# High Resolution Image Generation Algorithm for Archaeology Drawings

X. Zeng, L. Cheng, Z. Li, X. Liu

***Abstract*—**Aiming at the problem of low accuracy and susceptibility to cultural relic diseases in the generation of high-resolution archaeology drawings by current image generation algorithms, an archaeology drawings generation algorithm based on a conditional generative adversarial network is proposed in this paper. An attention mechanism is added into the high-resolution image generation network as the backbone network, which enhances the line feature extraction capability and improves the accuracy of line drawing generation. A dual-branch parallel architecture consisting of two backbone networks is implemented, where the semantic translation branch extracts semantic features from orthophotographs of cultural relics, and the gradient screening branch extracts effective gradient features. Finally, the fusion fine-tuning module combines these two types of features to achieve the generation of high-quality and high-resolution archaeology drawings. Experimental results on the self-constructed archaeology drawings dataset of grotto temple statues show that the proposed algorithm outperforms current mainstream image generation algorithms in terms of pixel accuracy (PA), structural similarity (SSIM), and peak signal-to-noise ratio (PSNR) and can be used to assist in drawing archaeology drawings.

***Keywords*—**Archaeology drawings, digital heritage, image generation, deep learning.

## I. Introduction

THE archaeology drawings, drawn to scale according to archaeological standards, uses simple lines to delineate the shape and contours of cultural relics [1]. Archaeology drawing is an indispensable part of archaeological work, and every complete archaeological excavation report needs to be accompanied by drawing data [2]. The archaeology drawings provide a direct representation of the form, stratification, and layout of cultural artifacts, effectively addressing the challenge of accurately conveying the abstract features of relics through words and photographs. Consequently, it stands as a crucial resource for both relic restoration efforts and academic inquiry. Currently, the creation of archaeology drawings relied primarily on manual measurements and drawings by trained professionals, which was inefficient and accompanied by the risk of secondary damage to artifacts. Consequently, there has been a continuous emergence of research on algorithms for generating archaeology drawings in recent years. For instance, Li [3] and Wang et al. [4] have proposed an algorithm for generating archaeology drawings based on the detection of ridge line features to extract the contour lines of artifact 3D models. Liu [5] employs edge tangential flow fields and Gaussian difference filters to extract and generate line features, while Song [6] has introduced a cell ant colony algorithm for edge extraction across three-dimensional relic views, employing cubic B-spline curve fitting to produce the final relic drawings. The adoption of algorithmic methods for archaeology drawings generation represents a significant advancement over manual techniques, substantially enhancing the efficiency of relic workers and alleviating the burden of line drawing tasks. However, above methods still encounter challenges such as noise interference, loss of details, and the production of cluttered lines.

The rapid advancement of data-driven approaches and deep learning techniques has led to the emergence of numerous algorithms in image generation. Notably, Convolutional Neural Networks (CNNs) [7] and Generative Adversarial Networks (GANs) [8] have achieved remarkable success in the field of image generation. Among these, GANs have particularly excelled in image generation tasks. Initially proposed by Goodfellow et al. [8], GANs have undergone significant enhancements in subsequent research efforts. Mehdi et al. [9] extended GANs by incorporating supervised information as conditional constraints, introducing Conditional Generative Adversarial Networks (CGANs). The optimization significantly alleviated the instability problems present in the training process of the original GANs. Based on CGANs, Isola et al. [10] proposed pix2pix, a conditional image translation network. Pix2pix utilizes a U-Net [11] structure in its generator to integrate shallow and deep features. Moreover, its discriminator employs a Markov discriminator (PatchGAN) [14] to discern the authenticity of image division regions, enhancing the network's ability to recognize local high-frequency features while reducing model computation. Furthermore, deep learning algorithms have found effective application in the extraction of archaeology drawings. For instance, Peng et al. [12] combined the bidirectional cascade network (BDCN) [13] with U-Net to construct a hierarchical depth structure network for extracting archaeology drawings from murals. The application of these deep learning technologies has the potential to revolutionize the generation of archaeology drawings, offering new opportunities and inspiring future research directions.

Deep learning methods hold promise for generating realistic

Xiaolin Zeng* is with Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guangdong 518055, China (e-mail: zengxl22@mails.tsinghua.edu.cn).

Lei Cheng* and Xueping Liu are with Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guangdong 518055, China.

Zhirong Li is with Cultural Heritage Institute, Zhejiang University, Hangzhou, Zhejiang 310027, China.

*These authors contributed to the work equally and should be regarded as co-first authors.

archaeology drawings, yet the existing methods often fail to deliver satisfactory results for complex scenes. There are two main challenges. Firstly, existing deep learning methods are limited in their ability to handle high-resolution images. Large artifacts such as caves and murals can only be processed by downsampling or cropping to reduce image resolution to meet the requirements of the model input, which often leads to loss of details in the output line drawings. Secondly, cultural heritage diseases disrupt the inherent features of artifacts. However, the existing methods have limited capabilities in extracting artifact line features, leading to poor quality in generated line drawings [12].

To address these problems, this paper presents a high-resolution archaeology drawings generation algorithm based on CGAN. The algorithm enhances the feature extraction network by incorporating attention mechanism. Additionally, it introduces a gradient screening branch and a semantic translation branch to generate semantic feature maps and gradient feature maps, respectively. These branches aim to better capture both the semantic and gradient information essential for generating high-quality line drawings. Finally, a fusion module integrates the gradient and semantic information to produce superior quality line drawings. Experiments were conducted on a self-constructed archaeology drawings dataset of grotto temples. Comparative analysis against mainstream methods demonstrates that the proposed method generates higher-quality line drawings, thereby advancing the development of computer-aided line drawing technology.

## II. Related Work

### A. Conditional Generative Adversarial Networks

Generative Adversarial Networks [10], initially proposed by Goodfellow et al., operate by continually optimizing their parameters through adversarial training of generators and discriminators, ultimately achieving the generation of realistic images. However, GANs typically take random noise as input, which can sometimes lead to difficulties in training the network due to information loss. To address this limitation, Mehdi et al. [9] introduced supervisory information as a conditional constraint, proposing Conditional Generative Adversarial Networks (CGANs).

As illustrated in Fig. 1, CGANs incorporate not only random noise but also real tag information as input. This enhancement allows the generator to produce more targeted images, while enabling the discriminator to provide more directed feedback to the generator, thereby enhancing the overall performance of CGANs. The proposed method in this paper is based on the core idea of CGANs, introducing conditional constraints into the generative adversarial network framework. This enables the model to continuously make progress in the training process of generation, discrimination, and feedback optimization.

### B. Convolutional Block Attention Module

Woo et al. [16] present the Convolutional Block Attention Module (CBAM), which integrates spatial attention and channel attention. The structure of CBAM, depicted in Fig. 2, comprises the connection of the channel attention module and spatial attention module, augmented with a residual connection. Spatial attention evaluates pixels across the entire image, assigning higher weights to pixels in significant regions, while channel attention assesses the importance of various channel features, determining the allocation of weights for each channel. By combining both spatial and channel attention mechanisms, CBAM effectively captures informative spatial and channel-wise dependencies within the input features.

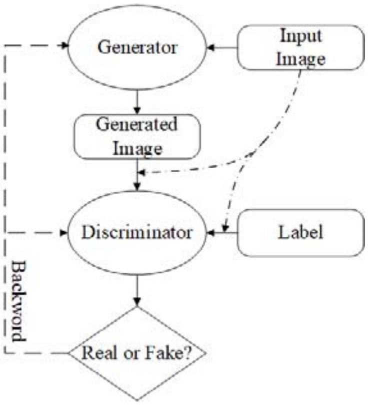

Fig. 1 CGAN structure

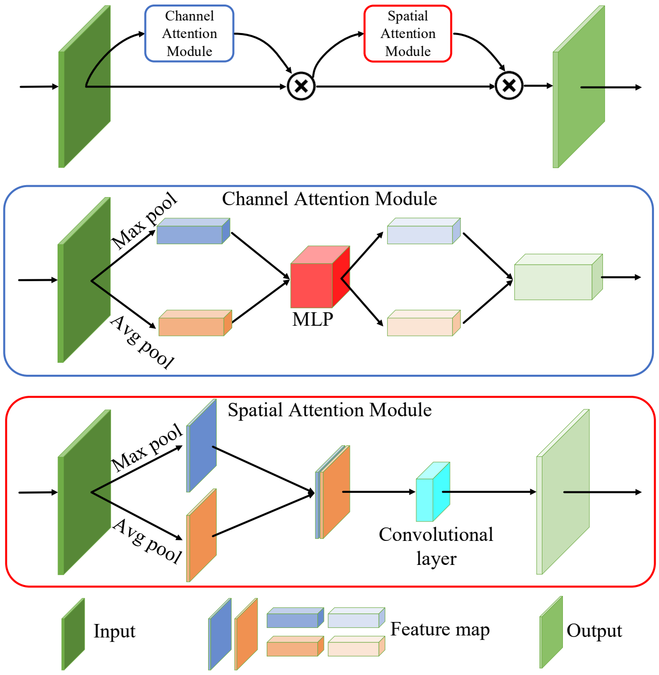

Fig. 2 CBAM structure

As depicted in Fig. 2, the channel attention module conducts both maximum pooling and average pooling operations on the input feature map along the spatial dimension. Subsequently, it extracts channel-wise features through a Multi-Layer Perceptron (MLP) [17] with shared weights. The outputs of both operations are then added pixel-wise, and then activated through a sigmoid function to obtain the channel attention feature map. This process can be mathematically represented by

(1), where $M_c(F)$ denotes the channel attention feature map, $\sigma$ represents the sigmoid function, $W_0$ and $W_1$ denote the weights of the two convolution layers respectively, $F^c_{avg}$ represents the output of average pooling, and $F^c_{max}$ represents the output of maximum pooling.

$$M_c(F) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \tag{1}$$

The spatial attention module conducts average and maximum pooling operations on the input feature map along the channel dimension. It then concatenates the resulting feature maps and convolves the output to obtain the spatial attention feature map. This process is mathematically expressed by (2), where $M_s(F)$ denotes the spatial attention feature map, $\sigma$ represents the sigmoid function, $f^{7\times7}$ denotes the convolution operation with a 7×7 convolution kernel, $F^s_{avg}$ represents the output of average pooling, and $F^s_{max}$ represents the output of maximum pooling.

$$M_s(F) = \sigma(f^{7\times7}([F^s_{avg}; F^s_{max}])) \tag{2}$$

Finally, CBAM combines the channel attention feature map and the spatial attention feature map to obtain the attention feature weight, which is then element-wise multiplied with the input feature map to produce the final output feature map.

In this paper, CBAM is incorporated into the archaeology drawings generation model to bolster the model's capacity in identifying local features within cultural relic orthophotographs. These features encompass image contours, decorative edges, and other intricate details crucial for producing higher-quality line drawings. With the help of CBAM's ability to capture both channel-wise and spatial dependencies, the model enhances sensitivity to relevant features, thereby facilitating the generation of line drawings with improved fidelity and accuracy.

## III. Method

For the design of archaeology drawing generation networks, it is important to consider two critical factors: image resolution and gradient information.

Firstly, cultural relic orthophotographs obtained from archaeological surveys typically possess high resolutions, which may exceed the input resolution limits of general deep learning models. As a result, preprocessing steps such as downsampling or cropping are often required to reduce image resolution to facilitate network training. However, downsampling leads to direct information loss, while cropping compromises image integrity and disrupts contextual coherence. Hence, network design should prioritize enhancing the model's ability to process high-resolution images effectively. Secondly, due to the focus on features such as contours and edges of artifacts in archaeology drawing generation tasks, obtaining gradient information from images is a crucial aspect of network design.

To generate high-resolution and detailed archaeological line drawings, it is necessary to fuse features extracted from images at different scales to ensure the integrity of both global and local features. Additionally, it is important to design a reasonable approach for extracting effective gradient information.

Based on CGAN architecture, this paper proposed a method that incorporated both image semantics and gradient information extraction, termed the Semantic Translation and Gradient Screening Dual-Branch Archaeology Drawings Generation Model. The network structure, depicted in Fig. 3, consists of two parallel branches: the Semantic Translation Branch (STB) and the Gradient Screening Branch (GSB). The feature maps output from both branches are subsequently combined by a fusion fine-tuning module to generate the archaeology drawings.

The primary concept underlying this method is to leverage the STB to learn texture, color, and other semantic features of cultural relic images. Concurrently, the GSB focuses on extracting contour and edge features from the image while filtering out redundant gradient information. Finally, the outputs of both branches are effectively integrated to generate an archaeology drawing with high resolution and clear details. This approach facilitates comprehensive feature extraction and synthesis, enhancing the model's ability to generate accurate and detailed line drawings.

The network calculation process unfolds as follows: Initially, the cultural relic image serves as the input to the STB, while the edge extraction algorithm generates the initial gradient map from the same cultural relic image, which acts as the input to the GSB. Dual branches independently produce the semantic feature map and the gradient feature map, respectively. Subsequently, the fusion fine-tuning module combines the feature maps to generate the predicted archaeology drawings.

During network training, the parameters of the network undergo updates through the adversarial game between the discriminator and generator of each branch. This adversarial training process facilitates the refinement of the model's ability to generate accurate archaeology drawings by iteratively optimizing the semantic translation and GSBes.

### *A. Gradient Screening Branch*

The extraction of gradient information can be considered as an edge detection task. However, gradient maps obtained solely through conventional edge detection operators often contain a significant amount of noisy artifacts. The key to generating high-quality line drawings lies in filtering out this intricate noise and retaining essential line information within the cultural relic gradient image.

Consequently, the task of the GSB is defined as the selective extraction of necessary gradients for generating line drawings from the extensive and intricate gradient information present in the cultural relic images, as depicted in Fig. 4.

The GSB is constructed based on CGAN, comprising the generator $G_2$ and the discriminator $D_2$. As illustrated in Fig. 3, the input of $G_2$ is the initial gradient image $I_g$ extracted from the cultural relic image $I_o$, and the output is the refined gradient image $g$ after screening. Discriminator $D_2$'s task is to compare the gradient feature map g with the gradient label of the line drawing $gt_g$, thereby compelling $G_2$ to generate a more realistic $g$. Essentially, generator $G_2$ learns the mapping relationship between the images from domain $I_g$ and $gt_g$. Therefore, it can focus more on distinguishing between noise and lines, thus achieving the purpose of filtering gradient information.

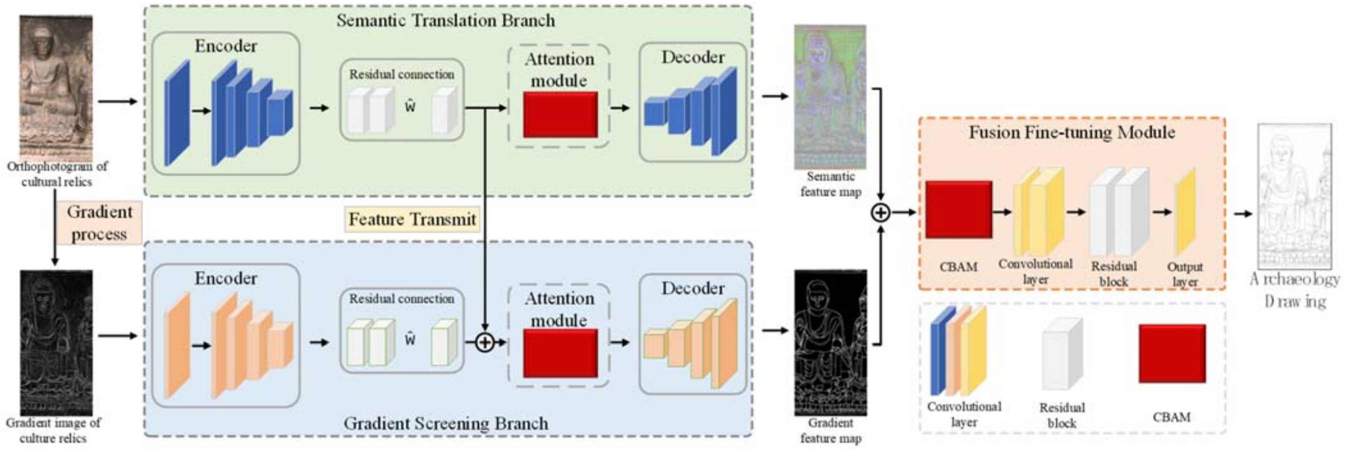

Fig. 3 Network structure

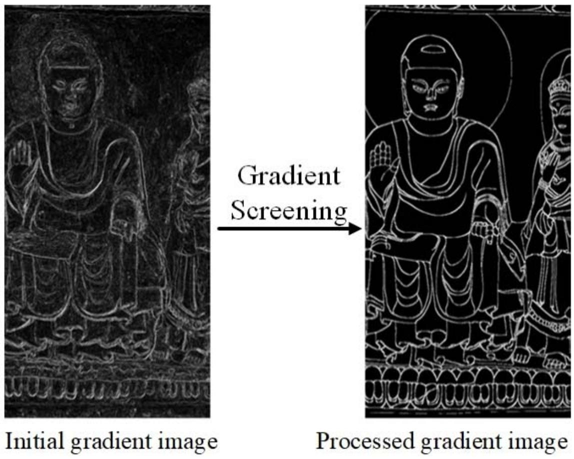

Fig. 4 Process of extracting gradient feature maps by GSB

The initial gradient map $I_g$ and the gradient label of the line drawing $gt_g$ are obtained using the Sobel operator. The Sobel operator is a first-order derivative edge detection operator, which calculates gradients separately in the $x$ and $y$ directions to obtain the final gradient. It acts on the artifact orthophotograph $I_o$ and the line drawing label $gt_g$.

In terms of specific network architecture design, discriminator $D_2$ has been improved based on a multi-scale Markov discriminator [10], as illustrated in Fig. 5. The core concept of the Markov discriminator is to partition the image into several regions and penalize based on the similarity of each region, allowing the network to focus on learning high-frequency information and local features. The discriminator adopts multi-scale architecture, creating an image pyramid by downsampling the line drawing results and labels, which are then separately fed into the Markov discriminator. In this paper, an attention mechanism is integrated into the discriminator to enhance its discriminative ability, specifically employing the CBAM attention module. By discriminating between the gradient feature maps and gradient labels at different scales, discriminator $D_2$ can effectively identify differences between them in terms of global perspective and local details, thereby guiding the generator $G_2$ to produce more accurate gradient maps.

The generator employs a full convolutional structure organized in an encoder-decoder architecture, as illustrated in Fig. 6. In the encoder section, the channel number is initially increased to 64 using a 7x7 convolution kernel with a large receptive field. Subsequently, downsampling is performed four times using 3x3 convolution kernels, and several residual blocks are incorporated to enhance feature extraction. Finally, the generator's feature extraction capability is bolstered by integrating the CBAM attention module.

In the decoder section, four symmetrical upsampling operations are sequentially conducted to decode the extracted deep features and generate the gradient feature map. This symmetrical convolutional structure is designed to output a high-resolution gradient map identical to the cultural relic image. By performing multiple downsampling and upsampling operations, the model's receptive field is expanded while preserving more spatial information.

The inclusion of an intermediate feature layer comprising multiple residual blocks serves to deepen the network's depth, enabling it to learn more abstract features. This mitigates issues such as gradient disappearance and explosion, thereby enhancing the training stability of the model. Overall, this architecture facilitates effective feature extraction and synthesis, ultimately contributing to the generation of high-quality gradient maps.

The generator is a fully convolutional structure, constructed in an encoder-decoder architecture, as depicted in Fig. 6. The encoder part first increases the number of channels to 64 using a 7x7 convolutional kernel with a larger receptive field, followed by four downsampling operations using 3x3 convolutional kernels. Subsequently, several residual blocks are connected, and finally, the CBAM attention module is employed to enhance the feature extraction capability of the generator. The symmetric convolutional structure is designed to output high-resolution gradient maps with the same resolution as the input artifact images. By incorporating multiple downsampling and upsampling operations, the model's receptive field can be increased while preserving more spatial information. The intermediate feature layers composed of multiple residual blocks deepen the network's depth, enabling it to learn more abstract features and effectively alleviate the vanishing and exploding gradient issues, thus improving the training stability of the model.

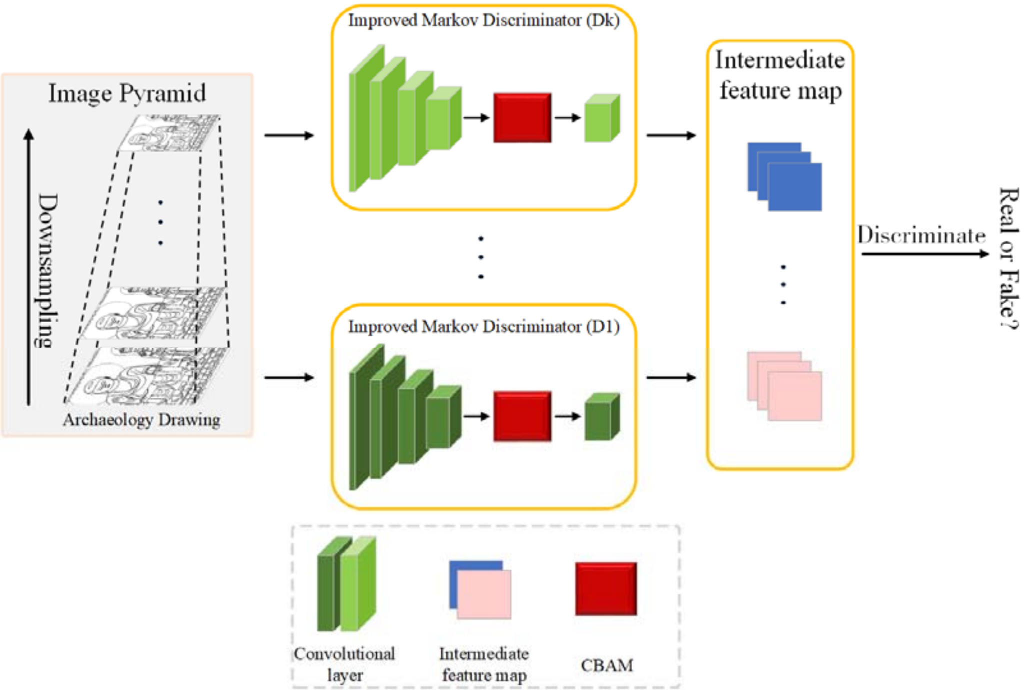

Fig. 5 Improved discriminator structure

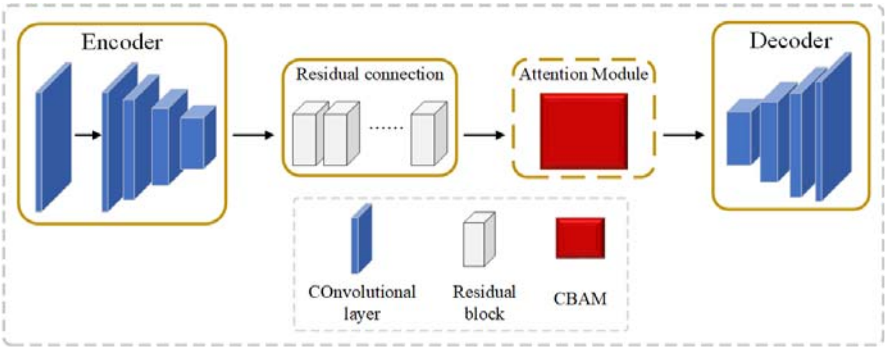

Fig. 6 Improved generator structure

### *B. Semantic Translation Branch*

Different from the GSB, the STB focuses on extracting semantic information from cultural relic images, including advanced features such as color and texture, as depicted in Fig. 7.

Cultural relics often exhibit various degrees of deterioration, such as weathering, erosion, and pollution. It is challenging to capture the advanced features of cultural relics solely through gradient information. Therefore, the STB is tasked with learning the mapping relationship between different cultural relic images and archaeology drawings. Its objective is to enrich the semantic information generated by the cultural relic images and reduce semantic loss caused by cultural relic deterioration as much as possible.

The structure of STB closely resembles that of GSB, comprising a generator $G_1$ and a multi-scale Markov discriminator $D_1$, both of which are augmented by the attention module CBAM to enhance their abilities. In the STB, the input to $G_1$ is the cultural relic orthophotographs $I_o$, and the output is the semantic feature map $s$ containing high-level semantic information. Similar to the GSB, the generator $G_1$ adopts a full convolutional structure, incorporating convolution layers with varying receptive fields and multiple intermediate residual block layers. Similarly, the discriminator $D_1$ in the STB also adopts a multi-scale Markov discriminator structure. It focuses on identifying local detail differences in cultural relics while considering the overall contextual information, which ensures that $G_1$ generates a semantic feature map *s* with comprehensive semantic information and clear details, guided by both local and global information provided by $D_1$. Overall, the STB is designed to effectively capture and represent the complex semantic characteristics of cultural relics, ensuring that the generated semantic feature map contains rich and accurate semantic information essential for generating high-quality line drawings.

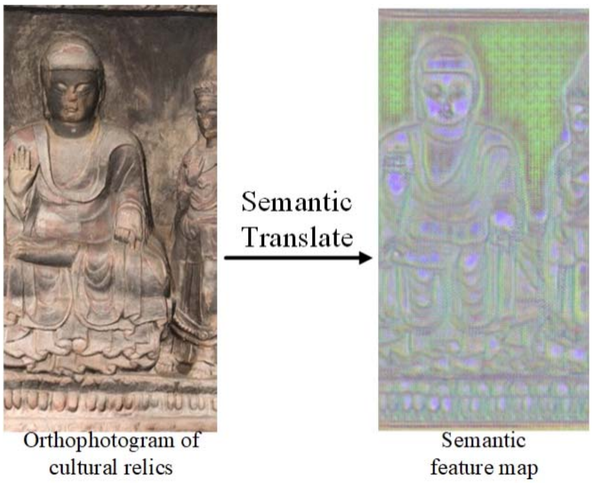

Fig. 7 Process of extracting semantic feature maps by STB

Compared to the advanced features such as color and texture that STB can learn from cultural relic images, GSB is limited to learning gradient information such as edges and contours. This limitation may result in the loss of details during the generation process. To address this problem, this paper proposed feature transmission to facilitate the transmitting and sharing of semantic information learned by STB. Specifically, during the training process of STB, intermediate semantic features obtained by the encoder are transmitted to the intermediate feature layer of GSB. These features are then decoded to output the gradient feature map and concatenated along the channel dimension. The transmission of intermediate semantic feature maps supplements the high-level semantic information learned by STB from the orthophotographs of cultural relics. It compensates for the information lost during the calculation process of the Sobel operator in GSB. As a result, GSB can converge faster and output more accurate gradient maps, thereby guiding the synthesis of archaeology drawings more effectively.

### C. *Fusion Fine-Tuning Module*

The gradient feature maps output by the GSB are grayscale images, primarily carrying high-frequency information such as the contours and edges of the artifact's orthophotographs. On the other hand, the STB outputs semantic feature maps as RGB images, mainly containing higher-level semantic features such as colors, textures, and objects present in the artifact's orthophotographs. An ideal archaeology drawing should combine both clear and accurate lines (from gradient information) and clear semantic representation of cultural relic objects (from semantic information). Therefore, the task of the Fusion Fine-tuning Module is to merge the semantic feature maps and gradient feature maps, supplementing the selected gradient information into the semantic feature maps, thus generating the archaeology drawings.

Since the semantic features and gradient information required for the archaeology drawings have already been extracted in the STB and GSB, the Fusion Fine-tuning Module needs to learn the process of merging these features. It is not reasonable to design overly complex network structures to avoid increasing the difficulty of network training and overfitting.

The structure of the Fusion Fine-tuning Module is illustrated in Fig. 8, comprising an attention module, fusion convolutional layer, residual connection layer, and output convolutional layer. The fusion convolutional layer consists of two layers of 3x3 convolutions, the residual connection layer consists of two residual blocks, and the output convolutional layer is a single 3x3 convolution. After inputting the semantic feature maps and gradient feature maps into the Fusion Fine-tuning Module in a 3:1 ratio, channel concatenation and feature fusion are performed, resulting in the output of the line drawings.

### D. *Gradient Propagation Strategy*

The proposed method in this paper consists of three different generators, namely $G_1$ for the STB, $G_2$ for the GSB, and $G_f$ for the fusion fine-tuning module.

During the network training process, it is necessary to adhere to a reasonable forward propagation strategy; otherwise, there may be null pointer errors due to undefined feature maps at certain moments, leading to training failures. Additionally, since this model involves two types of labels, namely gradient labels and line map labels, the design of the backpropagation strategy also needs to be rational. So, the propagation strategy designed in this paper is as follows:

(1) Forward Propagation Strategy:

1) Generator $G_1$ first conducts forward propagation to obtain semantic feature maps and intermediate features used for feature transmission, and then waits.
2) $G_2$ conducts forward propagation, incorporating intermediate features from $G_1$ into the calculation process to obtain gradient feature maps.
3) After both $G_1$ and $G_2$ complete forward propagation, $G_f$ conducts forward propagation to generate the final archaeology drawing.

(2) Backward Propagation Strategy:

1) After $G_2$ generates the gradient map, the loss of the gradient filtering branch is calculated based on the gradient labels, and then backpropagation is performed. Since the generation of the $G_2$ gradient feature map integrates the intermediate features from $G_1$, the backpropagation will update the overall network parameters of $G_2$ and the network parameters of the downsampling part of $G_1$.
2) After the fusion fine-tuning module finally generates the line drawing, the loss of the line drawing is calculated, and backpropagation is carried out. Since the input of the fusion fine-tuning module comes from both $G_1$ and $G_2$, $G_1$ and $G_2$

will undergo backpropagation again. Throughout the entire backpropagation process, both branches have undergone multiple parameter updates, enhancing the stability of network training.

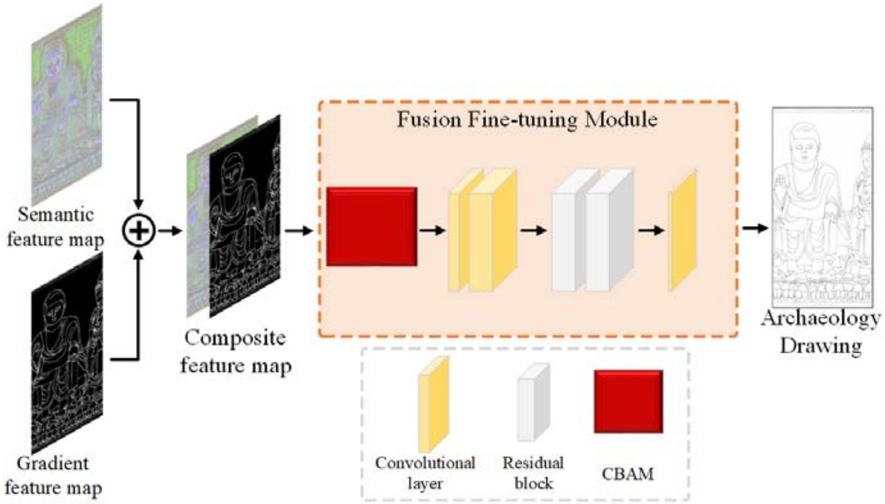

Fig. 8 Fusion Fine-tuning Module

### *E. Loss Function*

The loss function proposed in this paper includes adversarial loss, feature matching loss, and VGG loss. These losses measure the differences between the intermediate feature maps and the generated archaeological line maps and their corresponding labels from different dimensions to guide the overall training process of the network, aiming to improve the accuracy of the model in generating archaeology drawings as much as possible.

(1) Adversarial Loss

The STB and GSB each constitute a set of adversarial losses, calculated using mean squared error. For GSB, its adversarial loss is defined as (3):

$$L_{GAN\text{-}GSB}\left(G_2,D_2\right)=E_{I_g,gt_g}\left\{logD_2\left(I_g,gt_g\right)\right\}+ E_{I_g}\left\{log\left[1\text{-}D_2\left(I_g,G_2\left(I_g;\theta_{G_2}\right)\right)\right]\right\} \tag{3}$$

where $I_g$ and $gt_g$ represent the gradient labels of the orthophotograph and line drawing respectively, and $\theta_{G2}$ denotes the parameters of generator $G_2$.

The STB will input the line drawings generated by the fusion fine-tuning module into discriminator $D_1$ to calculate the loss. Thus, the adversarial loss definition for STB is defined as (4):

$$L_{GAN-\mathrm{STB}}\left(G_1,D_1\right)=E_{I_o,gt}\left\{\log D_1\left(I_o,gt\right)\right\}+ E_{I_o}\left\{\log[1-D_1(I,G_f(G_1(I_o;\theta_{G_1}),G_2(I_g;\theta_{G_2});\theta_{G_f}))]\right\} \tag{4}$$

where $I_o$ and $gt$ are the labels of the cultural relics orthophotographs and line drawings respectively, while $\theta_{G1}$, $\theta_{G2}$, and $\theta_{Gf}$ represent the parameters of generators $G_1$, $G_2$, and $G_f$ respectively.

(2) Feature Matching Loss

This paper introduces a feature matching loss to guide line drawings generation, using the extracted intermediate features for loss computation. The feature matching loss is computed using the L1 distance, where multiple intermediate feature maps obtained from the discriminator are compared with the archaeology drawing labels, and the losses are calculated separately. After weighted summation, the average pixel-wise value computes the final feature matching loss. The definitions of feature matching losses for GSB and STB are defined as (5) and (6):

$$L_{FM-GSB}\left(G_2,D_2\right)= E_{(I_g,gt_g)}\sum_{k=1}^{m}\sum_{i=1}^{n}\frac{1}{N_i}\left\|D_k^{(i)}\left(I_g,gt_g\right)-D_k^{(i)}\left(I_g,G_2\left(I_g\right)\right)\right\|_1 \tag{5}$$

$$L_{FM-STB}\left(G_1,D_1\right)= E_{(I_o,gt)}\sum_{k=1}^{m}\sum_{i=1}^{n}\frac{1}{N_i}\left\|D_k^{(i)}\left(I_o,gt\right)-D_k^{(i)}\left(I_o,G_f\left(G_1\left(I_o\right),G_2\left(I_g\right)\right)\right)\right\|_1 \tag{6}$$

where $m$ represents the number of multi-scale discriminators, $n$ represents the number of discriminator layers, $k$ represents the number of multi-scale discriminators, and $N_i$ is the size product of the output feature map of each layer.

(3) VGG Loss

Similar to the concept of feature matching loss, the intermediate feature maps extracted by the VGG model are used to compute the loss with respect to the line drawing labels. The VGG model serves as a powerful and generic feature extractor, capable of objectively assessing the quality of line map results. By pre-training the VGG19 model [15] to compute the L1 loss, the losses from various layers are weighted and summed, followed by averaging over pixels to obtain the final VGG loss.

The definitions of VGG losses for GSB and STB are defined as (7) and (8):

$$L_{\mathrm{VGG}-GSB}\left(G_2,V\right)= E_{\left(I_g,gt_g\right)}\sum_{i=1}^{r}\frac{1}{N_i}\left\|V^{(i)}\left(I_g,gt_g\right)-V^{(i)}\left(I_g,G_2\left(I_g\right)\right)\right\|_1 \tag{7}$$

$$L_{\mathrm{VGG}-STB}\left(G_1,V\right)= E_{\left(I_o,gt\right)}\sum_{i=1}^{r}\frac{1}{N_i}\left\|V^{(i)}\left(I_o,gt\right)-V^{(i)}\left(I_o,G_f\left(G_1\left(I_o\right),G_2\left(I_g\right)\right)\right)\right\|_1 \tag{8}$$

where *r* represents the number of layers in VGG19, $N_i$ denotes the product of the dimensions of the output features for each layer, and V represents the pre-trained VGG19 model.

In summary, the loss function of the network model proposed in this paper is defined as (9):

$$L_{loss} = \lambda_1 L_{GAN-GSB} + \lambda_2 L_{GAN-STB} + \lambda_3 L_{FM-GSB} + \lambda_1' L_{FM-STB} + \lambda_2' L_{\mathrm{VGG}-GSB} + \lambda_3' L_{\mathrm{VGG}-STB} \tag{9}$$

where $\lambda$ represents the weight of each loss function, determining the degree of influence of the three losses on the network training.

## IV. Experimental Setup

### A. Dataset

In this paper, a portion of the cultural relics data is sourced from relevant institutions, while the remainder is extracted from publicly available publications. The completed paired dataset is depicted in Fig. 9. Additionally, certain orthophotographs are derived from 3D models provided by relevant units. Furthermore, some orthophotographs and lines are obtained from various sources including *Yungang Grottoes: Volume I* [18], *Research Report of the Institute of Human Sciences of Kyoto University: Yungang Grottoes* [19], and *Archaeological Report of Xumishan Grottoes: Yuanguangsi District* [20].

The dataset obtained, comprising orthophotographs and archaeology drawing labels, undergoes preprocessing and pairing to being fed into the network for training. The dataset production process in this paper is outlined as follows:

1. *Background Removal and Adjustment:* The background of the line drawing and irrelevant labels are eliminated, while brightness and contrast are adjusted to enhance line visibility.
2. *Height Alignment Adjustment:* Based on cultural relic images, the alignment between the line and picture content is adjusted to ensure coherence.
3. *Multi-scale Cropping and Resolution Adjustment:* Cultural relic images and line drawings are cropped based on the multi-scale concept, followed by resolution standardization to ensure uniformity.
4. *Data Augmentation:* Techniques such as rotation and symmetry are employed to further expand the dataset.

The dataset is then divided into training, validation, and test sets as outlined in Table I.

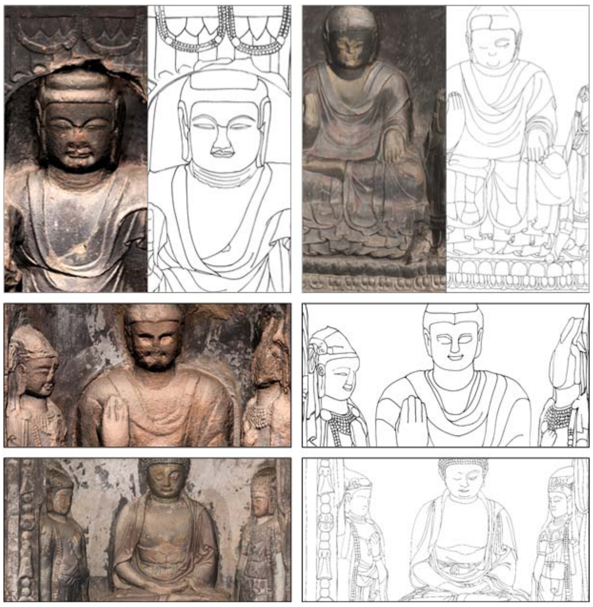

Fig. 9 Examples of archaeology drawing dataset

TABLE I
Statistics of the Dataset

| Dataset | Number of images |
|---|---|
| Training | 1536 |
| Validation | 240 |
| Test | 240 |
| Total | 2016 |

### B. Evaluation Indexes

Archaeology drawings are grayscale images composed of white backgrounds and black lines. Taking into account commonly used image quality assessment indexes as well as the specific characteristics of archaeology drawings, we selected three metrics — PA, SSIM, and PSNR — to establish a comprehensive evaluation framework for archaeology drawings.

The pixel accuracy is defined as (10):

$$PA=\frac{1}{h\times w}\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}\left\|x_{i,j}\odot y_{i,j}\right\|_1 \tag{10}$$

where *x* and *y* represent the two images being compared, *i* and *j* denote the indices of pixels, *h* and *w* are the numbers of pixels in the image's height and width directions, respectively.

The SSIM is defined as (11):

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x{}^2+\mu_y{}^2+C_1)(\sigma_x{}^2+\sigma_y{}^2+C_2)} \tag{11}$$

where $\mu_x$ represents the mean of *x*, $\sigma_x$ represents the standard deviation of *x*, $\sigma_{xy}$ represents the covariance of *x* and *y*. $C_1$ and $C_2$ are constants to prevent denominators from being zero.

Considering that the backgrounds of archaeology drawings are predominantly pure white, resulting in similar brightness attributes across different line drawings, the evaluation ability of SSIM among line drawings generated by different models is not distinct enough, making it unsuitable as a direct evaluation index for archaeology drawings. Therefore, in this paper, we adjusted the calculation of SSIM by removing the brightness attribute, defining a new evaluation index termed SSIM-woL (SSIM without luminance), as shown in (12):

$$SSIM - woL(x,y) = \frac{(2\sigma_{xy} + C_2)}{(\sigma_x{}^2 + \sigma_y{}^2 + C_2)} \tag{12}$$

The peak signal-to-noise ratio formula is defined as (13):

$$PSNR = 10\log_{10}\left(\frac{MAX_I{}^2}{MSE}\right) \tag{13}$$

where $MAX_I$ represents the maximum pixel value of the image, which is typically 255 in the case of 8-bit grayscale images, and *MSE* stands for mean square error.

### C. Experimental Settings

The experiments were conducted on a deep learning platform equipped with an NVIDIA GeForce RTX 3090Ti GPU (24GB). The server system runs on the Ubuntu 18.04 LTS operating system, and the network model was built using the PyTorch 1.12 deep learning framework.

The method proposed in this paper, referred to as HD-dual, was trained using a dataset of archaeology drawings from the grotto's temple, comprising 2016 pairs. During the model training phase, the network batch size was set to 1, and instance normalization was applied. The Adam optimization algorithm [21] was employed to iteratively update the generators and discriminators of the two branches. The initial learning rate (*lr*) was set to 0.0002, with a momentum decay factor ($\beta_1$) of 0.5 and an infinite norm decay factor ($\beta_2$) of 0.999. In the experiments, the model was initially trained for 150 epochs with the specified learning rate and then further trained for 100 epochs with linearly decaying learning rates. The training process was monitored using the TensorBoard visualization tool.

## V. Result and Discussion

### A. Comparison with Other Methods

To evaluate the line drawing generation performance of the proposed module in this paper, we conducted comparative experiments with mainstream image generation models, including HED, BDCN, pix2pix, and pix2pixHD [14]. Comparative experiments were performed on a self-built multi-scale archaeological line graph dataset, with experimental parameters set consistent with the "Experimental Setup" section.

Fig. 10 presents a comparison of the results of line drawing generation from five models. From the figure, it can be observed that the lines generated by HED are thick and blurry, resulting in significant overlapping in areas with dense anomalies and noise. The lines generated by BDCN are thinner compared to HED, but discontinuities in the lines are prominent, leading to the loss of numerous detailed features. Pix2pix partially addresses the issues of high noise and discontinuous lines; however, it introduces many randomly scattered false lines, resulting in poor line accuracy. Pix2pixHD mitigates some of the false line issues compared to pix2pix but still exhibits a noticeable gap from the ground truth labels. In contrast, the proposed method demonstrates improvements in noise reduction and line continuity, approaching closer to the ground truth labels. Overall, the proposed method presents optimal performance.

The results of the line drawing generation from the five methods were quantitatively evaluated using the evaluation indexes listed in Table II. Analysis of the results in the table reveals that the proposed method achieved the best results across all evaluation metrics. In terms of PA, HD-dual improved by 5.9% compared to pix2pix and by 2.5% compared to pix2pixHD, indicating a significant enhancement in the accuracy of line drawing generation using our method. Regarding SSIM-woL, HD-dual showed increase of 16.7% and 0.8% compared to pix2pix and pix2pixHD, respectively, demonstrating a higher similarity between the archaeology drawings generated by HD-dual and the ground truth. Furthermore, the PSNR increased by 8.4% and 2.3% compared to pix2pix and pix2pixHD, respectively, indicating a stronger noise suppression ability of HD-dual. In conclusion, the line drawing generation performance of our method surpasses that of the current mainstream image generation models.

TABLE II
Comparison of Evaluation Indexes for Generating Archaeology Drawings by 5 Models Above

| Network | PA | SSIM-woL | PSNR |
|---|---|---|---|
| HED | 70.67% | 48.90% | 10.70 |
| BDCN | 81.64% | 54.89% | 11.94 |
| Pix2pix | 83.29% | 61.49% | 12.89 |
| Pix2pixHD | 86.04% | 71.20% | 13.66 |
| HD-dual | 88.21% | 71.78% | 13.97 |

### B. Effect of Feature Transmission Module

To evaluate the effect of the feature transmit module in HD-dual, the performance of the HD-dual without the feature transmit module (referred to as HD-dual-Notrans) was tested on the self-constructed archaeology drawings dataset. The influence of the absence of the feature transmission module on the output feature maps of GSB and STB was analyzed.

(1) Influence on GSB Output Feature Map

Fig. 11 presents the GSB output of HD-dual-NoTrans and HD-dual. Comparing Figs. 11 (a) and (b), it is evident that the GSB output feature map of HD-dual-Notrans exhibits many line breakpoints and lacks smoothness. For instance, the line texture at the skirt of Fig. 11 (a) is not clearly defined, and the outline on the left side of the figure lacks key edges.

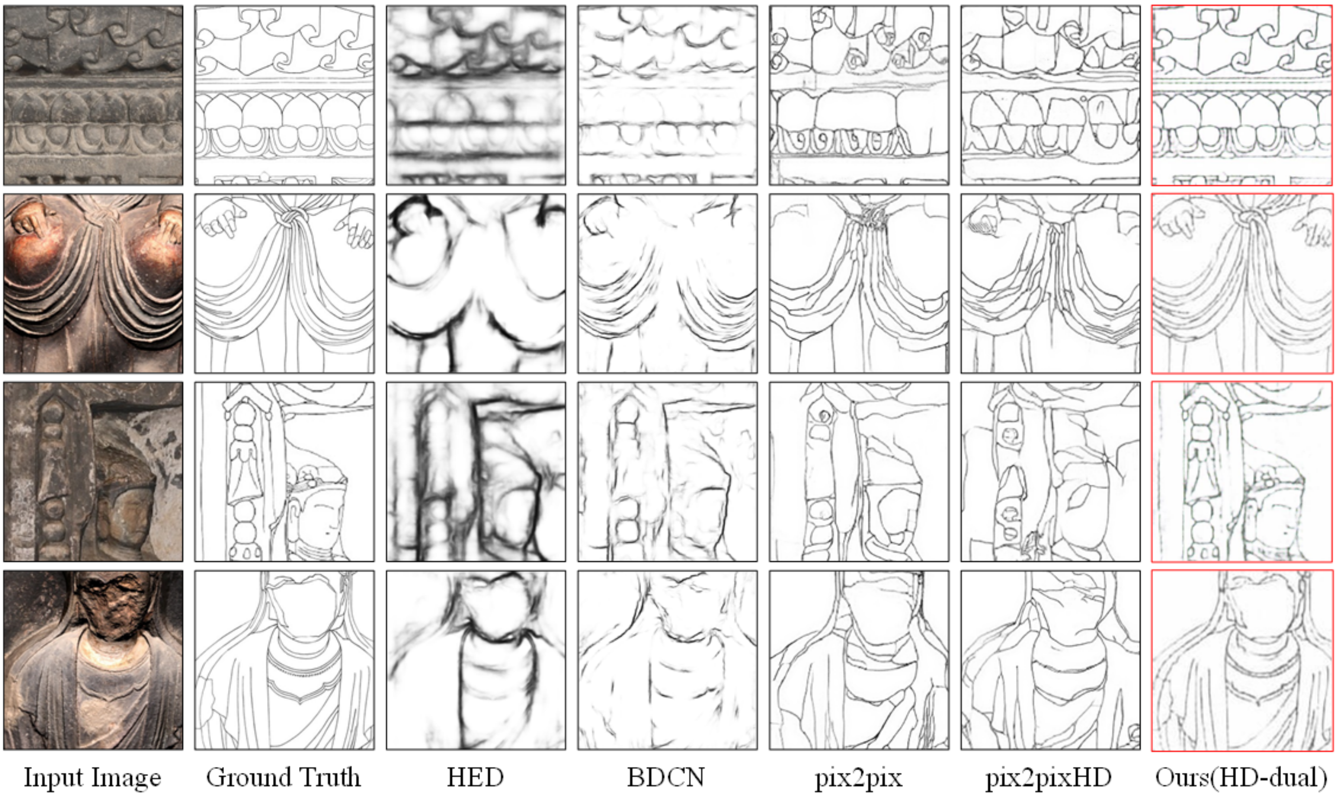

Fig. 10 Comparison of HD-dual and traditional models to generate line drawings

Due to the fact that the GSB of HD-dual-NoTrans only learns the mapping relationship between two gradient domains, it focuses more on edge information while lacking semantic information, resulting in incomplete features in the generated line drawing. By comparison, the output of the GSB in HD-dual in Fig. 11 (c) exhibits smoother lines with almost no noise. With the assistance of semantic information from the STB, it overcomes problems such as discontinuities and uneven lines, thereby generating higher-quality gradient lines. In summary, introducing feature transmission can enhance the output image quality of GSB.

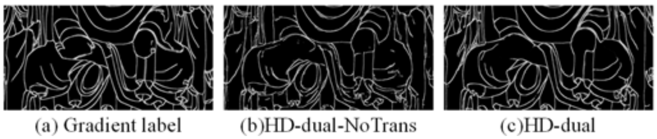

Fig. 11 Effect of feature transmit module on GSB output

(2) Influence on STB Output Feature Map

The feature transmit module may also affect the output feature maps of the STB because part of the features of the GSB generator $G_2$ comes from the STB generator $G_1$, and during gradient backward propagation, $G_1$ undergoes parameter optimization as well. To find the impact of the transmit module on the output feature maps of the STB, we compared the STB outputs of HD-dual and HD-dual-NoTrans, as shown in Fig. 12.

The STB output feature maps of HD-dual-NoTrans appear more blurred, with less distinct expression of semantic information, whereas the STB semantic feature maps of HD-dual exhibit better quality (Fig. 12 (c)), with clearer garment textures and an improvement in the extraction of different features.

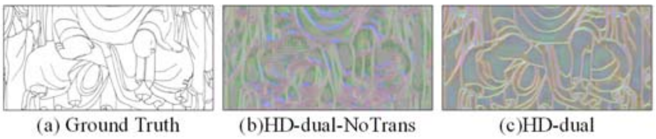

Fig. 12 Effect of feature transmit module on STB output

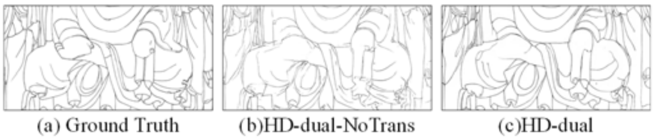

Fig. 13 Effect of the feature transmit module on the generated line drawings

Additionally, we compared the generated results of HD-dual and HD-dual-NoTrans line drawings. As depicted in Fig. 13, the lines of HD-dual-NoTrans appear messy and contain more noise, while the lines of HD-dual are clearer with fewer breakpoints. For instance, as illustrated in Table III, HD-dual exhibits 3.3% increase in PA, 1.3% increase in SSIM-woL, and 6.3% increase in PSNR compared to HD-dual-NoTrans. These findings further evaluate the effectiveness of the feature transmit module and affirm that the semantic features from STB complement the gradient information from GSB, thereby facilitating HD-dual in generating line drawings with superior quality.

TABLE III
COMPARISON OF EVALUATION INDEXES FOR ABLATION EXPERIMENTS OF FEATURE TRANSMIT MODULE

| Network | PA | SSIM-woL | PSNR |
|---|---|---|---|
| HD-dual-NoTrans | 85.36% | 70.86% | 13.1385 |
| HD-dual | 88.21% | 71.78% | 13.9742 |
| Improvement | 3.3% | 1.3% | 6.3% |

### C. *Effect of Attention Module*

In order to evaluate the impact of attention mechanisms on the quality of line drawing generation, we incorporated the CBAM attention module into the generator, discriminator, and fusion fine-tuning module of the HD-dual model. The generated line drawings were then compared with those generated by the HD-dual-NoAtt model, which does not include attention module, as shown in Fig. 14.

After incorporating CBAM, the contours of the line drawings are closer to the labels. The positions of facial lines and robe texture lines on the Buddha statue are more accurate, and there are fewer erroneous lines in the damaged areas. This suggests that the influence of cultural relic damage on line drawing generation is minimized with the use of attention mechanisms. Additionally, attention mechanisms increase the weight allocation for features such as contour edges, leading to significant improvements in the generation of these details. This effectively evaluates the effectiveness of the attention module.

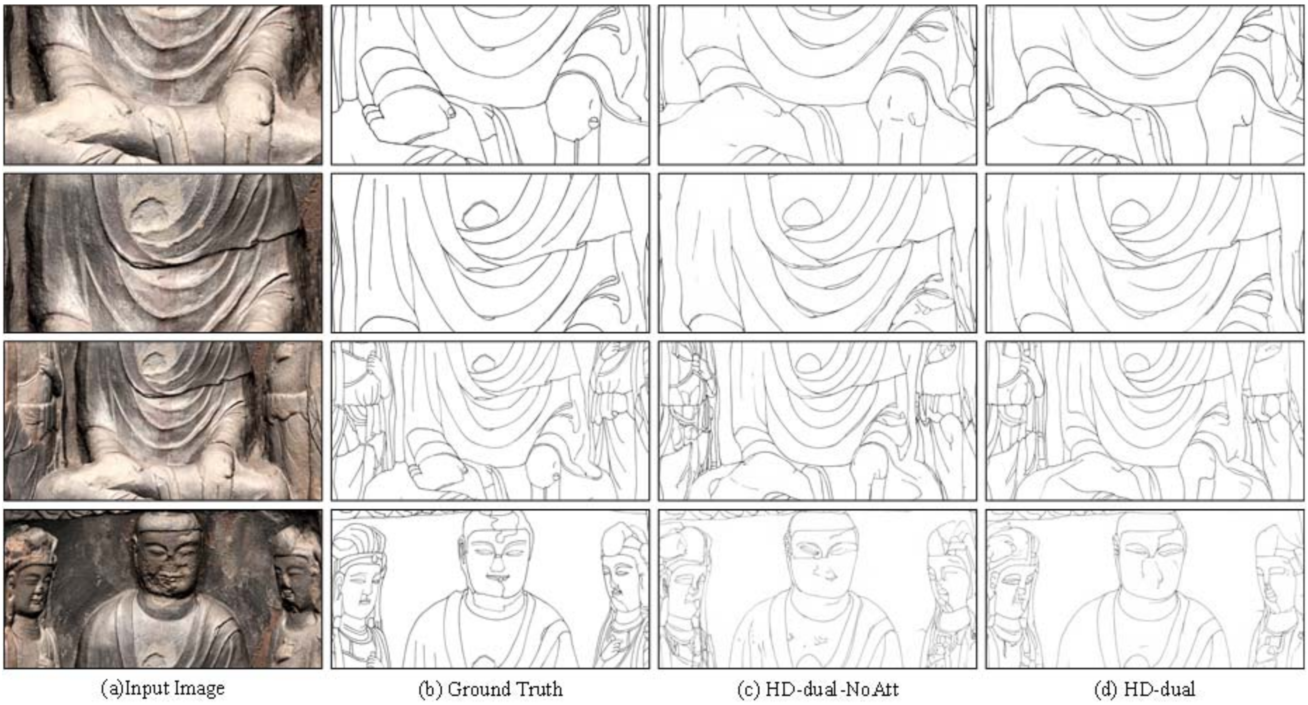

Fig. 14 Comparison of generated line drawings before and after optimization of CBAM

TABLE IV
COMPARISON OF EVALUATION INDEXES OF GENERATED LINE DRAWINGS BEFORE AND AFTER OPTIMIZATION OF CBAM

| Network | PA | SSIM-woL | PSNR |
|---|---|---|---|
| HD-dual-NoAtt | 88.21% | 71.78% | 13.97 |
| HD-dual | 88.64% | 71.86% | 14.10 |
| Improvement | 0.5% | 0.1% | 0.9% |

The quantitative indicators for the generated line drawings before and after incorporating attention mechanisms are shown in Table IV. PA, SSIM-woL, and PSNR, show improvements, indicating that the HD-dual model integrated with CBAM performs better in archaeology drawing generation tasks and can generate higher-quality archaeological line graphs.

### D. *Generalization Verification*

In order to assess the generalization of the proposed method on different archaeology drawings datasets, we conducted generalization verification on a self-constructed dataset of Dunhuang mural archaeological line drawings.

TABLE V
COMPARISON OF EVALUATION INDEXES OF DUNHUANG MURAL LINE DRAWINGS GENERATED BY THE MODELS ABOVE

| Network | PA | SSIM-woL | PSNR |
|---|---|---|---|
| Pix2pixHD | 88.81% | 75.08% | 12.60 |
| HD-dual | 88.97% | 75.69% | 12.66 |
| Improvement | 0.2% | 0.8% | 0.5% |

The comparison between the line drawings generated by HD-dual and pix2pixHD is illustrated in Fig. 15. It can be observed that the lines generated by HD-dual are more continuous with fewer noise artifacts, and the hand ring and fingers are clearer. This demonstrates that HD-dual can better capture edge information at different scales, enriching the local details. The quantitative comparison indexes are presented in Table V, indicating improvements across all three indexes. This confirms that the proposed method exhibits a certain degree of generalization capability.

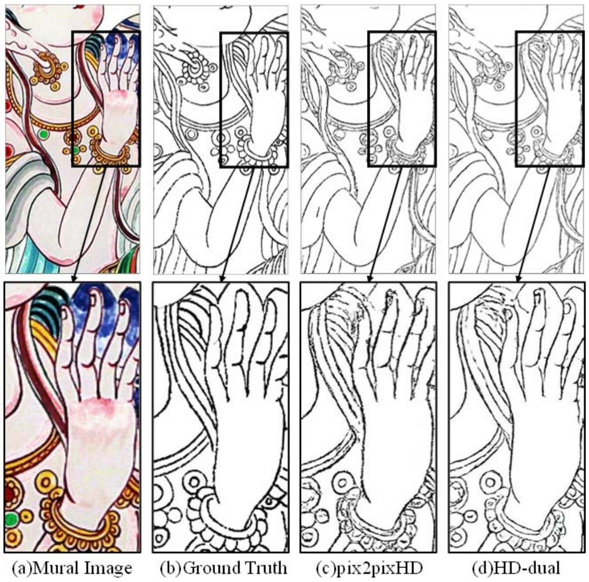

Fig. 15 Comparison of models generating Dunhuang mural line drawings

## VI. Conclusion

Archaeology drawings serve as integral components of archaeological reports and play a vital role in cultural relics preservation. The application of deep learning models in archaeology drawings generation significantly alleviates the workload burden on archaeologists. In order to address challenges such as low resolution in line drawings produced by current mainstream algorithms and susceptibility to cultural relic diseases, this paper proposes a high-resolution image generation algorithm for archaeological line drawings. Based on CGAN, we present a dual-branch archaeology drawings generation model comprising semantic translation and GSBes, each designed to extract distinct features. Further enhancement of line graph generation quality is achieved through feature transmission and attention modules.

We constructed an archaeology drawing dataset of grotto temples based on publicly available publications and data from collaborating institutions, and further expanded this dataset using multi-scale cropping and data augmentation strategies. Experimental comparisons were conducted between our proposed model and mainstream image generation models on self-constructed dataset to validate the line drawing generation performance. The results show that the PA, SSIM-woL, and PSNR of the line drawings generated by the proposed model are improved, reaching 88.21%, 71.78%, and 13.97 dB, respectively. Additionally, we also designed ablation experiments to independently verify the functionality of the feature transmit module and the attention module. The results demonstrate superior quality in line drawing generation by proposed model compared to the comparison model lacking these modules, proving their effectiveness. The application of our method to the Dunhuang mural dataset has also achieved good results, which verifies that the model has a certain generalization.

Our method facilitates the automatic generation of archaeology drawings, thereby aiding archaeological workers in line drawing tasks. However, the proposed method has a large number of parameters, and the training period is somewhat lengthy and challenging. Additionally, to integrate with archaeological practices, it is necessary to design a suitable vectorization algorithm to be applied to the line drawings generated by the model, so that the archaeological experts can easily edit and modify the lines in sections.

## Acknowledgment

Financial support of this work from the Shenzhen Stable Support Natural Fund (Grant No. WDZC20200821140447001).

## References

[1] Ma H Z. Field Archaeological Mapping (M). Beijing: Peking University Press, 2010: 10-14.
[2] Ma H Z. Drawing of Archaeological Artifacts (M). Beijing: Peking University Press, 2008: 1-4.
[3] Li X F. Research and Implementation of Line Graph Generation and Measurement Method of 3D Model of Cultural Relic (D). Northwest University, 2013.
[4] Wang X, Geng G, Li X, et al. A cultural relic line drawings generation algorithm based on explicit ridge line,2015 International Conference on Virtual Reality and Visualization (ICVRV),2015.
[5] Liu F J. Generation and Rendering of Non-photorealistic Line Graph based on Two-dimensional Images (D). Northwest University, 2013.
[6] Song Q N. Research and implementation of Cultural Relics Line Drawing method based on Cellular Ant Colony Edge Extraction (D). Northwest University, 2019.
[7] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition (J). Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[8] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks (J). Communications of the ACM, 2020, 63(11): 139-144.
[9] Mirza, S Osindero. Conditional Generative Adversarial Nets 2014. arXiv preprint 2014 (J), arXiv:1411.1784.
[10] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks (C)//Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2017: 1125-1134.
[11] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation (C)//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.
[12] Peng J, Wang J, Wang J, et al. A relic sketch extraction framework based on detail-aware hierarchical deep network. Signal Processing, 2021, Vol.183.
[13] He J, Zhang S, Yang M, et al. Bi-directional cascade network for perceptual edge detection (C)//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 3828-3837.
[14] Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional GANs (C)//Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2018: 8798-8807.
[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition (J). International Conference on Learning Representations, 2015: 1-14.
[16] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module (C)//Proceedings of the European conference on computer vision. 2018: 3-19.
[17] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors (J). nature, 1986, 323(6088): 533-536.
[18] Zhang Z, Wang H, Zhao K Y. Yungang Grottoes: Volume I (M).

Shandong: Qingdao Publishing House, 2017: 1-331.

[19] Mizuno S, Nagahiro T. Yun-Kang: the Buddhist cave-temples of the fifth century A.D. in North China: Detailed report of the archaeological survey carried out by the Mission of the Tōhōbunka kenkyūsho 1938-45. Volume X Caves Thirteen and Outside Wall of Caves XI-XIII (M). Kyoto: Kyoto University Yungang Press, 1953: 1-45.

[20] Ningxia Institute of Cultural Relics and Archaeology, Zhejiang University Cultural Heritage Institute, Xumu Mountain Grottoes Cultural Relics Management Institute. Archaeological Report of Xumishan Grottoes: Yuanguangsi District (M). Beijing: Cultural Relics Press, 2020: 1-130.

[21] Kingma D P, Ba J. Adam: A method for stochastic optimization (J). arXiv preprint 2014. arXiv:1412.6980.