Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.066842

ARTICLE

An Image Inpainting Approach Based on Parallel Dual-Branch Learnable Transformer Network

Rongrong Gong[1,#], Tingxian Zhang[2,#], Yawen Wei[2], Dengyong Zhang[2] and Yan Li[3,*]

[1]School of Software, Changsha Social Work College, Changsha, 410004, China
[2]School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, 410076, China
[3]Department of Computer Engineering, INHA University, Incheon, 22201, Republic of Korea

*Corresponding Author: Yan Li. Email: leeyeon@inha.ac.kr

[#]These authors contributed equally to this work

Received: 18 April 2025; Accepted: 30 June 2025; Published: 29 August 2025

ABSTRACT: Image inpainting refers to synthesizing missing content in an image based on known information to restore occluded or damaged regions, which is a typical manifestation of this trend. With the increasing complexity of image in tasks and the growth of data scale, existing deep learning methods still have some limitations. For example, they lack the ability to capture long-range dependencies and their performance in handling multi-scale image structures is suboptimal. To solve this problem, the paper proposes an image inpainting method based on the parallel dual-branch learnable Transformer network. The encoder of the proposed model generator consists of a dual-branch parallel structure with stacked CNN blocks and Transformer blocks, aiming to extract global and local feature information from images. Furthermore, a dual-branch fusion module is adopted to combine the features obtained from both branches. Additionally, a gated full-scale skip connection module is proposed to further enhance the coherence of the inpainting results and alleviate information loss. Finally, experimental results from the three public datasets demonstrate the superior performance of the proposed method.

KEYWORDS: Artificial intelligence; image inpainting; transformer network; dual-branch fusion; gated full-scale skip connection

1 Introduction

With the rapid advancement of artificial intelligence (AI) technology, we are entering a new era driven by intelligent systems that are not only capable of understanding complex data patterns, but also adapting to changing environmental demands. The proliferation of AI-driven adaptive systems brings unprecedented capabilities for managing distributed computation, fostering innovation in a variety of fields. In this context, image restoration techniques under deep learning are particularly prominent. This technique utilizes existing information to reconstruct damaged images through deep learning models, especially Convolutional Neural Networks (CNNs), which not only improves the accuracy of restoration, but also speeds up the processing speed, making it an important tool for processing image data in distributed computing environments.

Deep learning-based methods for image inpainting have demonstrated substantial advancements in practical applications, largely due to their strengths in automatic learning, contextual modeling, advanced feature representation, and improved inpainting results. In our study, we focus on three key categories of deep learning techniques employed for image inpainting.

The first category encompasses CNN-based approaches, which have laid the groundwork for image inpainting with their ability to capture local patterns and features effectively. Notable examples of CNN-based methods include LeNet [1] and the Neocognitron [2], both of which have contributed significantly to the evolution of image processing techniques. The second category includes generative adversarial network (GAN)-based methods. GANs, further refined by Miyato et al., are designed to generate realistic image content through the interaction between the generator and the discriminator. This adversarial process enhances the quality and realism of the inpainted images. The third category features the Transformer model, originally proposed by Vaswani et al., which has been adapted for image inpainting tasks to leverage its strength in handling complex contextual information. By examining these three categories, our study aims to provide an in-depth analysis of advanced techniques in deep learning for image inpainting, highlighting their specific contributions and advancements in enhancing image inpainting quality.

CNN is suitable for image inpainting tasks. By utilizing shared weights and local connections, CNN reduces the model parameters and computational complexity, leading to significant achievements in various domains [3]. However, CNN has a relatively weak grasp of global information, which may result in a lack of contextual consistency in the inpainting results under certain circumstances.

Generative Adversarial Networks (GANs) are adept at producing high-quality restored images by leveraging adversarial training to improve the realism of inpainting results. Despite their advantages, GANs face challenges during training, including instability and the need for careful balance between the generator and discriminator. Issues such as non-convergence, mode collapse [4], and potential artifacts or blurriness in the generated images can also arise, affecting the final output quality.

The Transformer model demonstrated excellent performance in image inpainting. It possesses a self-attention mechanism that models the global correlations in images. Furthermore, the interpretability of Transformers provides valuable information for model optimization, allowing us to understand the attention levels of the model at each position towards other positions [5]. However, its computational complexity increases quadratically.

Overall, traditional CNNs effectively capture local features but struggle with large missing regions, often causing structural and semantic inconsistencies. In contrast, Transformers model global dependencies well but lack fine-grained detail, leading to blurred textures. Most existing methods combine the two superficially, making it difficult to achieve a balanced synergy between local feature extraction and global semantic reasoning.

To address this, we propose the Parallel Dual-Branch Learnable Transformer Network (PDT-Net), which features a dual-encoder architecture combining CNNs and transformers for joint local-global feature extraction. A dual-branch decoder and feature fusion strategy further preserve fine textures and spatial coherence. Skip connections enhance low-level detail propagation throughout the network. PDT-Net represents a significant advancement in image inpainting technology, incorporating both convolutional and transformer-based techniques. We conducted a comprehensive series of experiments using three distinct datasets (Paris Street View, CelebaA, and Places2) to thoroughly assess the network's performance [6–8]. These datasets provide a broad evaluation of the network's capabilities across various types of image data. The results of these experiments, which highlight the network's impressive performance and advantages, are visually represented in Fig. 1.

Figure 1: Figure showcasing the inpainting results of PDT-Net model

Our contributions include the following:

- An innovative image inpainting model that combines CNN and Transformer is proposed: parallel adoption of CNN blocks and Transformer blocks for feature extraction via downsampling.
- A dual-branch fusion module is proposed to integrate the globally and locally extracted features, thereby improving the quality of image inpainting.
- Through extensive validation on three datasets, the proposed model outperformed existing image inpainting methods across the board.

2 Related Work

2.1 Deep Learning Image Inpainting

Deep learning-based image inpainting methods offer notable advantages due to their sophisticated automatic learning and contextual modeling capabilities. Unlike traditional techniques that depend on predefined rules and basic algorithms, deep learning models use advanced neural networks to automatically learn complex patterns, textures, and features from extensive datasets. This capability allows the models to capture extensive visual information without manual intervention. Moreover, these models are proficient in contextual modeling, taking into account not just the nearby area but the entire image context. This comprehensive approach results in inpainting that integrates seamlessly with the surrounding content, ensuring high visual coherence. Consequently, deep learning methods achieve more accurate and realistic inpainting results for damaged images, even in cases of irregular damage and varied textures. Their enhanced ability to deliver visually satisfying results marks a significant improvement over traditional inpainting techniques.

Initially, Pathak et al. [9] attempted to use CNN for image inpainting. Subsequently, in order to propel the progress of deep learning-based image inpainting, researchers have introduced a range of cutting-edge and inventive methodologies. For instance, Liu et al. [10] proposed a probabilistic diversity Generative Adversarial Network called PD-GAN, which introduces random noise and probabilistic sampling to generate high-quality and diverse image inpainting results. Although the introduction of noise can increase diversity, it may sometimes result in inpainting outputs with noise or blurriness, which can impact the quality and authenticity of the inpainting results.

Furthermore, certain approaches integrate attention mechanisms and context modeling into image inpainting networks, elevating their ability to restore crucial regions and intricate details within the image. In conclusion, deep learning-based image inpainting techniques offer an efficient and precise approach to

image restoration tasks through the utilization of automatic learning and contextual modeling features of neural networks, along with their ability to manage multi-scale image structures [11].

2.2 Transformer

The Transformer was first introduced by Vaswani et al. [12] in 2017 to address natural language processing (NLP) tasks. Building on the Transformer, Dosovitskiy et al. [13] introduced the Vision Transformer (ViT) in 2020, successfully applying it to computer vision tasks. ViT segments images into a series of patches, treats each patch as an input sequence, and uses the Transformer model for feature extraction and classification.

With the successful application of Transformer, researchers have made further improvements and extensions to it. For example, Liu et al. [14] have opened up new research directions and expanded the application scope of visual Transformer mechanisms by addressing the challenges of processing large-scale images and providing efficient computational methods. Yuan et al. [15] have improved the performance of Transformer networks by adopting more efficient training strategies that leverage the structural information reconstruction of images.

These advancements have resulted in considerable advancements in applying the Transformer mechanism to a broader spectrum of visual tasks. These studies have also demonstrated the powerful potential of the Transformer mechanism in addressing image-related problems. However, despite the Transformer's advantage in capturing long-range dependencies and generating diverse structures. Nevertheless, the Transformer significantly increases computational complexity, making it challenging to handle high-resolution images.

2.3 Model Combining CNN and Transformer

Recent studies have shown significant performance improvement in computer vision tasks by combining Convolutional Neural Networks (CNNs) and Transformers. CNNs excel at image feature extraction, while Transformers have advantages in handling long-range dependencies and capturing global relationships. CNNs excel at image feature extraction, while Transformers have advantages in handling long-range dependencies and capturing global relationships. Due to the expertise of Transformers in sequence modeling and CNNs in local feature extraction, the dual-branch encoder-decoder applies to various tasks.

For example, recent studies [16–20] have also explored methods for image inpainting, where Transformers are employed to reconstruct complex coherent structures and rough textures, while CNNs enhance local texture details guided by the rough restored images. The effective fusion of global contextual information and local features during the inpainting stage has greatly enhanced the overall results. Such a combination of CNNs and Transformers has achieved significant progress in image inpainting tasks.

Models that combine Convolutional Neural Networks (CNNs) and Transformers showcase exceptional abilities in both feature extraction and sequence modeling. CNNs are effective at capturing intricate spatial details and patterns in image. In contrast, Transformers are adept at modeling sequences and understanding long-range dependencies, which is essential for applications like natural language processing and time-series analysis. By integrating CNNs with Transformers, these hybrid models benefit from the strengths of both approaches: CNNs for detailed spatial feature extraction and Transformers for handling sequential data efficiently. This combination enhances versatility, allowing these models to excel in various deep learning tasks, including the mechanism has superior capabilities to capture the local and nonlocal dependencies on face image in the face reconstruction [21]. Consequently, the fusion of CNN and Transformer architectures is becoming increasingly prevalent and valuable in advancing deep learning technologies.

3 Our Approach

3.1 Network Architecture

The parallel dual-branch learnable Transformer network (PDT-Net) proposed by us for image inpainting tasks is illustrated in Fig. 2. In our network, the generator adopts an encoder-decoder architecture, comprising two concurrent encoders. Each encoder utilizes different feature extraction methods and model architectures. The objective of this architecture is to maximize the benefits of distinct encoders, enhancing both the expressive capability and performance of the model.

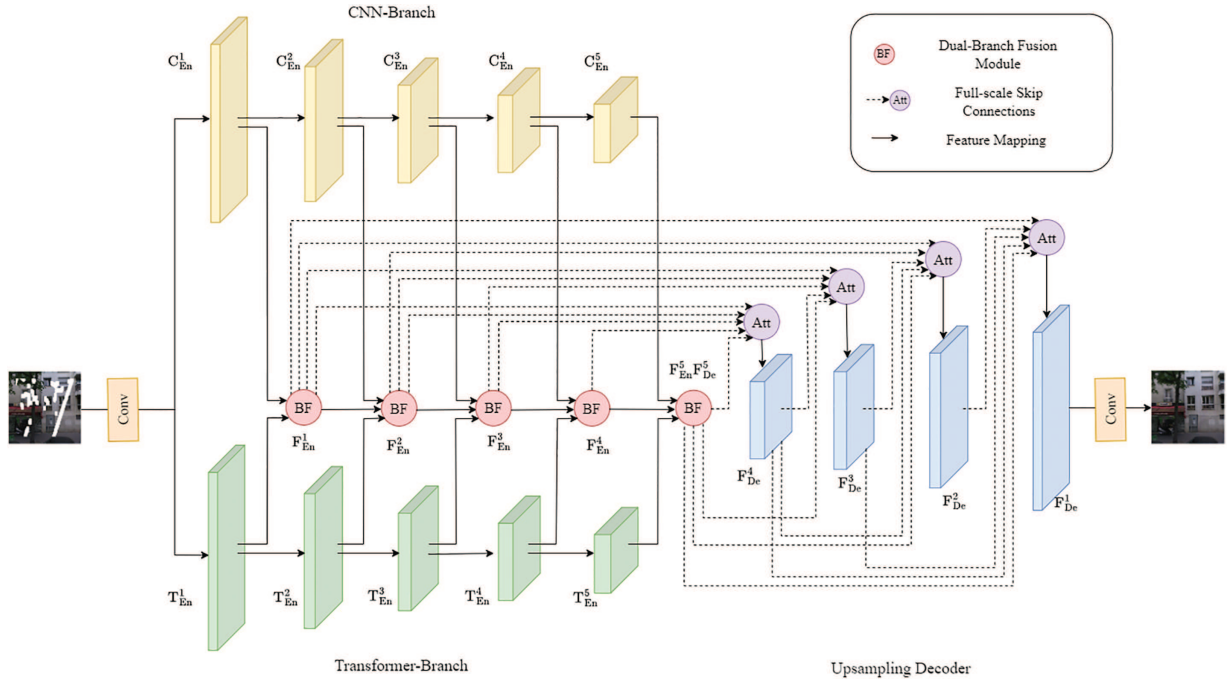

Figure 2: Inpainting results of the proposed PDT-Net model

The features extracted by the dual-branch encoders are denoted as $C_{En}^{i} \in \mathbb{R}^{\frac{H}{2^{(i-1)}} \times \frac{W}{2^{(i-1)}} \times 2^{(i-1)}C'}$ and $T_{En}^{i} \in \mathbb{R}^{\frac{H}{2^{(i-1)}} \times \frac{W}{2^{(i-1)}} \times 2^{(i-1)}C'}$, where i represents the number of encoding layers, and $i = 1, \cdots, 5$. The outputs of these two encoders are fused through the dual-branch fusion module to integrate the feature representations from both branches, resulting in the fused feature of the encoding layers, denoted as $F_{En}^{i} \in \mathbb{R}^{\frac{H}{2^{(i-1)}} \times \frac{W}{2^{(i-1)}} \times 2^{(i-1)}C'}$. Then, by utilizing full-scale skip connections, we obtain rich fusion multi-scale feature information with global and local feature interactions, represented as the decoding layer feature map $F_{De}^{i} \in \mathbb{R}^{\frac{H}{2^{(i-1)}} \times \frac{W}{2^{(i-1)}} \times 2^{(i-1)}C'}$. Finally, the image details and textures are restored through the decoder.

3.1.1 CNN-Transformer Encoder

The dual-branch encoder architecture integrates two distinct branches to capture both local and global dependencies in data. The first branch utilizes Convolutional Neural Networks (CNNs) for encoding spatial or local features through convolutional operations. This approach excels in identifying patterns and structures within specific areas of the input data. Meanwhile, the second branch uses Transformer models to model extended dependencies and sequential relationships through self-attention mechanisms. Together,

these two branches complement each other, allowing the model to effectively process and integrate both fine-grained local details and broader, global contexts. This dual-branch setup enhances the overall capability of the encoder to handle diverse types of information.

The CNN branch encoder extracts local features and texture information from the input data. The convolutional layers capture local features at different positions. This feature extraction approach enables the CNN branch encoder to effectively capture spatial locality in images and exhibit certain robustness to image translation and scale variations.

The main feature of the Transformer encoder is its ability to capture relationships between different positions in a sequence. We introduce a learnable Transformer module to further enhance the feature representation and context modeling capabilities during the inpainting process. In the self-attention mechanism, linear transformations are used to compute attention weights and context representations, replacing the traditional dot product operation with matrix multiplication [12]. The introduction of more parameters and non-linear transformations can enhance the expressive power of the model [22]. Fig. 3 illustrates the execution process of a Transformer module.

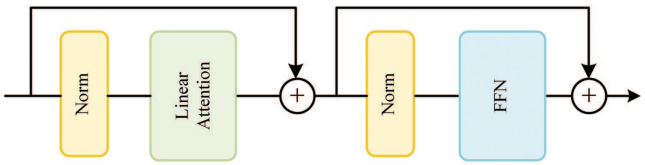

Figure 3: Linear transformer module

3.1.2 Dual-Branch Fusion Module

The detailed structure of the dual-branch fusion module is illustrated in Fig. 4. This module plays a crucial role in combining the outputs from the CNN branch encoder and the Transformer branch encoder. By merging their extracted features into a unified set of fusion features with consistent dimensions, the dual-branch fusion module facilitates effective interaction between the two branches during the inpainting process. The ability of the module to harmonize these different types of information contributes to improved image fidelity and overall quality in the inpainting process.

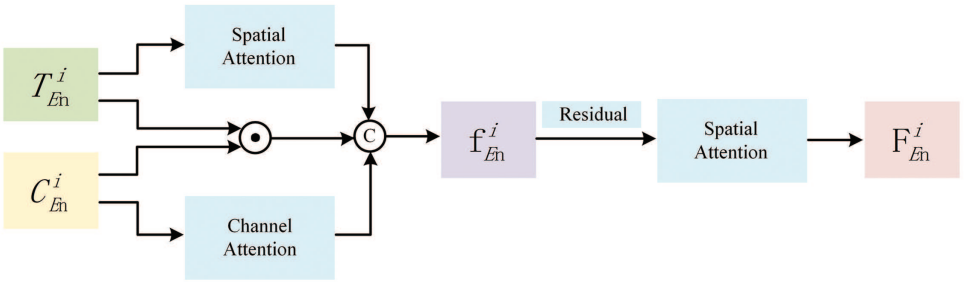

Figure 4: Dual-branch fusion module

Specifically, we obtain the fused feature F_{En}^{i} of the encoding layers through the following operations, as shown in Eqs. (1) and (2).

$$f_{En}^{i} = \left[\left(T_{En}^{\mathrm{i}} \odot C_{En}^{\mathrm{i}}\right), SA\left(T_{En}^{\mathrm{i}}\right), CA\left(C_{En}^{\mathrm{i}}\right)\right] \tag{1}$$

$$F_{En}^{\mathrm{i}} = GA\left(Res\left(f_{En}^{i}\right)\right) \tag{2}$$

Among them, f_{En}^{i} represents the intermediate feature obtained by simple channel concatenation. $SA(\cdot)$ represents spatial attention, $CA(\cdot)$ represents channel attention, $[\cdot]$ represents channel concatenation, $Res(\cdot)$ represents residual block, and $GA(\cdot)$ represents gate mechanism.

3.1.3 Full Scale Skip Connection

Fig. 5 shows the detailed structure diagram of the full-scale skip connection. We incorporate a gated attention mechanism at the end of the traditional full-scale skip connection. The full-scale skip connection with a gated attention mechanism achieves feature transmission and fusion across different levels. It selectively combines the source features and target features through weighted fusion, transferring low-level detail information to higher levels while preserving high-level semantic information.

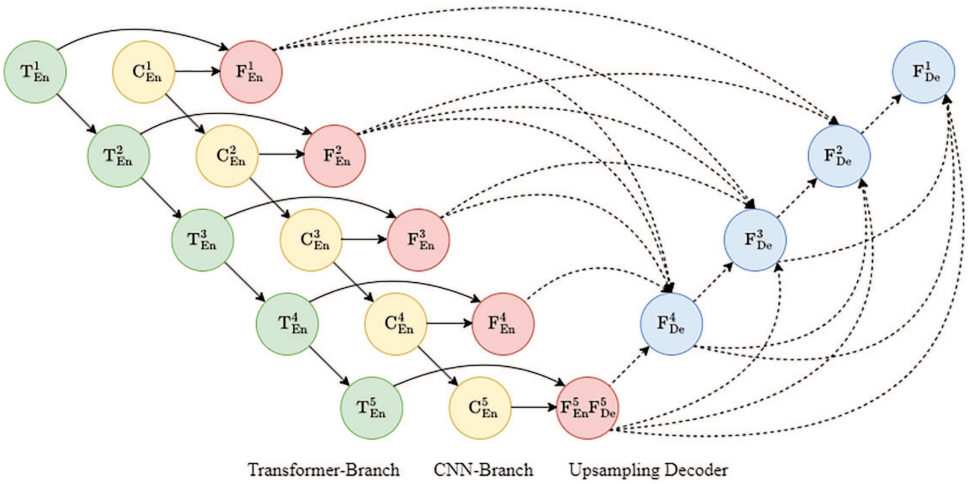

Figure 5: Full-scale skip connection

In the formula, the calculation of the feature map stack for the decoding layers represented by C_{De}^{i} is as follows, as shown in Eqs. (3)–(6):

$$c_1 = Att\left(Conv\left(Down\left(F_{En}^{k}\right)\right)_{k=1}^{i-1}\right) \tag{3}$$

$$c_2 = Att\left(Conv\left(F_{En}^{i}\right)\right) \tag{4}$$

$$c_3 = Att\left(Conv\left(Up\left(F_{De}^{k}\right)\right)_{k=i+1}^{N}\right) \tag{5}$$

$$F_{De}^{i} = \begin{cases} F_{En}^{i}, i = N \\ Conv\left(Concat\left(c_1, c_2, c_3\right)\right), i = 1, \cdots N-1 \end{cases} \tag{6}$$

Among them, the parameters C_1, C_2, and C_3 represent the larger size, same size and smaller size of the feature compared to the current size after simple processing for fusing full-scale features, respectively.

The function $Conv(\cdot)$ represents the operation of convolution. $Att(\cdot)$ represents the feature fusion mechanism achieved through operations like spatial and channel attention and convolution. Furthermore, $Down(\cdot)$ and $Up(\cdot)$ represent the operations of upsampling and downsampling, respectively.

3.2 Loss Function

We optimize our PDT-Net using a joint loss function L, which includes multiple components: the reconstruction loss ℓ_{re}, the adversarial loss ℓ_{adv} [23], the perceptual loss ℓ_{p} [24], and the style loss ℓ_{s} [25]. These loss functions are commonly employed in various image inpainting methods [18] and are defined as shown in Eqs. (7)–(11):

$$\ell_{\text{re}} = \left\| I_{out} - I_{gt} \right\|_1 \tag{7}$$

$$\ell_{adv} = \mathrm{E}_{I_{gt}} \left[logD\left(I_{gt}\right) \right] + \mathrm{E}_{I_{out}} log\left[1 - D\left(I_{out}\right)\right] \tag{8}$$

$$\ell_p = \mathrm{E}\left[\sum_i \frac{1}{N_i} \left\| \phi_i\left(I_{out}\right) - \phi_i\left(I_{gt}\right) \right\|_1 \right] \tag{9}$$

$$\ell_s = \mathrm{E}_i \left[\left\| \phi_i(I_{out})^T \phi_i\left(I_{out}\right) - \phi_i\left(I_{gt}\right)^T \phi_i\left(I_{gt}\right) \right\| \right] \tag{10}$$

$$L = \lambda_{re}\ell_{re} + \lambda_{adv}\ell_{adv} + \lambda_p\ell_p + \lambda_s\ell_s \tag{11}$$

where D represents the PatchGAN discriminator [26] with spectral normalization. ϕ_i refers to the i-th layer activation function of the VGG19 [27] network pre-trained on the ImageNet [28] dataset. N_i represents the number of elements in ϕ_i. λ_{re}, λ_{adv}, λ_p, and λ_s are the weight ratios of the corresponding loss functions. We set $\lambda_{re} = 1$, $\lambda_{adv} = 0.1$, $\lambda_p = 1$, and $\lambda_s = 250$.

4 Experiments

4.1 Datasets

As shown in Table 1, the original image dataset consists of three public datasets. Our experiments fully utilized the training and testing sets of the Paris street view dataset, which is composed of a large number of images collected from the streets, for evaluation purposes and strictly followed its original split configuration. For the CelebA and Places2 datasets, we divided the training, validation, and testing sets with a ratio of 8:1:1 for our experiments. The CelebA dataset primarily focuses on human faces and consists of 202,599 face images. We selected 50,000 images from the CelebA dataset for training and 6250 images from the testing set for evaluation, images were selected sequentially from the beginning of the dataset. The Places2 dataset covers various scenes and environments, containing millions of images. We selected twenty scene categories from the Places2 dataset, including attics, airports, arches, campuses, and more. Each category has 5000 training images, making a total of 100,000 training images. We selected 12,500 images from the testing set for evaluation, 5000 images were selected from each relevant subset.

Table 1: Dataset details for image analysis experiments

Dataset name	Paris street view	CelebA	Places2
Main content	A vast collection of images gathered from the streets	Primarily focuses on human faces, 202,599 in total	Encompasses various scenes and environments
Training set size	14,900	50,000 (Randomly Selected)	100,000 (Randomly Selected)
Test set size	100	6250 (Randomly Selected)	12,500 (Randomly Selected)

4.2 Implementation Details

The experiments were conducted on a system running Ubuntu 18.04, equipped with a single NVIDIA A30 Tensor Core GPU that has a memory capacity of 24 GB. For implementing the experiments, we used the PyTorch framework, which is well-suited for deep learning tasks due to its flexibility and efficiency. To optimize the model parameters, we employed the Adam optimizer [29], known for its effectiveness in handling large datasets and complex neural network structures. The Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.9$ was used to train the model, the learning rate was set to 10^{-4}, and later it was adjusted to 10^{-5} to fine-tune the model. The choice of hardware and software tools was aimed at ensuring robust performance and accurate results, with the NVIDIA A30 GPU providing the necessary computational power and the PyTorch framework facilitating smooth implementation and experimentation.

4.3 Comparative Experiment

To offer a clearer evaluation of our model's performance in image inpainting, we compared it against five leading image inpainting models across three different datasets. We utilized three well-established metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [30], and the ℓ_1 loss function. Higher PSNR and SSIM values indicate better reconstruction quality and visual similarity, while lower ℓ_1 loss values suggest better pixel-level accuracy. We ensured consistency in our experimental setup by using the same hardware and software environment for all models, and we maintained a uniform mask coverage rate throughout the experiments. This approach allowed for a fair and comprehensive assessment of our model's performance relative to its peers.

- PC [31]: Introduces a method for repairing irregular holes in images using partial convolution techniques.
- RFR [32]: Introduces a progressive image inpainting network based on recurrent feature reasoning.
- AOT [33]: Enables context reasoning by capturing information-rich remote contexts and diverse patterns of interest.
- CTSDG [34]: A new dual-branch network for image inpainting that seamlessly combines structurally constrained texture synthesis with texture-guided structural reconstruction.
- T-former [22]: A Transformer network for image inpainting is proposed, incorporating an attention mechanism linearly correlated with the resolution.

The comparative experimental results presented in Table 2 clearly indicate that our model surpasses all baseline models across the three evaluation metrics. This performance improvement is further reflected in the enhanced visual coherence of the inpainted images, as shown in Fig. 6. Additionally, Fig. 6 provides a detailed visual comparison: it features the original corrupted image, the results from PC, RFR, AOT, CTSDG,

T-former, our PDT-Net, and the ground truth image. This sequence of images highlights the improved fidelity and visual quality achieved by PDT-Net compared to the other methods.

Table 2: We present the quantitative comparison results obtained from conducting experiments on three publicly available datasets using our PDT-Net model

Datasets		Paris street view			CelebA			Places2		
Mask ratio		0%–20%	20%–40%	40%–60%	0%–20%	20%–40%	40%–60%	0%–20%	20%–40%	40%–60%
PSNR↑	PC	32.12	25.56	21.26	31.85	26.55	21.34	29.39	24.26	21.88
	RFR	32.67	26.31	22.41	33.33	27.62	22.63	29.90	24.96	22.14
	AOT	32.80	26.36	22.66	33.58	27.73	22.80	29.96	25.05	22.29
	CTSDG	32.95	27.51	22.89	33.97	27.81	22.94	30.18	25.52	22.58
	T-former	33.12	27.83	22.97	34.10	27.94	23.08	30.36	25.74	22.87
	Ours	33.35	27.95	23.16	34.27	28.06	23.22	30.45	25.99	23.11
SSIM↑	PC	0.897	0.690	0.500	0.897	0.749	0.557	0.887	0.731	0.527
	RFR	0.920	0.773	0.569	0.917	0.781	0.602	0.899	0.751	0.554
	AOT	0.923	0.776	0.572	0.919	0.783	0.603	0.901	0.757	0.559
	CTSDG	0.924	0.777	0.574	0.921	0.787	0.610	0.905	0.760	0.566
	T-former	0.926	0.779	0.578	0.923	0.788	0.614	0.907	0.769	0.569
	Ours	0.927	0.781	0.579	0.926	0.792	0.616	0.911	0.774	0.570
ℓ_1(%)↓	PC	0.057	0.133	0.272	0.044	0.120	0.220	0.064	0.132	0.281
	RFR	0.041	0.113	0.233	0.031	0.090	0.185	0.049	0.100	0.238
	AOT	0.041	0.112	0.231	0.030	0.089	0.183	0.047	0.098	0.237
	CTSDG	0.038	0.105	0.228	0.028	0.080	0.178	0.041	0.094	0.226
	T-former	0.035	0.103	0.224	0.025	0.079	0.175	0.038	0.091	0.223
	Ours	0.034	0.103	0.223	0.025	0.078	0.173	0.037	0.090	0.223

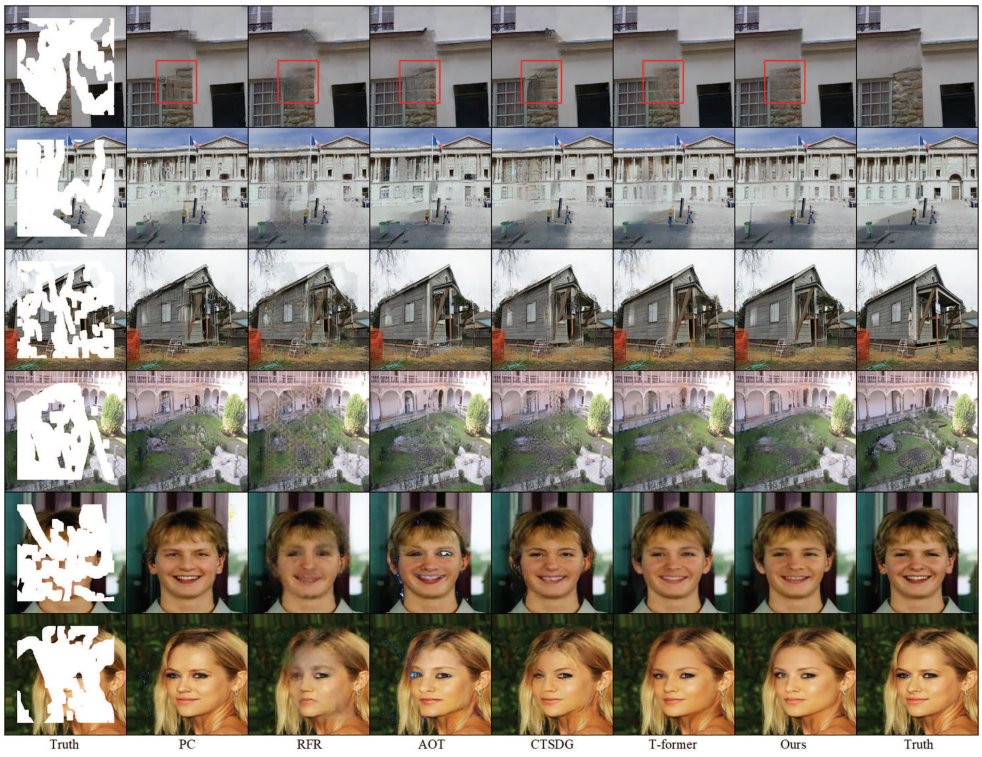

Figure 6: Inpainting results of the proposed PDT-Net model

4.4 Ablation Study

We performed a series of ablation experiments on the dataset of Paris street view. By individually eliminating the crucial components of our proposed method, we reevaluated the results of inpainting. Net_1 represents the removal of the linear attention module in the Transformer encoder, Net_2 represents the replacement of the dual-branch fusion module with regular convolutions, and Net_3 represents the replacement of the full-scale skip connections with regular skip connections. The experimental outcomes showcased in Table 3 substantiate the vital role and efficacy of the incorporated components within the method. Removing these key components noticeably deteriorated the quality of the inpainting results, further affirming their significance. The comprehensive network model, as depicted in Fig. 7, excels in reinstating intricate texture details and enhancing the coherency of the reconstructed structures.

Table 3: Quantitative ablation study of our PDT-Net model on the paris street view dataset

Models	Full-scale skip connections	Dual-branch fusion module	Linear Transformer	PSNR↑	SSIM↑	ℓ_1(%)↓
Net_1		✓	✓	28.73	0.878	0.162
Net_2	✓		✓	29.54	0.879	0.162
Net_3	✓	✓		29.58	0.881	0.158
Ours	✓	✓	✓	29.60	0.882	0.156

Figure 7: Ablation experiment results of our proposed PDT-Net model on the Paris street view dataset

Through extensive comparative and ablation studies, we have rigorously confirmed the advantages and effectiveness of our proposed approach for image inpainting tasks. Our rigorous analysis of the experimental data reveals that our approach consistently delivers higher quality results and enhanced performance compared to existing methods. Additionally, our method outperforms variants of other approaches where key components have been removed. The proposed PDT-Net has approximately 52.1 million parameters and requires 172.1 GFLOPs, reflecting a reasonable computational cost given its dual-branch architecture.

5 Conclusion

In the field of image inpainting, we present an innovative parallel dual-branch learnable Transformer network. This advanced network architecture incorporates a CNN-Transformer dual encoder, which excels at extracting features from both local and global perspectives. By integrating these two types of encoders, our approach significantly enhances the effectiveness and precision of the inpainting process.

Furthermore, we introduce a sophisticated dual-branch fusion module along with a comprehensive skip connection mechanism. These components work together to propagate more detailed information throughout the network while preserving the structural integrity of the image. As a result, our method produces inpainting outcomes that are both more natural and realistic, demonstrating improved performance and visual coherence compared to existing approaches.

Our empirical analysis confirms that this method not only advances image inpainting capabilities but also effectively upholds the visual consistency of the inpainted images. The superior quality of the inpainting outcomes highlights a greater sense of naturalness and realism, offering valuable insights for ongoing research and development in image inpainting and related domains. This framework also shows potential for practical applications such as photo restoration and object removal.

In future work, we plan to explore real-time performance, extend the method to high-resolution scenarios, and investigate how to improve the anti-forensic robustness of inpainted images to prevent easy detection.

Acknowledgement: Not applicable.

Funding Statement: This paper was supported by Scientific Research Fund of Hunan Provincial Natural Science Foundation under Grant 2023JJ60257, Hunan Provincial Engineering Research Center for Intelligent Rehabilitation Robotics and Assistive Equipment under Grant 2025SH501, Inha University and Design of a Conflict Detection and Validation Tool under Grant HX2024123.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization and methodology: Rongrong Gong, Tingxian Zhang; data curation and investigation: Yawen Wei; writing—original draft preparation: Rongrong Gong, Tingxian Zhang, Yawen Wei; funding acquisition: Rongrong Gong; writing—review and editing: Dengyong Zhang; resources and supervision: Yan Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All data included in this study are available upon request by contact with the corresponding author.

Ethics Approval: Not applicable.

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

References

1. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324. doi:10.1109/5.726791.
2. Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern. 1980;36(4):193–202. doi:10.1007/bf00344251.
3. 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.
4. Wu Q, Chen Y, Meng J. DCGAN-based data augmentation for tomato leaf disease identification. IEEE Access. 2020;8:98716–28. doi:10.1109/access.2020.2997001.

5. Liu Q, Tan Z, Chen D, Chu Q, Dai X, Chen Y, et al. Reduce information loss in transformers for pluralistic image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 11347–57.
6. Doersch C, Singh S, Gupta A, Sivic J, Efros A. What makes paris look like paris? ACM Trans Graph. 2012;31(4):101–9. doi:10.1145/2185520.2185597.
7. Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile. p. 3730–8.
8. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell. 2017;40(6):1452–64. doi:10.1109/tpami.2017.2723009.
9. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and PATTERN Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2536–44.
10. Liu H, Wan Z, Huang W, Song Y, Han X, Liao J. PD-GAN: probabilistic diverse GAN for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 9371–81.
11. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS. Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 4471–80.
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv: 1706.03762. 2017.
13. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
14. 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 IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, QC, Canada. p. 10012–22.
15. Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang ZH, et al. Tokens-to-Token ViT: training vision transformers from scratch on ImageNet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, QC, Canada. p. 558–67.
16. Wan Z, Zhang J, Chen D, Liao J. High-fidelity pluralistic image completion with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, QC, Canada. p. 4692–701.
17. Zheng C, Cham TJ, Cai J, Phung D. Bridging global context interactions for high-fidelity image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 11512–22.
18. Dong Q, Cao C, Fu Y. Incremental transformer structure enhanced image inpainting with masking positional encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 11358–68.
19. 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.
20. Jin Z, Qiu Y, Zhang K, Li H, Luo W. MB-TaylorFormer V2: improved multi-branch linear transformer expanded by taylor formula for image restoration. arXiv:2501.04486. 2025.
21. Shi J, Wang Y, Yu Z, Li G, Hong X, Wang F, et al. Exploiting multi-scale parallel self-attention and local variation via dual-branch transformer-CNN structure for face super-resolution. IEEE Trans Multimed. 2023;26:2608–20. doi:10.1109/tmm.2023.3301225.
22. Deng Y, Hui S, Zhou S, Meng D, Wang J. An efficient transformer for image inpainting. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisboa, Portugal. p. 6559–68.
23. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. IEEE Signal Process Mag. 2018;35(1):53–65. doi:10.1109/msp.2017.2765202.
24. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision-ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands. p. 694–711.

25. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2414–23.
26. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy. p. 2223–32.
27. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
28. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA. p. 248–55.
29. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv:1412.6980. 2014.
30. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12. doi:10.1109/tip.2003.819861.
31. Liu G, Reda FA, Shih KJ, Wang TC, Tao A, Catanzaro B. Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. p. 85–100.
32. Li J, Wang N, Zhang L, Du B, Tao D. Recurrent feature reasoning for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 7760–8.
33. Zeng Y, Fu J, Chao H, Guo B. Aggregated contextual transformations for high-resolution image inpainting. IEEE Trans Vis Comput Graph. 2023;29(7):3266–80. doi:10.1109/tvcg.2022.3156949.
34. Guo X, Yang H, Huang D. Image inpainting via conditional texture and structure dual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, QC, Canada. p. 14134–43.