Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063547

ARTICLE

Multi-Scale Dilated Attention-Based Transformer Network for Image Inpainting

Jinrong Li[1,2], Chunhua Wei[2], Lei Liang[2,3,*] and Zhisheng Gao[1,*]

[1]School of Computer and Software Engineering, Xihua University, Chengdu, 610039, China
[2]China Aerodynamics Research and Development Center, Mianyang, 621000, China
[3]College of Aerospace Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China

*Corresponding Authors: Lei Liang. Email: liangl94@nuaa.edu.cn; Zhisheng Gao. Email: gzs_xihua@mail.xhu.edu.cn

Received: 17 January 2025; Accepted: 06 May 2025; Published: 03 July 2025

ABSTRACT: The Pressure Sensitive Paint Technique (PSP) has gained attention in recent years because of its significant benefits in measuring surface pressure on wind tunnel models. However, in the post-processing process of PSP images, issues such as pressure taps, paint peeling, and contamination can lead to the loss of pressure data on the image, which seriously affects the subsequent calculation and analysis of pressure distribution. Therefore, image inpainting is particularly important in the post-processing process of PSP images. Deep learning offers new methods for PSP image inpainting, but some basic characteristics of convolutional neural networks (CNNs) may limit their ability to handle restoration tasks. By contrast, the self-attention mechanism in the transformer can efficiently model non-local relationships among input features by generating adaptive attention scores. As a result, we propose an efficient transformer network model for the PSP image inpainting task, named multi-scale dilated attention transformer (D-former). The model utilizes the redundancy of global dependencies modeling in Vision Transformers (ViTs) to introduce multi-scale dilated attention (MDA), this mechanism effectively models the interaction between localized and sparse patches within the shifted window, achieving a better balance between computational complexity and receptive field. As a result, D-former allows efficient modeling of long-range features while using fewer parameters and lower computational costs. The experiments on two public datasets and the PSP dataset indicate that the method in this article performs better compared to several advanced methods. Through the verification of real wind tunnel tests, this method can accurately restore the luminescent intensity data of holes in PSP images, thereby improving the accuracy of full field pressure data, and has a promising future in practical applications.

KEYWORDS: Pressure-sensitive paint technology; deep learning; image inpainting; vision transformer; self-attention mechanism

1 Introduction

Pressure Sensitive Paint (PSP) technology, as a non-touch optical measurement technique for full-field pressure distribution on surfaces, has been significantly developed and widely applied since the late 20th century, driven by disciplines such as biochemistry, optics, information technology, and image processing technology. The technique works by coating the surface of the object with a special pressure-sensitive material and using a laser or ultraviolet lamp as an excitation light source to induce the paint to emit fluorescence or phosphorescence, and then taking the fluorescent intensity image, and using the Stern-Volmer formula to calculate the surface pressure distribution after image processing [1]. PSP technology has the benefits of wide detection range, low cost, short preparation time, and can better solve the flow field interference problem

caused by traditional detection technology. Currently, PSP technology has matured and has been used in many aerospace pressure measurement tests.

However, during the post-processing of PSP images, due to various factors such as paint peeling, setting of artificial markers, and noise from image acquisition equipment, defects such as black holes and noise often appear in the images, which can seriously affect the subsequent calculation and analysis of pressure distribution. Therefore, image inpainting is particularly important in the post-processing of PSP images. Effective image inpainting techniques can improve the image quality and ensure the precision and reliability of subsequent pressure distribution calculations, thereby further promoting the application and development of PSP technology in the aerospace field.

The current PSP technology mainly relies on traditional algorithms for image inpainting, which can be categorized into two types: diffusion-based [2] and patch-based methods [3]. Although these methods perform well when handling simple scenes or images with small damaged regions, they often struggle to generate semantically consistent and visually reasonable restoration results in complex scenarios due to their lack of understanding of high-level semantic information in images. Consequently, their practical application effectiveness is limited. In recent years, the rapid development of deep learning has opened new research directions for PSP image inpainting, however, relevant research papers in this field remain relatively scarce. We expect to achieve new breakthroughs in this field through in-depth research.

PSP image inpainting differs from regular image inpainting in that its defects are typically discretely distributed, varying in shape and size, and exhibit significant differences from surrounding pixels [4]. Since the color distribution of PSP images directly reflects the surface pressure distribution, the inpainting task requires not only visual rationality and semantic consistency but also strict adherence to the physical laws of the pressure flow field, ensuring that each luminescence intensity data point accurately represents the pressure value of the corresponding location. PSP images are characterized by texture consistency, local continuity, and the lack of complex structural information, so the inpainting model must balance local texture continuity with the precision of global pressure distribution. This high-precision requirement makes PSP image inpainting a complex task that integrates visual and physical consistency. Convolutional Neural Networks (CNNs), due to their inherent localized receptive fields, struggle to fully capture non-local relationships between features [5], limiting their applicability in PSP image inpainting. Visual Transformers (ViTs), due to their own structure, can obtain global contextual information and have the ability to acquire and store long-range dependency information, enabling them to achieve overall perception and macro understanding of images [6].

For the special requirements of PSP image inpainting, we propose a transformer network based on a multi-scale dilated attention mechanism (MDA), termed D-former. Specifically, to effectively capture key feature responses in PSP images and acquire global contextual information, we propose a dilated attention mechanism (DA) based on the locality and sparsity characteristics of the global attention mechanism in ViTs [7]. By performing self-attention calculations between sparsely selected image patches within neighboring regions, this mechanism not only effectively expands the receptive field and retains the ability to capture dependencies among relevant features but also significantly reduces the computational complexity of global attention.

To further enhance the model's performance, we introduce the multi-scale dilated attention module (MDA). This module is capable of simultaneously capturing local details and global contextual information by setting different dilation rates for DA at various attention heads, thereby efficiently acquiring the texture characteristics and global pressure distribution of PSP images. We integrate the MDA module into a transformer block and create a U-Net style-based [8] transformer architecture (D-former). This design not only

achieves a receptive field that covers the entire image at the shallow layer but also models the global dependencies between multi-scale relevant features, enabling accurate and reasonable restoration of PSP images. Experiments demonstrate that D-former exhibits higher restoration accuracy and stronger robustness in the PSP image inpainting task compared to existing inpainting methods, while achieving efficient global feature modeling with lower computational and memory resources. Additionally, we conduct extensive ablation experiments on D-former to validate the effectiveness of the MDA module and demonstrate its advantages in balancing local inpainting accuracy and the consistency of global pressure distribution.

The main contributions of our work are summarized as follows:

- We propose a novel multi-scale dilated attention (MDA) mechanism to capture critical features and global contextual information in images, while significantly reducing the computational overhead of global attention.
- The MDA module is integrated into the transformer block, and we employ the improved transformer module to model global pixel relationships.
- We integrate the proposed transformer module with a U-Net style network [8] to propose a novel inpainting network D-former, which performs superior performance over existing state-of-the-art methods.
- We specifically create a PSP image dataset and conduct extensive comparison experiments on this dataset and two public datasets, validating the restoration performance and generalization capability of our proposed model for PSP images.

2 Related Work

2.1 Image Inpainting

Before deep learning, because of the inability to understand the semantics of images, non-learning methods could only reconstruct missing regions based on local neighborhood information [2] or fill pixels based on all observed regions [3]. These methods often have better results when dealing with small damaged regions or basic background padding, while their effectiveness is restricted when facing images with complex patterns. To make the network capable of outputting semantic results, Pathak et al. [9] introduce the Generative Adversarial Network (GAN) [10] and train a conditional image generation model using CNN. Some researchers choose to use additional image information (such as edges [11], structures [12], and semantics [13]) to guide the model in completing the image. Nevertheless, several of these approaches involve multi-stage or multi-model architectures, which pose difficulties for end-to-end training. For example, CTSDG [12] studies texture and structure from each other on a two-stream network. RFR [14] gradually completes the image through multiple iterations.

CNNs have demonstrated good performance in content generation, but their inherent characteristics limit the ability of image inpainting. Specifically, (a) CNNs mainly focus on local features and are difficult to capture long-range dependencies. (b) Its convolutional kernel coefficients are spatially fixed, affecting the adaptability of the network for different inputs. Such characteristics make CNNs perform poorly in restoring large damaged images. To address this problem, Reference [15] has combined attention operation and fast Fourier convolution to enhance the capture of global dependencies, but these methods are only applied to a few low-resolution layers, leading to possible distortions when reconstructing complex scenes. PConv [16] and GatedConv [17] enhance the restoration performance by optimizing the convolution operation, but may cause color distortion and texture blurring due to the lack of global dependencies.

Recently, the remarkable achievements of the transformer model [18] in the field of natural language processing (NLP) have inspired the exploration of its potential for applications in computer vision [6,7].

For example, MFViT [19] proposes a multimodal data fusion framework that can significantly improve the classification accuracy of remote sensing images. Driven by this, efforts have been made to apply transformers to image inpainting [20,21]. These methods model long-range dependencies by employing multiple spatial self-attention modules. Despite the advantages of the attention mechanism over CNNs in certain respects, their computational complexity grows quadratically with the spatial resolution, which makes them challenging when dealing with high-resolution images.

2.2 Visual Transformer

The transformer model was initially developed for sequence processing in NLP [18]. Carion et al. [22] pioneered its introduction into computer vision for object detection. It has now been applied in many visual tasks, such as image recognition [7], super-resolution [23], and segmentation [24,25]. However, the computational complexity of attention in transformers increases exponentially with the number of image patches, which limits its application in high-resolution images. To make the transformer more suitable for visual tasks, recent studies have adopted different strategies to reduce complexity. For example, the PVT [25] network adopts a hierarchical pyramid structure to decrease computational complexity. Reference [26] substitutes spatial self-attention with channel self-attention with linear complexity. Swin Transformer [6] is designed to apply self-attention in non-overlapping and shifted windows to reduce complexity. These methods either apply attention to patches unfolded from the image, or divide the image into regions that do not overlap each other and independently calculate the attention of each region, which reduces the computational effort but limits the model's ability to capture global spatial dependencies among pixels in the image.

To address this issue, T-former [27] innovatively designs a self-attention mechanism with linear complexity by leveraging the Taylor formula. Lingle et al. [28] achieve efficient attention computation through vector quantization techniques. Spa-former [29] introduces a sparse self-attention mechanism, reducing computational costs by computing global pixel interactions across channel features. HINT [21] proposes a spatially activated channel attention mechanism, effectively alleviating memory pressure due to high computational complexity. In contrast to these methods, we propose a simple and efficient attention mechanism that not only enables multi-scale modeling of global context but also maintains low computational complexity.

2.3 Dilated Convolution

Traditional CNNs often expand the receptive field and reduce computational complexity by downsampling or convolving in large strides. Still, these methods reduce the resolution of feature maps, which in turn negatively affects the model performance for tasks such as object detection and semantic segmentation. To overcome this limitation, Cohen et al. [30] introduce dilated convolution [31], a technique that can increase the receptive field while maintaining resolution, and capture feature information at different scales by adjusting the dilation rate. Furthermore, DDC [32] enhances the flexibility of feature extraction by utilizing the entire feature map to generate convolution kernel parameters with data specificity.

In contrast to existing methods, we propose an innovative dilated attention operation that integrates multiple dilation rates into a single self-attention mechanism to model multi-scale feature interactions more flexibly. The difference from DDC is that our method executes self-attention computation for sparse keys and values within a shifted window with the query patch at the center. In addition, although a similar study DiNAT [33] exists, it only uses a single-scale and fixed dilation rate in each stage block and therefore lacks multi-scale interaction. By contrast, the proposed D-former implements a multi-scale operation within each block, that is, different dilation rates are assigned to each head to capture and integrate multi-scale semantic features effectively.

3 Method

In this section, we describe the overall architecture of a visual transformer image inpainting network based on multi-scale dilated attention. Firstly, we describe in detail the composition of the transformer block, which is the central part of the D-former model. Then, we present the DA operation, which effectively models long-range dependencies in feature maps. MDA module is further proposed, which concurrently captures contextual semantic dependencies at different scales to fully utilize information within patches.

3.1 Overall Architecture

We construct an efficient transformer network called D-former. The framework of D-former is illustrated in Fig. 1. First, the D-former is designed as a multi-stage encoder-decoder network using the U-Net structure [8]. The input to the network is the damaged image $I_{in} \in R^{H\times W\times 3}$ and the output is the restored image with the same resolution. The encoder consists of four stages, each containing multiple layers of Transformer blocks. While the features propagate progressively in the encoder, operations to decrease the spatial resolution and increase channel count are performed at each stage. Specifically, the spatial dimension of the input features decreases by a factor of 2, and the channel dimension increases by a factor of 2 at each encoder stage. The decoder has a symmetrical structure, where the multilevel features of the encoder stages are transferred to the corresponding stages of the decoder through skip connection. Each stage in the decoder performs the opposite operation of the encoder: the spatial resolution is progressively recovered, and the number of channels is progressively reduced. Lastly, the decoder outputs a restored result with the same resolution as the input image.

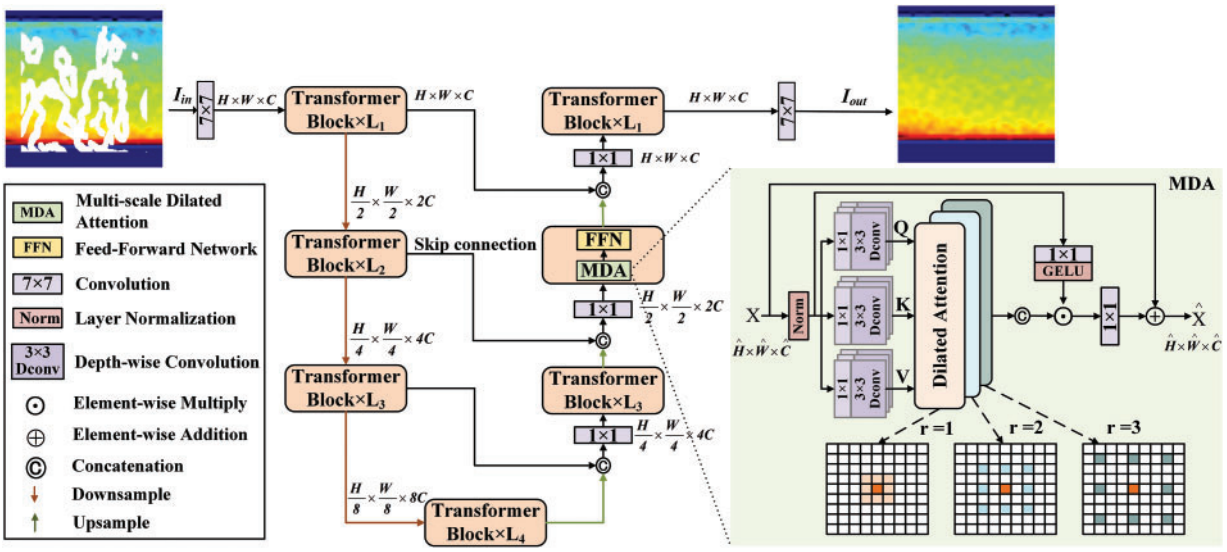

Figure 1: The framework of we proposed D-former. The damaged image I_{in} is sent into the generator, which is a U-Net style network composed of multiple transformer blocks. The Transformer block consists of two core modules: (1) multi-scale dilated attention (MDA). (2) feed-forward network (FFN)

3.1.1 Encoder

The encoder consists of a stack of several transformer blocks with a hierarchical structure, which enables feature extraction from damaged images $I_{in} \in R^{H\times W\times 3}$. The transformer block includes a MDA module and a feed-forward network (FFN) module. The encoder first sends the input I_{in} into a 7 × 7 convolutional layer

and obtains low-level feature maps $E_0 \in R^{H\times W\times C}$, where $H \times W$ is the spatial dimension and C is the number of channels. Next the feature map E_0 is sent into four stages of the encoder, each stage consisting of 1, 2, 3, and 4 transformer blocks. Each stage uses the attention of 1, 2, 4, and 8 heads, respectively, with corresponding dilation rates of $r = \{[1], [1, 2], [1, 2, 3, 4], [1, 2, 3, 4]\}$. Between the two stages, a 3×3 convolutional layer with stride 2 is used to downsample for features. The output feature map for each stage is $E_i \in R^{\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times C2^{i-1}}$, $i = 1, 2, 3, 4$. Such that the output of the final stage (4th stage) of the encoder is $E_4 \in R^{\frac{H}{8}\times\frac{W}{8}\times 8C}$.

3.1.2 Decoder

The decoder uses the final output of the encoder E_4 as input and gradually restores the complete image $I_{out} \in R^{H\times W\times 3}$. The decoder has 3 decoding stages, each containing 3, 2, and 1 transformer blocks, respectively. In each stage, the features are first upsampled through nearest neighbor interpolation, then connected to the corresponding encoding stage features, and the channels are compressed using 1×1 convolution to obtain feature maps $D_i \in R^{\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times C2^{i-1}}$, where $i = 1, 2, 3$. Next, these features are input into the transformer block of the corresponding decoder level to obtain dimensionally invariant output features $D_i^{'}$. Finally, the feature map $D_3^{'} \in R^{H\times W\times C}$ is transformed into the restored image $I_{out} \in R^{H\times W\times 3}$ through a 7×7 convolutional layer.

3.2 Transformer Block

As shown in Fig. 1, we propose a new transformer block to construct our network, which consists of two modules: the MDA module and the FFN module.

The MDA module is made of dilated attention with multi-headed attention, a layer normalization (LN), and a gated operation, as shown in detail in the MDA module in Fig. 1. In the MDA module of a transformer block at a certain level, the input feature $X \in R^{\hat{H}\times\hat{W}\times\hat{C}}$ is first fed into the LN layer, and then the linearly transformed feature is fed into dilated attention with multiple heads to obtain the attention result X_{MDA}. The dilated attention calculates the attention of each head and concatenates the attention maps of all heads. In addition, we introduce a spatial gated mechanism that utilizes 1×1 convolution and the GELU activation function to calculate gated values for each spatial position. These gated values are then multiplied element by element with the attention results to guide spatial features. Finally, the input features are combined with the adjusted attention results through residual shortcuts to obtain the output feature $\hat{X}$:

$$X' = X + DA\left(LN\left(X\right)\right) \odot GELU\left(Conv\left(LN\left(X\right)\right)\right) \tag{1}$$

where X and X' are the input and output features of the MAD module, respectively.

For the design of FFN, we refer to recent transformers [26], which use residual connections and gated mechanisms. As shown in Fig. 2, specifically, the gated mechanism is expressed as the unit product of two parallel path linear transformation layers, which consist of a 1×1 convolution and a 3×3 deep convolution, one of which is nonlinearly activated by GELU. Finally, the features obtained from the gated mechanism are channelized by a 1×1 convolutional layer for dimensionality reduction, and the processed results are residually concatenated with the original input for effective fusion and transfer of information. FFN can be formulated as:

$$\hat{X} = X' + Conv\left[Convs\left(LN\left(X'\right)\right) \odot GELU\left(Convs\left(LN\left(X'\right)\right)\right)\right] \tag{2}$$

where $\hat{X}$ is the output of the FFN in a transformer block at a certain layer, and the Convs include 1×1 convolution and 1×1 depth-wise convolution.

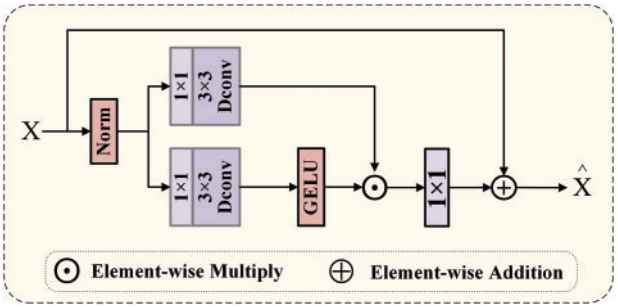

Figure 2: The framework of the feed-forward network

3.3 Multi-Scale Dilated Attention

3.3.1 Dilated Attention

To effectively capture key feature responses in PSP images and obtain global contextual information, we leverage the sparsity and locality of patch interactions of the global attention mechanism in ViTs to propose a dilated attention (DA) operation. This operation expands the receptive field of the model without increasing additional computational overhead, enabling it to capture longer-range contextual information. This not only facilitates the model to learn local texture features in PSP images but also improves the understanding of its global distribution characteristics. Specifically, the key and value matrices are sparsely selected in a shifted window centered around the query patch, then self-attention operation is conducted on these relevant patches. Our DA operation can be described as follows:

$$X = DA(Q, K, V, r) \tag{3}$$

where Q, K, and V denote query, key, and value matrices, respectively. Each row of the three matrices represents a single eigenvector corresponding to the query, key, or value. For the query with position (i, j) in the original feature map, DA performs self-attention operation by sparsely selecting keys and values in a $w \times w$ shifted window centered on (i, j) position. In addition, we also design a dilation rate $r \in N^+$ to control sparsity. Specifically, for position (i, j), the component x_{ij} of the output X from the DA operation is described as follows:

$$\begin{aligned} x_{ij} &= Attention\left(q_{ij}, K_r, V_r\right) \\ &= Soft\max\left(\frac{q_{ij}K_r^T}{\sqrt{d_k}}\right)V_r, \quad 1 \le i \le W, 1 \le j \le H, \end{aligned} \tag{4}$$

where H and W are the height and width of the feature map, respectively. K_r and V_r denote the keys and values selected from the feature maps K and V.

Given the query location at (i, j), the keys and values located at (i', j') will be selected to perform the self-attention operation:

$$\left\{(i', j') \,|\, i' = i + p \times r, j' = j + q \times r\right\}, \quad -\frac{w}{2} \le p, q \le \frac{w}{2}. \tag{5}$$

DA performs self-attention operations on all query patches in a shifted window manner. For the query feature on the edge of the feature map, we employ the zero padding method to maintain the size of the feature map. The proposed DA explicitly satisfies locality and sparsity properties by sparse selection of query-centered keys and values, which can effectively model long-range dependencies.

3.3.2 Multi-Scale Dilated Attention

To further address defects of different sizes in PSP images, we apply dilated attention to the multi-head attention mechanism, proposing multi-scale dilated attention (MDA). By setting different dilation rates for different attention heads, the MDA module expands the receptive field and at the same time enhances the model's comprehension of the global structure, and the module can simultaneously capture multi-scale semantic information, enabling the model to more flexibly handle defect regions of different sizes in PSP image, and strengthens the network's ability to extract multi-scale features, thereby significantly improving the accuracy and robustness of image inpainting. The introduction of multi-scale dilated attention not only increases the model's sensitivity to local details but also strengthens its understanding of global structures, ensuring that the restored image maintains consistency both locally and globally.

Specifically, as seen in the MDA module in Fig. 1, for the feature map X, we get the corresponding queries, keys, and values through linear mapping. Then, we split the channels of the feature map into n different heads, and conduct DA operation between the colored patches in the window around the orange query patch, using different dilation rates in different heads, when the dilation rates are 1, 2 and 3 using 3×3 kernel size, the corresponding receptive field sizes are 3×3, 5×5 and 7×7, respectively. Our MDA module is described as follows:

$$h_i = MDA(Q_i, K_i, V_i, r_i), \quad 1 \le i \le n \tag{6}$$

$$X_{MDA} = Concat\left([h_1, \ldots, h_n]\right) \tag{7}$$

where r_i is the dilation rate of the i-th head, Q_i, K_i and V_i denote the feature map slices of the i-th head. The outputs $\{h_i\}_{i=1}^{n}$ are connected in series to obtain the result X_{MDA} of the multi-scale DA. The final output is then obtained by using the spatial gated mechanism and residual connection.

According to previous research [33], we employ fixed dilation rates of {[1], [1,2], [1,2,3,4], [1,2,3,4]} at the four stages of the encoder. The early stages primarily focus on local details, thus utilizing smaller dilation rates, and as the network depth increases, subsequent stages gradually introduce larger dilation rates to capture broader contextual information. Through defining different dilation rates for each head, our MDA module effectively converges semantic information at different scales within the receptive field of attention, effectively decreasing the redundancy of the attention mechanism without requiring complex operations and additional computational costs.

3.4 Loss Function

The overall loss function L_{all} for training our D-former model is as follows:

$$L_{all} = \lambda_r L_r + \lambda_a L_a + \lambda_p L_p + \lambda_s L_s \tag{8}$$

where L_r denotes reconstruction loss, L_a denotes adversarial loss, L_p denotes perceptual loss, and L_s is style loss. We set $\lambda_r = 1$, $\lambda_a = 0.1$, $\lambda_p = 1$, $\lambda_s = 250$.

Reconstruction loss: the reconstruction loss L_r is the L_1-distance between the output I_{out} and the ground truth I_{gt}, which can be defined as:

$$L_r = \| I_{out} - I_{gt} \|_1 \tag{9}$$

Adversarial loss: The formula for adversarial loss L_a is:

$$L_a = \mathrm{E}_{I_{gt}}\left[\log D\left(I_{gt}\right)\right] + \mathrm{E}_{I_{out}}\left[\log\left[1 - D\left(I_{out}\right)\right]\right] \tag{10}$$

here, D is the discriminator network Patch GAN [34], which can generate natural image details.

Perception loss: The perception loss L_p compares the difference between the depth feature map of the generated image and the real image, and the formula is:

$$L_p = \mathrm{E}\left[\sum_i \frac{1}{N_i} \parallel \phi_i^{gt} - \phi_i^{out} \parallel_1\right] \tag{11}$$

where ϕ_i is the feature map from the i-th pooling layer of VGG-16 pre-trained on ImageNet, and N_i is the number of elements in ϕ_i.

Similarly, the style loss is defined as follows:

$$L_s = \mathrm{E}_i\left[\parallel G\left(\phi_i{}^{gt}\right) - G\left(\phi_i{}^{out}\right) \parallel_1\right] \tag{12}$$

where G represents the Gram matrix on the feature map, $G(\phi) = \phi^T\phi$.

4 Experiments

We describe the datasets, experimental details, and comparison models in our experiment setting. We compare the inpainting results of our proposed D-former and advanced methods on two public datasets, and evaluate the inpainting effect of the D-former model on PSP images on the PSP image dataset. Finally, we perform an ablation study of the modules and loss functions in the D-former model.

4.1 Experimental Settings

4.1.1 Dataset

We evaluated the restore performance of the D-former model on three datasets.

1. PSP dataset: Real pressure-sensitive paint test images from the China Aerodynamics Research and Development Center (CARDC) are cropped to create a training dataset of 15,000 images and a testing dataset of 1000 images. The training set is expanded to 20,000 samples by means of panning, scaling, rotating and adding noise to improve the generalization ability of the model under different working conditions. As shown in Fig. 3.
2. Paris StreetView dataset [35]: This dataset mainly includes city buildings collected from Google StreetView in Paris. It contains 14,900 training images and 100 test images and is suitable for training inpainting models of various buildings.
3. CelebA-HQ dataset [36]: This dataset focuses on the faces of some celebrities and is a high-resolution version of the CelebA dataset. It includes 30,000 images, and we use the first 2000 images to test the model and the remaining 28,000 images for training. It is commonly used in model training for face editing and restoration.
4. Irregular mask dataset: Following existing methods, we use the irregular mask dataset of Liu et al. [16] as the test mask to evaluate all trained models. The irregular mask data contains 6 categories with different hole ratios, each with 2000 masks. We evaluated all models using masks with mask ratios of 10%–20%, 20%–30%, 30%–40%, and 40%–50%.

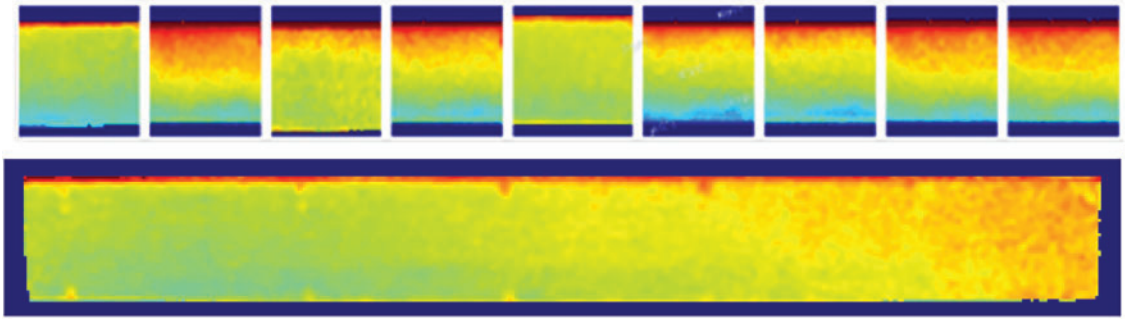

Figure 3: PSP image dataset

4.1.2 Experimental Details

Our D-former was programmed using PyTorch and trained on a single RTX 6000 GPU with batch size 6. We optimized the network using AdamW with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ [11]. We first trained the model from scratch with a learning rate of 10^{-4}, and then fine-tuned it with a learning rate of 10^{-5}. For the PSP dataset, 300,000 and 50,000 iterations were used for training and fine-tuned, respectively, and for the Paris StreetView dataset, 400,000 and 200,000 iterations were used for the training from scratch and fine-tuned stages. For CelebA-HQ, the training and fine-tuned stages used 450,000 and 200,000 iterations, respectively. All input images were resized to 256×256.

4.1.3 Comparison Model

To show the performance of our D-former, we compare it with five advanced models. These models include RFR [14], LGNet [37], T-former [27], Spa-former [29], and HINT [21]. To be fair, a part of the data is obtained by testing directly using the officially released pre-trained models, and the rest of the data is reproduced and tested according to the paper.

RFR [14]: A recurrent inpainting method with special contextual attention that recursively restores missing regions and gradually strengthens the results.

LGNet [37]: a multilayer network architecture for image inpainting that combines networks with different receptive fields, considering the complexity of missing regions.

T-former [27]: A U-net style image inpainting network constructed by the proposed linear attention.

Spa-former [29]: An efficient transformer-based image inpainting network that utilizes a novel attention mechanism with linear complexity.

HINT [21]: An end-to-end high-quality inpainting transformer that utilizes a novel mask-aware pixel-shuffling downsampling module can extract visible information from corrupted images.

4.2 Qualitative Analysis

We selected PSP images with pressure taps for the qualitative evaluation of the models. In PSP technology, it is crucial to restore luminescent intensity data at these locations accurately, hence, we chose 30%–40% masks that could cover the pressure taps. Fig. 4 demonstrates the inpainting results of various models on the PSP dataset, while Figs. 5 and 6 show the qualitative evaluation results on the Paris StreetView and CelebA-HQ datasets, respectively. It depicts the overall quality of inpainting results at 30%–40% mask rates.

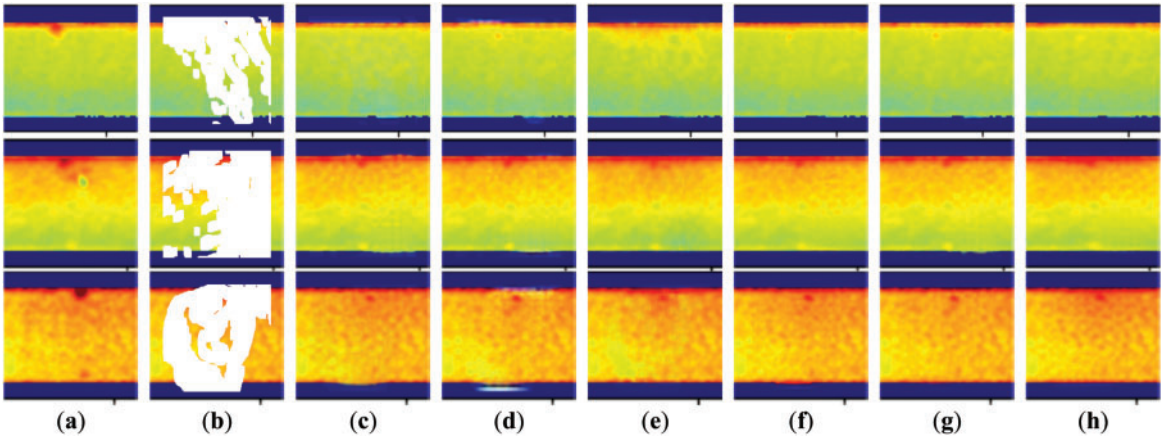

Figure 4: Qualitative results of different methods on the PSP Image dataset: (**a**) Input Image; (**b**) Masked Image; (**c**) RFR; (**d**) LGNet; (**e**) T-former; (**f**) Spa-former; (**g**) HINT; (**h**) Ours

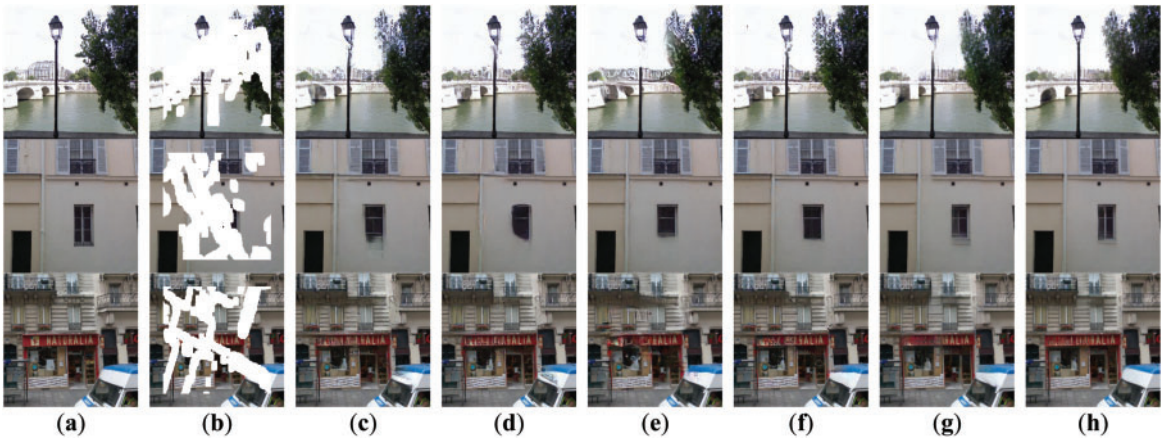

Figure 5: Qualitative results of different methods on the Paris StreetView dataset: (**a**) Input Image; (**b**) Masked Image; (**c**) RFR; (**d**) LGNet; (**e**) T-former; (**f**) Spa-former; (**g**) HINT; (**h**) Ours

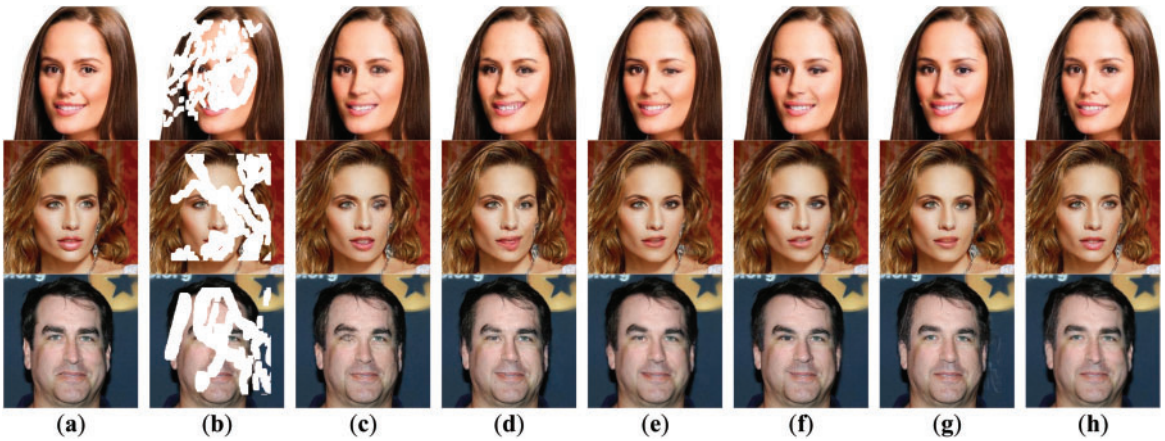

Figure 6: Qualitative results of different methods on the CelebA-HQ dataset: (**a**) Input Image; (**b**) Masked Image; (**c**) RFR; (**d**) LGNet; (**e**) T-former; (**f**) Spa-former; (**g**) HINT; (**h**) Ours

Through visual comparison, the inpainting results obtained using Spa-former and HINT methods are relatively satisfactory, but they are still prone to problems such as blurring, artifacts, and semantic inconsistencies. For example, the holes are not fully restored in Fig. 4f,g and Fig. 5f,g shows artifacts, and Fig. 6f,g shows distorted eyes and inconsistent colors. Similarly, other methods also exhibit noticeable problems, such as distorted boundaries (second and third rows in Fig. 4c,d) and mask traces (first row in Fig. 4c,e), artifacts (first row in Fig. 5c–e) and unrestored window structures (second row in Fig. 5c,d), as well as blurring and artifacts of the eyes in Fig. 6c,d. Compared with these models, our restored images can present more reasonable and realistic visual effects in most cases.

4.3 Quantitative Analysis

4.3.1 Evaluation of Inpainting Effectiveness

We chose several common metrics for image inpainting tasks to evaluate the model performance [27]: PSNR (peak signal-to-noise ratio) and SSIM (structural similarity index) scores, which are evaluated based on low-level pixel values. FID (Fréchet Inception Distance) score, which calculates distance based on deep high-level representations, measures image quality similar to human perception. The performance of different models on three datasets under different mask ratios is shown in Table 1, tests with different ratios of masks under the same dataset use the identical set of test images. Compared with a set of comparative models, including RFR [14], LGNet [37], T-former [27], Spa-former [29], and HINT [21], our proposed D-former exhibits excellent performance on all three datasets.

Table 1: Quantitative results of different methods on the PSP Image, Paris StreetView, and CelebA-HQ datasets, bolded in the table indicates optimal results

Dataset		PSP image				Paris street view				CelebA-HQ			
Masks		10%–20%	20%–30%	30%–40%	40%–50%	10%–20%	20%–30%	30%–40%	40%–50%	10%–20%	20%–30%	30%–40%	40%–50%
PSNR↑	RFR	43.39	39.58	36.79	34.71	31.75	27.76	25.99	24.25	30.93	28.94	26.85	24.87
	LGNet	44.37	40.21	37.07	34.83	29.59	26.48	24.22	22.44	32.55	29.28	27.11	25.21
	T-former	45.16	40.70	38.02	35.16	32.09	29.52	27.27	25.06	33.28	30.10	27.48	25.54
	Spa-former	45.21	41.21	38.07	35.70	32.54	29.63	27.33	25.21	33.65	30.37	27.83	25.78
	HINT	45.20	40.99	38.04	35.36	32.40	29.57	27.31	25.08	33.29	30.11	27.75	25.74
	Ours	**45.27**	**41.25**	**38.21**	**35.83**	**32.88**	**29.65**	**27.35**	**25.31**	**34.71**	**31.11**	**28.38**	**26.00**
SSIM↑	RFR	0.992	0.984	0.974	0.962	0.933	0.891	0.838	0.778	0.969	0.943	0.926	0.889
	LGNet	0.994	0.987	0.977	0.965	0.924	0.863	0.798	0.725	0.976	0.955	0.935	0.909
	T-former	0.995	0.989	0.978	0.969	0.959	0.921	0.876	0.820	0.981	0.961	0.939	0.913
	Spa-former	0.995	0.989	0.981	0.971	0.961	0.921	0.878	0.822	0.983	0.964	0.944	0.918
	HINT	0.995	0.988	0.980	0.970	0.960	0.918	0.878	0.820	0.982	0.961	0.942	0.914
	Ours	**0.995**	**0.990**	**0.982**	**0.971**	**0.969**	**0.938**	**0.888**	**0.828**	**0.988**	**0.971**	**0.949**	**0.921**
FID↓	RFR	2.36	4.84	9.16	17.63	18.33	27.93	39.84	48.96	5.17	4.06	4.89	6.11
	LGNet	1.70	4.17	7.62	15.95	15.53	24.39	36.81	47.18	1.87	3.21	4.57	5.80
	T-former	1.59	3.85	7.33	13.71	13.45	23.78	35.79	46.36	1.60	2.70	3.97	5.66
	Spa-former	1.48	3.18	6.83	13.24	13.23	23.67	35.29	46.15	1.44	2.63	3.76	5.23
	HINT	1.55	3.70	7.14	13.55	13.36	23.71	35.73	46.28	1.53	2.65	3.87	5.41
	Ours	**1.38**	**3.02**	**6.63**	**10.81**	**11.39**	**19.43**	**28.49**	**39.61**	**1.29**	**2.25**	**3.59**	**4.91**

Additionally, to further analyze the computational complexity of the model, we compare the performance of D-former with other comparative models and baseline models in Table 2 across multiple metrics, including multiply-accumulate operations (MAC), floating point operations per second (FLOPs), number of

parameters, inference time for a single image, and GPU memory consumption. The results in Table 2 clearly demonstrate that, compared to other methods, our model has the most streamlined parameter count and delivers the best inference time. Moreover, our approach also exhibits lower computational complexity in terms of FLOPs.

Table 2: Complexity measurement of different models on the PSP Image dataset. Bolded in the table indicates optimal results

Model	MAC	FLOPs	Params	Inference time	GPU_Memory_Usage
RFR	206.1 G	176.1 G	30.6 M	38.48 ms	**8458 MB**
LGNet	69.6 G	55.4 G	115.0 M	78.74 ms	12,218 MB
T-former	84.5 G	51.3 G	14.8 M	83.58 ms	21,378 MB
Spa-former	46.8 G	44.4 G	13.2 M	52.66 ms	22,062 MB
HINT	75.9 G	52.4 G	139.0 M	74.85 ms	9232 MB
Swin-B [6]	**27.6 G**	15.4 G	88.0 M	54.93 ms	21,630 MB
DiNAT [33]	28.4 G	**13.7 G**	90.0 M	46.94 ms	20,162 MB
Ours	44.4 G	15.7 G	**12.4 M**	**33.36 ms**	10,876 MB

Combining the results in Tables 1 and 2, it can be concluded that, thanks to the long-range dependency capture capability and dynamic parameters brought by the proposed MDA, our D-former can provide relatively higher-quality restored images with fewer parameters and lower computational complexity when facing different scenarios (datasets) and encountering different damage situations.

4.3.2 Evaluation of Inpainting Accuracy

(1) Real Wind Tunnel Test Validation

To evaluate the accuracy of the model in restoring the luminescent intensity data of PSP images, a set of experiments was designed in this part for verification. The PSP tests were conducted in CARDC's ∅0.7 m wind tunnel using a flying wing model under strictly controlled environmental conditions including constant exposure time, temperature, aperture size, and atmospheric pressure, and at a wind speed of 70 m/s, the experimental model was subjected to image capture at 0° and 10° to the horizontal plane, respectively.

The same image processing (denoising, etc.) was first performed on both sets of image data. Fig. 7a shows the unrestored PSP image, with black dots along the model boundary as marker points, and smaller holes on the back and lower half of the wing model as pressure taps. Fig. 7 shows the inpainting effects of different methods. It can be observed that other methods have some holes that have not been restored, and the restored area is not coherent with the surrounding area, which clearly does not conform to the pressure distribution. Our restored images have smoother results with consistent context.

In order to quantify the performance of the model in restoring the luminescent intensity data of PSP images, we use the actual pressure data obtained from pressure taps as a benchmark and compare it with the pressure data calculated by restoring the luminescent intensity data, and calculate the relative error between them. The relationship between luminescent intensity data and pressure can be described by the Stern-Volmer formula:

$$\frac{I_0}{I} = A + B\frac{P}{P_0} \tag{13}$$

here, A and B are constants (usually determined by calibration experiments), I_0 and P_0 are the luminescent intensity value and air pressure at a reference condition, I is the luminescent intensity value predicted by the model, and the pressure P at any position on the surface of the model can be calculated by this formula.

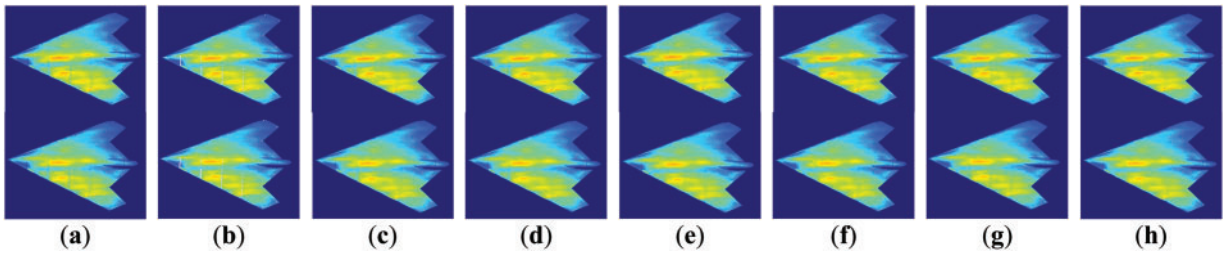

Figure 7: Comparison of the effect of PSP images of flying wings restored by different methods. The top is 0–70 (angle-wind speed), and the bottom is 10–70. Zoom in for a better view. (**a**) Original; (**b**) Masked Image; (**c**) RFR; (**d**) LGNet; (**e**) T-former; (**f**) Spa-former; (**g**) HINT; (**h**) Ours

Using the above formula to obtain pressure data at 68 pressure taps, the relative error is calculated by comparing them with the actual pressure values, and the results are then averaged. The formula for the average relative error is as follows:

$$\delta = \frac{\sum_{i=1}^{N} \Delta / T}{N} \times 100\% \tag{14}$$

where Δ denotes the absolute error between true value and measured value, T denotes the true value, and N denotes the number of pressure taps.

After calculating the pressure data of the PSP images restored using different methods, we derived the results. This is shown in Table 3. It can be observed that the restoration results of RFR and LGNet have a large deviation with a relative error of over 0.2%, while our restoration results have a higher accuracy of around 0.1%. This indicates that the method in this article can effectively restore the luminescent intensity data of the PSP image and conforms to the flow field pressure distribution.

Table 3: Average relative error results of pressure data on PSP images restored by different methods. 0–70 means that the wind speed is 70 m/s and the model is at 0° to the horizontal. Bolded in the table indicates optimal results

	Methods	0–70	10–70
$\delta \downarrow$	RFR	0.3081%	0.2450%
	LGNet	0.2601%	0.2271%
	T-former	0.1854%	0.2012%
	Spa-former	0.1494%	0.1659%
	HINT	0.1647%	0.1805%
	Ours	**0.1049%**	**0.1225%**

To visualize the restoration accuracy of the model, we conduct a comparative analysis between the calculated pressure values from the 15 pressure taps in the fourth column of the flying wing model and the true pressure values obtained using pressure scanner instrument (PSI), and plot the corresponding line graphs. It can be observed from Fig. 8 that although the calculated results of each model generally follow the

trend of the true value curve, they all exhibit significant errors. In contrast, our calculated results are closer to the actual values, with noticeably smaller errors. This comparison visually demonstrates the advantages of our model in restoring pressure data.

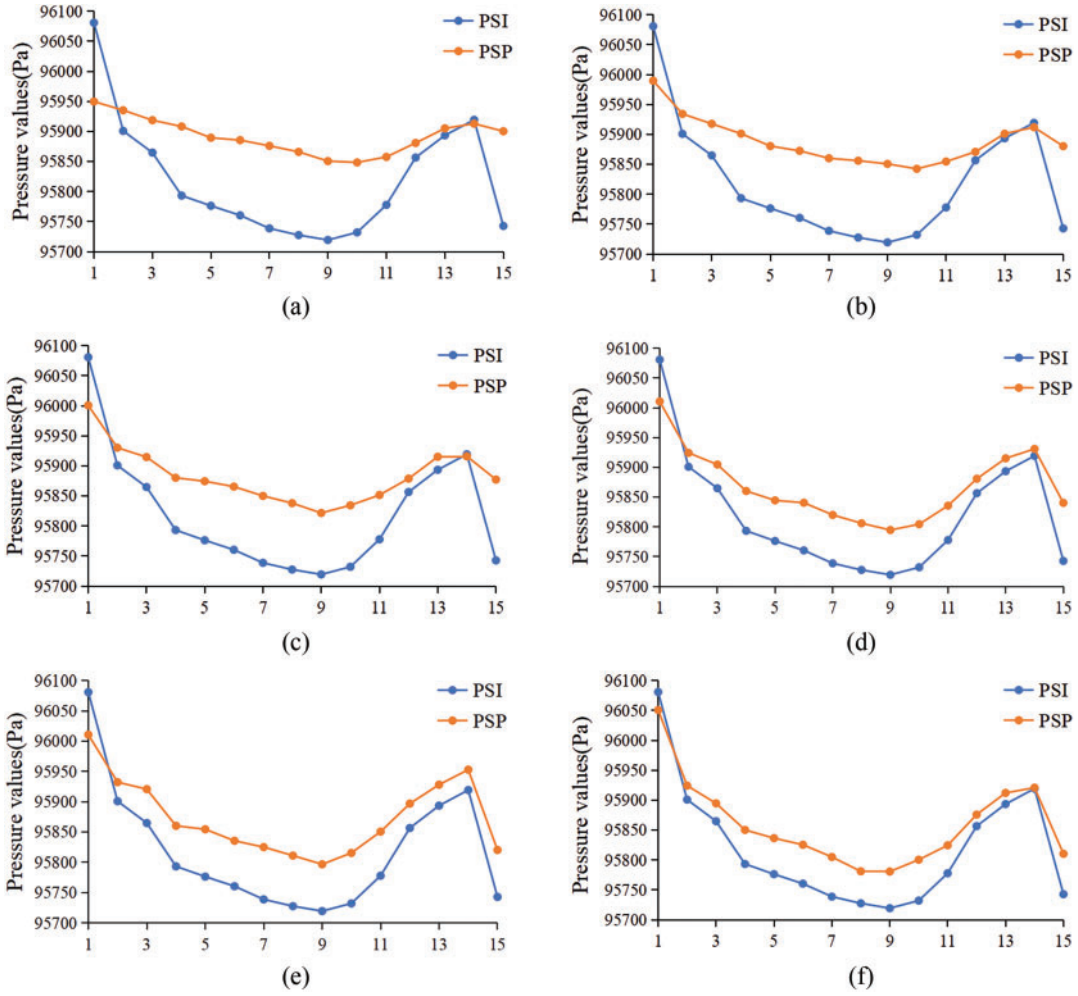

Figure 8: Comparison of pressure data obtained using PSP and PSI techniques at the fourth column of pressure taps on the flying wing under 10–70 conditions. The orange and blue dots represent the PSP and PSI data, respectively. (**a**) RFR; (**b**) LGNet; (**c**) T-former; (**d**) Spa-former; (**e**) HINT; (**f**) Ours

(2) Large-Area Defect Evaluation

To validate the model's restoration capability under extreme PSP damage scenarios, we select PSP images with pressure taps and simulate large-area damage conditions by covering them with 40%–50% masks. The D-former and various comparative methods were employed for inpainting, and the qualitative results are shown in Fig. 9, where it can be observed that our method not only fully restores the damaged regions and the unique pressure taps, but also maintains excellent pressure consistency. Based on Eq. (13), we calculate the pressure values at the locations of two pressure taps on each image and combine the true pressure values to calculate the relative errors using Eq. (14). The results are presented in Table 4. The experimental results demonstrate that even under extreme damage conditions, our model can achieve high precision inpainting effects, highlighting its potential for practical applications.

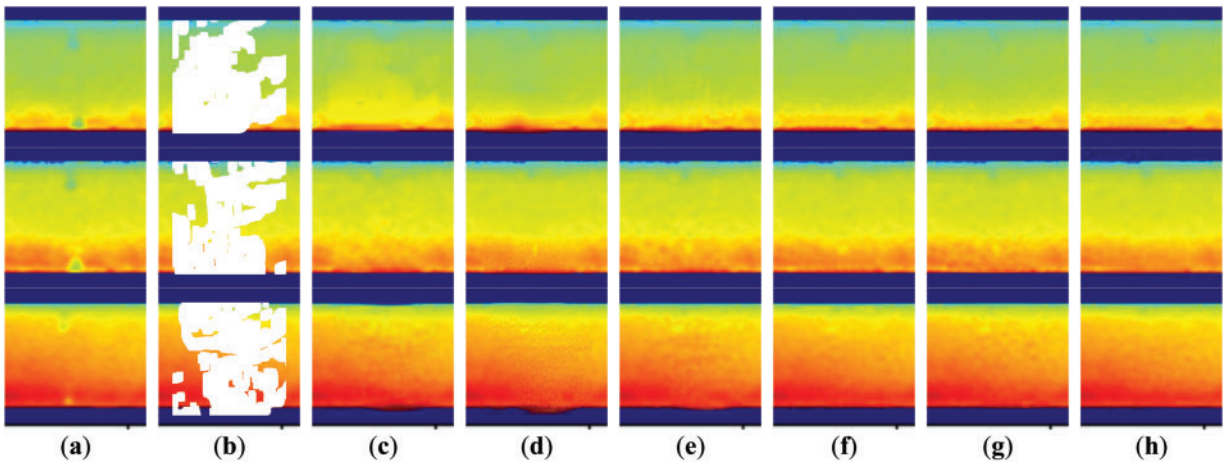

Figure 9: Qualitative comparison of PSP image with pressure taps and large damaged areas restored by different methods: (**a**) Input Image; (**b**) Masked Image; (**c**) RFR; (**d**) LGNet; (**e**) T-former; (**f**) Spa-former; (**g**) HINT; (**h**) Ours

Table 4: Average relative error results at pressure taps on large damaged PSP images restored by different methods, Group 1 represents the first image above. Bolded in the table indicates optimal results

	Methods	Group 1	Group 2	Group 3
$\delta\downarrow$	RFR	0.0128%	0.0284%	0.0506%
	LGNet	0.0102%	0.0242%	0.0486%
	T-former	0.0841%	0.0216%	0.0402%
	Spa-former	0.0067%	0.0129%	0.0349%
	HINT	0.0060%	0.0148%	0.0341%
	Ours	**0.0056%**	**0.0101%**	**0.0297%**

4.4 Ablation Study

4.4.1 Effectiveness of the MDA Module

- Experiment I (G + G + G + G): Global attention is used in all four stages [7].
- Experiment II (D + G + G + G): Using MDA module only in the first stage.
- Experiment III (D + D + G + G): Using MDA module in the first and second stages.
- Experiment IV (D + D + D + G): Using MDA module in stages 1, 2, and 3.
- Experiment V (D + D + D + D): Using MDA module in all four stages.

We use MDA module in all transformer blocks in our D-former. To demonstrate the effectiveness of our proposed MDA, we explore the performance of using MDA modules at different stages. The baseline model uses multi head self-attention (MHA) [7] instead of the MDA module in all stages. Over the four stages of the model, we progressively replace the global MHA module with MDA module at each stage. The ablation experiment is conducted on the PSP Image dataset, and Fig. 10 shows the qualitative results of each experiment. Table 5 shows the quantitative results of models with different structures. The results show that the performance of the model using MDA module in stage 1 is much better than that using only global attention. As the proportion of MDA module increases in the model stage, the performance of the model reaches a small peak using MDA modules in stages 1 and 2, and then declines. The best model performance is achieved when all stages use MDA module. This further demonstrates the effectiveness of the proposed

locality and sparse attention mechanism, as well as the redundancy of modeling dependencies between all image patches.

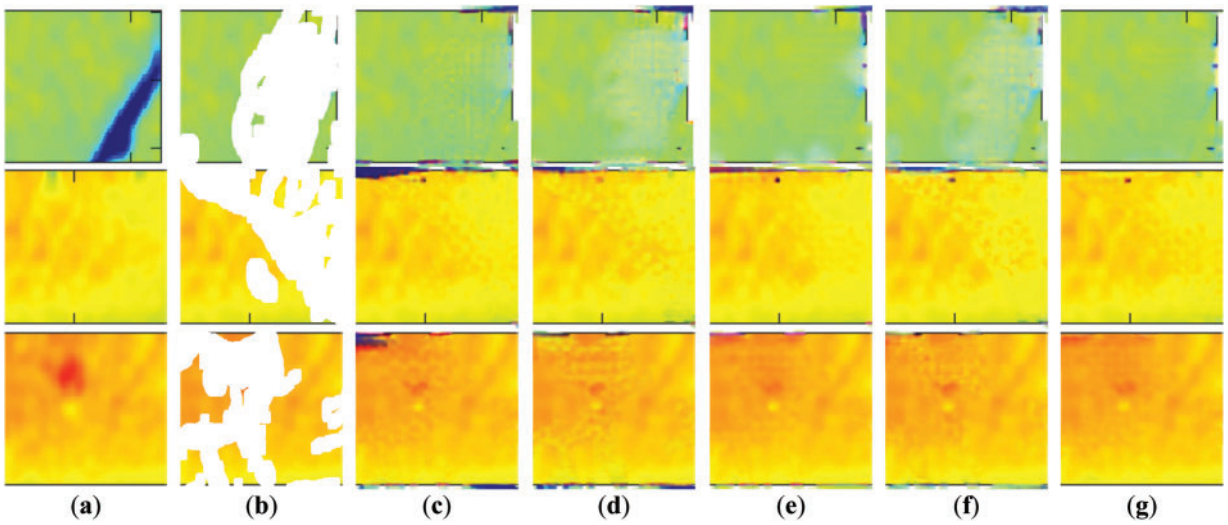

Figure 10: Qualitative comparison results of ablation experiments: (**a**) Input Image; (**b**) Masked Image; (**c**) Experiment I; (**d**) Experiment II; (**e**) Experiment III; (**f**) Experiment IV; (**g**) Experiment V

Table 5: Experimental results for different stages of the MDA module analyzed on the PSP image dataset. Bolded in the table indicates optimal results

Dataset		PSP image			
Masks ratio		10%–20%	20%–30%	30%–40%	40%–50%
PSNR↑	Experiment I	36.76	35.70	34.42	32.96
	Experiment II	39.42	38.54	36.77	35.00
	Experiment III	40.64	39.05	36.99	35.18
	Experiment IV	39.15	38.15	36.79	34.97
	Experiment V	**45.27**	**41.25**	**38.21**	**35.83**
SSIM↑	Experiment I	0.980	0.978	0.972	0.962
	Experiment II	0.987	0.984	0.976	0.966
	Experiment III	0.989	0.985	0.977	0.967
	Experiment IV	0.987	0.984	0.977	0.967
	Experiment V	**0.995**	**0.990**	**0.982**	**0.971**
FID↓	Experiment I	5.59	4.85	7.33	12.87
	Experiment II	4.62	4.12	6.73	11.76
	Experiment III	2.98	4.00	6.62	11.23
	Experiment IV	4.01	4.23	6.76	12.15
	Experiment V	**1.38**	**3.02**	**6.63**	**10.81**

To evaluate the capability of the MDA module in capturing global context and local details, we conduct a visualization analysis of attention maps for the above versions of the model. Specifically, we applied the

grad-CAM technique [38] to the last layer of the model's encoding stage, generating heat maps that were overlaid on the original images, thus allowing us to intuitively observe the key regions the model focuses on during image inpainting. As shown in Fig. 11, the model without the MDA module (Fig. 11c) exhibits a more uniform weight distribution, primarily capturing texture features. Experiment II (Fig. 11d), which introduces the MDA module at the shallow layers, begins to focus on information surrounding the missing regions. As the proportion of the MDA module increases in the encoder stage, the model gradually focuses on the global features of the image. Ultimately, our model (Fig. 11g) demonstrates the ability to simultaneously capture large-scale features and local details, showcasing enhanced contextual understanding.

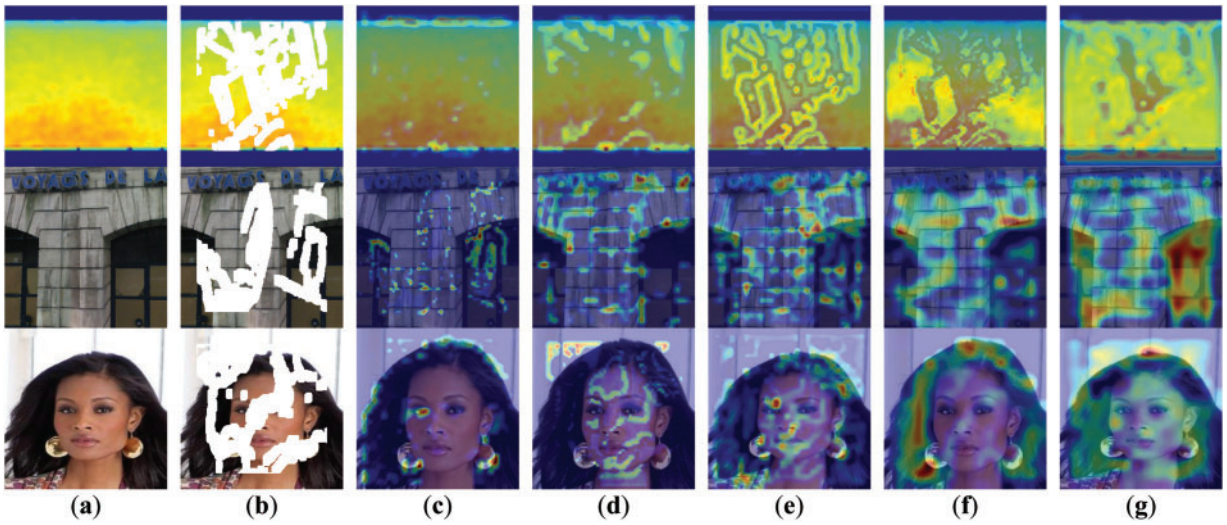

Figure 11: Heat map results. Darker colors indicate more attention. (**a**) Input Image; (**b**) Masked Image; (**c**) Experiment I; (**d**) Experiment II; (**e**) Experiment III; (**f**) Experiment IV; (**g**) Experiment V

4.4.2 Effects of Attention Head Count and Dilation Rate

To investigate the effects of two key parameters in the MDA module on model performance, we design a series of experiments for analysis. Based on the settings of the number of heads in the multi-head attention mechanism from previous studies [6,29], we select three combinations of attention head numbers as shown in Table 6, and evaluate the performance of models with different dilation rates on the PSP dataset. As shown in Table 6, when the dilation rate of the MDA module is moderate, it can effectively leverage both the locality and sparsity of attention without causing redundant information modeling due to an excessively large receptive field (e.g., global attention), thereby achieving optimal model performance. Accordingly, we set the dilation rates of the model to {[1],[1,2],[1,2,3,4],[1,2,3,4]}.

4.4.3 Ablation Study of Loss Function

To validate the rationality of our selected loss functions and hyperparameter settings, this section conducts systematic ablation experiments. The experimental design includes two aspects: first, the contribution of each loss function to model performance is evaluated by sequentially removing the reconstruction loss L_r, adversarial loss L_a, perceptual loss L_p, and style loss L_s; second, based on the optimal hyperparameter combination ($\lambda_r = 1$, $\lambda_a = 0.1$, $\lambda_p = 1$, $\lambda_s = 250$) determined through Optuna search [39], we compare and analyse it with other weight combinations that perform well.

Table 6: Analyze the number of attention heads and the dilation rate of the MDA module on the PSP dataset. Bolded in the table indicates optimal results

Head num in 4 stage		Dilation rate	PSNR↑	SSIM↑	FID↓
1	[1,2,4,8]	[1],[1,2],[1,2],[1,2]	38.64	0.966	6.58
		[1],[1,2],[1,2],[1,2,3,4]	39.52	0.976	5.84
		Ours {[1],[1,2],[1,2,3,4],[1,2,3,4]}	**40.90**	**0.983**	**4.18**
2	[2,4,8,12]	[1,2],[1,2],[1,2,3,4],[1,2,3,4]	37.14	0.965	10.53
		[1,2],[1,2,3,4],[1,2,3,4],[1,2,3,4,5,6]	35.12	0.954	12.99
3	[3,6,12,24]	[1],[1,2,3],[1,2,3,4],[1,2,3,4,5,6]	38.34	0.961	6.75
		[1,2,3],[2,3,4],[1,2,3,4],[1,2,3,4]	36.81	0.957	10.86

The experimental results are shown in Table 7, which indicates that each loss function improves the performance of the model to varying degrees, proving their necessity in the overall loss function, with the removal of the adversarial loss having the most pronounced impact on model performance. In the hyperparameter comparison experiments, our parameter combination exhibits the best performance, confirming the rationality of the parameter selection. Ablation experiments are performed on the PSP image dataset, the mask is the whole test mask dataset.

Table 7: Ablation study of loss functions and hyperparameter settings. Bolded in the table indicates optimal results

Methods	PSNR↑	SSIM↑	FID↓
w/o L_r	38.22	0.971	9.40
w/o L_a	36.71	0.969	11.51
w/o L_p	38.83	0.974	7.58
w/o L_s	39.25	0.978	5.43
A [1,2,0.5,1]	40.83	0.982	4.35
B [1,0.01,1,60]	40.79	0.982	4.32
Ours ($\lambda_r + 0.1\lambda_a + \lambda_p + 250\lambda_s$)	**40.90**	**0.983**	**4.18**

5 Conclusion

This article introduces an efficient PSP image inpainting network D-former, which is based on a multi-scale dilated attention mechanism. For the characteristics of PSP images and practical inpainting requirements, the MDA module we proposed fully leverages the localization and sparsity properties of self-attention mechanism in networks, by expanding the self-attention range, the module can capture receptive fields of different scales in each attention head, thereby efficiently integrating multi-scale semantic information and effectively reducing the redundancy of self-attention mechanism without requiring complex operations or additional computational costs. By capturing local details and global contextual information in PSP images through the MDA module, the D-former model can more accurately restore the luminescent intensity data at the holes of PSP images, generating luminescent intensity images with local continuity and global distribution consistency. Qualitative and quantitative experiments verify that D-former outperforms the other five advanced models and has the lowest complexity. In addition, ablation experiments further confirm that the MDA module can enhance the focus of attention, effectively filter out irrelevant information

or noise, thus improving the effectiveness of image inpainting. Notably, physical experiments validate the advantages of D-former in PSP image processing: compared to other methods, the network can restore the luminescent intensity data of PSP images more precisely, obtain the surface pressure distribution of the wind tunnel model, and also performs well in large-area damage scenarios.

Although our method has achieved certain success in PSP image inpainting tasks, it still faces challenges of structural and semantic inconsistencies when dealing with large damaged areas in complex scenes. To address this issue, we plan to explore multimodal learning approaches in future research, integrating textual and structural information to enhance inpainting performance. Additionally, D-former method is limited by computational resources when restoring high-resolution images, resulting in constrained performance. Therefore, we will also focus on designing more efficient attention mechanisms and lightweight models to better handle the inpainting of high-resolution images.

Acknowledgement: The authors gratefully acknowledge the technical support provided by the China Aerodynamic Research and Development Centre.

Funding Statement: This research was partly supported by the National Natural Science Foundation of China under Grant 12202476, author Chunhua Wei, https://www.nsfc.gov.cn/.

Author Contributions: Conceptualization, Jinrong Li; methodology, Jinrong Li and Zhisheng Gao; resources, Chunhua Wei; supervision, Chunhua Wei and Lei Liang; writing—original draft preparation, Jinrong Li; writing—review and editing, Lei Liang and Zhisheng Gao. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data supporting the reported results are available upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Gregory JW, Asai K, Kameda M, Liu T, Sullivan JP. A review of pressure-sensitive paint for high-speed and unsteady aerodynamics. Proc Inst Mech Eng Part G J Aerosp Eng. 2008;222(2):249–90. doi:10.1243/09544100jaero243.
2. Bertalmio M. Strong-continuation, contrast-invariant inpainting with a third-order optimal PDE. IEEE Trans Image Process. 2006;15(7):1934–8. doi:10.1109/TIP.2006.877067.
3. Buyssens P, Daisy M, Tschumperlé D, Lézoray O. Exemplar-based inpainting: technical review and new heuristics for better geometric reconstructions. IEEE Trans Image Process. 2015;24(6):1809–24. doi:10.1109/TIP.2015.2411437.
4. Jiang T, Li Q, Chen S, Chang Y, Zhang Q. Image inpainting of wind tunnel test based on Otsu method and FMM. Acta Aeronaut Et Astronaut Sin. 2020;41(2):123293. (In Chinese). doi:10.7527/S1000-6893.2019.23293.
5. Deng Y, Hui S, Zhou S, Meng D, Wang J. Learning contextual transformer network for image inpainting. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021 Oct 20–24; Online. doi:10.1145/3474085.3475426.
6. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/ICCV48922.2021.00986.
7. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
8. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015; 2015 Oct 5–9; Munich, Germany. doi:10.1007/978-3-319-24574-4_28.

9. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. Context encoders: feature learning by inpainting. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. doi:10.1109/CVPR.2016.278.
10. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. arXiv:1406.2661v1. 2014.
11. Nazeri K, Ng E, Joseph T, Qureshi F, Ebrahimi M. EdgeConnect: structure guided image inpainting using edge prediction. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW); 2019 Oct 27–28; Seoul, Republic of Korea. doi:10.1109/iccvw.2019.00408.
12. Guo X, Yang H, Huang D. Image inpainting via conditional texture and structure dual generation. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/ICCV48922.2021.01387.
13. Liao L, Xiao J, Wang Z, Lin CW, Satoh S. Image inpainting guided by coherence priors of semantics and textures. arXiv:2012.08054v1. 2020.
14. Li J, Wang N, Zhang L, Du B, Tao D. Recurrent feature reasoning for image inpainting. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. doi:10.1109/cvpr42600.2020.00778.
15. Suvorov R, Logacheva E, Mashikhin A, Remizova A, Ashukha A, Silvestrov A, et al. Resolution-robust large mask inpainting with Fourier convolutions. arXiv:2109.07161v2. 2021.
16. Liu G, Reda FA, Shih KJ, Wang TC, Tao A, Catanzaro B. Image inpainting for irregular holes using partial convolutions. arXiv:1804.07723v2. 2018.
17. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang T. Free-form image inpainting with gated convolution. arXiv:1806.03589v2. 2018.
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA.
19. Yang B, Wang X, Xing Y, Cheng C, Jiang W, Feng Q. Modality fusion vision transformer for hyperspectral and LiDAR data collaborative classification. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17(21):17052–65. doi:10.1109/JSTARS.2024.3415729.
20. Bai J, Fan Y, Zhao Z, Zheng L. Image inpainting technique incorporating edge prior and attention mechanism. Comput Mater Contin. 2024;78(1):999–1025. doi:10.32604/cmc.2023.044612.
21. Chen S, Atapour-Abarghouei A, Shum HPH. HINT: high-quality INpainting transformer with mask-aware encoding and enhanced attention. IEEE Trans Multimed. 2024;26:7649–60. doi:10.1109/TMM.2024.3369897.
22. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proceedings of the Computer Vision—ECCV 2020; 2020 Aug 23–28; Glasgow, UK. doi:10.1007/978-3-030-58452-8_13.
23. Yang F, Yang H, Fu J, Lu H, Guo B. Learning texture transformer network for image super-resolution. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. doi:10.1109/cvpr42600.2020.00583.
24. Liu S, Chi J, Wu C, Xu F, Yu X. SGT-net: a transformer-based stratified graph convolutional network for 3D point cloud semantic segmentation. Comput Mater Contin. 2024;79(3):4471–89. doi:10.32604/cmc.2024.049450.
25. Wang W, Xie E, Li X, Fan DP, Song K, Liang D, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/ICCV48922.2021.00061.
26. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH. Restormer: efficient transformer for high-resolution image restoration. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/CVPR52688.2022.00564.
27. Deng Y, Hui S, Zhou S, Meng D, Wang J. T-former: an efficient transformer for image inpainting. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisboa, Portugal. doi:10.1145/3503161.3548446.
28. Lingle LD. Transformer-VQ: linear-time transformers via vector quantization. arXiv:2309.16354. 2023.

29. Huang W, Deng Y, Hui S, Wu Y, Zhou S, Wang J. Sparse self-attention transformer for image inpainting. Pattern Recognit. 2024;145(3):109897. doi:10.1016/j.patcog.2023.109897.
30. Cohen TS, Welling M. Group equivariant convolutional networks. Proc Mach Learn Res. 2016;48:2990–9.
31. Ma YJ, Shuai HH, Cheng WH. Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation. IEEE Trans Multimed. 2021;24:261–73. doi:10.1109/TMM.2021.3050059.
32. Chen Y, Dai X, Liu M, Chen D, Yuan L, Liu Z. Dynamic convolution: attention over convolution kernels. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. doi:10.1109/CVPR42600.2020.01104.
33. Hassani A, Shi H. Dilated neighborhood attention transformer. arXiv:2209.15001. 2022.
34. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. doi:10.1109/ICCV.2017.244.
35. Doersch C, Singh S, Gupta A, Sivic J, Efros AA. What makes Paris look like Paris? Commun ACM. 2015;58(12):103–10. doi:10.1145/2830541.
36. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196. 2017.
37. Quan W, Zhang R, Zhang Y, Li Z, Wang J, Yan DM. Image inpainting with local and global refinement. IEEE Trans Image Process. 2022;31:2405–20. doi:10.1109/tip.2022.3152624.
38. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. doi:10.1109/ICCV.2017.74.
39. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019 Aug 4–8; Anchorage, AK, USA. doi:10.1145/3292500.3330701.