2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Two-stage occlusion giant panda image inpainting based on partial convolutions, multi-scale contextual attention and a new PatchGAN with two discriminators

Xingchen Dong[1], Zhiwu Liao[1, ✉], ChenPeng[2, ✉]

[1] School of Computer Science, Sichuan Normal University, No. 1819, Section 2, Chenglong Avenue, Longquan District, Chengdu, 610101
20060097@sicnu.edu.cn (Zhiwu Liao)

[2] Chengdu Research Base of Giant Panda Breeding, Sichuan Key Laboratory of Conservation Biology for Endangered Wildlife, Chengdu 610086, China

capricorncp@163.com (Peng Chen)

Abstract. Image-based individual recognition of giant pandas in real wild scenes suffers from difficulties such as occlusion and multiple postures. Image painting is an important preprocessing step in solving the occlusion of giant panda images. A two-stage inpainting model of giant panda images based on partial convolutions and multi-scale features is proposed. In the coarse inpainting, the structural information of the occluded part is restored, while the fine inpainting focuses on restoring details such as textures and edges based on the coarse inpainting. A new PatchGAN with two discriminators in the coarse inpainting balances two-scale information guided by the WGAN-GP loss and L1 norm. The generator of the new PatchGAN uses partial-convolutions to avoid the propagation and influence of misinformation in the occluded area. In the fine inpainting, a new proposed module fusing multi-scale feature by contextual attention is added to the PatchGAN. The fine inpainting model learns and enhances the texture and details of the image output by the coarse inpainting through the global searching for multi-scale similar image patches by multi-scale context attention. Thus, it can strengthen the semantic connection between occluded and real image regions. In order to achieve better multi-scale inpainting results, the perceived loss and style loss are added to the adversarial loss. Compared with state-of-art methods, the proposed method can effectively restore image textures and details while suppressing noise and artifacts from visual effects. The PSNR, SSIM and FIN of proposed method can achieve 35.69, 0.971 and 5.22 respectively, indicating that proposed method can obtain satisfied inpainting results.

Keywords: image inpainting, dual discriminator, multi-scale context attention, partial convolution, PatchGAN.

1 Introduction

Deep learning can significantly improve the accuracy and efficiency of animal individual identification in animal conservation [1]. However, most of existing animal identification technologies are only carried out on animal faces under ideal conditions [1-4], and cannot be applied to identify wild animals in the natural environment. Individual identification of wild giant pandas in wild environments presents many difficulties such as the complexity and variability of panda postures, occluded regions of panda by natural environments, and environmentally induced light changes and shadows etc.

Especially for occluded Giant panda images, the image inpainting for occluded regions while preserving important features is a key and difficult important preprocessing step. This study aims to address the challenges in inpainting panda images due to their color-changed fur and features that are inconspicuous and with different sizes but are useful in individual identification, should be recovered precisely and correctly.

The specific features on the panda's face and body, such as the shape and size of the eye circles and ears, as well as their specific layout and the delicate textural features of fur, place highly specific demands on image restoration tasks, see Fig 1. When these key features at different levels are occluded, it becomes very difficult to infer the occluded parts only based on existing data and a single learning model.

Fig. 1. The images of Giant Pandas (GPs). Each GP has specific features on its face and body that are important for individual identification, such as the shape and size of the eye circles and ears, as well as their specific layout and the delicate textural features of fur. Thus, image inpaiting for occluded GPs is a highly specific image restoration task, which must maintain and correctly restore GP's complex facial and body features.

In computer vision, many efforts have been proposed recently in image inpainting. Deepak Pathak et al. [5] proposed Context Encoder, which trained a Convolutional Neural Network (CNN) to generate the arbitrary image regions based on the image surroundings by context-based pixel prediction. Yang et al. [6] used multi-scale patches for high-resolution image restoration and proposed a framework combining context encoder and style transfer to deal with high-resolution restoration in a three-layer pyramid approach. In 2017, Satoshi Iizuka et al. [7] introduced local and global discriminators into the fully convolutional network to fine-tune the texture, and achieved the semantic consistency between the restored area and the overall image. Yan et al. [8] proposed a Shift-Net network to inpaint images by introducing a special skip connection layer in

the U-Net network [9]. However, the repainted texture details of above methods may be less than ideal, and the ability to fill in missing parts in complex scenes needs to be improved.

At the ICCV 2019, Yu et al. proposed GateConv [10], which adopts gated convolutions. However, the computational complexity of gated convolution is relatively high, which may limit its application in some scenarios that are limited to computational resources, and the accuracy of free-form restoration for complex scenarios should be further improved.

Recently, some studies have introduced attention mechanisms to capture remote information in images. Zeng et al. developed an image inpainting technique based on pyramidal attention, PEN-NET [11], which guides shallow feature filling through deep semantic features to ensure coherence between textures. However, the pyramid structure may lead to an increase in the number of model parameters and more difficult training. Yu et al. [12] and Zeng et al. [13] proposed methods of attention mechanism and Transformer model for long-distance information capture and image inpainting in 2022, respectively. In order to restore high-resolution images, scholars have proposed new frameworks, such as capturing remote context information by Zeng et al. in 2022 [14], and the GAN-based restoration technology proposed by Yang et al. [15]. But aforementioned methods tend to generate over-smoothed inpainting regions or artifacts.

Stable diffusion [16] are proposed in 2022 and soon became an important tool for image inpainting [17-18]. Since the inpainting of stable diffusion is performed only on local information, stable diffusion-based methods can't deal with large occluded regions.

Chen et al [19] proposed a new convolution operator, Partial Convolution (PConv), to extract spatial features more efficiently by reducing computation cost. PConv selects the features of a portion of the channels for regular convolution and leaves the features of the remaining portion of the channels unchanged, to reduce the computational complexity.

VideoWorld [20] uses the Vector Quantized Variational Autoencoder (VQ-VAE) and Autoregressive Transformer architectures to generate high-quality video frames and infer task-relevant operations from these frames.

However, above one-stage image painting by deep learning models can't achieve satisfied results when inpainting panda images because of *error recovered texture* and *missing shape structure features*. Two-stage methods that progressively extrapolated information from coarse to fine becomes a viable alternative and showed more satisfied inpainting performance compared with one-stage methods.

Nazeri K et al. proposed a two-stage inpainting model, CM-GAN [21] in 2021. CM-GAN embedded pre-trained GAN into a U-shape DNN, which is helpful to deal with some invalid features and better inject global context into the spatial domain. But CM-GAN will generate some small artifacts and blur edges and textures.

Li [22] proposed a two-stage model to inpaint regular shape regions using Channel-Coordinate attention mechanism and multi scale features. But small targets may be missed in global pooling due to feature dilution. In addition, when there are a large number of densely distributed small targets in the image, the attention maps in two

directions may not be able to accurately distinguish the positions and directions of neighboring targets, leading to attention confusion and many artifacts.

To improve the performance of existing methods, recovering the big features and small details correctly is an essential measure. Thus, two discriminators are adopted to judge the coarse inpainting both in small-scale and large-scale simultaneously. In order to avoid the spread of misinformation in occluded areas, partial convolutions are used to replace the convolutions in U-net and a large weight is set to the pixel reconstruction loss of the occluded region to ensure recovering the information of occluded regions firstly.

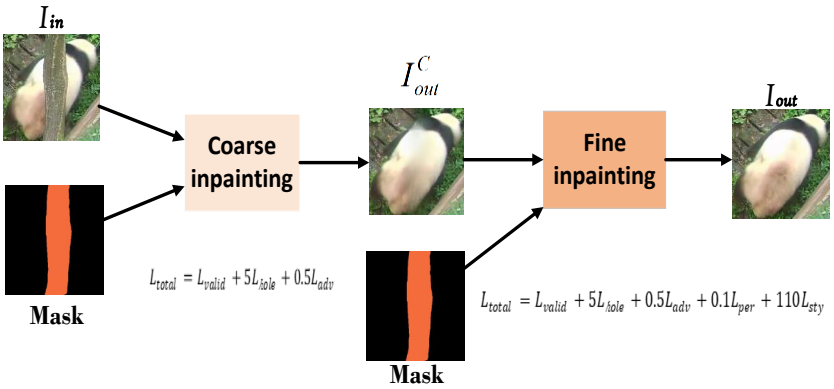

Fig. 2. Structure of proposed two-stage inpainting model. The occluded giant panda image I_{in} and its occluded mask are fed into the coarse inpainting model to output I^{C}_{out}. The superscript "C" of I^{C}_{out} indicates that it is the inpainting image in the coarse inpainting stage. Then, the fine inpainting model are get by the I^{C}_{out} and the same mask as the coarse inpainting stage to get fine inpainting image I_{out}. The total losses of coarse inpainting and fine inpainting are shown under the stages, indicating that the objective of the coarse inpainting is to recover structure features in occluded regions while fine inpainting focuses on recovering textures and details in occluded regions by setting big weights to reconstructed loss L_{hole} in the occluded regions and style loss L_{sty}.

In fine inpainting, in order to preserve small features, unlike fusing multi-scale contextual by concatenation and Channel-Coordinate attentions in [22], multi-scale contextual in our framework are fused by concatenation and Squeeze-and-Excitation (SE) [23] attention, which preserves details through feature recalibration in the channel dimension: enhancement of detail-related channels to suppress background noise and highlight channels corresponding to details such as textures and edges, and cross-layer circulation of detail information to preserve semantic cues of details and mitigating information loss due to down-sampling.

The main contributions include:

1. Construct a GP's occluded image set, iPanda-50-occlusion, which contains 1000 images of giant pandas with various occlusion situations and is constructed from the iPanda-50 dataset.

2. Propose a new framework for coarse inpainting to recover structures of occluded panda images. The framework includes a partial convolutional-Uet generator to avoid the propagation and influence of misinformation in the occluded areas; two discriminators to judge whether large-scale patch and small-scale patch are genuine or fake; WGAN-GP and pixel reconstruction loss with large weight in occluded regions.
3. Proposed a fine inpainting model based on multi-scale contextual attention and PatchGAN. The multi-scale contextual information is fused through the SE module to preserve details. The cost of multi-scale information is pixel reconstruction loss, adversarial loss, perceptual loss and style loss with far large weighs to focus on the recovering the textures and details of occluded regions.

The remain of this paper is: the proposed method is introduced in section 2; the experiments and discussions are presented in section 3; Finally, it is the conclusions and acknowledgement.

2 Method

2.1 Model Framework

A two-stage inpainting model is proposed to inpainting the occluded parts of the giant panda images. The network consists of two parts: coarse inpainting and fine inpainting, both of which adopt generative adversarial networks (GANs), and the overall network structure is shown in Figure 2.

The mask is a matrix whose elements are zeros or ones and with the same dimensions as the panda image. The occluded parts in M are ones while the other parts are set to zeros. After coarse inpainting, the generated image and its mask are fed into the image fine inpainting to further restore details and textures of occluded areas.

2.2 Coarse inpainting based on dual PatchGAN

2.2.1 Network structure of coarse inpainting

The data are fed into the coarse inpainting network includes: an image of a giant panda with occlusion and a mask of the occluded region. The generator is a partial convolution Unet, whose convolutions are replaced by the partial convolutions. The network structure of coarse inpainting is shown in Figure 3. The generator includes three key parts: encoder, decoder and long-skip connections. The encoder consists of six convolutional layers, each down-sampled using a 3×3 convolutional kernel with stride=2. The convolution kernel of the decoder is 3×3 and the stride=1 to keep the size of the feature map. The inputs to the decoder are the channels of aggregating the feature map of the current decoder layer with the feature map of the corresponding size in the encoder.

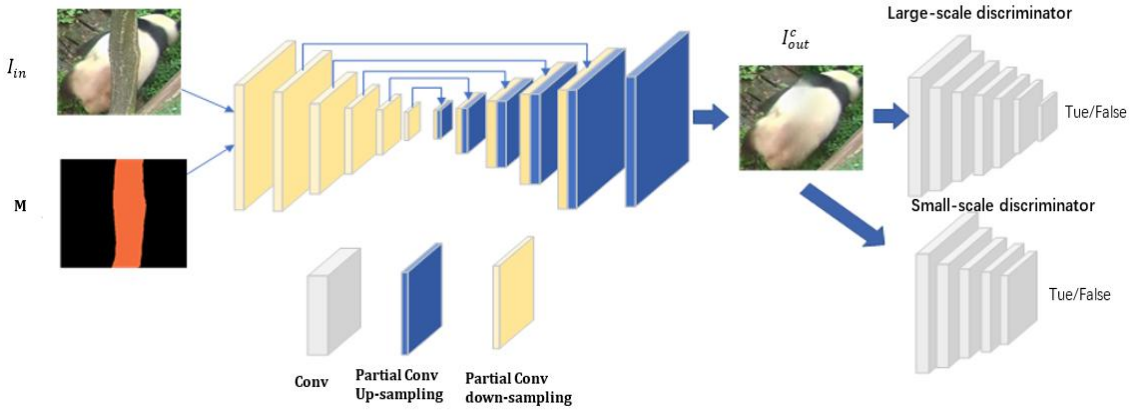

Fig. 3. Network structure of coarse inpainting. The generator includes an encoder, a decoder, and long skip connections. The encoder consists of six convolutional layers, each using a 3×3 convolutional kernel with a stride of 2. The decoder uses 3×3 convolutional kernels with strides of 1 to maintain the feature map size. Two discriminators: a large-scale PatchGAN discriminator $\boldsymbol{D_L}$ with seven convolution layers transforms the dimension of the image from 224×224×3 to 7×7×256; while a small-scale PatchGAN discriminator $\boldsymbol{D_S}$ with five convolution layers transforms the image size to 14×14×256.

2.2.2 Dual PatchGAN discriminators.

Traditional discriminators in GAN focus on the global information of the generated image, and judge the authenticity of the whole image by outputting a scalar value 0 or 1. Thus, GAN is clearly not sufficient for the tasks that need to ensure the repaired boundaries are not broken or jumped.

Unlike traditional GAN discriminators, PatchGAN does not attempt to assess the authenticity of the entire image, but rather local regions (or "patches") of the image. The core idea of PatchGAN is to determine the authenticity of an entire image by segmenting the input image into small chunks or "patches", and then determining whether each chunk is authentic or fake separately.

Although single-scale PatchGAN can achieve visually detailed restoration, it tends to *ignore contextual information in whole images*. At the same time, the overlaps of local image blocks may also lead to the inability to effectively discriminate the structure of the image.

In order to solve the above problems, we propose a new PatchGAN with dual discriminators, which simultaneously uses large-scale D_L and small-scale discriminators D_S to analyses the image. The discriminator network structure is shown in Figure 3, the right top is the large-scale discriminator D_L used to capture the global information, while the other is the small-scale discriminator D_S focusing on local features. These two discriminators work together, so that we make judgements based on information from two scales and ensure reliability both of global structures and local details.

2.2.3 Loss

In coarse inpainting, we use pixel reconstruction L1 loss and adversarial loss. The L1 loss is a direct measure of the pixel-level difference between the inpainting image and the real image, emphasizing the importance of accurate reconstruction for each pixel. By minimizing this loss, the network learns how to accurately repair the occluded regions while preserving the original pixel values of the un-occluded regions as much as possible. There are two parts of the pixel reconstruction loss:

$$L_{valid} = \frac{1}{\sum(\mathbf{1}-M)} \left\| \left(I_{out}^{c} - I_g\right) \odot (\mathbf{1} - M) \right\|_1 \tag{1}$$

$$L_{hole} = \frac{1}{\sum(M)} \left\| \left(I_{out}^{c} - I_g\right) \odot M \right\|_1 \tag{2}$$

where I_{out}^{c} is the coarse inpainting image and C means "coarse", I_g is the real image, M is the occlusion mask whose elements are zeros or ones, $\odot$ represents the element-wise multiplication, $\sum$ represents the summation of the elements of M, $\mathbf{1}$ is a matrix whose elements are ones and with the same size of M, and the $||\cdot||_1$ represents the L1-norm. Thus, L_{valid} is the pixel reconstruction loss of background regions (un-occluded regions) while L_{hole} is the pixel reconstruction loss of foreground regions (occluded regions).

The adversarial loss Wasserstein GAN with Gradient Penalty (WGAN-GP) [24] was used to adversarial loss. WGAN-GP limits the range of gradient variation by adding a gradient penalty term to the loss function. The gradient penalty is achieved by computing the gradient and applying the penalty at points interpolated between the real and generated data. This approach allows the inpainting network to not only reconstruct the image at the pixel level, but also learn the distribution of the real giant panda image at the distribution level, thus generating visually more realistic and natural restoration results. The WGAN-GP is defined as

$$L_{adv} = E_{I_{out}^{c} \sim P_{out}^{L}}[D_L(I_{out}^{c})] - E_{I_g \sim P_g^{L}}[D_L(I_g)] + E_{I_{out}^{c} \sim P_{out}^{S}}[D_S(I_{out}^{c})] -$$
$$E_{I_g \sim P_g^{S}} D_S(I_g)] + \lambda_1 E_{\hat{I} \sim P_{\hat{I}}^{L}}(\left\|\nabla_{\hat{I}} D_L(\hat{I})\right\|_2 - 1)^2 + \lambda_2 E_{\hat{I} \sim P_{\hat{I}}^{S}}(\left\|\nabla_{\hat{I}} D_S(\hat{I})\right\|_2 - 1)^2 \tag{3}$$

where D_L is the output of the large-scale PatchGAN discriminator and D_S represents the output of the small-scale PatchGAN discriminator. λ_1 and λ_2 are the weight coefficients of the gradient penalty, and $\hat{I}$ represents the sample points randomly inserted between the real image and the inpainting image. ∇ is the gradient operator.

The overall cost function:

$$L_{total} = L_{valid} + \lambda_h L_{hole} + \lambda_{adv} L_{adv} \tag{4}$$

where λ_{hole} and λ_{adv} are the weight parameters that control the importance of different losses, are set to 5 and 0.5 respectively. Note that λ_{hole} is five times as large as the weight of L_{valid} and ten times as large as the weight of L_{vad}. Thus, accurately restoring pixel values of the occluded regions is the most important task in model training.

2.3 Fine inpainting

2.3.1 Network structure of fine inpainting.

The fine inpainting network first performs feature extraction on the input image through a series of convolutional layers for down-sampling. These features are then fed into a multi-scale contextual attention module. Finally, the network uses deconvolution to up-sample the feature map back to the size of the original image. The discriminator is PatchGAN. The network structure of fine inpainting is shown in Figure 4.

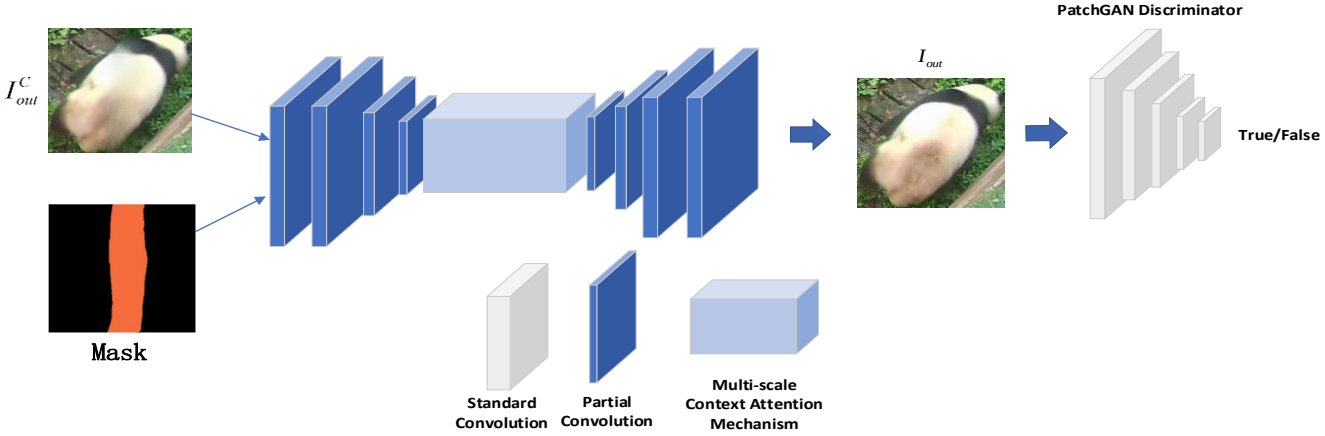

Fig. 4. Network structure of fine inpainting. The generator includes an encoder, a decoder and a multi-scale context attention module. The discriminator is a PatchGAN discriminator.

2.3.2 Fusion module of multi-scale contextual attention.

In the fine inpainting stage, we want to create an image that blends the occluded portion seamlessly with the background region. One strategy to achieve this aim is patch matching at the feature level such as Transformer, where missing parts are reconstructed by borrowing or copying features from known background areas. However, choosing the right size of matching patch becomes challenging due to the different detail and style. In general, large patch sizes are more suitable for style preservation, while small patch sizes provide more flexibility in reusing background features. Single-scale patch matching limits the application of the model to different scenes. We propose a multi-scale contextual attention module that flexibly selects and utilizes contextual information based on the overall style and contents of the image.

In the fusion module of multiscale contextual attention, the input feature map is directed to two parallel branches to generate an attention map and two sets of attention features, one of branch which uses a 3x3 size patch and the other uses a 1×1 size patch respectively. The two groups of attention features are fused to form fused multi-scale attention features, see Figure 5.

The feature map is divided into two regions before processing: the foreground (i.e., the occluded pixel area) and the background area (i.e., the un-occluded pixel area). Achieving the recovery of the foreground region using distance information requires

the identification of the background region that is critical to recover the foreground region.

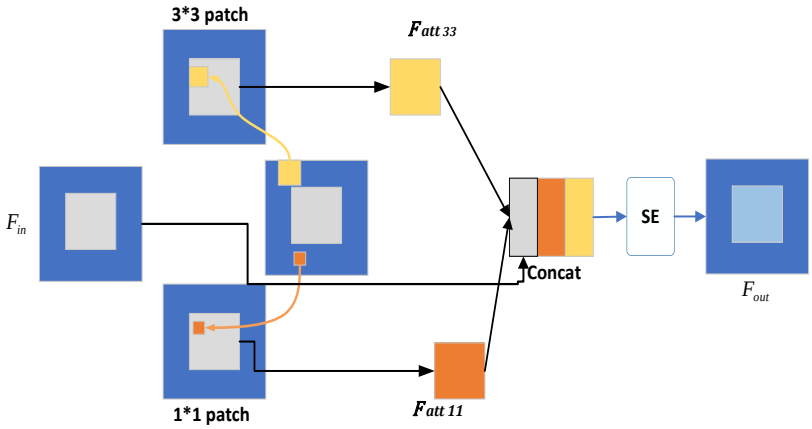

Fig. 5. Proposed fusion module of Multi-scale Context Attention. The input feature $\boldsymbol{F_{in}}$ is divided into foreground regions (foreground regions are represented by the inner gray squares and are the occluded regions restored by coarse inpainting) and background regions (background regions are represented by the blue parts and are the unoccluded regions). 3×3 and 1×1 patches of background regions are matched with the same scale patches within the foreground region and the similarity are calculated. The feature maps $\boldsymbol{F_{att11}}$ and $\boldsymbol{F_{att33}}$ based the attention scores of the pairs of patches are reconstructed from the convolution kernel. The foreground of $\boldsymbol{F_{in}}$, $\boldsymbol{F_{att11}}$ and $\boldsymbol{F_{att33}}$ are fused using SE to obtain the output feature maps $\boldsymbol{F_{out}}$.

The cosine similarity between each foreground pixel and a background patch are defined as the similarity between the image patch centered on the foreground pixel and the background patch. Thus, patches with the same size of the foreground patch are extracted in the background region, and then are used as filters and convolved with the foreground region to derive the similarity of each pixel of the foreground region with respect to the background patches.

The softmax function is applied to the channel dimension to obtain the attention score of each background patch with respect to the foreground position.

Finally, the patch with the highest score is selected as a filter for the inverse convolution operation to reconstruct the foreground region, and the overlap is averaged. The similarity between the foreground pixel (x, y) and background patches is calculated as follows:

$$S_{x,y,x',y'} = \frac{f_{x,y} \cdot b_{x',y'}}{\|f_{x,y}\| \cdot \|b_{x',y'}\|} \tag{5}$$

where $S_{x,y,x',y'}$ represents the similarity between the location (x, y)-centered foreground patch and the location (x', y')-centered background patch. $f_{x,y}$ represents the

foreground patch centered at (x, y), $b_{x',y'}$ represents the background patch centered at (x', y'). The attention score is calculated as follows:

$$S^*_{x,y,x',y'} = soft\,max\ \ (rS_{x,y,x',y'}) \tag{6}$$

where $S^*_{x,y,x',y'}$represents the score of attention, r is a constant, $softmax$ is the softmax function.

The propagation attention score of the $(2k+1) \times (2k+1)$ patch size is calculated as follows:

$$S^*_{x,y} = \sum_{i=-k}^{k} S^*_{x+i,y+i,x'+i,y+i} \tag{7}$$

We use attention scores to recover feature maps from patches at different scales by deconvolution. The feature map obtained by the convolutional kernel 1×1 is denoted as F_{att11}, and the feature map obtained by the convolutional kernel 3×3 is denoted as F_{att33}.

The feature maps of these two scales are fused with F_{in}. Finally, aa SE module [21] is used to process the fused feature map, and the final optimized output feature map is obtained by

$$F_{out} = SE(\mathrm{M} \cdot F_{in}, F_{1\times1}, F_{3\times3}) \tag{8}$$

where M is the Mask, $M \cdot F_{in}$ is the occluded region of the panda image.

2.3.3 Loss of fine inpainting.

We use L1-nom. There are two parts of the pixel reconstruction loss:

$$L_{valid} = \frac{1}{\sum(\mathbf{1}-M)} \left\| (I_{out} - I_g) \odot (\mathbf{1} - M) \right\|_1 \tag{9}$$

$$L_{hole} = \frac{1}{\sum(M)} \left\| (I_{out} - I_g) \odot M \right\|_1 \tag{10}$$

where I_{out} is the generated image by the fine inpainting, $I_g$is the real image, M is the occlusion mask, $\odot$ represents the element-wise multiplication, $\sum$ represents the summation of M's elements, and the $||\cdot||_1$ represents the L1-norm.

The adversarial loss is WGAN-GP with the following adversarial loss:

$$L_{adv} = E_{I_{out}\sim P_{out}} D(I_{out}) - E_{I_g\sim P_g} D(I_g) + \lambda E_{\hat{I}\sim P_{\hat{I}}} (\left\| \nabla_{\hat{I}} D(\hat{I}) \right\|_2 - 1)^2 \tag{11}$$

In the fine inpainting, the perceptual loss and style loss are introduced to further improve the inpainting quality. The perceptual loss is:

$$L_{per} = \sum_l \left\| \Phi_l(I_{out}) - \Phi_l(I_g) \right\| \tag{12}$$

where $\Phi_l(\cdot)$ represents extracted features from the lth layer of VGG.

The style loss is:

$$L_{sty} = \sum_l \left\| G_l(I_{out}) - G_l(I_g) \right\| \tag{13}$$

where $G_l(\cdot)$ is the Gram matrix of extracted features from the lth layer of VGG. The Gram matrix is the inner product of the vectorized feature mappings, which captures the tendency of features to appear simultaneously in different parts of the image. Thus, Gram matrix is an important method to describing the textures in images.

The overall loss is：

$$L_{total} = L_{valid} + \lambda_{hole} L_{hole} + \lambda_{adv} L_{adv} + \lambda_{per} L_{per} + \lambda_{sty} L_{sty} \tag{14}$$

The hyperparameters of the loss function are set to λ_{hole}=5, λ_{adv}=0.5, λ_{per}=0.1, λ_{sty}=110. Note that, λ_{sty} is far larger than other hyperparameters, indicating that the model training focuses on recovering the textures in images.

3 Experiments and discussions

3.1 Data

Most datasets for occlusion restoration are based on predefined Mask datasets, such as NVIDIA irregular Mask Dataset [25], and QD-IMD dataset [26]. Although the inpainting model can perform well on these non-real occlusion datasets with predefined masks, its performance is often difficult to meet demands in practical application scenarios.

Moreover, although under ideal conditions, animal individual recognition techniques, especially for animal facial recognition algorithms [8-11], have achieved high accuracy, especially under the training of large amounts of data. However, in real wild environments, GP individual recognition still face multiple challenges, such as changes in the posture of GPs, uncertainty in lighting conditions, and interference from occlusion.

Thus, a dataset iPanda-50-occlusion specialized for giant panda occlusion restoration are constructed. The iPanda-50-occlusion contains 1000 images of giant pandas with various occlusion situations and is constructed from the iPanda-50 dataset, which is an open dataset to prompt researches on fine-grained panda identification. It provides 6,874 high-quality images covering 50 different pandas.

In order to simulate the occlusion that giant pandas may encounter in their natural living environment, we pay special attention to the natural elements closely related to the daily life of giant pandas, including trees, bamboo, flowers and grass etc. The extent and location of the occlusion reflected the real life situation as much as possible. This includes the occlusion of the giant panda's back, eyes, face and other critical and non-critical parts, simulating various degrees of occlusion from mild to severe.

To enhance the usability of the dataset, the occlusions of each image in the dataset are carefully labeled and segmented, and the corresponding occlusion segmentation map is generated.

We divided the iPanda-50-occlusion dataset in a ratio of 2: 1 between the training set and the test set. In the preprocessing stage, the horizontal flip plus rotation are used to enhance the data.

In the model training stage, a certain round of pre-training was carried out on the coarse inpainting network, and then end-to-end joint training was carried out. The pre-training was set at 20 rounds and the end-to-end training was done with 200 rounds.

In the model training phase, we chose the Adam optimizer to optimize the network parameters, which β_1 were set to 0.9, β_2 were set to 0.999. For the generator parameters, we used a batch size of 8 and a base learning rate of 10-4 for iterative updating, while the learning rate of the discriminator was set to one-tenth of that of the generator.

3.2 Evaluation index

The evaluation indexes used in include Peak Signal to Noise Rate (PSNR), Structural Similarity Index Measure (SSIM) and Fréchet Inception Distance (FID). PSNR evaluates the image quality based on the ratio between the maximum possible pixel value and the errors of the image. Higher PSNR indicates better quality of the restored image. The PSNR is defined as:

$$PSNR = 20 \cdot \boldsymbol{log}_{10}(\frac{MAX_I}{\sqrt{MSE}}) \tag{15}$$

where MAX_I is the maximum possible pixel value of the image (usually 255 for 8-bit images), and MSE is the mean square error between the original image and the restored image. MSE is defined as:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I_g(i,j) - I_r(i,j)]^2 \tag{16}$$

where I_r is the real image, I_g is the image generated by proposed method. m and n are the length and width of the image respectively.

SSIM takes into account the structural information, contrast and brightness of an image, and therefore more fully reflects the human eye's perception of image quality. The value of SSIM lies between -1 and 1, with higher values indicating better image quality. The formula for SSIM is

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{17}$$

where x and y are the original and restored images respectively, μ_x, μ_y, σ_x^2, σ_y^2, are their means and variances respectively, σ_{xy} is their covariance, and c_1 and c_2 are small constants added to avoid the denominator being zero.

FID is a metric for evaluating GAN networks, reflecting the distance between two images, the smaller the value the closer the generated image is to the real image. FID is:

$$FID = \left\|\mu_r - \mu_g\right\|^2 + Tr(\sum_r + \sum_g - 2(\sum_r\sum_g)^{\frac{1}{2}}) \tag{18}$$

where μ_r and μ_g are means of real image and generated image respectively, Tr is trace of the matrix，$\sum_r$ and $\sum_g$ is the covariance matrixes of the real image and the generated image.

Table 1. Quantitative comparison of ablation experiments

	Occlude images	Model1	Model2	Proposed method
PSNR	22.06	33.25	30.26	**34.49**
SSIM	0.910	0.964	0.955	**0.967**
FID	21.94	5.10	10.27	**4.22**

3.3 Ablation experiments

Models performed ablation experiments include: models without fusion module of multi-scale contextual attention mechanisms are denoted as Model 1, models without dual PatchGAN discriminators and using ordinary discriminators are denoted as Model 2. SSIM, PSNR, and FID were used as evaluation indexes. The results of the ablation experiment are shown in Table 1.

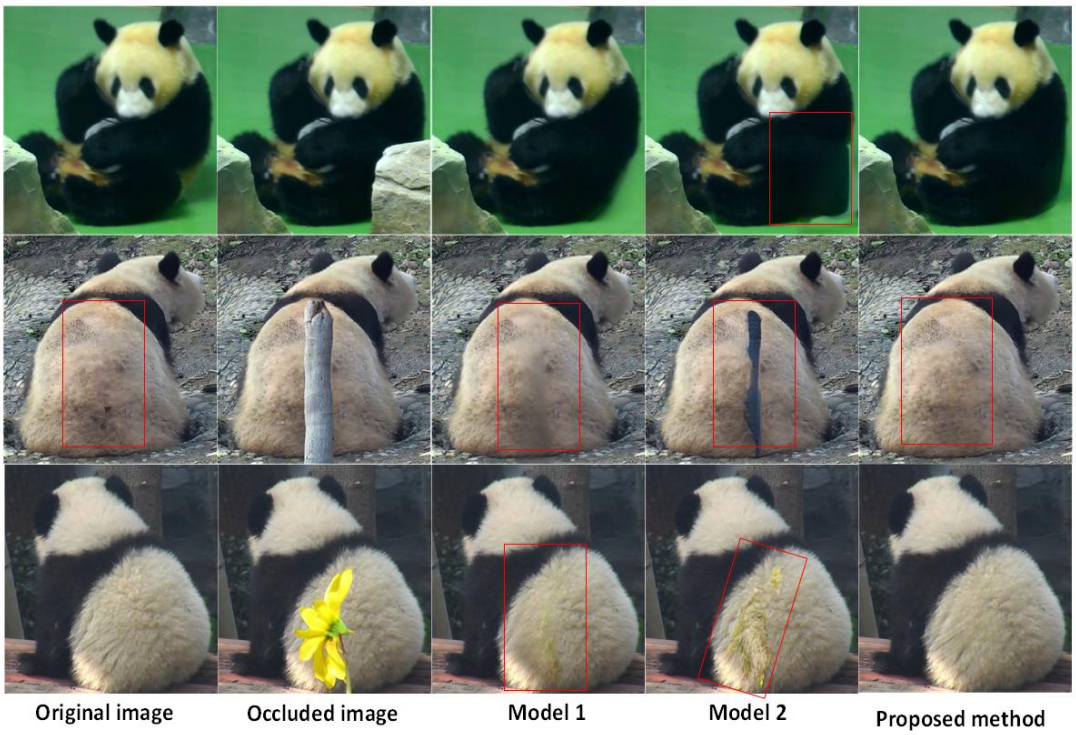

Fig. 6. The visual inpainting results of the ablation experiments.

From the experimental results, the model with two-stage inpainting performs the best with the highest PSNR 34.49, the highest SSIM 0.967 and the lowest FID 4.22. The PSNR and SSIM of proposed method is higher than Model1 by 1.24 and 0.003 respectively while FID of proposed method is lower than Model1 0.88. Small difference between the proposed method and Model1 indicated that multi-scale contextual fusion is a limited role in restoring image structure.

Model2, which is missing the two-discriminator PatchGAN, performs worse in all evaluation indexes than Model1. That is, the PSNR and SSIM of Model1 are higher than Model2 by 2.99 and 0.009 respectively while FID of Model1 is lower than Model2 by 5.17. The SSIM of Model1 are higher than Model2 by 0.009 indicating that the dual discriminator is very important in inpainting structures of images.

Figure 6 shows the visual effects comparison of the ablation experiments. When the multi-scale contextual attention fusion module is missing, the output images of Model1 appears blurry in occlusion areas and the details are not well restored. When the dual PatchGANs are missing, there are part of occluded objects in the inpainting area. Thus, Model2 fails to restore the normal structure of the image.

3.4 Comparison experiments

We compared proposed method with existing state-of-art image inpainting methods, including: GateConv [10] proposed in 2019, PEN [11] proposed in 2019 and PConv [19] proposed in 2023, CM-GAN [20] proposed in 2021, Stable diffusion [16] proposed in 2022 and VideoWorld [24] proposed in 2025.

Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Fréchet Inception Distance (FID) were selected to quantitatively evaluate the quality of the restored giant panda images. Tab 2 shows the comparison results.

Table 2. Comparison results of different methods in terms of PSNR, SSIM, and FID.

	Occluded images	**PConv** [19] 2023	**PEN** [11] 2019	**GateConv** [10] 2019	**CM-GAN** [20] 2021	**Videoworld** [24] 2025	**Stable diffusion** [16] 2022	**Proposed method**
PSNR	20.35	30.99	31.94	29.97	35.59	34.71	26.97	**35.69**
SSIM	0.905	0.951	0.961	0.963	**0.972**	0.921	0.891	0.971
FID	21.21	12.53	8.52	7.47	5.33	5.35	5.80	**5.22**

From Tab 2, all deep learning models dramatically improved the quality of Occluded images on three metrics. Among them, PConv is with the worst performers but also improved PSNR, SSIM and FID of occluded images by 10.64, 0.046 and 8.68 respectively while proposed method improved occluded image 15.36, 0.066 and 15.99 respectively, demonstrating the validation of deep learning in image painting. The PSNR, SSIM and FID of our proposed method were 35.69, 0.971 and 5.22 respectively, which were better than PConv, PEN, GateConv, Videoworld and Stable diffusion. And in the comparison with CM-GAN, we achieved better performance in PSNR and FID, and slightly weaker than it on SSIM. These results show that proposed method has satisfied image inpainting performance and can restore the structure and details of the image well.

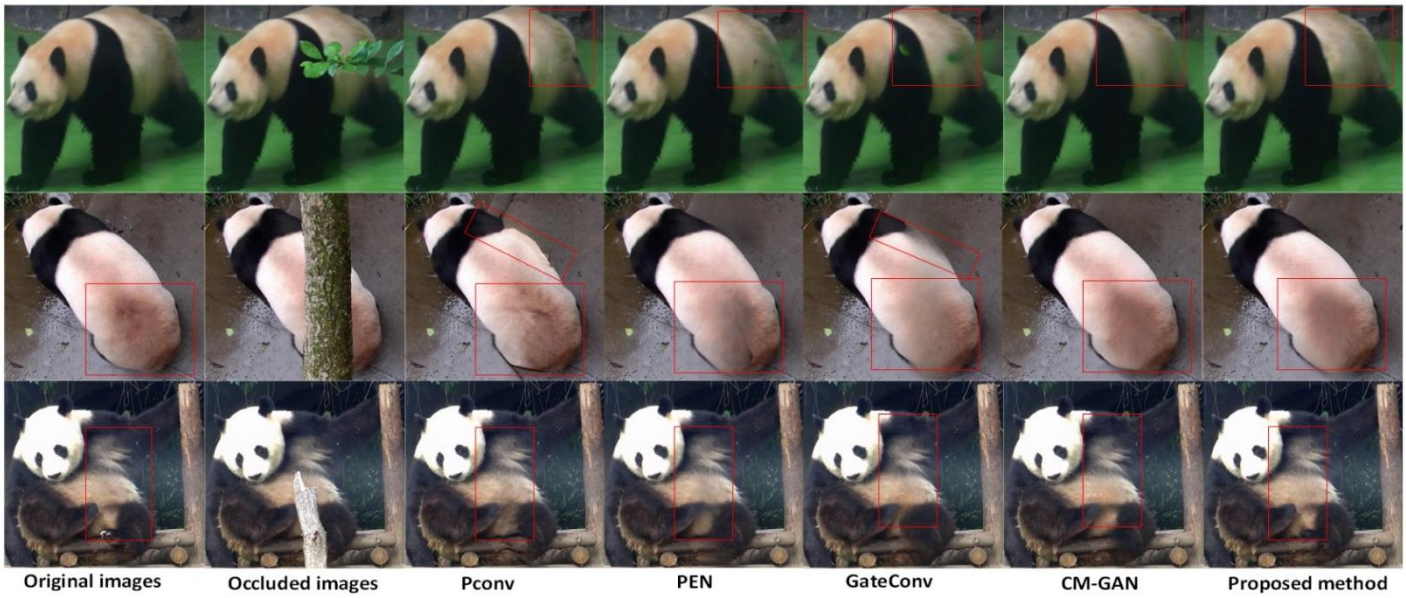

Fig. 7. Visual comparison of traditional inpainting models.

Except for the proposed method, CM-GAN has the best performance in other six models showing that power of two-stage inpainting. The new Videoworld proposed in 2025 has the best performance in all one-stage inpainting methods exhibiting promising power of large language models in image inpainting.

Since textures and details of panda images are very important in individual panda identification, correctly recovered the textures and details of occluded images becomes indispensable stage in individual panda identification. In order to show the textures and details more clearly, the visual comparison is divided into two groups: one group is traditional methods including PConv, PEN, GateConv and CM-GAN; the other group is new methods including Videoword and Stable diffusion. Each group contains original images, occluded images, images recovered by the proposed method.

As shown in Figure 7, the PConv makes the most serious artifacts in all method indicated by the square regions. The PEN and GateConv shows some blurriness and artifacts. The GateConv has better visual perceptions than the PEN with few artifacts and less smoothing. Two slanted rectangles indicated the over-recovered body by PConv and uncovered body by GateConv.

The CM-GAN coexists blurry and artifacts: smoothing out the textures on the back of the body in the first image but enhancing the furs in the last image. Its ability to maintain most of the detail and produce a small number of artifacts makes it the most visually appealing in comparison methods except for the proposed method. The proposed method effectively restores the structure and details of the image.

Observing Figure 8, two new models have created some illusions: Stable diffusion and VideoWorld creates two ‘perfect’ feet in the first image, a big leg in the second

image; Stable also creates two feet with clear nails and imaginary postures in the third image. They also make different fur styles to the original image: Stable diffusion makes yellow fine curl furs while VideoWord strengths the textures of furs and makes furs with wild styles. These illusions change many important features of panda which will lead to fail to recognize individual panda.

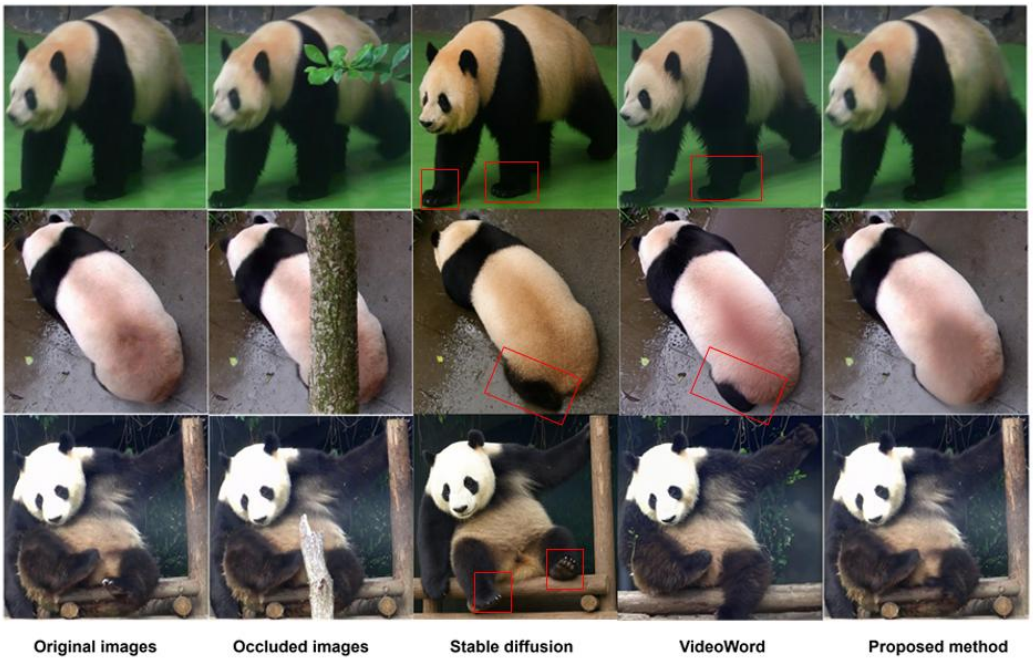

Fig. 8. Visual comparison results of new proposed methods: Stable diffusion and VideoWord.

4 Conclusion

In this paper, we propose to inpainting the occluded part of the giant panda image use a two-stage method. The method composed by coarse inpainting and fine inpainting, achieves the accurate inpainting the occluded region of the giant panda images. Compared with the SOTA method, the structure, texture and details of the occluded region are restored better. Through ablation experiments, we found:

1. Dual PatchGAN is important in giant panda occluded region inpainting. The PSNR, SSIM and FID of model2 without dual PatchGAN are much worse than the proposed method, while model1, which only removes the multi-scale information, is not much different from the proposed method in the three indexes.
2. The performance of two-stage inpainting model is better than the one-stage method. The two-stage inpainting method can part the inpainting objective to two steps and each step can only focus on one target.

Acknowledgments. This research was funded by Sichuan Science and Technology Program (grant numbers: 24GJHZ0388), the Chengdu Research Base of Giant Panda Breeding (grant numbers: 2020CPB-C09, 2024CPB-B08).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Junjie Zhong, Bing Niu, Qin Chen, Xiang Chen, Yan Wang. 2023, The Application of Deep Learning in Wildlife Conservation [J]. Journal of Animal Science, 43(6): 734-744.
2. Shukla A, Cheema G S, Anand S, et al. Primate face identification in the wild[C] PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26-30, 2019, Proceedings, Part III 16. Springer International Publishing, 2019: 387-401.
3. Brust C A, Burghardt T, Groenenberg M, et al. Towards automated visual monitoring of individual gorillas in the wild[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 2820-2830.
4. Hansen M F, Smith M L, Smith L N, et al. Towards on-farm pig face recognition using convolutional neural networks[J]. Computers in Industry, 2018, 98: 145-152.
5. Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2536-2544.
6. Yang C, Lu X, Lin Z, et al. High-resolution image inpainting using multi-scale neural patch synthesis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6721-6729.
7. Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-14.
8. Yan Z, Li X, Li M, et al. Shift-net: Image inpainting via deep feature rearrangement[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 1-17.
9. 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.
10. Yu J, Lin Z, Yang J, et al. Free-form image inpainting with gated convolution. [C]// Proceedings of the IEEE/CVF international conference on computer vision. 2019: 4471-4480.
11. Zeng Y, Fu J, Chao H, et al. Learning pyramid-context encoder network for high-quality image inpainting[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1486-1494.
12. Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5505-5514.
13. Zeng Y, Fu J, Chao H, et al. Aggregated contextual transformations for high-resolution image inpainting[J]. IEEE Transactions on Visualization and Computer Graphics, 2022.
14. Zheng C, Cham T J, Cai J, et al. Bridging global context interactions for high-fidelity image completion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11512-11522.

15. Yang T, Ren P, Xie X, et al. Gan prior embedded network for blind face restoration in the wild[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 672-681.
16. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695, 2022.
17. Duque, J.J.L., Marcillo-Vera, A., Carranco, F., Casa, D., Cajamarca, G. (2024). Reviewing Inpainting Techniques Using Diffusion Models: A Comprehensive Analysis and Evaluation. In: Vizuete, M.Z., et al. Applied Engineering and Innovative Technologies. AENIT 2023. Lecture Notes in Networks and Systems, vol 1134. Springer, Cham. https://doi.org/10.1007/978-3-031-70760-5_30
18. Yikai Wang, Chenjie Cao, Yanwei Fu (2023). Towards Stable and Faithful Inpainting. https://arxiv.org/html/2312.04831v1/#S7
19. Jierun Chen, Shiu-hong Kao, Hao He Weipeng Zhuo, Song Wen, Chul-Ho Lee, S.-H. Gary Chan. (2023) Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023.
20. Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, Xiaojie Jin (2025). VideoWorld: Exploring Knowledge Learning from Unlabeled Videos. 5 March 2025. https://arxiv.org/abs/2501.09781
21. Nazeri K, Ng E, Joseph T, et al. Edgeconnect: Generative image inpainting with adversarial edge learning[J]. arXiv preprint arXiv:1901.00212, 2021.
22. Linfan Li. RESEARCH AND SYSTEM IMPLEMENTATION OF IMAGE INPAINTING BASED ON DEEP LEARNING [D]. Southwest Jiaotong University Master Degree Thesis, 2022.
23. Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Enhua Wu, Squeeze-and-Excitation Networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.2018.
24. Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans [J]. Advances in neural information processing systems, 2017, 30.
25. Liu G, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 85-100.
26. Iskakov K. Semi-parametric image inpainting[J]. arXiv preprint arXiv:1807.02855, 2018.