Date of publication December 15, 2022.

A Diffusion Model with A FFT for Image Inpainting

Yuxuan Hu[1], Hanting Wang[2], Cong Jin[3], Bo Li[4] and Chunwei Tian[2, 5]
[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[2]School of Software, Northwestern Polytechnical University, Xi'an, 710129, China
[3]School of Information and Communication Engineering, Communication University of China, Beijing, 100024, China
[4]School of Electronics and Information, Northwestern Polytechnical University, Xi'an, 710072, China
[5]Research & Development Institute, Northwestern Polytechnical University, Shenzhen, 518057, China

Corresponding author: Chunwei Tian (e-mail: chunweitian@nwpu.edu.cn).

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110079, in part by the Youth Science and Technology Talent Promotion Project of Jiangsu Association for Science and Technology under Grant JSTJ-2023-017, in part by the Key Project of NSFC under Grant 61836016.

ABSTRACT Some convolutional neural networks cannot extract robust structural information for image inpainting under complex scenes. In this paper, we propose a diffusion model with an FFT (FFT-DM) to generate content that matches missing region texture and semantics to inpaint damaged images. Specifically, FFT-DM contains two components: a Denoising Diffusion Probabilistic Model (DDPM) and a Convolutional Neural Network (CNN). The DDPM is used to extract global features and generate image prior and the CNN is employed to capture more fine-grained details and predict the parameters in the reverse process of the diffusion model. Additionally, a Fast Fourier Transform (FFT) is fused into a diffusion model to enhance perception ability to improve expressive ability for image inpainting. The mentioned techniques can improve performance in image inpainting. Extensive experiments demonstrate that FFT-DM outperforms current state-of-the-art inpainting approaches in terms of qualitative and quantitative analysis.

INDEX TERMS Convolutional Neural Network, Diffusion model, Fast Fourier Transform, Image Inpainting.

I. INTRODUCTION

Image inpainting, is a practical and essential issue in image editing and image restoration. To restore the original image with high fidelity, image inpainting employs the known portion of the image as a prior to infer the missing region and generate content that is consistent with the structure, texture, and semantics of the surrounding pixel region[1]. Traditional methods of image inpainting attempt to address this issue through diffusion-based methods or patch-based techniques. Diffusion-based methods propagate information from adjacent regions to fill in the missing ones[2, 3]. However, this method is only suitable for small-scale mask inpainting, and it requires significant computational resources as the missing area grows. Patch-based methods define a distance metric between pixel patches and employ various search strategies to find similar patches to fill in the missing areas. Among these methods, texture-patch based[4, 5] methods can be vulnerable to pixel discontinuity due to the complexity and limitations of texture. Additionally, although the structure-patch based method[6, 7] performs well in regions with distinct boundaries, there still exhibits pixel discontinuity in complex backgrounds. Thus, due to the lack of high-level vision understanding, traditional inpainting methods face challenges in generating visually realistic and semantically plausible patches that are consistent with the original image when encountering large mask areas or complex image semantics.

By extracting the shallow features and deep semantic information from the image, deep learning-based inpainting methods effectively address the challenges that plague traditional methods. One such method applies a Convolutional Neural Network (CNN) to automatically extract deep image features, enhancing the accuracy and robustness of the process. Inspired by human visual perception connectivity, Alilou et al.[8] developed a general regression neural network (GRNN) for predicting missing regions in images by employing adjacent pixels to make accurate predictions. He et al.[9] established a masked autoencoder with an asymmetric encoder-decoder architecture to improve restoration in scenarios where a significant amount of information is missing. Nevertheless, CNN can learn features within localized regions but is limited in its ability to incorporate contextual information about the image at a global level.

Another category of image inpainting integrates the diffusion model, which relies on partial differential

equations (PDEs), to generate diverse restored images. In 2020, the Denoising Diffusion Probabilistic Model (DDPM)[10] gained prominence as an emerging paradigm of image generation. Dhariwal et al.[11] demonstrated that the performance of DDPM can outperform Generative Adversarial Network (GAN)-based methods in image synthesis. Moreover, the diffusion model updates the pixel values based on the gradients of the neighboring pixels, thereby achieving image smoothing while preserving the global structural information[12, 13]. Whilst the diffusion model can generate images with exceptional quality, its effectiveness may be limited when applied to complex images. This is because complex images typically contain intricate details and structural information that the diffusion model can only preserve at a global level. It is also worth noting that this approach also results in longer inference time due to the additional computations required at each iteration.

In this paper, we present a diffusion model with a Fast Fourier Transform (FFT) (FFT-DM) for image inpainting. Specifically, FFT-DM leverages the diffusion model to preserve the global structural information while utilizing a convolutional neural network to extract a maximal amount of fine-grained details. To further strengthen the effectiveness and efficiency of this model, we incorporate the Fast Fourier Transform (FFT) into the diffusion model to extract frequency domain information from images and remove the high-frequency noise and artifacts. Additionally, comprehensive experiments have illustrated the superiority of our proposed FFT-DM over existing state-of-the-art methods, including EdgeConnect[14], DeepFill v2[15], RegionWise[16], Aggregated COntextual-Transformation GAN (AOT-GAN)[17], and Co-Modulated GAN (CoModGAN)[18].

The main merits of our FFT-DM can be delineated as follows.

(1) A diffusion method is used to address image inpainting.

(2) FFT is exploited to be embedded into a diffusion to extract more frequency information to improve the performance of image inpainting.

The subsequent sections of this paper are organized as follows: Section 2 provides a literature review of the proposed method, including non-learning approaches to image inpainting, image inpainting based on deep learning, and diffusion model for image inpainting. The specifics of our proposed method are outlined in Section 3. In Section 4, we present a quantitative and qualitative analysis of the experimental results, focusing on the evaluation metrics and visual comparisons between the restored images and the originals. Finally, Section 5 offers a summary of this paper.

II. Related Work

A. Non-learning Approaches to Image Inpainting

In the field of image inpainting, there are non-learning methods that estimate the content of missing regions by analyzing the correlation between pixels or the similarity of content from the edge to the center. These methods comprise algorithms based on sparse representation and algorithms based on external data search. Sparse representation techniques[19, 20] utilize image patches for sparse decomposition, with the resulting information subsequently used for reconstructing the image restoration through signal reconstruction methods. Despite the favorable outcomes in image inpainting, the algorithm based on sparse representation exhibits high computational complexity during iterative training with a large dictionary. In situations where it is unsuitable to procure a dictionary, the processed image may exhibit visual incoherence, such as blurring or block effect. Moreover, due to the restricted texture information available in the damaged image, inpainting techniques that rely on external data search can enhance the pre-existing knowledge[21]. Also, Hays et al.[22] employed a large dataset to identify comparable images to the impaired image and subsequently restored the missing regions. Inspired by this, we believe that sufficient image prior knowledge can guide the process of image inpainting and improve its effectiveness and robustness.

B. Image inpainting based on deep learning

The emergence of deep learning-based methods and adversarial training has enabled significant progress in image inpainting. Compared to Non-learning algorithms, deep inpainting models can generate reasonable content and realistic fine-grained textures for complex scenes. Specifically, there are two kinds of deep learning-based image inpainting methods. The first category utilizes GAN to improve visual effects. As an unsupervised network, GAN consists of a generator and discriminator[23]. Specifically, the generator generates deceptive samples to defraud the discriminator, while the discriminator attempts to distinguish them. The two learn through playing against each other. Pathak et. al[24] proposed the Context Encoder (CE), which combined autoencoder and tailored loss with a Fully Convolutional Network (FCN) as a generator to boost the quality of inpainted images. Since then, an increasing number of enhanced strategies, such as contextual attention[25], partial convolution[26], and gated convolution[15], have been proposed. However, due to the training strategy of adversarial learning, it is extremely challenging to train a GAN network to its optimal state, even though GAN can produce photo-realistic results with minimal computational overhead. Additionally, most GAN-based methods transform deterministically[27], making it difficult to produce diversified results. The second category fuses CNN to fill texture details with promising results. Shift-Net[28], based on the U-Net[29] architecture, exhibits a high level of precision in the restoration of missing patches, particularly regarding fine texture and structure. To achieve better performance,

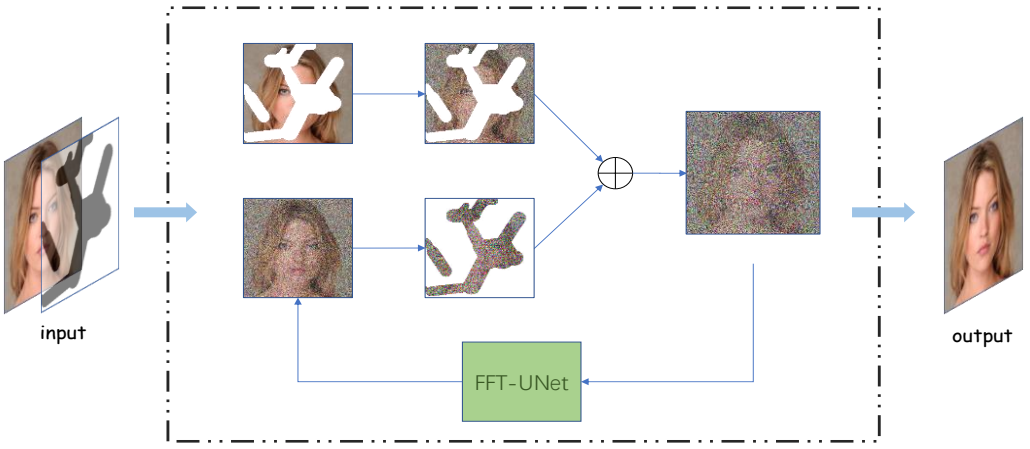

FIGURE 1. Overview of our FFT-DM. FFT-DM modifies the reverse process in the diffusion model in order to restore the missing region of the given image and leverages a U-Net with FFT to predict the posterior distribution parameter for reverse diffusion.

Zeng et al.[30] established a Pyramid-context Encoder Network (PEN-NET) to extract multiscale features. Additionally, Cai et al.[31] reached the goal of semantic object removal with CNN architecture. Inspired by this, we use a CNN to capture the complex and fine-grained texture and structure information of the original images and improve the accuracy and fidelity of the inpainting results.

C. Diffusion Model for Image Inpainting

Diffusion models have their roots in non-equilibrium thermodynamics. Initially, a Markov chain comprising diffusion steps is established, followed by a gradual introduction of random noise to the image. This process continues until the image is fully transformed into a state of random noise. Subsequently, a neural network is employed to learn the reverse diffusion mechanism and then generate images from the noise. In contrast to GAN, diffusion models exhibit superior training efficiency in generating samples through the acquisition of noise. Lugmaryr et al.[27] introduced the Repaint, which leveraged a pre-trained unconditional DDPM model as an image generation prior and modified the iterative process of reverse diffusion by sampling the uncovered area, ultimately leading to high-quality restoration results. Nichol proposed Glide[32], which investigated the distinction between Clip and non-classifier methods for directing conditional diffusion models. Saharia et al. established Palette[33], which employed a multi-task diffusion model to accomplish the image-to-image conversion, i.e., filling, coloring, cropping, and JPEG restoration. Despite producing images of exceptional quality, diffusion models have prolonged inference times due to the iterative generation employed. Motivated by this, we employ a diffusion model to effectively capture the statistical properties of image content and generate a prior distribution that matches the original image characteristics.

III. Method

A. Network architecture

Inspired by Repaint[27], we propose a diffusion model with an FFT (FFT-DM) for image inpainting, as shown in Figure 1. It is notable that Repaint only employs a pre-trained unconditional DDPM model as the image generation model. Thus, we follow the training strategy proposed by the literature [11], then use the trained diffusion model to generate prior from known regions in damaged images. Subsequently, we leverage a U-Net[34] to predict the Gaussian distribution parameter for reverse diffusion. However, many CNN networks suffer from a limited receptive field. Motivated by LaMa[35], we enhance the perception ability of the U-Net by embedding a Fast Fourier Convolution (FFC)[36] to extract frequency domain information, which provides essential global contextual information about the input image. This allows the network to better understand the overall structure of the image and generate higher-quality results.

B. Loss Function

DDPM describes the diffusion process as transforms from an image x_0 into white Gaussian noise $x_T \sim \mathcal{N}(0,1)$ in T time steps during training. In the forward direction, each stride is described as

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I}\right), \tag{1}$$

where x_t denotes the sample image at timestep t while x_{t-1} indicates the previous sample. Moreover, according to a variance schedule, intermediate image x_t is obtained by adding independent and identically distributed (i.i.d.) Gaussian noise with variance β_t, and scaling x_{t-1} with a factor of $\sqrt{1-\beta_t}$.

We define $\alpha_t = 1-\beta_t$ and then calculate the total noise variance $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$, so Eq. (1) can be rewritten as

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathrm{I}\right). \tag{2}$$

Combining Eq. (1) and (2), we can calculate the joint posterior $q(x_{t-1}|x_t, x_0)$ by implementing Bayes theorem as follows.

$$q\left(x_{t-1}|x_t,x_0\right)=\mathcal{N}\left(x_{t-1};\tilde{\mu}\left(x_t,x_0\right),\tilde{\beta}_t\mathrm{I}\right), \tag{3}$$

where mean value $\tilde{\mu}_t\left(x_t,x_0\right)=\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0+\frac{\sqrt{\alpha_t\left(1-\bar{\alpha}_{t-1}\right)}}{1-\bar{\alpha}_t}x_t$ and variance $\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

Based on the construction of the forward noising process, we assume that the distribution $p_\theta\left(x_{t-1}|x_t\right)$ of the reverse process, as defined in Eq. (4), also follows Gaussian distribution that is similar in nature.

$$p_\theta\left(x_{t-1}|x_t\right)=\mathcal{N}\left(x_{t-1};\mu_\theta\left(x_t,t\right),\sum\nolimits_\theta\left(x_t,t\right)\right), \tag{4}$$

where $\mu_\theta\left(x_t,t\right)$ denotes the mean value and $\sum_\theta\left(x_t,t\right)$ represents the diagonal covariance matrix. Both the two parameters are always predicted by a neural network. Besides, this neural network could also predict the noise ε added to the image x_0, which can be predicted via

$$x_0=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon\right) \tag{5}$$

Ho et al.[10] proposed a simplified training objective with Mean-Squared Error (MSE) loss as follows:

$$L_{simple}=E_{t,x_0,\varepsilon}\left[\left\|\varepsilon-\varepsilon_\theta\left(x_t,t\right)\right\|^2\right] \tag{6}$$

Furthermore, this objective can be seen as a reweighted form of variational lower bound (VLB)[34] L_{vlb} as shown in the following equations.

$$L_{vlb}=L_0+L_1+\ldots+L_{T-1}+L_T, \tag{7}$$

$$L_0=-\log p_\theta\left(x_0|x_1\right), \tag{8}$$

$$L_{t-1}=D_{KL}\left(q\left(x_{t-1}|x_t,x_0\right)\middle\|p_\theta\left(x_{t-1}|x_t\right)\right), \tag{9}$$

$$L_T=D_{KL}\left(q\left(x_T|x_0\right)\middle\|p\left(x_T\right)\right), \tag{10}$$

where $D_{KL}(\ \|\ )$ expresses the Kullback–Leibler divergence.

We choose the hybrid objective[34] for training both $\sum_\theta\left(x_t,t\right)$ and $\varepsilon_\theta\left(x_t,t\right)$. Thus, the loss function of FFT-DM can be represented as:

$$L=L_{simple}+\lambda L_{vlb}, \tag{11}$$

where λ is a constant used to prevent L_{vlb} from overwhelming L_{simple}, and $\lambda=0.001$.

Subsequently, we interpolate the output v of this neural network to obtain the variance $\sum_\theta\left(x_t,t\right)$ as shown in Eq. (12).

$$\sum\nolimits_\theta\left(x_t,t\right)=\exp\left(v\log\beta_t+\left(1-v\right)\log\tilde{\beta}_t\right) \tag{12}$$

Additionally, the mean value $\mu_\theta\left(x_t,t\right)$ could be predicted through the following Eq. (13).

$$\mu_\theta\left(x_t,t\right)=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\left(x_t,t\right)\right) \tag{13}$$

C. Diffusion model

As we know, the forward process could be defined by a Markov Chain of adding Gaussian noise. This allows us to sample the known region x_t^{known} of x_t through Eq. (2) and the unknown region $x_t^{unknown}$ of x_t using Eq. (4) at any point in time. Thus, we obtain the ensuing equation for one reverse step,

$$x_t^{known}\sim\mathcal{N}\left(\sqrt{\bar{\alpha}_t}x_0,\left(1-\bar{\alpha}_t\right)\mathrm{I}\right), \tag{14}$$

$$x_t^{unknown}\sim\mathcal{N}\left(\mu_\theta\left(x_t,t\right),\sum\nolimits_\theta\left(x_t,t\right)\right). \tag{15}$$

For the ground truth image x, we assume that $mask\odot x$ indicates the unknown pixels while $\left(1-mask\right)\odot x$ denotes the known pixels. Repaint[27] leverages Gaussian sampling on the known region to generate the image x_t^{known}, which is then combined with the currently generated unknown region $x_t^{unknown}$ to obtain the input x_t of the reverse process at the timestep t. This technique enables us to incorporate the known regions and other generative priors in a manner that enhances the overall quality and consistency of generated images. That is, the generated images that are not only visually appealing but also coherent and consistent with the underlying context and structure of the input image. This process could be illustrated by Eq. (16). Furthermore, FFT-DM also incorporates the resampling strategy of Repaint[27] to harmonize the generated regions with the known region in the aspect of semantics and boundaries.

$$x_t=mask\odot x_t^{known}+\left(1-mask\right)\odot x_t^{unknown} \tag{16}$$

D. FFT-UNet

As mentioned before, a neural network is used to model the Gaussian distribution $p_\theta\left(x_{t-1}|x_t\right)$ to predict its parameters $\sum_\theta\left(x_t,t\right)$ and $\varepsilon_\theta\left(x_t,t\right)$, then performs Gaussian sampling on the intermediate image x_t to obtain the output x_{t-1} of the reverse process at the timestep t. Additionally, in image inpainting tasks, the generation prior to the model is derived from the damaged input images. To accurately extract relevant image features from these inputs, we employ U-Net[34] with strong representation capabilities that can capture local changes and texture variations in images. It is notable that Isola et al.[37] constructed their generator based on a U-Net architecture, which has been shown to effectively capture contextual information and possess a power generation ability. Besides, Suvorov et al.[35] proved that incorporating the FFT mechanism within the CNN framework can increase the global receptive field and improve generalization in the task of image inpainting. Motivated by this, we select an improved U-Net, FFT-UNet, as a predictor in the reverse process. The structure of FFT-UNet is depicted in Figure 2.

Specifically, the input of FFT-UNet is a colored noisy image I_n with 3 channels. The first layer is a convolution layer with a kernel size of 3×3 and 64 output channels, which converts image information into feature space. Then, the second and third layers apply a resblock with time

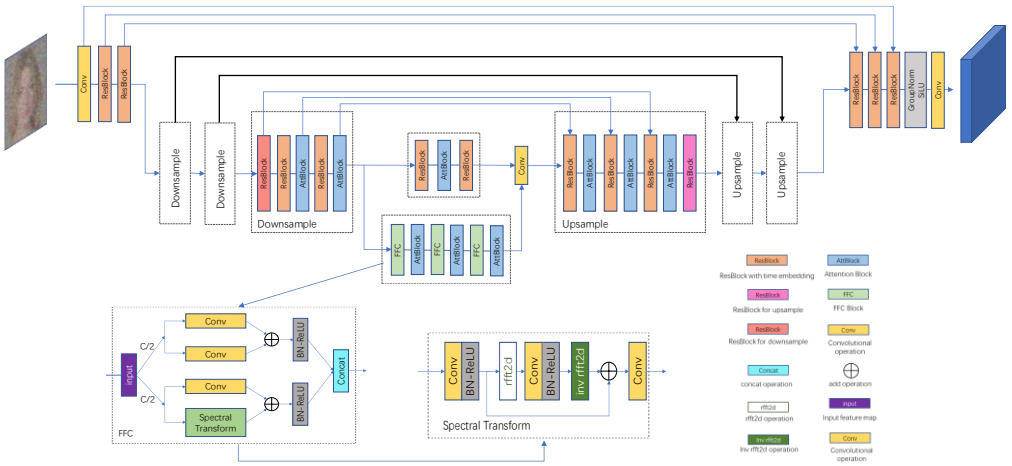

FIGURE 2. **The architecture of the proposed FFT-UNet, which utilizes FFC to capture essential global contextual information of the input image.**

embedding, which can not only enhance its representation ability to capture long-distance dependencies in sequence data, but also improve its overall performance and robustness. The output of the 1st, 2nd, and 3rd layer O_{1st}, O_{2nd}, O_{3rd} can be presented as the following equations, respectively.

$$O_{1st} = C(I_n), \tag{17}$$

$$O_{2nd} = Rs_{time}(O_{1st}), \tag{18}$$

$$O_{3rd} = Rs_{time}(O_{2nd}), \tag{19}$$

where C expresses a convolutional operation and Rs_{time} denotes the resblock with time embedding.

Subsequently, we implement a 3-layer of combination residual layers and average pooling for downsampling, followed by a 3-layer of combination residual layers with nearest neighbor interpolation for upsampling, which connect layers with the same spatial size by skip connections. We describe the 6th layer and 8th layer concretely. That is, the third downsample block and the first upsample block, which are symmetrical in position. The input of the 6th layer can be represented as

$$I_{6th} = 2Downsample(O_{3rd}), \tag{20}$$

where *Downsample* denotes a series of resblock for downsampling, resblock with time embedding, attention block, resblock with time embedding, and attention block.

Then the output of the resblock for downsampling $O_{res-down}$ and the output of the first attention block O_{atten1} are defined as the following equations.

$$O_{res-down} = Rs_{down}(I_{6th}), \tag{21}$$

$$O_{atten1} = Atten\left(Rs_{time}\left(Rs_{down}(I_{6th})\right)\right), \tag{22}$$

where Rs_{down} and *Atten* represents the resblock for downsampling and Legacy QKVAttention[11], respectively. Thus, the output of the 6th layer O_{6th} is expressed as

$$\begin{aligned} O_{6th} &= 3Downsample(O_{3rd}) \\ &= 3\left(Atten\left(Rs_{time}\left(Atten\left(Rs_{time}\left(Rs_{down}(I_{6th})\right)\right)\right)\right)\right). \end{aligned} \tag{23}$$

Next, we should introduce the neck part of FFT-UNet, which consists of two branches and a convolutional operation. The output of the upper branch O_{upper} is expressed as

$$O_{upper} = Rs_{time}\left(Atten\left(Rs_{time}(O_{6th})\right)\right). \tag{24}$$

The output of the lower branch, which comprises a 3-group combination of FFC and attention block, is defined as Eq. (25). This design has been optimized to achieve three key objectives: extracting global contextual information, enhancing the network's perception ability, and improving the efficiency of the training process.

$$O_{lower} = Atten\left(FFC\left(Atten\left(FFC\left(Atten\left(FFC(O_{6th})\right)\right)\right)\right)\right), \tag{25}$$

where *FFC* indicates a fast Fourier convolution operation, whose structure is shown in Figure 2. Followed by a convolutional operation with the size of 1×1, the output of the neck, i.e. the 7th layer, is shown as

$$O_{7th} = Conv\left(\left[O_{upper}, O_{lower}\right]\right), \tag{26}$$

where $[\ ,\ ]$ denotes the concatenate operation.

For FFC[35], it first splits the input with c channels into two parts with $c/2$ channels to not only extract two diverse features, i.e., local features and global features, but also enlarge the receptive field. The first part is followed by two paralleled convolutional operations, add operation, batchnormalization (BN), and Rectified Linear Unit (ReLU), where BN normalizes data to accelerate network speed and ReLU transforms linear features into non-linear features. Meanwhile, the process of the second part differs from the first in that it follows a parallel combination of a convolutional operation and a spectral transform. The outputs of these two processed parts are then concatenated to form a new feature map that contains frequent domain information. Specifically, the spectral transform is structured by three convolutional operations, two combinations of BN and ReLU BR, an FFT operation FFT, an inverse FFT

operation $rFFT$, and an add operation $\oplus$, as defined in Eq. (27).

$$O_{ST} = C\left(\begin{array}{l} BR\left(C\left(\frac{C}{2}O_{7th}\right)\right) \\ \oplus rFFT\left(BR\left(C\left(FFT\left(BR\left(C\left(\frac{C}{2}O_{7th}\right)\right)\right)\right)\right)\right) \end{array}\right), \tag{27}$$

where $\frac{C}{2}O_{7th}$ expresses the half channel of the neck input and O_{ST} denotes the output of spectral transform.

Thus, the output of FFC can be denoted as

$$\begin{aligned} O_{FFC} &= FFC\left(O_{7th}\right) \\ &= \left[\left(C\left(\frac{C}{2}O_{7th}\right) \oplus C\left(\frac{C}{2}O_{7th}\right)\right), \left(C\left(\frac{C}{2}O_{7th}\right) \oplus O_{ST}\right)\right]. \end{aligned} \tag{28}$$

Subsequently, we introduce the structure of the first 7-layer upsample block $Upsample$, which consists of three resblocks with time embedding, three attention blocks, and a resblock for upsampling. The output of this upsample block, i.e., the 8th layer, O_{8th} is defined as the following equation.

$$\begin{aligned} O_{8th} &= Upsample\left(O_{7th}\right) \\ &= Rs_{up}(Atten([Rs_{time}(Atten[\\ &\quad Rs_{time}(Atten[Rs_{time}(O_{7th}), O_{6th}]), O_{atten1}]), O_{res-down}])), \end{aligned} \tag{29}$$

where Rs_{up} expresses the resblock for upsampling.

Assume that the output of the third similar structured upsample block, that is, the 10th layer, is O_{10th} . Then we exploit a stack of three resblocks with time embedding, which is connected with the 1st, 2nd, and 3rd by skip connection, to refine the extracted features. Subsequently, a combination of groupnormalization (GN) and SiLU is applied. Concretely, GN reduces the dependence of model performance on batch size and improves the training stability while SiLU enhances the nonlinear fitting ability and training speed. Then, a single convolutional layer transforms the obtained feature map into predicted parameters and then constructs x_{t-1} . Thus, the output of FFT-Unet $O_{FFT-UNet}$ with 6 channels is formulated as follows:

$$\begin{aligned} &O_{FFT-UNet} = \\ &C\left(GSi\left(\left[Rs_{time}\left(\left[Rs_{time}\left(O_{10th}\right), O_{3rd}\right]\right), O_{2nd}\right]\right), O_{1st}\right]\right)\right), \end{aligned} \tag{30}$$

where GSi denotes the combination of GN and SiLU.

IV. Experimental Analysis and Results

A. Datasets

The CelebA-HQ dataset[38] is a large-scale collection of high-resolution face images, consisting of 30,000 samples. The dataset encompasses a wide range of pose variations, including tilts and rotations, as well as diverse background clutter, such as indoor and outdoor scenes. As a result, the dataset has been widely used for various computer vision tasks, including face attribute recognition[39], face detection[40], facial part localization[41], and face editing & synthesis[42]. In this work, we validate the performance of the proposed FFT-DM on the publicly available CelebA-HQ dataset at different resolutions, and we choose the image size of 64×64 and 256×256 for our experiments.

B. Implementation details

All experiments were conducted using Python 3.8.5 and PyTorch 1.13 on Ubuntu 20.04. The experimental setup consisted of a computer with an Intel Xeon Gold 6330 CPU @ 2.00 GHz, 128GB of RAM, and an NVIDIA GeForce RTX 3090 GPU. The running speed of the GPU was accelerated utilizing NVIDIA CUDA 11.7 and cuDNN 8.5.0. Also, FFT-DM employed a timestep of $T=250$ and resampled the data $r=10$ times with a jumpy size of $j=10$, which is in line with the approach taken in Repaint[27]. Additionally, we selected two commonly used metrics, Frechet Inception Distance Score (FID)[43] and Learned Perceptual Image Patch Similarity (LPIPS)[44], to evaluate the quality of the inpainted images in terms of semantic fidelity, texture consistency, and structural coherence. Lower FID and LPIPS scores indicate a higher degree of similarity between the restored image and the original image.

C. Network analysis

In this section, we verify the rationality and validity of the proposed FFT-DM, which is composed of a diffusion model and FFT-UNet.

Diffusion model: As we know, CNN is constrained by its localized inductive bias, which makes it difficult to capture long-distance information and comprehend global image semantics. In contrast, the diffusion model is a probabilistic model that can learn the data distribution of real image data, thereby preserving global details and structural information. Motivated by this, we combine U-Net and the diffusion model to generate realistic images. The diffusion model is utilized to model image data distribution and facilitate the generation of realistic images, while U-Net is leveraged to extract image features and predict the Gaussian distribution parameter for reverse diffusion. Table 1 demonstrates the effectiveness of the aforementioned combination, showing that 'U-Net with diffusion model' outperforms 'U-Net' not only at the mask ratio of 20%-40% but also at the mask ratio of 40%-60% in terms of generating more intricate pixel details in the inpainted images.

FFT: It is known to us the fast Fourier transform converts spatial domain information in images into the frequency domain space, thus enhancing the feature representation ability of CNN. Additionally, the FFT mechanism speeds up the computational speed of convolution operations, improving the efficiency of CNN. Inspired by this, we introduce the integration of FFT into the U-Net architecture to enhance the performance and efficiency of image inpainting tasks. The effectiveness of FFT is demonstrated in Table 2, where 'FFT-DM' exhibits a lower FID score than 'U-Net with diffusion model', not only for an image size of 64×64 or 256×256 , but

also for train steps of 5×10^4 or 1×10^5. Notably, the diffusion model that directly handles high-resolution images is characterized by a significant level of computational complexity, prolonged inference duration, and substantial time and computing resources required. To address this issue, we first concentrate on the dataset with a resolution of 64×64, and subsequently expanded the investigation to include the dataset with a resolution of 256×256 after examining an appropriate framework.

TABLE 1. RESULTS OF DIFFERENT METHODS FOR RANDOM MASK RATIOS WITH 1,000 IMAGE SAMPLES.

Mask Ratio	20%-40%		40%-60%	
Model	FID↓	LPIPS↓	FID↓	LPIPS↓
U-Net	17.98	0.0936	28.58	0.150
U-Net with diffusion model	14.47	0.0502	15.68	0.083

TABLE 2 IMAGE SYNTHESIS RESULTS FOR DIFFERENT METHODS WITH DIFFERENT IMG_SIZES AND TRAIN_STEPS, USING 1,000 IMAGE SAMPLES.

Model	Img_size	Channels	Train_steps $(\times10^4)$	FID↓
U-Net with diffusion model	64	64	5	84.50
FFT-DM (Ours)				62.00
U-Net with diffusion model	64	64	10	80.67
FFT-DM (Ours)				59.33
U-Net with diffusion model	256	64	5	89.40
FFT-DM (Ours)				47.86

TABLE 3. MULTI-SCALE COMPARISON RESULTS OF SEVERAL NETWORKS WITH 3,000 IMAGE SAMPLES.

Method	#Params $(\times10^6)$	Narrow masks		Wide masks	
		FID↓	LPIPS↓	FID↓	LPIPS↓
EdgeConnect[14]	22	9.61	0.099	9.02	0.120
DeepFill v2[15]	4	12.5	0.130	11.2	0.126
RegionWise[16]	47	11.1	0.124	8.54	0.121
AOT-GAN[17]	15	6.67	0.081	10.3	0.118
CoModGAN[18]	109	16.8	0.079	24.4	0.102
LaMa[35]	27	7.26	0.085	6.96	0.098
FFT-DM (Ours)	37.5	4.06	0.036	7.25	0.121

D. Comparisons with the state‐of‐the‐art inpainting methods

To evaluate the performance of FFT-DM, we conducted a comprehensive set of quantitative and qualitative tests on CelebA-HQ[38] with 3,000 images. Several state-of-the-art inpainting methods, including EdgeConnect[14], DeepFill v2, RegionWise[16], AOT-GAN[17], CoModGAN[18], and LaMa[35], were compared to FFT-DM. Quantitative results are demonstrated in Table 3. For narrow masks, FFT-DM outperforms all other methods in FID and LPIPS. For wide masks, although LaMa[35], the best GAN method, exhibits better global consistency, FFT-DM still achieved competitive performance in FID. Furthermore, we used parameter counts to verify their efficiency. Generally, a larger number of parameters indicates that the model is more expressive and fits the training data more accurately. Nevertheless, larger models typically require more computing resources and a longer training period. The parameter counts of FFT-DM rank in the middle of Table 3, elaborating a trade-off between model complexity and computational capacity.

To vividly illustrate the effectiveness of FFT-DM, we provide visual results in Figures 3 and 4. For each image in Figure 3, the first column shows the original image, the second column displays the image with different masks applied, and the third column presents the restored image. The results demonstrate that FFT-DM can handle masks with arbitrary shapes and sizes, and can perform well even for masks with large dimensions. Moreover, Figure 4 shows that FFT-DM can generate diverse filling area content consistent with the texture and semantics of the surrounding region.

V. Conclusion

In this paper, we propose a novel approach called FFT-DM for image inpainting, which generates content that is consistent with the surrounding area not only in texture but also in semantics. FFT-DM uses a diffusion model to increase the degree of freedom of masks and generate image prior that matches the semantic and texture characteristics of the original image. Then, FFT-DM leverages CNN architecture to capture more texture features and detailed information in the reverse process of the DDPM. Besides, we fused the FFT mechanism into the diffusion model to mine frequent features and boost perception ability. Extensive experiments demonstrate that FFT-DM can not only generate inpainted images with high visual quality but also balance a trade-off between effectiveness and efficiency. Next, we attempt to extend FFT-

DM to handle multiple low-level vision tasks, such as image denoising and deblurring, in the future.

FIGURE 3. Visual results for different mask ratios on CelebA-HQ[38].

FIGURE 4. Visual results for diverse generated images.

References

[1]. O. Elharrouss, N. Almaadeed, S. Al-Maadeed and Y. Akbari, "Image inpainting: A review," *NEURAL PROCESS LETT*, vol. 51, 2020, pp. 2007-2028.

[2]. M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," Proc. SIGGRAPH, New Orleans, Louisiana, USA, 2000, pp. 417-424.

[3]. C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," IEEE TIP, vol. 10, no. 8, 2001, pp. 1200-1211.

[4]. A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," Proc. ICCV, London, UK, 1999, pp. 1033-1038.

[5]. A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," Proc. SIGGRAPH, Los Angeles, CA, USA, 2001, pp. 341-346.

[6]. W. H. Cheng, C. W. Hsieh, S. K. Lin, C. W. Wang, and J. L. Wu, "Robust algorithm for exemplar-based image inpainting," Proc. CGIV, Beijing, China, 2005, pp. 64-69.

[7]. A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," Proc. CVPR, Madison, WI, USA, 2003, pp. II-II.

[8]. V. K. Alilou and F. Yaghmaee, "Application of GRNN neural network in non-texture image inpainting and restoration," Pattern Recognit. Lett., vol. 62, 2015, pp. 24-31.

[9]. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," Proc. CVPR, New Orleans, Louisiana, USA, 2022, pp. 16000-16009.

[10]. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," NIPS, vol. 33, 2020, pp. 6840-6851.

[11]. P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," NIPS, vol. 34, 2021, pp. 8780-8794.

[12]. Y. Chen, W. Yu, and T. Pock, "On learning optimized reaction diffusion processes for effective image restoration," Proc. CVPR, Boston, Massachusetts, USA, 2015, pp. 5261-5269.

[13]. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proc. CVPR*, New Orleans, Louisiana, USA, 2022, pp. 10684-10695.

[14]. K. Nazeri, E. Ng, T. Joseph, F.Z. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.

[15]. J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T.S. Huang, "Free-form image inpainting with gated convolution," *Proc. ICCV*, Seoul, Korea (South), 2019, pp. 4471-4480.

[16]. Y. Ma, X. Liu, S. Bai, L. Wang, A. Liu, D. Tao, and E.R. Hancock, "Regionwise generative adversarial image inpainting for large missing areas," *IEEE T. CYBERNETICS*, 2022.

[17]. Y. Zeng, J. Fu, H. Chao, and B. Guo, "Aggregated contextual transformations for high-resolution image inpainting," *T-VCG*, 2022.

[18]. S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I. Chang, and Y. Xu, "Large scale image completion via co-modulated generative adversarial networks," *arXiv preprint arXiv:2103.10428*, 2021.

[19]. M. J. Fadili and J. L. Starck, "Em algorithm for sparse representation-based image inpainting," Proc. ICIP, Genoa, Italy, 2005, pp. II-61.

[20]. Z. Xu and J. Sun, "Image inpainting by patch propagation using patch sparsity," *IEEE TIP*, vol. 19, no. 5, 2010, pp. 1153-1165.

[21]. O. Whyte, J. Sivic, and A. Zisserman, "Get Out of my Picture! Internet-based Inpainting," *Proc. BMVC*, London, UK, 2009, pp. 5.
[22]. J. Hays and A. A. Efros, "Scene completion using millions of photographs," ACM ToG, vol. 26, no. 3, 2007, pp. 4-es.
[23]. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, 2014, pp. 139-144.
[24]. D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *Proc. CVPR*, Las Vegas, USA, 2016, pp. 2536-2544.
[25]. J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T.S. Huang, "Generative image inpainting with contextual attention," *Proc. CVPR*, Salt Lake, Utah, USA, 2018, pp. 5505-5514.
[26]. G. Liu, F.A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," *Proc. ECCV*, Munich, Germany, 2018, pp. 85-100.
[27]. A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," *Proc.CVPR*, New Orleans, Louisiana, USA, 2022, pp. 11461-11471.
[28]. Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," *Proc. ECCV*, Munich, Germany, 2018, pp. 1-17.
[29]. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Proc. MICCAI*, Munich, Germany, 2015, pp. 234-241.
[30]. Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," *Proc. CVPR*, Long Beach, CA, USA, 2019, pp. 1486-1494.
[31]. X. Cai and B. Song, "Semantic object removal with convolutional neural network feature-based inpainting approach," MULTIMEDIA SYST., vol. 24, 2018, pp. 597-609.
[32]. A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," arXiv preprint arXiv:2112.10741, 2021.
[33]. C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," Proc. ACM SIGGRAPH, Vancouver, Canada, 2022, pp. 1-10.
[34]. A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," Proc. ICML, Lugano, Switzerland, 2021, pp. 8162-8171.
[35]. R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," Proc. WACV, Waikoloa, HI, USA, 2022, pp. 2149-2159.
[36]. L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," NIPS, vol. 33, 2020, pp. 4479-4488.
[37]. P. Isola, J. Y. Zhu, T. Zhou, and A.A. Efros, "Image-to-image translation with conditional adversarial networks," Proc. CVPR, Honolulu, HI, USA, 2017, pp. 1125-1134.
[38]. Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," Proc. ICCV, Santiago, Chile, 2015, pp. 3730-3738.
[39]. H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C.C. Loy, "CelebV-HQ: A large-scale video facial attributes dataset," *Proc. ECCV, Tel Aviv, Israel*, 2022, pp. 650-667.
[40]. Z. Liu, X. Qi and P.H. Torr, "Global texture enhancement for fake face detection in the wild," *Proc. CVPR*, 2020, pp. 8060-8069.
[41]. C. Nederhood, N. Kolkin, D. Fu, and J. Salavon, "Harnessing the conditioning sensorium for improved image translation," *Proc. ICCV*, 2021, pp. 6752-6761.
[42]. J. Sun, Q. Deng, Q. Li, M. Sun, M. Ren, and Z. Sun, "Anyface: Free-style text-to-face synthesis and manipulation," *Proc. CVPR*, 2022, pp. 18687-18696.
[43]. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," NIPS, vol. 30, 2017.
[44]. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," Proc. CVPR, Salt Lake, UT, USA, 2018, pp. 586-595.

AUTHOR BIO/IMAGE

Yuxuan Hu, received her B.S degree in Electrical Engineering and Automation from China University of Mining and Technology and M.S. degree in Measuring and Testing Technologies and Instruments from the China University of Mining and Technology (Beijing). She is currently a Ph.D. student in Computer Science and Technology at Central South University, Changsha, China. Her main research direction includes image denoising, image restoration, and deep learning.

Hanting Wang, is currently an undergraduate in Software Engineering at Northwestern Polytechnical University, Xi'an, China. His main research direction includes image inpainting and deep learning.

Cong Jin, received the B.S. degree and M.S. degree in Communication and Information System in 2010 and 2013 from the Communication University of China, respectively, and a Ph.D. degree in Communication and Information System from the Communication University of China, Beijing, P.R. China in 2019. She is currently an Associate Professor with the School of Information and Communication Engineering, Communication University of China. Her research interests focus on reinforcement learning, music AI, and audio digital twins. She is presiding over and undertaking the youth, general and key projects of the National Natural Science Foundation, Xiaomi joint fund of Beijing Natural Science Foundation, and National Key Research and Development Projects. She has published a total of more than 30 academic papers in IEEE, Springer, and other international journals and conferences, including more than 10 SCI-retrieved journals. She has served as AE or reviewer for several leading journals and the session chair or PC Member for several major international conferences.

Bo Li, received his B.S. degree in electronic information technology, M.S. and Ph.D. degrees in systems engineering from Northwestern Polytechnical University, Xi'an, China. He is currently an associate professor with the School of Electronics and Information, Northwestern Polytechnical University. His current research interests include intelligent command and control, deep reinforcement learning, and uncertain information processing.

Chunwei Tian, received his Ph.D. degree in Computer Application Technique from the Harbin Institute of Technology, China in Jan. 2021. He is currently an Associate Professor with the School of Software, Northwestern Polytechnical University, Xi'an, China. Also, he is a member of the National Engineering Laboratory for Integrated Aerospace Ground Ocean Big Data Application Technology. His research interests include image restoration and deep learning. He has published over 50 papers in academic journals and conferences, including IEEE TNNLS, IEEE TMM, IEEE TSMC, NN, Information Sciences, Pattern Recognition, KBS, PRL, ICASSP, ICPR, ACPR, and IJCB. He has four ESI highly-cited papers, three homepage papers of the Neural Networks, one homepage paper for the TMM, and one excellent paper in 2018 and 2019 for the CAAI Transactions on Intelligence Technology. Also, his three codes are rated as the contribution codes of GitHub 2020. His two paper techniques are integrated into the iHub and Profillic. He has obtained 2021 Shenzhen CCF Excellent Doctoral Dissertation and 2022 Harbin Institute of Technology Excellent Doctoral Dissertation. Besides, he is an associate editor of the Journal of Electrical and Electronic Engineering, International Journal of Image and Graphics, Journal of Artificial Intelligence and Technology, youth editor of CAAI Transactions on Intelligence Technology, Defence Technology, Data Science and Management and Ordnance Industry Automation, a guest editor of Mathematics, Electronics, Mathematical Biosciences, and Engineering, International Journal of Distributed Sensor Networks, Frontiers in Robotics and AI, Drones and Applied Sciences, Topical Advisory Panel Member of Electronics and Drones, PC Chair and Workshop Chair of MLCCIM 2022, SI Chair of ACAIT 2022, Special Session Co-Chair and Workshop Chair of ICCSI 2022, Publicity Chair of AMDS 2022, a workshop chair of ICCBDAI 2021, a reviewer of some journals and conferences, such as the IEEE TIP, the IEEE TII, the IEEE TNNLS, IEEE TCYB, the IEEE TSMC, the IEEE TMM, the NN, the CVIU, the information fusion, information sciences, etc.