Journal Pre-proof

AGD-GAN: Adaptive Gradient-Guided and Depth-supervised generative adversarial networks for ancient mural sketch extraction

Zhe Yu, Shenglin Peng, Shuyi Qu, Qunxi Zhang, Jun Wang, Jinye Peng

PII: S0957-4174(24)01506-9
DOI: https://doi.org/10.1016/j.eswa.2024.124639
Reference: ESWA 124639

To appear in: *Expert Systems With Applications*

Received date : 16 January 2024
Revised date : 28 April 2024
Accepted date : 25 June 2024

Please cite this article as: Z. Yu, S. Peng, S. Qu et al., AGD-GAN: Adaptive Gradient-Guided and Depth-supervised generative adversarial networks for ancient mural sketch extraction. *Expert Systems With Applications* (2024), doi: https://doi.org/10.1016/j.eswa.2024.124639.

Revised manuscript (with changes marked)

AGD-GAN: Adaptive Gradient-Guided and Depth-Supervised Generative Adversarial Networks for Ancient Mural Sketch Extraction

ZheYu[a], Shenglin Peng[a], Shuyi Qu[a], Qunxi Zhang[b], Jun Wang[a,c,*] and Jinye Peng[a,c,*]

[a]Northwest University, No.1, Xuefu Avenue, Chang'an District, Xi'an, Shaanxi Province, P.R. China
[b]Shaanxi History Museum, No. 91, Xiaozhai East Road, Yanta District, Xi'an, Shaanxi Province, P.R. China
[c]Shaanxi Province Silk Road Digital Protection and Inheritance of Cultural Heritage Collaborative Innovation Center, No.1, Xuefu Avenue, Chang'an District, Xi'an, Shaanxi Province, P.R. China

ARTICLE INFO

Keywords:
Sketch extraction
Deep learning
Image translation
edge extraction
Mural digital protection

ABSTRACT

To address the overlooked issues of multi-scale detail feature extraction and disease noise suppression in mural sketch extraction, we proposed a novel generative adversarial network with Adaptive Gradient-guided and Depth-supervised (AGD-GAN), which can generate high-quality mural sketches in an unsupervised manner. AGD-GAN first enriches mural feature details at various scales by introducing a cross-channel residual attention module, significantly improving the detail extraction effects. The Adaptive Gradient-guided strategy is based on the gradient attention maps, which can adaptively adjust the weights between gradient information and detail feature preservation according to the degree of damage in the mural, further balancing the preservation of mural detail features and the suppression of disease noise. Finally, the Depth-supervised reinforces constraints on the position and shape of the sketches by introducing a depth-predicted loss function, thereby reducing background noise interference and controlling the shape of the generated sketches. We compared with eight state-of-the-art algorithms quantitatively and qualitatively, experimental results demonstrate the promising capability of the proposed AGD-GAN to extract clear, coherent, and comprehensive mural sketches. We release the source code at https://github.com/Alice77bai/AGD-GAN.

1. Inroduction

Sketch has the great function of creating artistic beauty. It is the most fundamental visual language in Chinese painting. Sketch primarily employs lines to outline the contours of objects, which not only shows the shape of the object but also presents its posture (Han et al., 2015). Sketch finds extensive application in colored mural paintings, such as the Dongping colored murals from the Han Dynasty, in which the lifelike portrayal of characters is inseparable from the use of lines. Therefore, the distinctive artistic value of sketches holds significant guiding implications for mural copying and archaeological research.

In the past few decades, mural sketch copying mainly relied on manual replication due to its extensive history and on-site limitations, which is not only extremely inefficient but also causes secondary damage to cultural relics during processing (Delaney et al., 2017; Fu et al., 2017; Zhou et al., 2022). Simultaneously, numerous unearthed murals have suffered from various degrees of disease such as pigment peeling and pattern missing, making it impossible to present the original appearance of the murals. Consequently, computer-aided sketch extraction plays a pivotal role in the realm of digital heritage protection (Liu et al., 2006; He et al., 2013; Jiang et al., 2022).

In recent years, the digital protection of cultural heritage has attracted the attention of computer vision researchers. Several studies have successfully employed edge detection techniques to generate sketches from digital mural paintings. Conventional methods (Kittler, 1983; Lim et al., 2013; Martin et al., 2004a) for image edge detection rely on low-level local cues, such as color, gradient, and texture. Benefiting from the superiority of Convolutional Neural Networks (CNNs) in learning high-level features, various deep learning-based edge detection techniques have been significantly developed (Bertasius et al., 2015b; Kokkinos, 2015). As the network becomes deeper and the receptive field expands, CNN inevitably loses many important details while gradually capturing global and semantic features. To compensate for the low-level details lost during the training process of deep networks, methods in (He et al., 2019; Liu et al., 2017; Liu & Lew, 2016; Xie & Tu, 2015; Xu et al., 2017) aggregate features from both deep and shallow layers. However, these shallow features mainly focus on local intensity variations and may not capture the larger context or semantic information effectively, resulting in the generated edges containing noise or blur.

To our knowledge, extracting sketches from mural scenes presents a significant challenge for several reasons: 1) The varial style of mural. current sketch extraction techniques are mainly used in natural images, while ancient murals have specific artistic styles and cultural backgrounds. Existing extraction techniques frequently struggle to reproduce precise artistic sketches without taking different stylistic of murals into consideration. 2) Diseases of murals. Influenced by the environment, many unearthed murals exhibit varying degrees of damage. It is more difficult to extract refined sketches from diseased murals. 3) Scarcity of Data. Due to the scarcity of ancient mural datasets, some deep learning methods trained on large paired datasets may not be perfectly suited for the

*Corresponding author.
yuzhe77@stumail.nwu.edu.cn (ZheYu); pengshenglin@nwu.edu.cn (Shenglin Peng); syqu@nwu.edu.cn (Shuyi Qu); d202110716@xs.ustb.edu.cn (Qunxi Zhang); jwang@nwu.edu.cn (Jun Wang); pjyxida@nwu.edu.cn (Jinye Peng)

extraction of ancient mural sketches.

To address the above-mentioned issues, this study draws inspiration from the concept of image-to-image translation (Richardson et al., 2021; Alotaibi, 2020; Isola et al., 2017a), considering murals and sketches as distinct domains with varying stylistic characteristics. Our goal is to achieve style transfer between the two domains using an unsupervised approach to alleviate the need for large paired datasets. In this work, we develop an adaptive gradient-guided and depth-supervised generative adversarial network (AGD-GAN) to improve the efficiency of sketch extraction from ancient murals in an unsupervised training manner. To fully learn the unique detail information in murals, we first incorporate the cross-channel attention residual module to enhance the mining of fine-grained detail information, thereby enriching the features of murals. Then, to effectively alleviate the limitations in existing methods when applied to diseased murals, we proposed an adaptive guidance strategy, which employed learnable parameters and guided by gradient attention maps to different treat mural feature maps affected by disease and those that are not. Finally, we employ a depth prediction supervision loss to effectively generate sketches at the correct spatial locations, compensating for the lack of constraints on the geometric shapes of input mural patterns in existing unsupervised networks. The proposed AGD-GAN is expected to extract clear, coherent, and clean sketches of ancient murals. In summary, our main contributions can be summarized as follows:

(1) We designed a novel AGD-GAN model for ancient mural sketch extraction, which is trainable with unpaired sketches and mural images in an unsupervised manner. It not only incorporates the distinctive mural style into the sketch extraction process, effectively tackling the issue of limited data for ancient murals, but it also demonstrates applicability to damaged mural images.

(2) A cross-channel attention residual module is introduced to better exploit the intrinsic features of mural images via multi-scale spatial information, cross-channel attention, and split skip connections. This module can adeptly extract multi-scale spatial information at a granular level. Experimental results prove their effectiveness for extracting intricate details from mural feature maps.

(3) We propose an adaptive guidance strategy based on gradient attention maps to effectively reduce background noise in diseased murals while retaining crucial details in non-diseased murals. This strategy allows for learnable parameter adjustments specifically tailored to both diseased and disease-free murals.

(4) We introduce a depth prediction loss function to impose constraints on the geometric shape and position of the sketch. Given the absence of direct geometric constraints in conventional GAN networks, we focus on predicting depth maps from murals and sketches that can highlight the foreground area (i.e., the geometry of the mural pattern), which aims to eliminate partial background noise and ensure sketches within the correct geometric boundaries.

2. Related Work

2.1. Edge Detection

Early research mainly focused on conventional edge detection operators. These operators derived edges from low-level local cues like intensity, color, or texture discontinuities (Roberts, 1963; Prewitt et al., 1970; Canny, 1986), which have shown effectiveness in detecting edges in natural images and facial outlines. However, their reliance solely on gradient calculations renders them susceptible to extracting mural image diseases while potentially overlooking weak lines. Consequently, when applied to mural images, issues such as background noise and discontinuous lines tend to manifest. Subsequent approaches design hand-crafted features derived from attributes such as intensity, gradient, or texture, which are combined with learning strategies to classify edge and non-edge pixels through sophisticated paradigms (Dollar et al., 2006; Hallman & Fowlkes, 2015; Zhang et al., 2016; Samma et al., 2019; Arbelaez et al., 2010; Zitnick & Dollár, 2014). Although significant advancements in hand-crafted approaches, many encounter issues such as artifacts, color degradation, and edge blurring when applied to murals. Moreover, directly employing the same hand-crafted features in various murals proves challenging due to the random and intricate distribution of mural diseases.

In recent years, deep convolutional neural networks (CNNs) have shown powerful advantages in various computer vision tasks including edge detection. The early edge detection algorithm used CNN architectures as the basic network structure to predict the edge probability of the input image by constructing a classifier (He et al., 2019; Liu et al., 2017; Xie & Tu, 2015). Bertasius et al. (Bertasius et al., 2015a) proposed DeepEdge using object-related features as high-level cues for contour detection. Shen et al. (Shen et al., 2015) introduced the positive-sharing loss into the DeepContour model, enabling each contour subclass to share the loss of the entire positive class, thereby facilitating more efficient parameter learning. Su et al. (Su et al., 2021) proposed pixel differential convolution, which integrates traditional edge detection operators into general CNN convolution operations, resulting in robust and accurate edge detection. While adept at extracting blurred edges in natural images through feature learning, these methods still face limitations when applied to processing mural images. Deep learning algorithms demand extensive labeled data for training, presenting difficulties in obtaining diverse mural datasets. Moreover, the distinct texture, color, and intricate details of murals hinder accurate edge information extraction when directly employing deep learning models.

2.2. Image-to-Image Translation

Image-to-image translation, a domain in computer vision and graphics, focuses on learning the mapping relationship between input and output images from paired or unpaired datasets. Supervised methods demand paired data for training the model, while unsupervised methods can be trained with unpaired data. Paired data, particularly authentic sketches drawn by artists, holds significant value but is often rare.

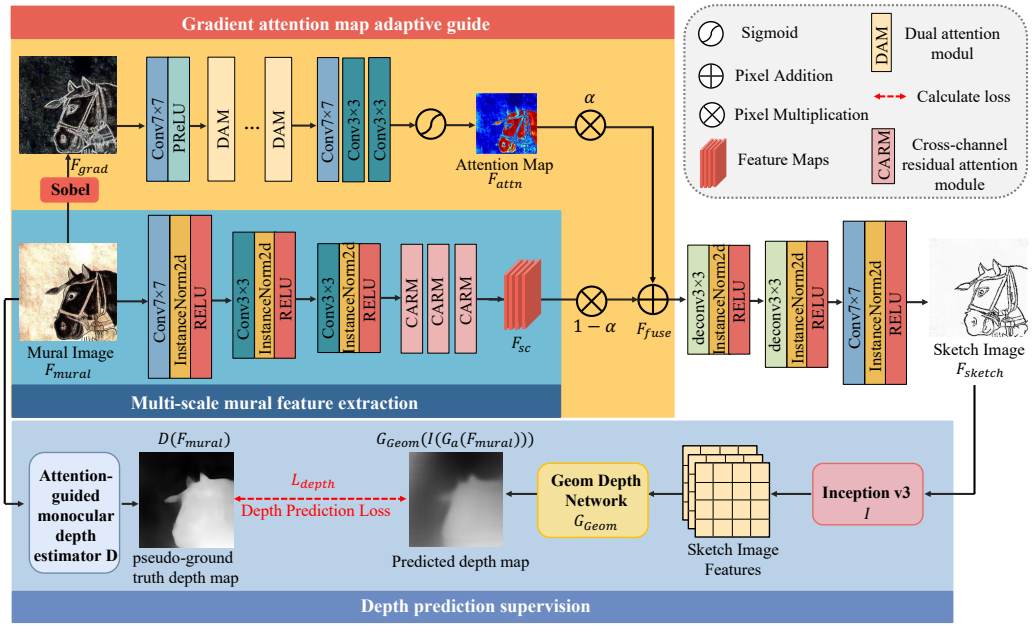


Fig. 1: Overview of our proposed AGD-GAN structure in mural sketch extraction problem.

Hence, unsupervised methods provide a more convenient approach to gathering datasets. This paper aligns with the unsupervised methods, as the conversion of ancient murals to sketch images lacks extensive datasets, making it unsuitable for supervised methods.

Zhu et al. (Zhu et al., 2017) introduced the CycleGAN framework, which aims to learn a mapping between different domains by leveraging a cycle-consistent constraint that ensures the consistency between the generated images and the original images. However, due to the constraints of unsupervised training, CycleGAN may not guarantee a complete match between the generated images and the real images particularly in terms of capturing fine-grained details and preserving semantic consistency. Kim et al. (Kim et al., 2019) propose a novel method for unsupervised image-to-image translation, which incorporates a new attention module and a new learnable normalization function in an end-to-end manner. Eskandar et al. (Eskandar et al., 2022) propose a new Unsupervised paradigm that makes use of a self-supervised segmentation loss and whole image wavelet-based discrimination. Yang et al. (Yang et al., 2022) present a generative prior-guided unsupervised image-to-image translation framework (GP-UNIT) employing a coarse-to-fine scheme for improved multi-level content translation quality and versatility. Chan et al. (Chan et al., 2022) present an unpaired method for generating sketches from photos, which maximizes the utilization of geometric and semantic information of lines, consequently reducing the reliance of Generative Adversarial Networks (GANs) on cycle-consistency loss. Park et al. (Park et al., 2023) present a LANguage-driven Image-to-image Translation model, dubbed LANIT, which addressed limitations in existing techniques by leveraging text-provided attributes for per-sample domain labels. Zhao et al. (Zhao et al., 2020) attempted to move away from the constraints of cycle consistency loss by proposing an adversarial consistency loss. This loss encourages the translated images to retain essential features of the source images. Torbunov et al. (Torbunov et al., 2023) enhanced the performance of the generator by equipping it with a Vision Transformer, surpassing more contemporary models without relaxing the cycle consistency constraints. Ashtari et al. (Ashtari et al., 2022) proposed a model that extracts a sketch from a colorized image in such a way that the extracted sketch has a line style similar to a given reference sketch while preserving the visual content identically to the colorized image. While all of the above methods have achieved success in various translation tasks, they fail to address the issues of disease noise and detail loss that occur when translating ancient murals into sketches.

3. Proposed method

3.1. Network Architecture

The pipeline of our proposed approach is depicted in Fig. 1. AGD-GAN is constructed upon the foundation of an encoder-decoder structure: multi-scale mural feature extraction, gradient attention map adaptive guided, and depth prediction supervision. For the discriminator, we employ the PatchGAN (Isola et al., 2017b) architecture with a 70×70 receptive field, the code of which is accessible for academic use under its license.

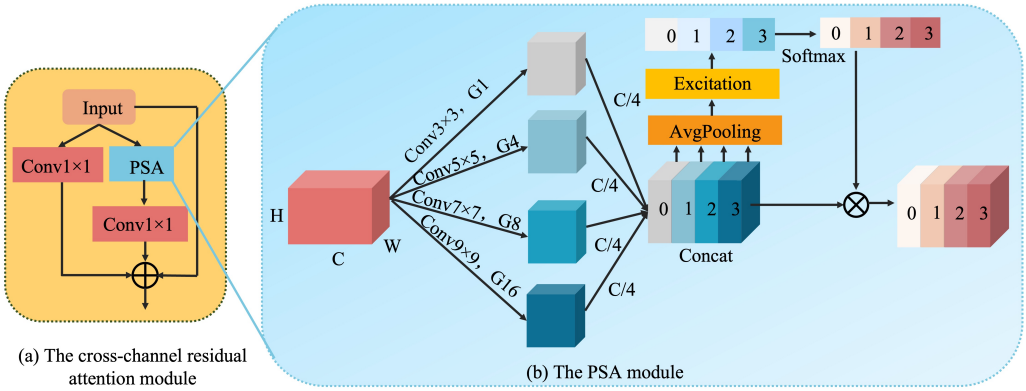


Fig. 2: Details of the proposed cross-channel attention residual model. (a) The CARM devised a shunt residual structure and efficiently embedded the PSA module, which encompasses two key branches: multiscale feature extraction and feature preservation. (b) PSA employs a multi-scale pyramid convolution structure to handle input tensors of multiple scales and construct interactions among different dimensions, which is beneficial for extracting multi-scale information from mural images.

3.2. Multi-scale mural feature extraction

When provided with an input mural image F_{mural}, our initial step involves converting it into a 64-channel representation using a 7×7 convolution operation to obtain the initial mural feature map

$$F_i = H_0(F_{mural}) \tag{1}$$

where $H_0(\bullet)$ represents a 7×7 convolutional operation. Subsequently, the mural feature map undergoes two downsampling operations, denoted as $H_{down}(\bullet)$, reducing its dimensions to one-fourth of the original size while quadrupling the number of channels

$$F_s = H_{down}(F_i) \tag{2}$$

Benefiting from the pyramid squeeze attention module (PSA (Zhang et al., 2022)), which employs a multi-scale pyramid convolution structure to handle input tensors of multiple scales and construct interactions among different dimensions, greatly aiding in the extraction of multi-scale information from mural images, as shown in Fig. 2(b). We incorporate the PSA module into ordinary residual blocks to maximize the learning of multi-scale information and fine-grained details within mural images, such as larger-scale figures and smaller-scale costume decorations. However, directly stacking multiple PSA modules can lead to overly abstract feature maps as the network deepens, resulting in the loss of many meaningful multi-scale detailed features. Considering this, we devised a shunt residual structure and efficiently embedded the PSA module to better mine the inherent characteristics of mural images while preserving multi-scale detailed features.

As depicted in Fig. 2(a), it is the cross-channel residual attention module (CARM) designed in this paper. The CARM encompasses two key branches: multiscale feature extraction and feature preservation. When given an input feature map, the multiscale feature extraction branch initially employs the PSA module to extract multiscale spatial information while concurrently capturing long-range channel feature dependencies. The feature preservation branch utilizes a 1x1 convolution to compress the features obtained from the preceding CARM module. Ultimately, the outputs from both branches are connected with the input feature map through a residual connection, which can prevent essential details from being lost as the network deepens. After performing calculations with the three cross-channel attention residual modules, the resulting feature map contains multi-scale spatial information and cross-channel attention

$$F_{sc} = H_{CARM}(F_s) \tag{3}$$

where $H_{CARM}(\bullet)$ represent three stacked cross-channel attention residual module.

3.3. Gradient attention map adaptive guide

We consider that directly upsampling the feature map of murals containing fine-grained detail information may not only introduce considerable background noise but also lose many precious details. According to our observations, gradient maps can significantly enhance the structured representation of edges, which is highly beneficial for highlighting patterned areas of murals. Theoretically, training attention mechanisms on mural gradient maps is more effective than training attention mechanisms directly on mural images. Therefore, before upsampling, we hope to enhance the detailed information by leveraging the attention of gradient maps while also guiding the feature learning for diseased and non-diseased murals. We employ the conventional gradient operator (Kittler, 1983) as a preprocessing step in gradient computing. The gradient map of the mural image is calculated via

$$F_{grad} = H_{Sobel}(F_{mural}) \tag{4}$$

where $H_{sobel}(\bullet)$ represents the Sobel operator. Subsequently, a 7×7 convolution operation is used on the gradient map to obtain a preliminary gradient feature map.

$$F_{pg} = H_0(F_{grad}) \tag{5}$$

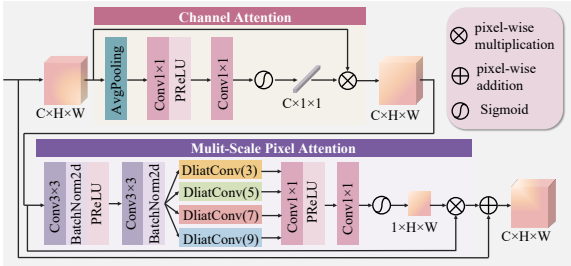


Fig. 3: The pipeline of the dual attention module (DAM).

Then, three dual attention modules are calculated based on the preliminary gradient feature map F_{pg}, enabling a greater emphasis on the content of the image rather than the background area. The primary objective of the dual attention module is to robustly extract the attention information of the gradient feature map, effectively highlighting the pattern areas of the mural image. The dual attention module comprises channel attention and multiscale pixel attention: channel attention enhances the feature responses of different channels in the gradient feature map, while multiscale pixel attention increases focus on different scales and positions within the gradient feature map. The combined utilization of these two attention mechanisms plays a crucial role in bridging the lack of interaction between feature information in the channel dimension and spatial position information at different scales.

The gradient feature map of the mural $F_{pg} \in R^{C\times H\times W}$ first passes through a global average pooling layer, two 1×1 convolutional layers, and a sigmoid function to obtain the channel attention weights $w_{CA} \in R^{C\times 1\times 1}$, as shown in Fig. 3. The obtained weights of each channel w_{CA} will be multiplied by the F_{pg} to get gradient feature maps $F_c \in R^{C\times H\times W}$ with channel attention. The formulation for calculating F_c is presented in Equation 7

$$F_c = \left(\left(Sigmoid\left(Conv_1\left(Conv_1\left(g_c\left(F_{pg}\right)\right)\right)\right)\right)\odot F_{pg}\right) \quad (6)$$

where $Conv_1$ represents the 1×1 convolutional. Subsequently, F_c will be processed through a multi-scale pixel attention module, which comprises two 3×3 convolutional layers, four dilated convolutional layers with varying dilation ratios $\in \{3,5,7,9\}$, two 1×1 convolutional layers, and a sigmoid function

$$F_m = Conv_1\left(Conv_1\left(Dliat_{\{3,5,7,9\}}\left(Conv_3\left(Conv_3\left(F_c\right)\right)\right)\right)\right) \quad (7)$$

where $Conv_3$ denotes the 3×3 convolutional and $Dliat_{\{3,5,7,9\}}$ denotes the dilated convolutions with dilation rates of 3, 5, 7, and 9 respectively. To enhance the perception of the spatial distribution of edges in gradient images, four dilated convolutions are specially used to capture feature information of multiple receptive fields. To obtain the dual attention F_{Dual}, we multiply $F_c \in R^{C\times H\times W}$ and $F_m \in R^{1\times H\times W}$ as follows in the equation

$$F_{Dual} = F_c \odot F_m \quad (8)$$

where $F_{Dual} \in R^{C\times H\times W}$. Finally, F_{Dual} is integrated with the input gradient feature map F_{pg} through element-wise addition to obtain the gradient feature map with dual attention

$$F_{gattn} = F_{pg} \oplus F_{Dual} \quad (9)$$

where $\oplus$ represents the element-wise addition operation. To match the size of multi-scale mural feature maps, we perform two downsampling and Sigmoid operations on $F_{gattn}(\bullet)$ to obtain the final gradient attention map

$$F_{attn} = Sigmoid\left(H_{down}(F_{gattn})\right) \quad (10)$$

While gradient information can emphasize pattern edges, it can also misinterpret defects such as plaster peeling and detachment on the background wall as edge information. Furthermore, most existing attention mechanisms typically lack adaptive guidance which can lead to severely diseased mural images being assigned higher attention weights and generating sketch results with cluttered backgrounds.. Theoretically, for relatively well-preserved murals, the gradient attention map can extract details more completely without introducing excessive noise. However, for severely damaged mural images, although their gradient attention maps contain rich details, they also retain a significant amount of defective edges, resulting in cluttered backgrounds and reduced accuracy of the extracted sketch. Therefore, for mural images with different characteristics, the weight of gradient attention introduced by the model should be different. Based on this, we adaptively weight the gradient attention map F_{attn} and the multi-scale mural feature map F_{sc} through a learnable parameter α, which can automatically adjust their weights according to the different degrees of disease. The feature map F_{fuse} weighted by the attention map can be given by

$$F_{fuse} = \alpha F_{attn} + (1-\alpha)\, F_{sc} \quad (11)$$

Through Equation 12, the network assigns smaller weights to the gradient attention maps of severely damaged areas, preserving more details while ensuring a clean background. On the other hand, more weight is allocated to the gradient attention maps of defect-free murals, enabling them to focus more on capturing rich details.

We adjust the size of F_{fuse} to the original dimensions using two upsampling operations followed by a 7×7 deconvolution operation, resulting in the reconstructed sketch image

$$F_{sketch} = H_1\left(H_{up}(F_{fuse})\right) \quad (12)$$

where H_{up} is the upsampling operations and H_1 stands for 7×7 deconvolutional operation.

3.4. Depth prediction supervision

Mural images are a type of artwork in which pigments are applied to wall surfaces. These can be viewed as consisting of foreground pigment areas and background wall areas. If the network could focus on the foreground during sketch extraction, it would not only impose geometric constraints on the placement of lines but also enhance the extraction of sketches from these foreground areas, thereby minimizing interference from background noise.

To emphasize the regions where pigment-drawn patterns, acquiring depth maps for mural images becomes imperative. In fact, the majority of existing datasets lack access to ground truth depth map labels, and depth estimation is typically achieved through stereo calibration methods. Recent methods (Bhattacharjee et al., 2022) have achieved outstanding results in generating depth maps from comic images, offering valuable technical guidance for us to estimate depth maps on mural images. We utilize the attention-guided monocular depth estimator from (Bhattacharjee et al., 2022) to obtain pseudo-depth maps for mural images as ground truth depth maps, which are employed during training to guide the prediction of depth maps from sketch images. Given an input mural image F_{mural}, which is fed into an attention-guided monocular depth estimator D, the resulting pseudo-ground truth depth map is expressed as $D(F_{mural})$.

Next, our objective is to predict the depth information from the sketch, aiming to align it as closely as possible with the pseudo-depth map. Due to the domain gap existing between mural images and sketch images, the depth estimator cannot be directly applied to sketch images. The study conducted by Chen et al. (Chan et al., 2022) demonstrated that using features from the middle of the Inception v3 (Szegedy et al., 2016) network can reduce the domain gap significantly, particularly suitable for predicting sketches. Based on this, given an input mural image F_{mural}, a sketch image $G_a(F_{mural})$ is obtained after passing through the generator G_a. Then, we utilize the pre-trained Inception v3 network to extract features $I(G_a(F_{mural}))$ from $G_a(F_{mural})$. Subsequently, the depth information from $I(G_a(F_{mural}))$ is predicted using the global generator G_{Geom} proposed in the pix2pixHD framework (Wang et al., 2018). The final depth map predicted from the sketch image is represented by $G_{Geom}(I(G_a(F_{mural})))$. Finally, the binary cross entropy loss function is employed to constrain the depth map predicted from the sketch images and the pseudo-ground truth depth maps.

$$L_{depth} = \left\| G_{Geom}(I(G_a(F_{mural}))) - D(F_{mural}) \right\| \quad (13)$$

3.5. Loss Function

Adversarial Loss. The adversarial losses (Goodfellow et al., 2014) encourage generated images to belong to their respective domains. The loss for the mural domain and sketch domain is formulated below

$$\begin{aligned} L_{GAN} = {} & E_{a\sim A}\left[D_A(a)^2\right] + E_{b\sim B}\left[\left(1 - D_A\left(G_B(b)\right)\right)^2\right] \\ & + E_{b\sim B}\left[D_B(b)^2\right] + E_{a\sim A}\left[\left(1 - D_B\left(G_A(a)\right)\right)^2\right] \end{aligned} \quad (14)$$

where A and B represent the mural domain and the sketch domain, respectively. a denotes the image from the mural domain, and b denotes the image from the sketch domain.

Cycle consistency loss. To alleviate the mode collapse problem, we apply a cycle consistency constraint (Zhu et al., 2017) to the generator. Given a mural image $a \in A$, after the sequential translations of a from A to B and from B to A, the image should be successfully translated back to the original domain

$$L_{Cycle} = \left\| G_B\left(G_A(a)\right) - a \right\| + \left\| G_A\left(G_B(b)\right) - b \right\| \quad (15)$$

Full objective. We jointly train the adversarial loss, cycle loss, and depth prediction loss to optimize the final objective:

$$L = \lambda_1 L_{GAN} + \lambda_2 L_{Cycle} + \lambda_3 L_{depth} \quad (16)$$

where $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 10$.

4. Experiment Settings

4.1. Implementation Details

Our AGD-GAN has totally 6.028M parameters and is trained with PyTorch1.4.0 and NVIDIA GeForce RTX 3090 on Linux operating system. The generator and discriminator were trained alternately. We use Adam (Kingma & Ba, 2014) to optimize with a fixed learning rate of 2×10^{-4} and train for at least 120 epochs with batch size 1 until no further performance increase is observed.

4.2. Datasets and evaluation metrics

The training dataset comprises both a mural domain dataset and a sketch domain dataset, primarily derived from on-site collections and electronic book extractions. We have collected mural images from Fengguo Temple, Qianling, and Dunhuang murals, featuring complex scenes including Buddhist statues and Tang dynasty tombs. Additional mural and sketch images were cropped from electronic books, specifically 'The Complete Collection of Chinese Excavated Murals' and 'Dunhuang Murals'. Overall, we have collected 3,000 original mural images, including 1,500 disease-free murals and 1,500 diseased murals (affected by mold, cracks, flaking, etc.). Furthermore, we have acquired a sketch domain dataset comprising 3,000 images with a style similar to the mural domain. In our experiments, the collected images were randomly cropped to a size of 256×256 after undergoing cropping, scaling, or high-definition scanning. To enhance the robustness of our model, we flipped each image horizontally and applied random rotations for data augmentation.

The test dataset contains 600 mural images of various styles cropped from e-books, including Dunhuang murals, Tang dynasty tomb murals, Indian murals, etc. Among these, 300 images feature clean backgrounds and clear murals, while the other 300 contain diseases such as mud stains and cracks. Additionally, we have collected 12 real diseased images obtained from on-site collections, with their ground truth sketches drawn by experienced experts from the Shaanxi History Museum. Mural images without ground truth sketches (such as

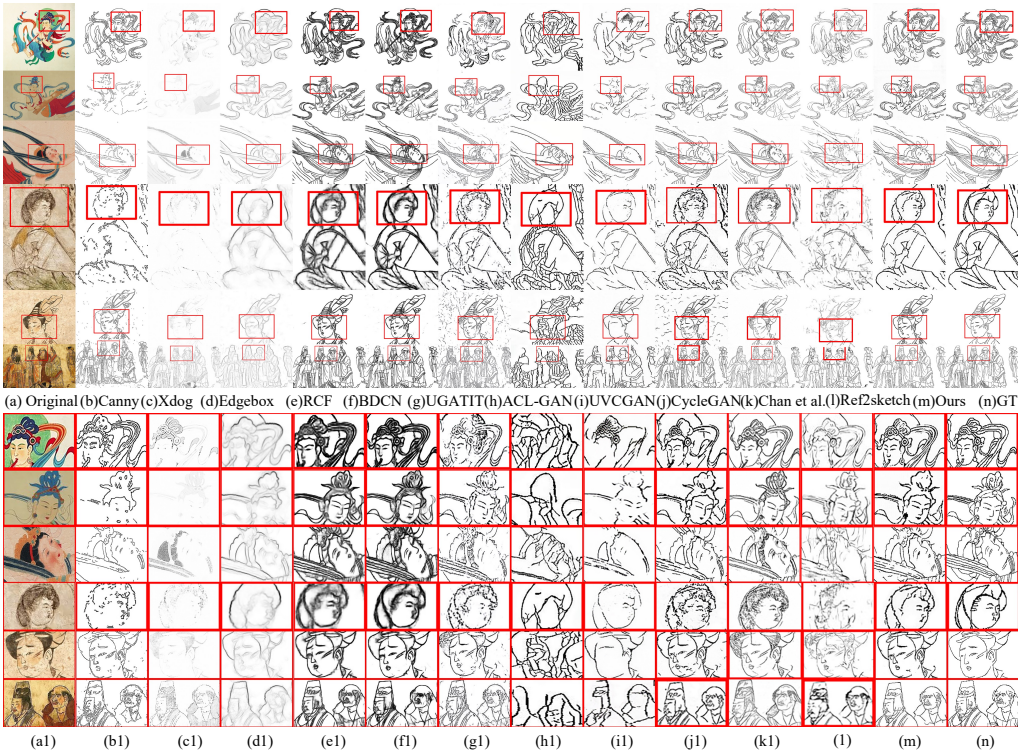


Fig. 4: Visual comparisons of eleven state-of-the-art sketches extracting approaches on Tang dynasty murals datasets. (a) The original mural images; (b) Canny (Canny, 1986); (c) Xdog (Winnemöller, 2011); (d) Edgebox (Zitnick & Dollár, 2014); (e) RCF (Liu et al., 2017); (f) BDCN (He et al., 2019); (g) UGATIT (Kim et al., 2019); (h) ACL-GAN (Zhao et al., 2020); (i)UVCGAN (Torbunov et al., 2023); (j)CycleGAN(Zhu et al., 2017);(k)Chan et al.(Chan et al., 2022); (l) Ref2sketch(Ashtari et al., 2022); (m) Ours; (n) Ground Truth. The figures below (a1-n1) are the corresponding partially enlarged details.

those of Indian murals) are used only for qualitative comparison. We use different metrics, including root mean squared error (RMSE) (Wang et al., 2004), structural similarity index (SSIM) (Wang et al., 2004), and average precision (AP) (Martin et al., 2004b) to evaluate our method and compare it with the other state-of-the-art methods.

5. Experimental results

5.1. Comparison with other works

Sketches extraction results of the proposed method along with state-of-the-art methods are shown in Fig. 4 - Fig. 5 and Table 1 - Table 2. The models include: the gradient-based methods (Canny (Canny, 1986) and Xdog (Winnemöller, 2011)), a learning-based algorithm (Edgebox (Zitnick & Dollár, 2014)), CNN-based algorithms (RCF (Liu et al., 2017) and BDCN (He et al., 2019)) and six GANs-based image translation algorithm (UGATIT (Kim et al., 2019), ACL-GAN (Zhao et al., 2020) ,UVCGAN (Torbunov et al., 2023), CycleGAN(Zhu et al., 2017), Chan et al.(Chan et al., 2022), and Ref2sketch(Ashtari et al., 2022)).

The numerical indexes including SSIM, RMSE, and AP are reported in Table 1 and Table 2. Among the array of existing edge detection methods, CNN-based methods demonstrate superior performance across SSIM, RMSE, and AP metrics in contrast to both traditional and learning-based methods. When compared to the learning-based edge detection method

Table 1
The average SSIM/RMSE/AP of existing eleven state-of-the-art sketches extracting approaches over disease-free murals datasetst

Method	SSIM(↑)	RMSE(↓)	AP(↑)
Canny (Canny, 1986)	0.8037	0.3640	0.3509
Xdog (Winnemöller, 2011)	0.3217	0.3480	0.2917
Edgebox (Zitnick & Dollár, 2014)	0.4805	0.3617	0.3022
RCF (Liu et al., 2017)	0.7941	0.3310	0.4093
BDCN (He et al., 2019)	0.8019	0.3267	0.4266
UGATIT (Kim et al., 2019)	0.9729	0.2903	0.6144
ACL-GAN (Zhao et al., 2020)	0.9241	0.4982	0.2640
UVCGAN (Torbunov et al., 2023)	0.9234	0.3954	0.5517
CycleGAN(Zhu et al., 2017)	0.9381	0.3710	0.5790
Chan et al.(Chan et al., 2022)	0.9397	0.3690	0.5832
Ref2sketch(Ashtari et al., 2022)	0.9341	0.3707	0.5701
Ours	**0.9808**	**0.2929**	**0.6117**

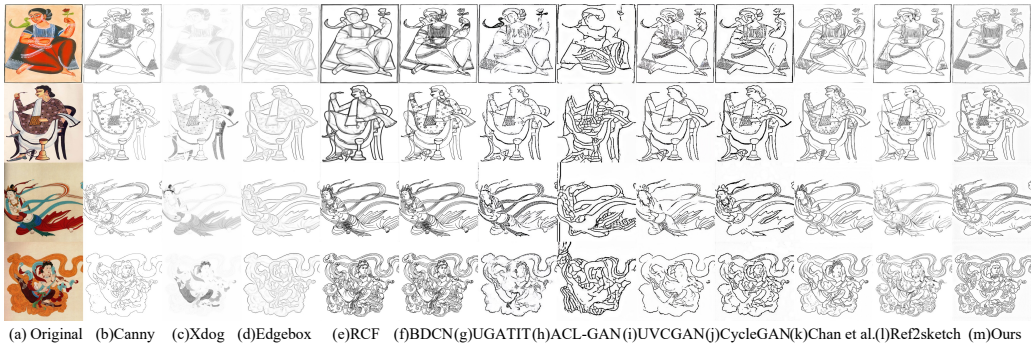


Fig. 5: Visual comparisons of eleven state-of-the-art sketches extracting approaches on Indian and Dunhuang murals datasets. (a) The original mural images; (b) Canny (Canny, 1986); (c) Xdog (Winnemöller, 2011); (d) Edgebox (Zitnick & Dollár, 2014); (e) RCF (Liu et al., 2017); (f) BDCN (He et al., 2019); (g) UGATIT (Kim et al., 2019); (h) ACL-GAN (Zhao et al., 2020); (i)UVCGAN (Torbunov et al., 2023); (j)CycleGAN(Zhu et al., 2017);(k)Chan et al.(Chan et al., 2022); (l) Ref2sketch(Ashtari et al., 2022); (m) Ours.

Table 2
The average SSIM/RMSE/AP of existing eleven state-of-the-art sketches extracting approaches over diseased murals datasets

Method	SSIM(↑)	RMSE(↓)	AP(↑)
Canny (Canny, 1986)	0.8152	0.4139	0.3004
Xdog (Winnemöller, 2011)	0.4195	0.4427	0.3410
Edgebox (Zitnick & Dollár, 2014)	0.3517	0.4689	0.2648
RCF (Liu et al., 2017)	0.8095	0.4256	0.4507
BDCN (He et al., 2019)	0.8170	0.4396	0.4900
UGATIT (Kim et al., 2019)	0.9257	0.4210	0.5452
ACL-GAN (Zhao et al., 2020)	0.8206	0.4609	0.1930
UVCGAN (Torbunov et al., 2023)	0.9093	0.4196	0.5082
CycleGAN(Zhu et al., 2017)	0.9134	0.3840	0.5190
Chan et al.(Chan et al., 2022)	0.9189	0.3775	0.5089
Ref2sketch(Ashtari et al., 2022)	0.9034	0.3954	0.5317
Ours	**0.9411**	**0.3897**	**0.5931**

(Edgebox), BDCN exceeds SSIM by approximately 32% and AP by about 12% on the disease-free dataset, while surpassing SSIM by approximately 46% and AP by approximately 22% on the diseased dataset. Consequently, CNN-based methods demonstrate exceptional advantages in pixel classification accuracy and structural similarity, exhibiting remarkable stability in performance on the diseased dataset. In contrast, most traditional methods experience decreased performance when handling diseased murals. The overall numerical indicators based on the unsupervised GAN method surpassed the CNN-based method, especially in terms of SSIM and AP indicators. This highlights the robust advantage of unsupervised GAN techniques in tasks with limited data and showcases visually more realistic extracted sketches. Significantly, the ACL-GAN employs adversarial consistency loss, leading to a lack of control over mural image content and the generation of inaccurate lines. With the lowest AP value among all methods, it achieved only 0.1930 on the diseased mural. This highlights the key role of loss function design, particularly for mural images characterized by intricate and variable content and style.

It is evident that our method has improved significantly improvements over other state-of-the-art methods in terms of different metrics on the disease-free and diseased datasets. This improvement is due to the proposed encoder architecture with multi-scale feature extraction and gradient attention map, and the employed depth prediction supervision. In comparison to the second-ranked UGATIT method, our approach demonstrates approximately a 2% enhancement in SSIM, along with an approximately 5% improvement in AP on the diseased dataset. The reported results demonstrate that the sketches of our method extract are not only closest to the real situation of the ground in terms of pixel accuracy and style and structure similarity but also maintain strong performance in the case of diseased murals.

The proposed method was also compared qualitatively. The sketch results on Tang dynasty murals datasets and partially enlarged details from different methods are shown in Fig. 4. The sketch results of the Indian and Dunhuang datasets are shown in Fig. 5. Our advantages are expressed in the following aspects: (a) Advantages in suppression of cluttered background noise. For example in Fig. 4, in the fourth image containing both light-colored cracks and holes in the wall, the sketches extracted by Canny and UGATIT are severely affected by the disease, especially in images with complex disease and dense gradient distribution, which produces a significant number of fake lines. The rationale behind this lies in Canny's reliance on gradient computation, where regions afflicted with drastic gradient alterations are prone to be misconstrued as edges within the murals. While UGATIT effectively employs the attention guidance model to dynamically regulate texture variations within the image, it lacks the capability to control background noise effectively. (b) Advantages of extracting detailed information. For example, in the detailed enlargements of Fig. 4, the sketch results extracted by UVCGAN lose a large number of detailed structures (e.g., human faces). Although the sketch extracted by Edge-Boxes

is slightly more robust to noise, they fall short of fully utilizing the extracted bounding box information to generate clear and detailed lines. Turning to our method, the boundaries are more accurate than other methods, yielding comprehensive extraction of detailed structures across both large and small scales. (c) Advantages of generating sharp edges. For example, in the fifth and sixth images of Fig. 4, it is difficult to extract details due to the large amount of information at different scales (the human contours generated by RCF and BDCN are blurred). The sketch results generated by our model consistently maintain clear edges without noticeable discrete points even amidst such complexities. (d) Advantages on different datasets. Fig. 5 shows the sketch extraction results tested on some unlabeled datasets. While UGATIT addresses the challenge of blurry sketches, its limitation lies in the absence of guided gradient prior knowledge, particularly evident in processing less obvious lines, as exemplified by the skirt pleats of the Flying Goddess in the third image of Fig. 5. UVCGAN, equipped with a Vision Transformer(Dosovitskiy et al., 2010) for the generator, better circumvents background noise by learning the relationship between global and local information. However, this approach sacrifices essential details, evident in the incomplete facial features and intermittent contour lines observed in Fig. 5. ACL-GAN employs adversarial consistency loss instead of cycle consistency loss, resulting in erratic sketches. The sketches generated by CycleGAN, Chan et al., and Ref2sketch have lost a great deal of detail. In contrast to the aforementioned GAN methodologies, our approach excels in generating sketches with clear lines and rich details.

Overall, the proposed method demonstrated significantly improved performance in both quantitative and qualitative evaluations for mural sketches compared to previous methods.

5.2. The effect of each module

To assess the efficiency of the model proposed in Section 3, we conduct a series of experiments as an ablation study to illustrate the contribution of the proposed components in our framework and validate our assumptions. We use CycleGAN as our baseline (Zhu et al., 2017), ablative experiments can be divided into four categories: cross-channel attention residual module performance evaluation, gradient attention map generation strategy evaluation, adaptive fusion performance evaluation of learnable parameters α, and depth prediction supervised performance evaluation.

5.2.1. The effect of cross-channel attention residual module

The cross-channel attention residual module (CARM) was designed on the pyramid squeeze attention (PSA). Specifically, we designed a shunt residual structure to mine the inherent characteristics of mural images while preserving multi-scale detailed features. As shown in Fig. 6, we designed three models to evaluate the effect of the proposed CARM, while the gradient attention map adaptive guide branch was removed from the main model individually to reduce interference terms. In the first model, we replace the CARM with the residual module given in (He et al., 2016) as the

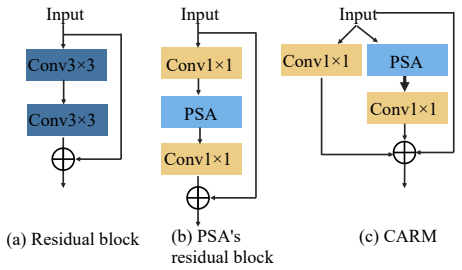


Fig. 6: Three models evaluated in an ablation study of cross-channel attention residual modules.

Table 3
Comparison of sketch generation results of "Residual block", "PSA's residual block", and "CARM" on diseased murals

Model structure	SSIM(↑)	RMSE(↓)	AP(↑)
Residual block	0.8983	0.3147	0.3703
PSA's residual block	0.9092	0.3164	0.3783
CARM	**0.9218**	**0.3079**	**0.4100**

'Residual block', and in the next model, we adopt the residual module integrated in the (Zhang et al., 2022) to replace the CARM. In the last model, we employ the shunt structure residual module CARM designed in this article. We call these models "Residual block", "PSA's residual block", and "CARM", respectively. Comparisons of sketch results directly trained from the above three models are conducted to validate the effectiveness of CARM, as shown in Table 3 and Fig. 7. In Fig. 7, we showed the qualitative comparison results of "Residual block", "PSA's residual block", and "CARM". Due to the diseases in the mural image, the results of the Residual block and PSA's residual block contain numerous scattered points and discontinuities. In the Residual block model, the original residual blocks learn features from mural images, including more texture information. However, these texture details lack a clear distinction between primary and secondary importance, leading to the incorporation of irrelevant background texture information in the image features (e.g., noticeable excess texture in the clothing of the depicted figure in Fig. 7(b)). PSA's residual block focuses on the learning of multi-scale features in both spatial and channel dimensions, significantly reducing background noise. Nevertheless, as feature learning accumulates, many shallow features fail to be preserved (e.g., the eyes of the depicted character in Fig. 7(f) remain inadequately represented). CARM can not only capture fine-grained details (e.g., all faces in the mural and the lady's bun) but also generate clearer edges (e.g., Residual block and PSA's residual block tend to generate blurry boundaries in detailed magnification images). This means that compared to the residual block, the PSA module is effective in capturing multi-scale spatial information of murals. The introduction of the shunt structure enhances the performance of the PSA module, enabling it to effectively mine fine-grained details while preserving essential details

(a) Mural image (b) Residual block (c) PSA's residual block (d) CARM (e) (f) (g)

Fig. 7: Visual comparisons of the effect of cross-channel attention residual module. (a) The original mural image; (b) Residual block; (c) PSA's residual block; (d) CARM. The right figures (e-g) are the corresponding partially enlarged details.

of mural images as much as possible.

Additionally, Table 3 shows the advantages of the sketches generated by the CARM in image error, style similarity, and completeness. In contrast to the Residual block and PSA's residual block, CARM demonstrates significant improvements in both SSIM and AP indicators, which indicates that the correct pixels in the sketch gradually increase as the model improves. Overall, integrating the PSA module into the shunt residual structure proves to be effective in preserving multi-scale detailed features, thereby significantly enhancing the performance of sketch generation.

5.2.2. The effect of gradient attention map generation strategy

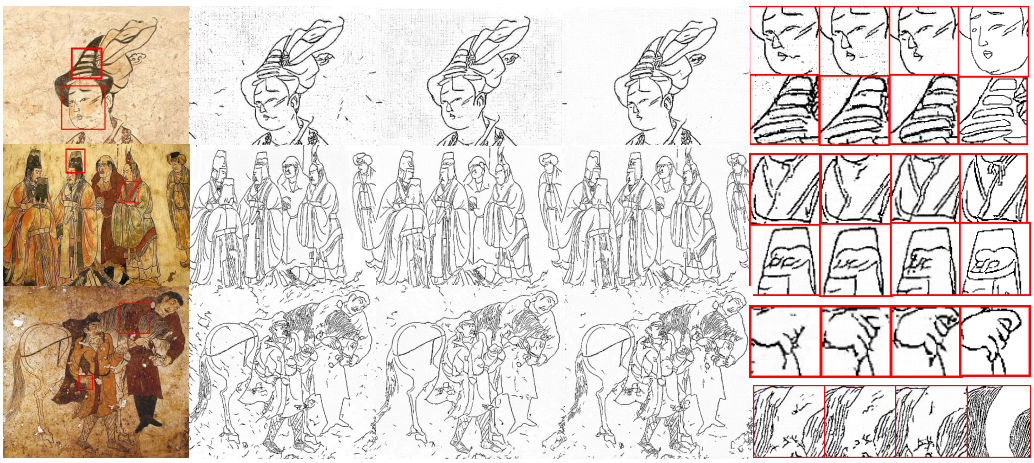
To demonstrate that training attention mechanisms on gradient maps of murals can significantly improve focus on patterned regions, and that compared to single attention, employing dual attention can better locate the salient pattern and suppress the background noise. We designed three other models to provide different combinations. Firstly, we evaluate a model that solely incorporates gradient information without introducing attention mechanisms. It directly adaptively fuses the mural image after gradient calculation with the multi-scale feature map extracted previously. Also, we designed another model that only adds channel attention after introducing gradient information. The last model introduces two types of attention mechanisms following the incorporation of gradient information, namely channel attention and multi-scale pixel attention. These models are called "Only gradient", " gradient + CA", and "gradient + DA", respectively.

The sketch extraction results produced by different models are shown in Fig. 8. The "Only gradient" model, guided by gradient information, fully exploits the edge structures and texture details contained in the gradient map, emphasizing the patterned regions from the perspective of the inherent edge information in mural images. It not only maintains the integrity of the target shapes in the mural images but also enhances details(eg. the collar, ponytail, and facial features of the character in Fig. 8(e)). However, direct extraction of gradient information from mural images also includes complex background noise and false edges caused by cracks. As shown in Fig. 8(b), the sketch background of the lady picture and the horse riding contain numerous messy lines, which are irrelevant information. Fig. 8(c) shows the sketch results generated by applying channel relation inference to the "gradient + CA" model. Compared to the model introducing only gradient information, it can be observed that calculating channel attention based on gradient information can reduce false lines caused by noise and artifacts, but some disturbances can still be detected (eg. patches near the maid's mouth in Fig. 8(f)). Nevertheless, channel relationship reasoning only constructs a channel position map, resulting in insufficient feature utilization and incomplete background information suppression. In contrast, the proposed "gradient+DA", incorporating both channel relationship reasoning and multi-scale pixel relationship reasoning, can significantly suppress background noise and discrete points. This effect may be attributed to the integration of the pixel attention module, which combines predictions from multiple scales, effectively alleviating the impact of noise and artifacts.

Table 4
Comparison of sketch generation results of "Only gradient", "gradient + CA", and "gradient + DA" on diseased murals

Model structure	SSIM(↑)	RMSE(↓)	AP(↑)
Only gradient	0.9267	0.3216	0.4069
gradient + CA	0.9119	0.3004	0.4447
gradient + DA	**0.9491**	**0.2998**	**0.4601**

In addition, we provide the sketch generation quantitative results of different models in Table 4. As seen from this table, the "gradient + CA" model exhibits an approximately 2%



(a) Mural image (b) Only gradient (c) gradient + CA (d) gradient + DA (e) (f) (g) (h) GT

Fig. 8: Visual comparisons of the effect of gradient attention map generation strategy. (a) The original mural image; (b) Only gradient; (c) gradient + CA; (d) gradient + DA. The right figures (e-h) are the corresponding partially enlarged details.

improvement in the RMSE metric and about 3.7% improvement in the AP metric than the "Only gradient" model, which showed computing the attention mechanism on the gradient map effectively reduces erroneous pixels (i.e., the background noise of the mural). However, the "gradient + CA" model experiences a decrease of approximately 1.4% in the SSIM metric, possibly due to the reduction of some correct fine edges while suppressing background noise. The "gradient + DA" model achieved SSIM values approximately 3.7% higher and AP values approximately 1.5% higher than the "gradient + CA" model, which indicated that multi-scale pixel attention is an effective supplement to sketch features, better enabled the gradient information to guide the generation of the sketch, and improved the accuracy and completeness of the final sketches.

In summary, experimental results proved that gradient maps containing dual attention offer superior guidance for generating sketches that are richer, more accurate, and more similar compared to only using gradient maps.

5.2.3. The effect of fusion performance of learnable parameter α

The state of preservation of murals in different regions and eras varies. If the attention map of the mural gradient image is directly fused with the multi-scale mural feature map, a large number of false edges may be retained when extracting sketches of severely damaged murals, which reduces the accuracy of sketch extraction. To evaluate the adaptive fusion effect of the proposed learnable parameter α on murals exhibiting varying degrees of deterioration, we conducted two sets of experiments pertaining to the fusion parameters for diseased and disease-free murals: (1) For disease-free murals, verifying the impact of different fixed α parameter settings on the sketch generation results. We utilized three fixed parameter groups: "$\alpha = 0.2, 1-\alpha = 0.8$", "$\alpha = 0.5, 1-\alpha = 0.5$", and "$\alpha = 0.8, 1-\alpha = 0.2$" respectively. Furthermore, we also demonstrate the results with α not fixed (automatically adjusted through network training). (2) For diseased murals, we set three groups of parameters: "$\alpha = 0.8, 1-\alpha = 0.2$", "$\alpha = 0.5, 1-\alpha = 0.5$", "$\alpha = 0.2, 1-\alpha = 0.8$", and α not fixed to verify the results of sketch extraction. Theoretically, a larger α value is preferable to acquire more gradient information in disease-free mural images. Conversely, the α value for diseased murals should be appropriately reduced to mitigate the residual background noise in gradient maps.

We selected a disease-free mural image with rich multi-scale details as an example to showcase the generated sketch results under different parameter settings, as depicted in Fig. 9. Here, α represents the weight of the gradient information at-

Table 5
Quantitative experimental results regarding fusion performance of learnable parameter α on disease-free murals

Model structure	SSIM(↑)	RMSE(↓)	AP(↑)
$\alpha = 0.2,\ 1-\alpha = 0.8$	0.9161	0.3105	0.4831
$\alpha = 0.5,\ 1-\alpha = 0.5$	0.9290	0.3180	0.5160
$\alpha = 0.8,\ 1-\alpha = 0.2$	0.9492	0.3043	0.5590
α	**0.9538**	**0.3019**	**0.5844**

Table 6
Quantitative experimental results regarding fusion performance of learnable parameter α on diseased murals

Model structure	SSIM(↑)	RMSE(↓)	AP(↑)
$\alpha = 0.8,\ 1-\alpha = 0.2$	0.8966	0.4597	0.4936
$\alpha = 0.5,\ 1-\alpha = 0.5$	0.9027	0.4510	0.5310
$\alpha = 0.2,\ 1-\alpha = 0.8$	0.9139	0.4268	0.5503
α	**0.9300**	**0.4096**	**0.5741**

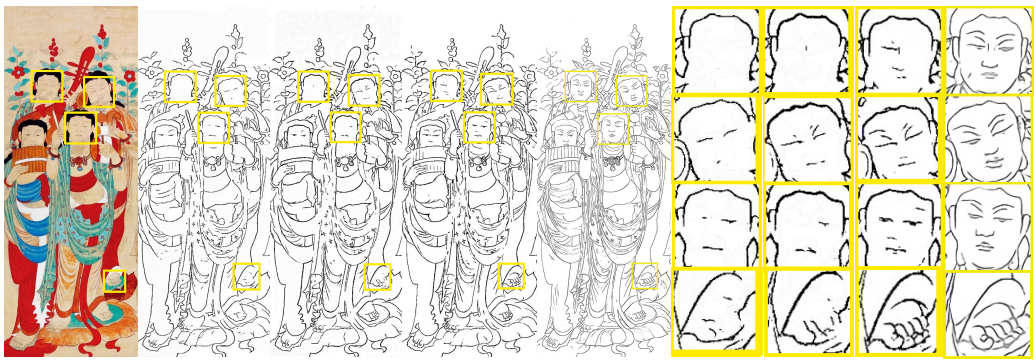


Fig. 9: Qualitative experimental findings regarding the fusion performance of the learnable parameter α on disease-free murals. The right figures (f-i) are the corresponding partially enlarged details.

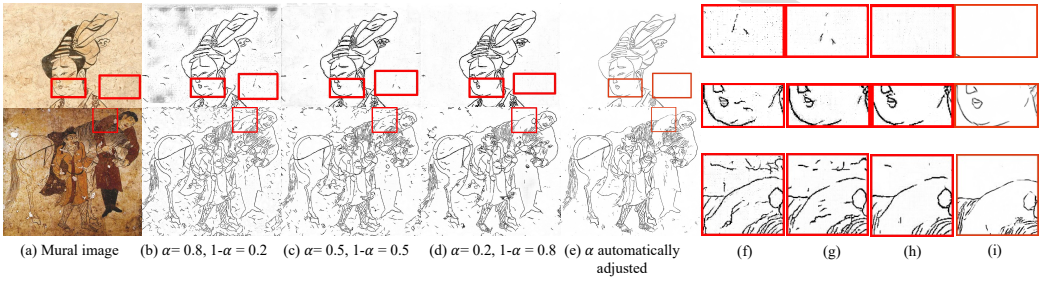


Fig. 10: Qualitative experimental findings regarding the fusion performance of the learnable parameter α on diseased murals. The right figures (f-i) are the corresponding partially enlarged details.

tention map, while $1-\alpha$ denotes the weight of the multi-scale mural feature map. In Fig. 9(b), when the weight attributed to the gradient information attention map is diminished, the extracted sketch exhibits a deficiency in capturing subtle features, such as the facial features and feet of the Buddha image. As the value of α increases, these details are progressively extracted effectively (refer to the detailed enlarged images in Fig. 9(f-h)). In reference to the three evaluation metrics displayed in Table 5, observations indicate a general improvement trend for SSIM, RMSE, and AP with the increment of α values. In particular, the AP value of $\alpha = 0.8$ is increased by approximately 7.6% compared to $\alpha = 0.2$, which means an augmentation in the count of accurately identified pixels. Furthermore, Fig. 9(e) demonstrates the results of α learned automatically, which achieved richer detail features. The metric results of α learned automatically in Table 5 are greater than those for $\alpha = 0.5$, indicating that the proportion of the value of α will be larger for murals without disease.

For damaged murals, we specifically selected two mural images with slight cracks and damaged holes. The sketch results generated under different parameter settings are shown in Fig. 10. Notably, when $\alpha = 0.8$ (Fig. 10(b)), the sketch background retains messy lines, especially in densely damaged areas like the second image. As can be seen from Fig. 10(b-d), as the α value continues to decrease, the resulting sketch is less affected by noise and disease. Table 6 presents the quantitative experimental results of the fusion performance of the learnable parameter α on diseased murals. Compared with $\alpha = 0.8$, the AP value of $\alpha = 0.2$ increases by approximately 5.6%, the RMSE value exhibits a decrease of around 3.2%, while the SSIM value also increased by approximately 1.7%. The above results show that appropriately diminishing the weight of gradient information attention maps for diseased murals can overcome the effects of noise and disease to a certain extent. When α is adjusted automatically, the generated sketch background becomes clearer but the lines are weaker (as shown in Fig. 10(e)). From the results in Table 6, it is evident that the performance metrics are better when α is adjusted automatically compared to $\alpha = 0.2$. Therefore, α tends to be adjusted lower when dealing with diseased murals.

5.2.4. The effect of depth prediction supervision

To prove the effectiveness of depth prediction, we designed a similar conceptually-driven supervision branch as a control group, namely saliency target supervision. We use the Stereoscopically Attentive Multiscale (SAM) given in (Liu et al., 2021) to replace the depth map prediction, aligning saliency feature maps of the mural and sketch to

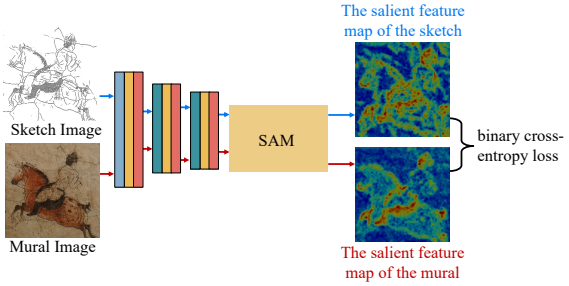


Fig. 11: The workflow of saliency target supervision. We employed a binary cross-entropy loss function to constrain the saliency feature map of the mural and the saliency feature map of the sketch, which can replace depth supervision theoretically.

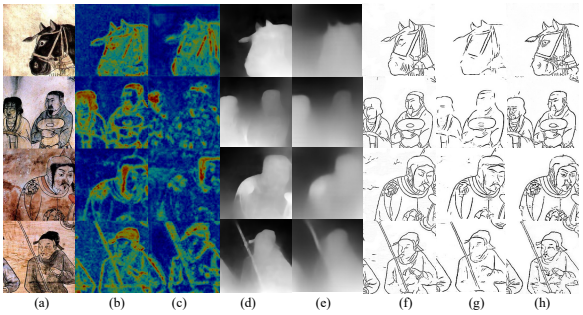


Fig. 12: Visualization results of depth prediction supervision and saliency target supervision during training. (a)The mural images; (b)The salient feature map of the mural; (c)The salient feature map of the sketch; (d) The Pseudo-ground truth depth map; (e)The depth map predicted from the sketch images; (f)The sketch results generated by baseline; (g)The sketch results generated by saliency target supervision; (h)The sketch results generated by deep prediction supervision.

constrain object positions. The SAM can adaptively adjust the information flow of feature branches at different scales, suppressing less informative branches while enhancing those with discriminative information. Theoretically, saliency target supervision can also emphasize the geometry of objects in murals by leveraging the stereo attention mechanism. Fig. 11 illustrates the workflow of saliency target supervision: First, the convolution and two downsampling operations in the multi-scale mural feature extraction branch are used to obtain the mural feature map Fe_{mural} and the sketch feature map Fe_{sketch} respectively. These are then input into SAM to obtain the salient feature maps for mural and sketch. Finally, similar to depth prediction loss, a binary cross-entropy loss function is employed to align the salient feature map of the mural $SAM\left(Fe_{mural}\right)$ and the salient feature map of the sketch $SAM\left(Fe_{sketch}\right)$. The saliency target supervision is computed as follows.

$$L_{saliency} = \left\| SAM\left(Fe_{sketch}\right) - SAM\left(Fe_{mural}\right) \right\| \quad (17)$$

where SAM denotes the Stereoscopically Attentive Multi-scale module.

To visually validate whether the depth prediction supervision and saliency target supervision during training can effectively focus on the patterned objects (foreground regions) in murals, we individually visualized the saliency attention maps and depth prediction maps generated during the training process, as depicted in Fig. 12(b-e). We observe that while the saliency attention feature map tends to focus on the main outline of the pattern, it places a stronger emphasis on conspicuous details within the patterns (such as the harness on the black horse head, the black hair, and the black hats). As a result, during the training process, it tends to generate sketches in attention-focused areas while ignoring other places, which will become distractors in training and lead to incomplete sketch results (such as the sketch results generated in Fig. 12(g)). As shown in Fig. 12(d-e), both pseudo-ground truth depth maps and predicted depth maps exhibit the ability to emphasize and precisely locate the subject of the object, converging in similarity as the training progresses. Due to the emphasis of depth images on the entire object rather than the salient information within the object, depth prediction supervision effectively preserves the geometric shape of patterns and constrains the position of lines compared to saliency target supervision. It not only promotes the training in a positive direction but also captures more image details(such as the sketch results generated in Fig. 12(h)).

Table 7
Quantitative experimental results for depth prediction supervision performance evaluation on diseased murals

Model structure	SSIM(↑)	RMSE(↓)	AP(↑)
Baseline	0.8591	0.3740	0.3727
Saliency target supervision	0.8017	0.4277	0.3110
Depth prediction supervision	**0.8933**	**0.3387**	**0.3965**

Table 7 presents the quantitative results of the performance evaluation for depth prediction supervision. We evaluate the quality of sketches generated by three models: baseline, Saliency target supervision, and Depth prediction supervision. Here, baseline means that no depth prediction supervision or saliency target supervision is used. The best result among the three previous models is obtained by model “Depth prediction supervision”, in which the AP value was approximately 8.5% higher than Saliency target supervision. The above results demonstrate the superior accuracy of sketches generated using the employed depth prediction supervision. It is notable that both the ’Depth prediction supervision’ and ’baseline’ models exhibited RSME values approximately 9% and 5% higher, respectively, compared to the ’Saliency target supervision’. This suggests the higher vulnerability of saliency target supervision to background noise in diseased murals, resulting in a higher count of erroneous pixels. In summary, we can understand the important role of depth prediction based on geometric constraints on mural images from the comparison of "baseline" and "Saliency target supervision" models.

6. Conclusion

In this paper, we developed a novel unsupervised network, AGD-GAN, for ancient mural sketch extraction. The proposed network not only considers the noteworthy fine-grained details and multi-scale features in mural characteristics but also thoroughly explores the balance between disease noise suppression and detail preservation. Experimental results indicate that compared to other methods, the proposed method achieves better sketch results and higher line precision on various styles of datasets, further proving its enhanced robustness. Additionally, the impact of each component within the AGD-GAN on sketch extraction performance is discussed and analyzed. The experimental results show that every component of the network plays a positive role in the accurate extraction of sketches. In conclusion, the mural sketch extraction method based on gradient guidance and deep supervision is more effective.

A main limitation of this work is that the proposed method cannot entirely eliminate the noise from severely damaged murals (such as wall cracks and gaps between bricks). This is because our method relies on gradient information for guidance, and murals with obvious cracks and large gradients will be extracted as sketches. We will take this issue as our future work and try to distinguish between damages and sketches as accurately as possible, which may have a better effect.

References

Alotaibi, A. (2020). Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*, *12*, 1705.

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2010). Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *33*, 898–916.

Ashtari, A., Seo, C. W., Kang, C., Cha, S., & Noh, J. (2022). Reference based sketch extraction via attention mechanism. *ACM Transactions on Graphics (TOG)*, *41*, 1–16.

Bertasius, G., Shi, J., & Torresani, L. (2015a). Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4380–4389).

Bertasius, G., Shi, J., & Torresani, L. (2015b). High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In *Proceedings of the IEEE international conference on computer vision* (pp. 504–512).

Bhattacharjee, D., Everaert, M., Salzmann, M., & Süsstrunk, S. (2022). Estimating image depth in the comics domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2070–2079).

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (pp. 679–698).

Chan, C., Durand, F., & Isola, P. (2022). Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7915–7925).

Delaney, J. K., Dooley, K. A., Radpour, R., & Kakoulli, I. (2017). Macroscale multimodal imaging reveals ancient painting production technology and the vogue in greco-roman egypt. *Scientific reports*, *7*, 15509.

Dollar, P., Tu, Z., & Belongie, S. (2006). Supervised learning of edges and object boundaries. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (pp. 1964–1971). IEEE volume 2.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2010). An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, .

Eskandar, G., Abdelsamad, M., Armanious, K., Zhang, S., & Yang, B. (2022). Wavelet-based unsupervised label-to-image translation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1760–1764). IEEE.

Fu, X., Han, Y., Sun, Z., Ma, X., & Xu, Y. (2017). Line-drawing enhanced interactive mural restoration for dunhuang mogao grottoes. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*, 99.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Hallman, S., & Fowlkes, C. C. (2015). Oriented edge forests for boundary detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1732–1740).

Han, X., Hou, M., Zhu, G., Wu, Y., & Ding, X. (2015). Extracting graphite sketch of the mural using hyper-spectral imaging method. *Tehnicki vjesnik/Technical Gazette*, *22*.

He, J., Wang, S., Zhang, Y., & Zhang, J. (2013). A computational fresco sketch generation framework. In *2013 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). IEEE.

He, J., Zhang, S., Yang, M., Shan, Y., & Huang, T. (2019). Bi-directional cascade network for perceptual edge detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3828–3837).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017a). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017b). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Jiang, C., Jiang, Z., & Shi, D. (2022). Computer-aided virtual restoration of frescoes based on intelligent generation of line drawings. *Mathematical Problems in Engineering*, *2022*.

Kim, J., Kim, M., Kang, H., & Lee, K. (2019). U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, .

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

Kittler, J. (1983). On the accuracy of the sobel edge detector. *Image and Vision Computing*, *1*, 37–42.

Kokkinos, I. (2015). Pushing the boundaries of boundary detection using deep learning. *arXiv preprint arXiv:1511.07386*, .

Lim, J. J., Zitnick, C. L., & Dollár, P. (2013). Sketch tokens: A learned mid-level representation for contour and object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3158–3165).

Liu, J., Lu, D., & Shi, X. (2006). Interactive sketch generation for dunhuang frescoes. In *Technologies for E-Learning and Digital Entertainment: First International Conference, Edutainment 2006, Hangzhou, China, April 16-19, 2006. Proceedings 1* (pp. 943–946). Springer.

Liu, Y., Cheng, M.-M., Hu, X., Wang, K., & Bai, X. (2017). Richer convolutional features for edge detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3000–3009).

Liu, Y., & Lew, M. S. (2016). Learning relaxed deep supervision for better edge detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 231–240).

Liu, Y., Zhang, X.-Y., Bian, J.-W., Zhang, L., & Cheng, M.-M. (2021). Samnet: Stereoscopically attentive multi-scale network for lightweight salient object detection. *IEEE Transactions on Image Processing*, *30*,

3804–3814.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004a). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE transactions on pattern analysis and machine intelligence*, *26*, 530–549.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004b). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE transactions on pattern analysis and machine intelligence*, *26*, 530–549.

Park, J., Kim, S., Kim, S., Cho, S., Yoo, J., Uh, Y., & Kim, S. (2023). Lanit: Language-driven image-to-image translation for unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 23401–23411).

Prewitt, J. M. et al. (1970). Object enhancement and extraction. *Picture processing and Psychopictorics*, *10*, 15–19.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D. (2021). Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2287–2296).

Roberts, L. G. (1963). *Machine perception of three-dimensional solids*. Ph.D. thesis Massachusetts Institute of Technology.

Samma, H., Suandi, S. A., & Mohamad-Saleh, J. (2019). Face sketch recognition using a hybrid optimization model. *Neural Computing and Applications*, *31*, 6493–6508.

Shen, W., Wang, X., Wang, Y., Bai, X., & Zhang, Z. (2015). Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3982–3991).

Su, Z., Liu, W., Yu, Z., Hu, D., Liao, Q., Tian, Q., Pietikäinen, M., & Liu, L. (2021). Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5117–5127).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Torbunov, D., Huang, Y., Yu, H., Huang, J., Yoo, S., Lin, M., Viren, B., & Ren, Y. (2023). Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 702–712).

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8798–8807).

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*, 600–612.

Winnemöller, H. (2011). Xdog: advanced image stylization with extended difference-of-gaussians. In *Proceedings of the ACM SIGGRAPH/eurographics symposium on non-photorealistic animation and rendering* (pp. 147–156).

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 1395–1403).

Xu, D., Ouyang, W., Alameda-Pineda, X., Ricci, E., Wang, X., & Sebe, N. (2017). Learning deep structured multi-scale features using attention-gated crfs for contour prediction. *Advances in neural information processing systems*, *30*.

Yang, S., Jiang, L., Liu, Z., & Loy, C. C. (2022). Unsupervised image-to-image translation with generative prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18332–18341).

Zhang, H., Zu, K., Lu, J., Zou, Y., & Meng, D. (2022). Epsanet: An efficient pyramid squeeze attention block on convolutional neural network. In *Proceedings of the Asian Conference on Computer Vision* (pp. 1161–1177).

Zhang, Z., Xing, F., Shi, X., & Yang, L. (2016). Semicontour: A semi-supervised learning approach for contour detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 251–259).

Zhao, Y., Wu, R., & Dong, H. (2020). Unpaired image-to-image translation using adversarial consistency loss. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16* (pp. 800–815). Springer.

Zhou, Z., Liu, X., Shang, J., Huang, J., Li, Z., & Jia, H. (2022). Inpainting digital dunhuang murals with structure-guided deep network. *ACM Journal on Computing and Cultural Heritage*, *15*, 1–25.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 391–405). Springer.

Jinye Peng

https://orcid.org/0000-0003-4286-2576

Employment (1)

Northwest University: Xi'an, CN

Employment

Source: Jinye Peng

Record last modified Apr 29, 2024, 1:32:03 AM

Credit Author Statement

Zhe Yu: Conceptualization, Methodology, Software, Writing- Original draft preparation.

Shenglin Peng: Data curation, Supervision.

Shuyi Qu: Investigation, Validation.

Qunxi Zhang: Resources.

Jun Wang: Formal analysis, Visualization.

Jinye Peng: Writing- Reviewing and Editing.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: