AIIVUS: Guidewire Artifacts Inpainting for Intravascular Ultrasound Imaging With United Spatiotemporal Aggregation Learning

Bangjun Li, Pengfei Zhang, Yankun Cao, Longkun Sun, Jianqin Feng, Yuezhong Zhang, Qi Yang, Yujun Li, and Zhi Liu

Abstract—The Intravascular Ultrasound (IVUS) technology is an important imaging modality used in realistic clinical practice, it is often combined with coronary angiography (CAG) to diagnose coronary artery disease. As the golden standard for in vivo imaging of coronary artery walls, it can provide high-resolution images of the artery wall. Generally, the IVUS acquisition device uses an ultrasonic transducer to acquire the fine-grained anatomical information of the cardiovascular tissue by means of pulse echo imaging. However, widely used mechanical rotating imaging system suffered from guidewire artifacts. The inadequate visualization caused by artifacts often caused huge trouble for clinical diagnosis and subsequent tissue structure evaluation, and there is no suitable way to solve this long-standing problem so far. In this paper, we conducted an exploratory study and proposed the first deep learning based network named AIVUS for repairing the corrupted IVUS images. The network has a novel generative adversarial architecture, the united of gated convolution and spatiotemporal aggregation structure has been introduced to enhance its restoration capability. The proposed network can handle large-scale, moving guidewire artifacts, and it can fully utilize spatial and temporal information hidden in sequence to recover the high-fidelity original content and maintain consistency between frames. Furthermore, we compared it with several latest restoration models, including both image restoration and video restoration models. Qualitative and quantitative comparison results on the collected IVUS datasets demonstrate that our method has achieved outstanding performance and its potential clinical value.

Index Terms—Deep learning, guidewire artifacts, intravascular ultrasound imaging, medical image inpainting.

I. INTRODUCTION

INTRAVASCULAR ultrasound (IVUS) is an important imaging technique used in clinical cardiovascular disease diagnosis and therapy. It is the modality that has been become the golden standard for in vivo imaging of the vessel wall of the coronary arteries [1]. As an important method used in surgical procedures, unlike other techniques such as MRI, CT, and coronary angiography (CAG), it is flexible and does not use ionizing radiation. Furthermore, it can provide high-resolution images of the entire artery wall and reveal more information about plaque buildup for percutaneous cardiovascular interventions (PCI) [2], [3].

However, different from the conventional ultrasound imaging test in vitro, intravascular ultrasound imaging is a kind of interventional therapy technique, it combines non-invasive ultrasound technology with invasive catheterization techniques. The flexible catheter first needs to reach the target position along the artery with the help of a guidewire, and then the transducer inside it captures the cross-sectional image of the blood vessel by means of pulse-echo imaging. Finally, the imaging system can present fine-grained anatomical information about the structure and geometry of the target cardiovascular tissue. At the moment, there are two major kinds of intravascular ultrasound transducers used to acquire IVUS images, one kind is a mechanical rotating transducer and the other is a phased array transducer [4]. The mechanical rotating device has a single rotating ultrasound crystal that rotates evenly to obtain cross-sectional images, while the phased array device has multiple piezoelectric transducers, each transducer is activated sequentially to produce a rotating ultrasound beam [5]. In these two types of equipment, the guidewire is designed on the side and center of the catheter, respectively, this can be illustrated in Fig. 1. Artifacts in IVUS imaging are an infrequently inevitable phenomenon due to the inherent limitations of ultrasound modality itself [6]. Different from the phased array device, images acquired by mechanical rotating
imaging devices are suffered from the guidewire artifacts, which is unique to IVUS imaging and makes it impossible to display the panoptic imaging signal. Specifically, the guidewire will intersect the ultrasound beam at a certain angle, resulting in the lack of captured anatomical information. In addition, the position of the artifact can be randomly shifted due to the pullback movement of the guidewire and the physiological activity of the artery. As can be seen from Fig. 2, insufficient visualization will interfere with clinical decision-making and poor quality due to artifacts in the images makes IVUS data are hard to analyze quantitatively [7].

In addition to requiring high cost and professional operation, in order to acquire the detailed information on interest of the cardiovascular anatomy image, it is often necessary to increase the number of catheter retractions to acquire clinical images multiple times. Effective and efficient acquisition of complete coronary inner wall information helps to save scarce medical resources, reduce patient pain, and greatly saves intraoperative time and reduce potential risks. Recently, artificial intelligence is gaining momentum in medical imaging, and repairing the missing regions with plausible and coherent textural has been of special interest for automatic image analysis. In this work, we attempt to apply deep learning-based inpainting technology to the task of medical ultrasound imaging recovery. Revealing the morphological structures and anatomical characteristics occluded by guidewire artifacts has magnificent significance for clinical diagnosis and post-processing tasks. Such as the delineation of lumen and media-adventitia borders in IVUS images [8], which takes an important place in quantitative assessment of coronary atherosclerosis and vulnerable plaques.

Automatic image completion is a practical and crucial issue, it has been widely used in color image processing [9]–[11]. Traditional image inpainting techniques such as patch-based methods [12]–[14] often search high similarity patches from known regions to pad the missing content. Despite some pleasing results, however, these methods are suffered from limited effectiveness and are susceptible to complex motions. Fortunately, deep learning-based image inpainting methods have shown us unprecedented benefits in processing images, and various learning based restoration methods have been proposed and dominated this domain. These methods could fill in corrupted areas based on the learned higher-level semantic features [15]. Moreover, it has presented an excellent performance on conventional image editing tasks, such as scratch restoration and undesired object removal. On the other hand, there are also some problems challenging it. First, for the additional temporal information brought by the video, it is hard to simply extend these methods to the video domain, it must consider the consistency between inter-frames. Second, the large scale irregular damaged region hinders the mining of effective information, especially for the complex texture and structural patterns. Currently, most of these works are committed to natural color image editing tasks, however, there are few reports in the literature on medical image restoration [16]–[18]. Armanious et.al [16] proposed ip-MedGAN for inpainting CT or MRI images with local distortions, however, not only it is restricted to square-shaped inpainting regions, it also demands exact localization of region of interest. Tukra et.al [18] proposed an unsupervised end-to-end deep learning framework STV-Net to infer the appearance of the surgical scene under occlusions. However, these works are mainly works for CT, MRI, and we have not found the corresponding works for ultrasound image domain.

Furthermore, medical imaging imposes higher and stricter standards for authenticity, high fidelity, and consistency, especially for subsequent diagnostic tasks, which makes the task of medical image restoration more meaningful and challenging. In this task, accurate image restoration not only contributes to the selection of appropriate stent size in cardiac stenting surgery, but also facilitates quantitative analysis of vulnerable plaques. Similarly, potential risks for medical imaging inpainting are still an unavoidable problem, since it is hard to completely guarantee that the restoration results are true to the original anatomical structure, any small important detail could have a significant impact on the final diagnosis. In an ideal case, the inpainting method can restore the fine-grained anatomical information on the structure and geometry of the target tissue. While in the worst case, it may generate wrong tissue structure, this has an important effect on therapy selection, e.g. the amount of medication, surgery time, stent size.

In this paper, we investigate the novel application of deep learning for intravascular ultrasound images inpainting. We have collected three kinds of intravascular ultrasound images dataset by professional physicians. Moreover, we present a novel end-to-end learning based network AIVUS to reveal the original information occluded by guidewire artifacts. Different from other methods, the proposed method can handle the large-scale artifacts, make full use of the correlation information between adjacent frames and improve the consistency in the time dimension. And a united of gated convolution and spatiotemporal aggregation structure has been proposed, the generator firstly adopted the gated convolution blocks to cope with the artifacts and obtained the coarse recovery content, the subsequent aggregation module merges the feature maps from the
spatial and temporal dimensions and further generates the refinement restoration details. In addition to the non-adversarial losses, a temporal patch adversarial loss has been incorporated into objective loss to improve the high-frequency detail features. We have conducted extensive experiments on the collected datasets, the experimental results have shown that the proposed method could achieve high quality inpainting results.

In summary, our main novelties and contributions of this paper are as follows:

- We adopt deep learning-based inpainting techniques to conduct pilot research to recover the original information disrupted by guidewire artifacts in intravascular ultrasound imaging. It can provide a reference for clinical decision making, especially for subsequent image processing tasks such as complex tissue structure analysis.
- We proposed a novel end-to-end generative network named AIVUS, which has powerful spatiotemporal modeling to improve the quality of repaired images. Up to our knowledge, it is the first work to provide a deep learning based network for repairing IVUS images. Furthermore, we designed three types of intravascular ultrasound data sets for this task.
- We quantitatively and qualitatively evaluate our proposed method on collected datasets and achieved superior results compared with several state-of-the-art natural image inpainting methods.

The remainder of the paper is organized as follows. In Section II, we first give a brief summary of previous related works, then we illustrate the proposed network together with its training losses in Section III. In Section IV, we present the experimental setting up and results. In Section V, we show the difficulties and limitations faced by current work. Finally, we summarize the whole work with some future directions in Section VI.

II. RELATED WORKS

A. Image Restoration

In the work related to natural image inpainting, numerous strategies and approaches have been proposed to processing the images [19]–[21], they can be used to process different image tasks, such as undesired object removal, scratch or damage restoration [22]. Traditional image completion techniques mainly include diffusion-based and patch-based approaches [23]–[26]. Diffusion-based methods can propagate local image information to the neighboring region from boundary to holes, while they can only handle small missing regions. Patch-based methods often iteratively find the most relevant patches from the remaining known image to repair the missing content. Like PatchMatch [27] utilizes the approximate nearest neighbor algorithm to search approximate matches and has shown magnitude practical values for image inpainting work. However, the large number of iterative procedures makes the patch-based methods computationally time-consuming, and it is difficult to recover images that have complicated scenes or missing sufficient available information. In this work, the task is to restore the original content to repair the guidewire artifact area, rather than simply replacing this missing information with similar content from the remaining image content. Especially in the domain of medical images where small structures could significantly alter the diagnostic decisions.

Recently, deep learning based methods have shown a significant advantage in the image inpainting domain, it could automatically learn semantic priors and meaningful hidden representations in an end-to-end fashion [28]. Moreover, GAN-based approaches also have emerged as a promising paradigm for image processing works, it using two networks to iteratively improve the model by a two-player minimax game. Context Encoders [29] firstly adopted deep neural networks to deal with the large image missing content by using an encoder-decoder architecture. It is constrained the network training by the $\ell_2$ reconstruction loss and the generative adversarial loss to fill the missing rectangle region centered on images. Izuka et al. [30] utilizes two adversarial losses, global and local discriminator adversarial loss to improve the inpainting quality. Yu et al. [22] proposed a coarse-to-fine generative image inpainting framework, which introduces a contextual attention layer to explicitly attend on related feature patches at distance spatial locations to synthesize visually realistic content. Furthermore, to deal with the irregular masks, several works have emerged to handle this problem [20], [28], [31]. Because the corrupted images are composed of valid pixel regions and corrupted pixel regions, conventional convolution is not effective in discriminating them due to its weight sharing property. To circumvent this problem, Liu et al. [28] proposed partial convolution to process respectively the valid pixels and invalid pixels, its masks followed a rule-based update step to recompute new ones layer by layer. Furthermore, Yu et al. [31] proposed gated convolution by generalizing partial convolution for image inpainting, it uses a gating mechanism to dynamically select the feature for each channel and each spatial location. Zhu et al. [32] proposed a novel mask-aware solution that inpainting arbitrary missing regions in a cascaded refinement with mask awareness fashion. Some other works adopt predicted structures to guide the network generates new content [21], [33]. Nazeri et al. [33] proposed a two-stage adversarial model EdgeConnect, where the edge generator firstly hallucinates edges of the missing region, and the image completion network generates the final output image using hallucinated edges as the priori condition.

B. Video Restoration

Compared to image restoration, video restoration tasks put forward higher requirements due to the introduction of the time dimension. Video restoration could be viewed as an extension of image restoration with temporal constraints. Meanwhile, like their counterpart methods used in image restoration, the patch-based methods [12], [14] also take an important role because there is more redundant information that can be mined in videos. These methods are prone to yield globally inconsistent results. Although to ensure maintain consistency on the time dimension, patch-based algorithms have been cast as a global optimization problem [14], [34]. However, these methods are still limited to the magnitude of computational time and need repetitive patterns in videos. They are not excelling at processing large
or long-lasting occlusions. To overcome the challenge faced by patch-based methods, deep learning based methods also have shown promising results in video inpainting. Wang et al. [35] firstly proposed a coarse-to-fine CombCN model to repair the corrupted region in videos. It comprises two sub-network modules, a 3D temporal structure prediction module, and 2D spatial detail recovering module, and we set it as a baseline in our work. Chang et al. [36] proposed a free-form video inpainting network with 3D gated convolution and temporal PatchGAN loss to handle the uncertainty of masks and improve the temporal consistency. Later, Chang et al. [15] integrate the Temporal Shift Module [37] into 2D convolution so that it could use 2D convolution to mimic 3D convolution to deal with the temporal information, but the higher efficiency is often achieved at the expense of accuracy. Flow-based methods [14], [34], [38], [39] also show a promising result in video processing tasks. However, these methods often depend upon the accurate optical flow computation, which is also a difficult problem to handle. In the work of flow-edge based guided video completion, Gao et al. first adopt FlowNet2.0 [40] to compute flow between adjacent frames, and then use Canny edge detector and EdgeConnect [33] to extract and complete the flow edge, after that the complete flow edge is used to guide the piecewise smooth flow completion. Kim et al. [9], [34] adopt recurrent feedback and a temporal memory module to enforce the temporally consistent. But the accurate flow estimation forces the flow-based methods cannot fit the work that images have large missing regions.

C. Temporal Feature Learning

To capture the temporal features hidden in the video sequence, the basic straightforward method is to replace the two dimensions convolution with the higher dimension convolution, the additional dimension enables it to exploit the time information. However, 3DCNN suffers from its limited capability of recovering motions in the video because the kernel size in 3DCNN limits the motion range it models [41]. In recent years, recurrent neural network (RNN) and its variants have shown their powerful capability in modeling spatiotemporal sequences. Shi et al. [42] firstly proposed the Convolutional Long Short Term Memory (ConvLSTM) for precipitation nowcasting. It replaces the fully connected operation in vanilla LSTM with the convolutional operation so as to the spatial correlation and temporal correlation can be built simultaneously. Specifically, ConvLSTM has been widely used in modeling long term temporal consistency for video related tasks, e.g. video segmentation, action recognition, and detection, video captioning.

III. METHODS

An overview of our proposed network AIVUS for inpainting intravascular ultrasound images is schematically illustrated in Fig. 3. The whole network mainly depends on the existing two works [31], [43]. It has a novel coarse-to-fine architecture generator $G$ and a Temporal PatchGAN [36] discriminator $D$. The generator $G$ is mainly used to restore the missing content caused by artifacts, while the discriminator $D$ is to discriminate between ground truth and the restoration images from the generator. The network is firstly trained and evaluated on the animal IVUS dataset, and to better verify its performance on realistic clinical application, we also conducted experiments on the clinical IVUS data set. The results show that the proposed model can exploit the temporal and spatial information effectively to recover the corrupted IVUS images.

A. Overview of the Framework

The ultimate goal is to remove the interference of guidewire artifacts and reveal the original morphological details lost in IVUS images. To simplify the task and restore high-fidelity original details, we cast our work as a supervised learning-based video inpainting problem. It takes the incomplete masked IVUS image sequence $V_{in} = \{V_{in}^0, V_{in}^1, \ldots, V_{in}^{T-1}\}$ and the corresponding mask sequence $M = \{M^0, M^1, \ldots, M^{T-1}\}$ as input. Finally, it generate the complete output sequence $V_{out} = \{V_{out}^0, V_{out}^1, \ldots, V_{out}^{T-1}\}$, where $T$ represent the number of consecutive video frames, and the outputs of network are expected to be consistent with the ground truth frames $V_{gt} = \{V_{gt}^0, V_{gt}^1, \ldots, V_{gt}^{T-1}\}$ both in spatial and temporal dimension. The repaired sequence not only provides an unobstructed view of the inner wall of artery, but also provides potential value for subsequent medical image processing tasks.

B. Generator Network

The generator network has a coarse-to-fine architecture to handle the large-scale guidewire artifacts. It repairs IVUS images in two stages, the stage1 is an image completion module $\mathcal{H}_c$, while the stage2 is a spatiotemporal aggregation network with two branches: temporal feature extracting module $\mathcal{H}_t$ and aggregation module $\mathcal{H}_a$. Specifically, image completion module $\mathcal{H}_c$ restore the input sequence frame by frame, and it has an encoder-decoder structure that comprised of a stack of gated convolution blocks to obtain the preliminary repair results. In addition, the gated convolution blocks with dilated factors 2, 4, and 8 are also embedded in the middle encode procedure. Different from the conventional convolution operation that views all the pixels as valid values, the gated convolution can efficiently distinguish valid pixels and invalid pixels in the spatial region. After obtaining the coarse restoration outputs, we further adopt the refinement processing to enhance the details and improve the consistency in temporal and spatial dimensions. To make the best use of the spatiotemporal information between successive frames, the stage2 includes two branches: $\mathcal{H}_t$ and $\mathcal{H}_a$. (a) The first branch $\mathcal{H}_t$ is a bidirectional ConvLSTM network. It comprises a series of bidirectional ConvLSTM layers, and each bidirectional ConvLSTM layer consists of a forward ConvLSTM branch and a backward ConvLSTM branch, these two branches enjoy the same structure. Furthermore, the two learned feature streams are merged together through a convolution operation and passed to the next layer. This structure ensures it can extract forward and backward temporal features simultaneously from the segments. The module $\mathcal{H}_a$ has a total of four layers and the number of hidden state channels were set to 32, 64, 128, 128. (b) While $\mathcal{H}_c$ composed of a stack of 2D conventional convolutional layers with $3 \times 3$ kernels. Moreover,
Fig. 3. The overall architecture of our proposed model. It consists of a coarse-to-fine structure generator network $G$ and a Temporal PatchGAN discriminator network $D$. The generator processes IVUS images in two stages, the first stage adopts an image completion module $H_c$, which comprises a stack of gated convolution operations; And the second stage is a spatiotemporal aggregation network with two branches, temporal feature extracting module $H_t$ and aggregation module $H_s$. Discriminative network $D$ composed of five vanilla 3D convolution followed by spectral normalization. In addition to the adversarial loss, we also adopt the reconstruction loss, perceptual loss, and style loss to further improve the detail texture.

the feature streams extracted from each layer of $H_t$ will be aggregated with the features from the counterpart of $H_s$. This combination method ensures that the network simultaneously utilizes the spatial and temporal features. Furthermore, through hierarchical skip connections from the respective layers, the reconstructed image can be refined at each forward pass. Without loss of generality, for all down-sampling and upsampling layers, we apply the convolution and deconvolution operation with the stride of 2. Each convolutional operation is followed by an instance normalization and we adopt LeakyReLU as the activation function. The entire training procedure has shown in Algorithm 1. In reality, we first train the module $H_c$ with the non-adversarial loss before the training of the whole model.

1) Gated Convolution Operation: For the purpose of distinguishing invalid pixels and valid pixels, gated convolution and partial convolution are developed for image inpainting tasks. However, different from the partial convolution that has a restrict mask update step after each process, the gated convolution employs a dynamic feature selection mechanism. It firstly used a gating convolution filter $W_g$ and a feature convolution filter $W_f$ for input features $I$ to obtain the gating value $Gating$ and output features $Features$ respectively. Then it applied a soft attention map on the output features $Features$. The whole gated convolution operation can be represented by the following mathematical formula [31]:

$$Gating_{x,y} = \sum_{i=-k'_h}^{k'_h} \sum_{j=-k'_w}^{k'_w} W_g(k'_h+i,k'_w) \cdot I_{x+i,y+j}$$  \hspace{1cm} (1)$$

$$Features_{x,y} = \sum_{i=-k'_h}^{k'_h} \sum_{j=-k'_w}^{k'_w} W_f(k'_h+i,k'_w) \cdot I_{x+i,y+j}$$  \hspace{1cm} (2)$$

$$Output_{x,y} = \sigma(Gating_{x,y}) \cdot \phi(Features_{x,y})$$  \hspace{1cm} (3)$$

where $x, y$ are the corresponding pixel spatial coordinates, the kernel size of $W_g$ and $W_f$ are both $(k_h, k_w)$, $k'_h = \frac{k_h}{2}$.
and + is the perceptual loss, φN are the number of elements in ∥ ⊛ + L ◦ f = and stride ⋆ H = and ⊛ = V + σ = H σ from the pretrained VGG19 = ⊛ + ⊛ = V )

The Convolutional Long Short-Term Memory (ConvLSTM) has a powerful capacity to extract long-range temporal cues, and it has been used to make full use of the correlation information hidden in the sequence. The detail operations can be seen follows:

\[
i_t = \sigma(W_{xi} \odot X_t + W_{hi} \odot H_{t-1} + b_i) \\
f_t = \sigma(W_{xf} \odot X_t + W_{hf} \odot H_{t-1} + b_f) \\
\tilde{c}_t = W_{xc} \odot X_t + W_{hc} \odot H_{t-1} + b_c \\
c_t = i_t \odot \tanh(\tilde{c}_t) + f_t \odot c_{t-1} \\
o_t = \sigma(W_{xo} \odot X_t + W_{ho} \odot H_{t-1} + b_o) \\
H_t = o_t \odot \tanh(c_t)
\]

where \(\sigma(\cdot)\) represents a sigmoid(\cdot) function applied to the input gate \(i(t)\), forget gate \(f(t)\) and output gate \(o(t)\), and a \(\tanh(\cdot)\) applied to the memory cell \(c(t)\) and cell state \(c(t)\). \(\odot\) is the convolution operator, \(\odot\) denotes pixel-wise product. The bidirectional ConvLSTM is the enhanced version of ConvLSTM, it comprise a forward and a backward branch to access long-range context in both directions of the time sequence.

C. Discriminator Network

Previous inpainting networks [22, 30] use a global GAN with an additional local GAN to improve the resulting outputs. To make the network adapt to the irregular form masks, Yu et al. proposed SN-PatchGAN, which directly applies GANs for each element in feature maps, formulating \(h \times w \times c\) number of GANs focusing on different locations and different semantics of input image [31]. Furthermore, to incorporate the time dimension into the GAN losses for better results, Chang et al. [36] proposed Temporal PatchGAN (T-PatchGAN), which has better spatial-temporal modeling performance by sufficiently extracting global and local features with additional temporal dimension information. In this work, we adopt the same discriminate strategy like it, a T-PatchGAN following with spectral normalization [44] to stabilize the training of the network. The discriminator comprises five 3D vanilla convolutional layers with kernel size \(3 \times 5 \times 5\) and stride \(1 \times 2 \times 2\). The restored frames from the generator are passed through the discriminator, and the discriminator generates a score map where each element corresponding to a local region of the input. Accordingly, the discriminator can be used to classify each spatiotemporal patch is real or fake.

D. Loss Functions

To improve the restoration quality of the IVUS images, optimization objectives should be helpful to ensure per-pixel reconstruction accuracy and spatiotemporal coherence. In this paper, we adopt the joint of pixel-wise reconstruction loss, perceptual loss, style loss, and adversarial loss function to optimize the proposed network. Total loss: The overall loss function utilities in our network is defined as:

\[
L_{\text{total}} = L_r + \lambda_p L_{\text{perc}} + \lambda_s L_{\text{style}} + \lambda_G L_G
\]

where \(L_r\) is the reconstruction loss, \(L_{\text{perc}}\) is the perceptual loss, \(L_{\text{style}}\) is the style loss, the first three are non-adversarial losses and the last \(L_G\) is the adversarial loss, \(\lambda_p\), \(\lambda_s\) and \(\lambda_G\) are the corresponding weights.

Reconstruction loss: To make the network produce consistent results, and improve the per-pixel reconstruction accuracy, we include \(L_r\) losses by calculating the mean absolute error (MAE) between output and target images, it comprises of two items, the valid pixels region loss \(L_{\text{valid}}\) and the hole region loss \(L_{\text{hole}}\), this can be represented as follows:

\[
L_{\text{hole}} = \frac{\|(1 - M) \odot (V_o - V_{gt})\|_1}{N_h}
\]

\[
L_{\text{valid}} = \frac{M \odot (V_o - V_{gt})\|_1}{N_v}
\]

\[
L_r = \lambda_h L_{\text{hole}} + \lambda_v L_{\text{valid}}
\]

where \(\odot\) indicates element-wise product operator, \(M\), \(V_{gt}\), and \(V_o\) indicating the mask, ground truth and output frames from generator separately, \(\lambda_h\) and \(\lambda_v\) are the weights for the corresponding loss, \(N_h\) and \(N_v\) are the number of elements in hole region and valid region, respectively.

Perception and Style Loss: Despite the reconstruction loss is helpful to improve the consistency of the results, but they are inevitably bringing blurry results [29]. As a result, the output of the network usually maintains a good global structures at the expense of distortion and loss of details [45]. Consequently, we also adopt perceptual loss and style loss to improve the high-frequency components for the recovered results. We use layer relu1_2, relu2_2, relu3_4, relu4_4 from the pretrained VGG19 network on ImageNet [46] for the loss calculation, and the
We have prepared three types of datasets for the work, which we referred to as Type I, Type II, and Type III. All the data are acquired by the Boston Scientific intravascular ultrasound system, and the system is equipped with a mechanical rotary transducer. The transducer frequency is 40 MHz and it rotates at approximately 1800 revolutions per minute, the system’s pullback speed is uniformly 5 millimeters per second, and the imaging speed is 30 frames per second. The first two are animal IVUS images datasets, which were collected by professional clinicians through in vitro experiments in the coronary arteries of experimental rabbits, and the third type dataset is collected from clinical IVUS data. As shown in Fig. 5, we have presented some samples from these three different datasets. For the first two animal datasets, we simulate the real coronary artery activity by injecting the physiological saline into the coronary vessels. Because the experiments are implemented in vitro, we can remove the guidewire from the flexible catheter to access complete IVUS imaging information. Therefore, the Type I dataset is not affected by the guidewire artifacts. However, the Type II dataset is collected without removing the device’s guidewire, which makes it closer to the real clinical scenario, and thus its panoptic information will be disrupted by guidewire artifacts. Moreover, the ultimate goal of this work is to repair images affected by real guidewire artifacts. Finally, after excluding obvious abnormal images caused by irregular operations, we selected continuous IVUS frames affected by guidewire artifacts. The animal data are collected from 4 experimental rabbits, the training dataset comes from the Type I dataset, includes 3000 IVUS images.

**Algorithm 1: Training of the Proposed Network.**

**Data:** Paired training dataset $(V_{gt}, M)$

**Result:** Complete intravascular ultrasound images without guidewire artifacts $V_{out}$

1. initialization;
2. while Generator $G$ has not converged do
   
   for $n = 1, 2, 3$ do
     
     Sample batch images $V_{gt}$ and the corresponding masks $M$ from training dataset;
     
     Construct inputs $V_{in} \leftarrow V_{gt} \odot M$;
     
     Getting raw restored images from generator $V_{o} \leftarrow G(V_{in}, M)$;
     
     Get composite IVUS images $V_{out} = M \odot V_{in} + (1 - M) \odot V_{o}$;
     
     Updating parameters $\theta_{D}$ of discriminator $D$ with loss $L_{D}$;
   
   end
   
   Sample batch images $V_{gt}$ and the corresponding masks $M$ from training dataset;
   
   Construct inputs $V_{in} \leftarrow V_{gt} \odot M$;
   
   Getting raw restored images from generator $V_{o} \leftarrow G(V_{in}, M)$;
   
   Getting composite IVUS images $V_{out} = M \odot V_{in} + (1 - M) \odot V_{o}$;
   
   Updating parameters $\theta_{G}$ of generator $G$ with loss $L_{r}, L_{style}, L_{perc}, L_{G}$;

end

### A. Datasets

1) **IVUS Data:** We have prepared three types of datasets for the work, which we referred to as Type I, Type II, and Type III. All the data are acquired by the Boston Scientific intravascular ultrasound system, and the system is equipped with a mechanical rotary transducer. The transducer frequency is 40 MHz and it rotates at approximately 1800 revolutions per minute, the system’s pullback speed is uniformly 5 millimeters per second, and the imaging speed is 30 frames per second. The first two are animal IVUS images datasets, which were collected by professional clinicians through in vitro experiments in the coronary arteries of experimental rabbits, and the third type dataset is collected from clinical IVUS data. As shown in Fig. 5, we have presented some samples from these three different datasets. For the first two animal datasets, we simulate the real coronary artery activity by injecting the physiological saline into the coronary vessels. Because the experiments are implemented in vitro, we can remove the guidewire from the flexible catheter to access complete IVUS imaging information. Therefore, the Type I dataset is not affected by the guidewire artifacts. However, the Type II dataset is collected without removing the device’s guidewire, which makes it closer to the real clinical scenario, and thus its panoptic information will be disrupted by guidewire artifacts. Moreover, the ultimate goal of this work is to repair images affected by real guidewire artifacts. Finally, after excluding obvious abnormal images caused by irregular operations, we selected continuous IVUS frames affected by guidewire artifacts.
We first construct masks for the Type I dataset, to continue to train the network. For the hyperparameters, we set \( \lambda = 1 \) until the losses plateau, and we lower the learning rate to \( lr = 10^{-5} \) to continue to train the network until convergence. The whole model is trained on one Nvidia Tesla V100 GPU with a batch size of 2. The sequence length is 5, and the size of input and output uniformly resized to be 256 \times 256 pixels. For the hyperparameters, we set the values \{ \lambda_h, \lambda_v, \lambda_{perc}, \lambda_{style}, \lambda_G \} = \{ 10, 1, 0.1, 100, 1 \}, all of these values are empirically based on the experimental observations in our model. The model was tested on the data from three types of IVUS images.

### Evaluation Metrics

Since there is no consensus regarding the best metric score for medical image inpainting, to evaluate the quality of the model outputs, we adopt Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [48], and Learned Perceptual Image Patch Similarity (LPIPS) [49] three complimentary evaluation metrics. These metrics can be used to evaluated at pixel and perceptual levels.

### Compared Methods

Since there has no existing work for inpainting the corrupted IVUS images, We compare our model with seven state-of-the-art inpainting models, where four image inpainting models PConv [28], GLCIC [30], GatedConv [31], and MADF [32], three video inpainting models Free-form [36], CombCN [35], and CPNet...
We present Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) three complimentary evaluation metrics. ↑ Means Higher is Better. ↓ Means Lower is Better.


table 1: quantitative comparison between different methods on the type i dataset

| Methods     | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|-------------|--------|--------|---------|
| PConv[28]   | 32.82  | 0.9205 | 0.0167  |
| Free-form[36]| 32.60  | 0.9737 | 0.0118  |
| CombCN[35]  | 31.56  | 0.9722 | 0.0197  |
| GLCIC[30]   | 32.52  | 0.9742 | 0.0147  |
| GatedConv[31]| 32.45  | 0.9818 | 0.0144  |
| CPNet[47]   | 32.14  | 0.9528 | 0.0135  |
| MADF[32]    | 33.16  | 0.9763 | 0.0152  |
| Ours        | 33.26  | 0.9832 | 0.0103  |

E. Experiment Results

1) Quantitative Comparisons: Different scores assess various aspects of the image generation process, a single score is unlikely to cover all aspects [50]. We report our evaluation results in terms of PSNR, SSIM, and LPIPS three metrics on the Type I animal dataset. The quantitative comparison results between different methods have shown in Table I. It can be seen from the results, under different measurement standards, the gap between various methods is not very obvious. However, compared with other state-of-the-art methods, our method gains better performance with PSNR of 33.26, SSIM of 0.9832, and LPIPS of 0.0103. The superior results show the effectiveness of our model in restoring the original content whether on pixel level or perceptual level. And for clinical data, since the construction of its training set is a trade-off solution, we pay more attention to its qualitative repair results rather than quantitative measures.

2) Qualitative Comparisons: The qualitative comparisons of the results on the Type I test dataset are shown in Fig. 8. The results have shown that all these methods can restore the general structure of the missing information in the appropriate position, but our method can achieve pleasing detailed information compared with other competitors. Particularly from the first two columns, we can see that the generated content is more consistent with the surrounding regions. Furthermore, to ensure that the network preserves envelope data statistical features, as shown in Fig. 9, we perform a histogram comparison of the predicted and ground truth regions. Specifically, the comparison images are from the third column of Fig. 8, and in order to make the histogram more statistically meaningful, the calculation area is selected on the more accurate arterial vessel wall rather than the entire mask area. In addition, we present the results of diverse methods tested on the Type II animal dataset, which has realistic guidewire artifacts, making it more consistent with the clinically collected data. The comparison restoration results can be seen in Fig. 10. Since it is difficult for us to know exactly the original information at the location of the artifacts, we can only provide the visual performance of the restored image without any accuracy measurement metrics. From the final results can be seen, there exist significant disparities between different methods. PConv [28], GatedConv [31], CPNet [47] and MADF [32] can generate relatively complete content for corrupted sequences, but the new content exists blurry effect compared with us. Unfortunately, the generalization performance of other models in repairing real guidewire artifacts is mediocre. Due to its excellent spatiotemporal modeling capabilities, the proposed network could effectively restore the corrupted areas with realistic and plausible tissue structure and details. Furthermore, to test the performance on clinical data, we have presented the restoration results on the Type III dataset. It can be seen from Fig. 11, the proposed method has achieved good performance to restore the damaged clinical IVUS images.

3) Subjective Evaluation: We perform a human subjective test for evaluating the quality of the inpainted IVUS images on Type II and Type III dataset, separately. We compare our method with seven compared methods and a total of eight clinicians participated in this study. Specifically, we randomly select two IVUS sequences from each dataset, and slow down the results to 10 FPS for better comparison, each participant is asked to give a grade between 1 to 10 (the higher the better) on each method. We specifically ask each participant to check for both image quality and temporal consistency. The study results are summarized in Fig. 12. Furthermore, we performed statistical comparisons between the compared methods with the proposed one using the Wilcoxon rank sum test. P-value < 0.05 was considered as a statistically significant level, and the difference between test methods is highly significant both in animal and clinical IVUS results.

F. Ablation Studies

We further conducted ablation experiments to assess the contribution of separate components in the proposed model. In this stage, we adopted the data from the Type I animal dataset. First, we remove the image completion module $H_c$, and feed the input sequence into the second stage of generator directly. Second, we remove the temporal feature extracting module $H_t$, and the remaining network structure is simplified to a two-stage image restoration network. In addition, we remove the $H_s$ while preserving the bidirectional ConvLSTM network, only use the last of the output of $H_t$ as the final restoration results. The statistical results of the ablation study are summarized in Table II. Experiment results demonstrate each element places an important role in the proposed model. Both the two-stage restoration method and the feature fusion mechanism we introduced facilitate restoring the original anatomical information with better reconstruction content.

V. LIMITATIONS

IVUS plays a vital role in assisting decision-making in percutaneous coronary intervention, including preoperative evaluation and postoperative optimization. Recovery the detailed
Fig. 8. Comparison results on Type I dataset between the different methods. (a) Input masked images with size of $256 \times 256$, (b) PConv [28], (c) Free-form [36], (d) CombCN [35], (e) GLCIC [30], (f) GatedConv [31], (g) CPNet [47], (h) MADF [32], (i) Ours, (j) Ground truth. It can be seen from the results that all methods have achieved relatively good repair results. However, compared with other methods, our method has an excellent recovery performance in terms of detailed structure.

TABLE II

| Methods | PSNR↑ | SSIM↑ | LPIPS↓ |
|---------|-------|-------|--------|
| w/o $H_c$ | 31.25 | 0.9416 | 0.0212 |
| w/o $H_t$ | 32.57 | 0.9665 | 0.0126 |
| w/o $H_\sigma$ | 31.18 | 0.9513 | 0.0164 |

We Conduct Different Ablation Experiments to Validate Our Network. ↑ Means Higher is Better. ↓ Means Lower is Better.

due to its feasibility, however, it is a novel but trade-off solution. Recently, Tehrani et al. [51] proposed a novel promising method to handle the domain shift, they introduced a domain adaptation stage and the reference phantom is adopted to CNN for domain transformation to further enhance the performance, meanwhile, a simple but effective data generation scheme is introduced.

Furthermore, medical image inpainting exists potential risks because data-driven models rely heavily on prior knowledge learned from training data, which is hard to be completely faithful to the anatomy. It can imagine that in the worst case, it may lead to missing a plaque or misjudging the size of a plaque. False recovery images may lead to the wrong decision of the clinicians if the information of that region is important. Since any changes on important anatomical structure could lead to a different therapeutic path, e.g. by taking the wrong stent size could have serious potential consequences, inaccurate quantitative calculation of coronary artery vulnerable plaque will be interfering with normal clinical analysis. Therefore, in anatomy has great significance while full of challenges, however, it is worth mentioning potential limitations of our proposed method. In realistic clinical operations, IVUS images acquired by the mechanical rotating device are inevitably corrupted by the guidewire artifacts, and it is difficult to acquire the paired dataset like Type I animal IVUS dataset for training our model as aforementioned. In this work, we have set the mask on non-artifact regions to construct paired clinical training datasets
Fig. 9. Comparison of histograms between predicted regions and ground truth regions for different methods. The first row: PConv [28], Free-form [36], CombCN [35], GLCIC [30]. The second row: GatedConv [31], CPNet [47], MADF [32], Ours method. The red line represents histogram of ground truth, while green line is histogram of predicted regions from different method.

Fig. 10. Comparison results on Type II dataset. (a) Input masked images with size of $256 \times 256$, (b) PConv [28], (c) Free-form [36], (d) CombCN [35], (e) GLCIC [30], (f) GatedConv [31], (g) CPNet [47], (h) MADF [32], (i) Ours, (j) Original animal IVUS images with real guidewire artifacts.
Fig. 11. Comparison results on the clinical dataset. (a) Input masked images with size of 256 × 256, (b) PConv [28], (c) Free-form [36], (d) CombCN [35], (e) GLCIC [30], (f) GatedConv [31], (g) CPNet [47], (h) MADF [32], (i) Ours, (j) Original clinical IVUS images with real guidewire artifacts.

Fig. 12. User study on results of inpainted IVUS images on Type II (animal) and Type III (clinical) dataset. The blue bar represents the average grade on the animal dataset, while the orange bar is the result on the clinical dataset. The difference between the proposed one with others is highly significant ($p < 0.05$) using the Wilcoxon Rank Sum test.

current stage, the recovered area may provide other artifacts and is not recommended for full replacement of the original image for clinical diagnosis. Nonetheless, it is more suitable for serving as a reference for computer-aided diagnosis and subsequent image processing tasks. In addition, its value in practical clinical applications requires further study.

VI. CONCLUSION

Intravascular ultrasound imaging plays an important role in the diagnosis of cardiovascular diseases, while it has been plagued by guidewire artifacts in clinical practice. In this paper, we adopt inpainting methodologies to conduct pilot research to better solve this long-standing problem. Here, we proposed the first end-to-end learning based network AIVUS, which can be effective to restore the original information in corrupted IVUS imaging. The whole network has a novel generative adversarial architecture. The generative part has a coarse-to-fine architecture, and the united of gated convolution and spatiotemporal aggregation learning structure take place a vital place in handling large-scale guidewire artifacts and mining redundancy information between frames. We first conducted experiments on animal IVUS datasets. Both the quantitative and qualitative results compared with other inpainting approaches have illustrated the superior performance of the proposed network, and the
restored results show it has pleasing performance in improving the consistency and high fidelity details.

Furthermore, to test the restoration performance on the realistic clinical IVUS images, we designed a novel clinical training dataset generation method and conducted experiments to recover the complete information for corrupted clinical images. The recovery results on the clinical dataset demonstrate that the proposed method has potential benefits for clinical diagnosis, especially the post-processing tasks for IVUS images.

Currently, there are more challenges for medical image inpainting techniques, and the best evaluate method is to regress the problem to the clinical application scenarios. In the future, we plan to conduct larger-scale clinical studies to thoroughly evaluate the proposed method. Furthermore, we plan to further evaluate the performance of the model in subsequent image analysis tasks, namely the impact of this task on complex tissue structure analysis, such as vulnerable plaques, calcified regions, and lumen segmentation. In addition, for the clinical dataset with guidewire artifacts, other potential learning strategies such as transfer learning and unsupervised learning also provide a promising solution.

REFERENCES

[1] G. S. Mintz et al., “American college of cardiology clinical expert consensus document on standards for acquisition, measurement and reporting of intravascular ultrasound studies (IVUS) a report of the american college of cardiology task force on clinical expert consensus documents developed in collaboration with the european society of cardiology endorsed by the society of cardiac angiography and interventions,” J. Am. Coll. Cardiol., vol. 37, no. 5, pp. 1478–1492, 2001.

[2] J. Tong, K. Li, W. Lin, X. Shudong, A. Anwar, and L. Jiang, “Automatic lumen border detection in ivus images using dictionary learning and kernel sparse representation,” Biomed. Signal Process. Control, vol. 66, 2021, Art. no. 102489.

[3] T. W. Johnson et al., “Clinical use of intracoronary imaging, part 2: Acute coronary syndromes, ambiguous coronary angiography findings, and guiding interventional decision-making: An expert consensus document of the European association of percutaneous cardiac interventions: Endorsed by the society of cardiology, the Hong Kong Society of transcatheater endovascularоlogists (HKSTENT) and the cardiac society of Australia and New Zealand,” Eur. Heart J., vol. 40, no. 31, pp. 2566–2584, 2019.

[4] A. Katouzian, “Quantifying atherosclerosis: IVUS imaging for lumen border detection and plaque characterization,” Ph.D. dissertation, Columbia Univ., York City, NY, USA, 2011.

[5] H. M. Garcia-Garcia, B. D. Gogas, P. W. Serruys, and N. Bruining, “IVUS-based imaging modalities for tissue characterization: Similarities and differences,” Int. J. Cardiovasc. Imaging, vol. 27, no. 2, pp. 215–224, 2011.

[6] H.-B. Park, Y.-H. Cho, and D.-K. Cho, “IVUS artifacts and image control,” in Coronary Imaging and Physiology. Berlin, Germany: Springer, 2018, pp. 9–17.

[7] A. Hammadou, G. Cloutier, J. C. Tardif, K. Hammouche, and J. Meunier, “Automatic IVUS lumen segmentation using a 3D adaptive helix model,” Comput. Biol. Med., vol. 107, pp. 58–72, 2019.

[8] M. Faraji, I. Cheng, I. Naudin, and A. Basu, “Segmentation of arterial walls in intravascular ultrasound cross-sectional images using extremal region selection,” Ultrasong, vol. 84, pp. 356–365, 2018.

[9] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, “Recurrent temporal aggregation framework for deep video inpainting,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 5, pp. 1038–1052, May 2020.

[10] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, “Deformable 3D convolution for video super-resolution,” IEEE Signal Process. Lett., vol. 27, pp. 1500–1504, 2020. doi: 10.1109/LSP.2020.3015518.

[11] R. Liu, Z. Weng, Y. Zhu, and B. Li, “Temporal adaptive alignment network for deep video inpainting,” in Proc. 29th Int. Joint Conf. Artif. Intell., 2020, pp. 927–933.
[40] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2462–2470.

[41] Y. Ding, C. Wang, H. Huang, J. Liu, J. Wang, and L. Wang, “Frame-recurrent video inpainting by robust optical flow inference,” 2019, arXiv:1905.02882.

[42] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 1. Cambridge, MA, USA: MIT Press, 2015, pp. 802–810.

[43] Q. Lyu et al., “Cine cardiac MRI motion artifact reduction using a recurrent neural network,” *IEEE Trans. Med. Imag.*, vol. 40, no. 8, pp. 2170–2181, Aug. 2021.

[44] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” in *Proc. Int. Conf. Learn. Representations*, 2018.

[45] K. Armanious et al., “MedGAN: Medical image translation using GANs,” *Computerized Med. Imag. Graph.*, vol. 79, 2020, Art. no. 101684.

[46] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

[47] S. Lee, S. W. Oh, D. Won, and S. J. Kim, “Copy-and-paste networks for deep video inpainting,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4413–4421.

[48] Z. Wang, A. C. H. R. Bovik Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[50] A. Borji, “Pros and cons of gan evaluation measures,” *Comput. Vis. Image Understanding*, vol. 179, pp. 41–65, 2019.

[51] A. K. Z. Tehrani, I. M. Rosado-Mendez, and H. Rivaz, “Robust scatterer number density segmentation of ultrasound images,” *IEEE Trans. Ultrason., Ferroelect., Freq. Control*, vol. 69, no. 4, pp. 1169–1180, Apr. 2022.