Should the requirement emerge to expand functionality and meet more complex experimental demands, the file management component of the cloud project is activated. By uploading the necessary code files related to the desired features into their respective directories, the capabilities of the cloud Web UI are bolstered. This advancement enables the experimental setup to manage a broader spectrum of tasks and inquiries that are crucial to the overarching research objectives, ensuring a smooth and comprehensive experimental process.

4. Experiments

4.1 Experimental Settings

Dataset: In the present study, we utilized super-resolution reconstruction datasets, namely DIV2K and Flickr2K. By adhering to a unified and consistent training configuration protocol, we incorporated 1,485 samples derived from the DIV2K dataset in conjunction with 700 samples from the Flickr2K dataset to form our comprehensive training set. This strategic combination of datasets was selected to ensure a diverse and representative sample pool, thereby enhancing the robustness and generalizability of our experimental results. We can also use more diversified

data sets to train this model. It is recommended to use BSD100, Urban100, and other multi-style and larger data sets for training. The resulting model will be improved in accuracy and style.

Evaluation Metrics: For comprehensively and precisely evaluating the performance of image phase consistency and feature extraction, a battery of eleven well-recognized and established metrics was adopted, namely BRISQUE, DISTS, DSS, CLIP-IQA, FSIM, GMSD, HaarPSI, IW-SSIM, LPIPS, MDSI, and PSNR. The higher the values of FSIM, HaarPSI, IW-SSIM, PSNR and other metrics, the clearer the image. The lower the values of BRISQUE, DISTS, DSS, CLIP-IQA, GMSD, LPIPS, MDSI and other metrics, the clearer the generated image.

Comparison Models: The trained tiled upscaling model was systematically benchmarked against five traditional and commonly used upscaling models, namely Latent, Lanczos, Nearest, SCuNET, and SwinIR. The training and testing procedures of these models were carried out in strict accordance with the default settings as meticulously described in their corresponding original research publications. In situations where the source code was not publicly accessible, we resorted to the published experimental outcomes and results for the sake of comparative analysis. This approach ensures a fair and objective comparison, allowing for a more accurate assessment of the relative strengths and weaknesses of the tiled upscaling model in relation to its counterparts.

Configuration: The experimental environment is configured with a Windows 11 x64 operating system. The Java development kit JDK 1.8 is installed. PyCharm is used as the development tool. The system is powered by an Intel(R) Core (TM) i7-9750H CPU @ 2.60GHz, and an NVIDIA GeForce GTX 1650 Ti graphics card. It has 8GB of memory available for the experiment.

Experimental Parameters: The experimental setup is configured with specific parameters. The GPU_IDS is set as [0, 1] to utilize multiple GPUs for enhanced computational performance. A scale of 2 is applied, which likely affects the size or resolution transformation of the data. The number of channels is 3, typically corresponding to the RGB color model. Keywords such as "Boy white shirt" and "friendly smile" are defined to guide or evaluate certain aspects of the experiment related to image generation or analysis. The random seed number 35467 is used to ensure the reproducibility of results. The sampling method DPM+2M Karras is selected, with a resampling amplitude of 0.7 to control the sampling process. The scaling factor of the amplification algorithm is 2, and the image width and height are both set to 512 to define the dimensions of the images involved in the experiment.

4.2 Experimental Results

4.2.1 Select magnification algorithm:

In Experiment One, the SwinIR framework was selected for code refinement due to its functional comprehensiveness. A tiling algorithm was

integrated into the original code to optimize model processing and training. For image segmentation, key parameters such as width, height, and aspect ratio were precisely calibrated. Rectangles represented as tuples were generated and sorted by width in descending order. Through an iterative loop, they were positioned to prevent image distortion. A verification function was developed to assess space availability at (x, y) for rectangles of specific dimensions. If any point exceeded the boundary or was occupied, a negative result was returned. An auxiliary function marked and validated occupancy status to calculate the unoccupied area. Given the low-memory computer, training parameters were minimized. The incorporation of the tiling algorithm enhanced image restoration efficiency and reduced memory demands for resolution improvement. Experimental results are shown in Fig. 6.

Fig. 6. Training model comparison chart.

The model with the best training effect is put into Stable Diffusion, and the fixed keywords and seeds are input. The result is a comparison like the one shown in Fig. 7.

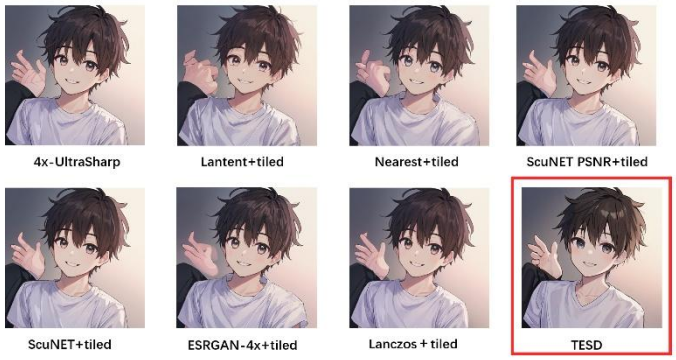

Fig. 7. Magnified model comparison diagram.

From the overall processing of the image, it can be observed that after incorporating tiled diffusion, the resolution of individual images becomes higher and clearer, meeting the basic requirements for generating detailed images on a small

computer with 4GB of memory. Regarding the overall style of the image, we notice that except for Latent [23], the pictures from the overall processing of the image, it can be observed that after incorporating tiled diffusion, the resolution of individual images becomes higher and clearer, meeting the basic requirements for generating detailed images on a small computer with 4GB of memory. Regarding the overall style of the image, we notice that except for Latent, the pictures processed by other upscaling algorithms appear more refined in style. Observing the clothing details of the characters, we find that the images upscaled by the 4x-UltraSharp algorithm lack detail and shading compared to those processed by Nearest [24], ScuNet [25], and ESRGAN [26] algorithms, which highlight details in clothing and shadow aspects. In terms of facial details, Lanczos [27] and ScuNETPSNR [28] show greater clarity. Examining the depiction of muscle lines in the hands, it is evident that ESRGAN suffers from significant blurring of the hand area, whereas the 4x-UltraSharp algorithm provides very realistic hand muscle lines.

We will evaluate the effects of different scale-up models from the following several indicators. LPIPS [29] is a perceptual similarity measurement method based on deep learning. It extracts deep representations in the image feature space through a pre-trained CNN to quantify the perceptual differences between two images. The core idea is to simulate the sensitivity of the human visual system (HVS) to image structure and high-order semantic features.

DSTS [30] is a no-reference (NR) image quality assessment index based on the statistics of the local structure direction of the image. This method quantifies the degree of structural confusion caused by distortion by calculating the Directional Statistics of the image gradient field. FSIM [30] is a full-reference image quality index based on phase consistency and gradient amplitude. Phase consistency characterizes the stability of structural features in an image that are not affected by illumination changes, and gradient amplitude describes the significance of local structures.

MS-SSIM [30] is a multi-scale extension of the Structural Similarity Index (SSIM), which improves the evaluation performance by simulating the multi-scale perception characteristics of the human visual system. HaarPSI [30] is a full-reference image quality index based on Haar wavelet coefficients, which assesses distortion by simulating the sensitivity of the visual system to edge information. GMSD [30] is a full reference image quality index based on the statistics of gradient amplitude similarity, quantifying distortion by calculating the spatial fluctuation of local gradient similarity. MDSI [30] is a full-reference index that integrates gradient, color and contrast information and comprehensively assesses distortion through multi-feature similarity bias. Evaluations of the images generated by the upscaling models are sequentially carried out, with specific evaluation data shown in Table 1 below.

Table 1

Model Data Comparison Table (The SWINIR that we use with the best results is already in bold)

Scale Algorithms		Latent	Lanczos	Nearest	Swinir	ScuNETPSNR
LPIPS		0.2912	0.2779	0.1769	**0.3609**	0.412
ContentLoss		1456.2516	840.1451	729.5243	**1131.7114**	1063.179
PSNR index		19.8546	23.414	23.4746	**22.3049**	22.6149
DISTS		0.1632	0.1386	0.0984	**0.1532**	0.1703
DSS	**index**	0.1665	0.2554	0.4141	**0.2104**	0.217
	loss	0.8335	0.7446	0.5859	**0.7896**	0.783
FSIM	**index**	0.8128	0.8886	0.9031	**0.8656**	0.8648
	loss	0.1872	0.1114	0.0969	**0.1344**	0.1352
GMSD	**index**	0.2299	0.1702	0.1565	**0.1865**	0.1874
	loss	0.2299	0.1702	0.1565	**0.1865**	0.1874
HaarPSI	**index**	0.3885	0.5575	0.5695	**0.5261**	0.5224
	loss	0.6115	0.4425	0.4305	**0.4739**	0.4776
IW-SSIM	**index**	0.6259	0.7853	0.8198	**0.7449**	0.7391
	loss	0.3741	0.2147	0.1802	**0.2551**	0.2609
MDSI	**index**	0.4316	0.3762	0.3663	**0.3989**	0.3949
	loss	0.4316	0.3762	0.3663	**0.3989**	0.3949
MS-SSIM	**index**	0.7718	0.8645	0.8947	**0.8337**	0.8304
	loss	0.2282	0.1355	0.1053	**0.1662**	0.1696
MS-GMSDc	**index**	0.2304	0.1684	0.1547	**0.186**	0.187
	loss	0.2304	0.1684	0.1547	**0.186**	0.187

4.2.2 Zero convolution layer combination

In the experimental paradigm of ControlNet, a sequence of meticulously designed configurations was executed. Specifically, the control weight parameter was systematically adjusted to a value of 1, while the start step and end step were deliberately set to 0 and 1, respectively. Concurrently, the Gamma Correction factor was precisely tuned to 1. Subsequently, the TESD model was incorporated into the experimental framework. The resultant restoration efficacy on black and white photographic images is visually demonstrated in Fig. 8.

Fig. 8. Black and white photo restoration comparison.

Through a comparative analysis of the figure, it becomes apparent that discernible coloring artifacts are present. When contrasted with the coloring manifestations in other figures, the coloring impact of the current model is more prominent in the context of landscape elements. In the case of small-scale images, a more precise and accurate coloration is attained. Following the incorporation of the TESD model, a substantial improvement in the image resolution was recorded, thereby leading to a more distinct and clearer visual rendition of the aged photographic materials.

4.2.3 Cloud computing

The TESD model was deployed in the Alibaba Cloud environment. The rationale behind this deployment lies in the model's unique approach of subdividing the target image into blocks, which effectively curtails the computational load on the cloud server. The integration of the Tiled module with the RSTB and STL modules within this cloud-based framework enables the generation of images with enhanced clarity. Moreover, this combination also leads to a notable improvement in the generation speed. The generation effect of the TESD model on the cloud computing interface is shown in Fig. 9. In the figure, it can be found that the efficiency and quality of generating pictures have been improved after adding the TESD model to the cloud framework.

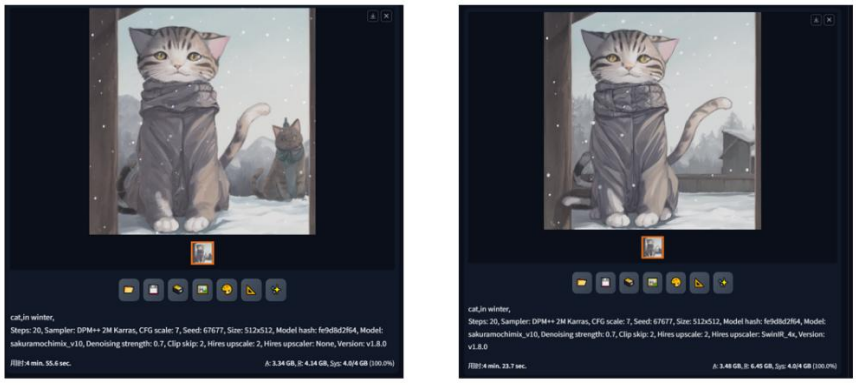

Fig. 9. Cloud computing generated effect comparison diagram.

5. Conclusion

In the current technological landscape, AIGC technology has emerged as a significant area of research and application. It holds great promise and potential for continuous growth and refinement. The versatility of AIGC technology, which encompasses a wide range of techniques and algorithms, allows it to permeate and make an impact in multiple industries and aspects of our lives. For instance, in the

field of content creation, it can assist in generating text, images, and even videos, streamlining the creative process and potentially opening up new avenues for artistic expression and communication.

When considering the specific application of image processing, the incorporation of tiling algorithms into the magnification model represents a crucial advancement. This innovation has led to a substantial elevation in the accuracy and visual fidelity of image repair. By subdividing the image into smaller, more manageable tiles, the model can more effectively allocate computational resources and handle complex image structures. As a result, not only is the quality of the restored images enhanced but also the memory requirements are significantly reduced. This achievement is of particular importance as it addresses one of the key challenges in modern computing, especially when dealing with large-scale image datasets or resource-constrained environments. TESD can be combined with other advanced technologies in the future. Such as processing high-definition video generation, improving video accuracy, or real-time AI image generation. These can be applied to future design work, generate a design drawing with high precision, and regenerate the corresponding video according to the design drawing.

This development in AIGC technology not only offers immediate benefits in terms of image processing but also paves the way for future exploration and innovation in the broader realm of artificial intelligence within the visual domain. It also encourages interdisciplinary collaborations between computer science, mathematics, and the visual arts, as the boundaries between these fields continue to blur in the pursuit of more advanced and intelligent image-processing techniques.

Acknowledgement

This research was supported by the Zhejiang Provincial. Philosophy and Social Science Planning Project under Grant. 22NDJC127YB, Ningbo City's 'Five Projects' - Research on the Talent Cultivation Model for Industry-Education Integration in the New Generation Information Technology Industry, Ningbo Science and Technology Fund under Grant. (2023Z228,2023Z213,2024Z296). College Level Research Project of College of Science and Technology, Ningbo University (YK202301). General Research Projects of Zhejiang Provincial Department of Education(Y202456101).

REFERENCE

[1] J. Kaleta, D. Dall'Alba, Szymon Płotka, and P. Korzeniowski, "Minimal data requirement for realistic endoscopic image generation with Stable Diffusion," International Journal of Computer Assisted Radiology and Surgery, vol. 19, no. 3, pp. 531–539, Nov. 2023, doi: https://doi.org/10.1007/s11548-023-03030-w.

[2] T. Lin, Z. Chen, Z. Yan, W. Yu, and F. Zheng, "Stable Diffusion Segmentation for Biomedical Images with Single-Step Reverse Process," Lecture notes in computer science, vol. 25, no. 26, pp. 656–666, Jan. 2024, doi: https://doi.org/10.1007/978-3-031-72111-3_62.

[3] Ahmed Imran KABIR, Limon MAHOMUD, Abdullah Al Fahad, and R. AHMED, "Empowering Local Image Generation: Harnessing Stable Diffusion for Machine Learning and AI," Informatică economică, vol. 28, no. 1/2024, pp. 25–38, Mar. 2024, doi: https://doi.org/10.24818/issn14531305/28.1.2024.03.

[4] Z. Luo, F. K. Gustafsson, Z. Zhao, Jens Sjölund, and T. B. Schön, "Photo-Realistic Image Restoration in the Wild with Controlled Vision-Language Models," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), vol. 24, no. 25, pp. 6641–6651, Jun. 2024, doi: https://doi.org/10.1109/cvprw63382.2024.00658.

[5] E. Merino-Gómez, F. M. Andrés, B. Querol, and P. R. Vasallo, "Stable Diffusion aprende de Sebastiano Serlio: dibujo de arquitectura con inteligencia artificial," EGA Revista de expresión gráfica arquitectónica, vol. 29, no. 51, pp. 258–267, Oct. 2024, doi: https://doi.org/10.4995/ega.2024.20332.

[6] W. Cai et al., "Hierarchical damage correlations for old photo restoration," Information Fusion, vol. 107, no. 25, p. 102340, Jul. 2024, doi: https://doi.org/10.1016/j.inffus.2024.102340.

[7] R. Li, X. Sheng, W. Li, and J. Zhang, "OmniSSR: Zero-Shot Omnidirectional Image Super-Resolution Using Stable Diffusion Model," Lecture notes in computer science, pp. 198–216, Oct. 2024, doi: https://doi.org/10.1007/978-3-031-72751-1_12.

[8] X. Yao, Y. Pan, and J. Wang, "An Omnidirectional Image Super-Resolution Method Based on Enhanced SwinIR," Information, vol. 15, no. 5, pp. 248–248, Apr. 2024, doi: https://doi.org/10.3390/info15050248.

[9] Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation," Feb. 2023, doi: https://doi.org/10.48550/arxiv.2302.08113.

[10] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer," arXiv.org, Aug. 23, 2021.

[11] B.-B. Gao et al., "AdaptCLIP: Adapting CLIP for Universal Visual Anomaly Detection," arXiv.org, 2025. https://arxiv.org/abs/2505.09926 (accessed Jun. 06, 2025).

[12] E. Chu, S.-Y. Lin, and J.-C. Chen, "Video ControlNet: Towards Temporally Consistent Synthetic-to-Real Video Translation Using Conditional Image Diffusion Models," arXiv (Cornell University), Jan. 2023, doi: https://doi.org/10.48550/arxiv.2305.19193.

[13] Z. Hu and D. Xu, "VideoControlNet: A Motion-Guided Video-to-Video Translation Framework by Using Diffusion Model with ControlNet," arXiv (Cornell University), Jan. 2023, doi: https://doi.org/10.48550/arxiv.2307.14073.

[14] J. Zhang, Y. Liu, Y.-W. Tai, and C.-K. Tang, "C3Net: Compound Conditioned ControlNet for Multimodal Content Generation," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 26876–26885, Jun. 2024, doi: https://doi.org/10.1109/cvpr52733.2024.02539.

[15] S. Xu, J. Zhang, and L. Yunqin, "Knowledge-Driven and Diffusion Model-Based Methods for Generating Historical Building Facades: A Case Study of Traditional Minnan Residences in China," Information, vol. 15, no. 6, pp. 344–344, Jun. 2024, doi: https://doi.org/10.3390/info15060344.

[16] B. Liu, C. Wang, T. Cao, K. Jia, and J. Huang, "Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7817–7826, Jun. 2024, doi: https://doi.org/10.1109/cvpr52733.2024.00747.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv (Cornell University), Dec. 2021, doi: https://doi.org/10.48550/arxiv.2112.10752.

[18] L. Zhang and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," Feb. 2023, doi: https://doi.org/10.48550/arxiv.2302.05543.

[19] T. Wu et al., "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development," IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 5, pp. 1122–1136, May 2023, doi: https://doi.org/10.1109/jas.2023.123618.

[20] X. Ren, L. Tong, J. Zeng, and C. Zhang, "AIGC scenario analysis and research on technology roadmap of Internet industry application," China Communications, vol. 20, no. 10, pp. 292–304, Oct. 2023, doi: https://doi.org/10.23919/jcc.fa.2023-0359.202310.

[21] D. B. J. Rao, V. Polepally, S. N. Prabhu, and P. Kalpana, "Deep recurrent neural network-based Hadoop framework for COVID prediction with applications to big data in cloud computing," International Journal of Bio-Inspired Computation, vol. 21, no. 1, p. 36, 2023, doi: https://doi.org/10.1504/ijbic.2023.130022.

[22] H. Gui, J. Liu, C. Ma, M. Li, and S. Wang, "Mist-edge-fog-cloud computing system for geometric and thermal error prediction and compensation of worm gear machine tools based on ONT-GCN spatial–temporal model," Mechanical Systems and Signal Processing, vol. 184, p. 109682, Feb. 2023, doi: https://doi.org/10.1016/j.ymssp.2022.109682.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv:2112.10752 [cs], Apr. 2022, Available: https://arxiv.org/abs/2112.10752.

[24] P. Langley, "Average-case analysis of a nearest neighbor algorithm," Core.ac.uk, 2018, doi: oai:CiteSeerX.psu:10.1.1.80.7122.

[25] Chen, Y., Zou, B., Guo, Z., Huang, Y., Huang, Y., Qin, F., ... & Wang, C. (2024). Scunet++: Swin-unet and cnn bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism ct image segmentation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 7759-7767)..

[26] X. Wang et al., "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," arXiv.org, 2018. https://arxiv.org/abs/1809.00219.

[27] Y. Saad, "On the Rates of Convergence of the Lanczos and the Block-Lanczos Methods," SIAM Journal on Numerical Analysis, vol. 17, no. 5, pp. 687–706, Oct. 1980, doi: https://doi.org/10.1137/0717059.

[28] B. Yan et al., "MT-SCUNet: A hybrid neural network for enhanced mode decomposition in optical fibers," Optical Fiber Technology, vol. 93, p. 104196, Mar. 2025, doi: https://doi.org/10.1016/j.yofte.2025.104196.

[29] M. Kettunen, E. Härkönen, and J. Lehtinen, "E-LPIPS: Robust Perceptual Image Similarity via Random Transformation Ensembles," arXiv.org, 2019. https://arxiv.org/abs/1906.03973?context=cs (accessed Jun. 06, 2025).

[30] J. Karotte and E. G. Sarma, "An evaluation of the effect of image down-sampling on performance indicators of IQA algorithms," ResearchGate, vol. 10, no. 17, pp. 7507–7513, 2015, Accessed: Jun. 06, 2025. [Online]. Available: https://www.researchgate.net/publication/283882739_An_evaluation_of_the_effect_of_image_down-sampling_on_performance_indicators_of_IQA_algorithms