VesselDiffusion: 3D Vascular Structure Generation Based on Diffusion Model

Zhanqiang Guo, Zimeng Tan, Jianjiang Feng, *Member, IEEE*, and Jie Zhou, *Fellow, IEEE*

***Abstract*— 3D vascular structure models are pivotal in disease diagnosis, surgical planning, and medical education. The intricate nature of the vascular system presents significant challenges in generating accurate vascular structures. Constrained by the complex connectivity of the overall vascular structure, existing methods primarily focus on generating local or individual vessels. In this paper, we introduce a novel two-stage framework termed VesselDiffusion for the generation of detailed vascular structures, which is more valuable for medical analysis. Given that training data for specific vascular structure is often limited, direct generation of 3D data often results in inadequate detail and insufficient diversity. To this end, we initially train a 2D vascular generation model utilizing extensively available generic 2D vascular datasets. Taking the generated 2D images as input, a conditional diffusion model, integrating a dual-stream feature extraction (DSFE) module, is proposed to extrapolate 3D vascular systems. The DSFE module, comprising a Vision Transformer and a Graph Convolutional Network, synergistically captures visual features of global connection rationality and structural features of local vascular details, ensuring the authenticity and diversity of the generated 3D data. To the best of our knowledge, VesselDiffusion is the first model designed for generating comprehensive and realistic vascular networks with diffusion process. Comparative analyses with other generation methodologies demonstrate that the proposed framework achieves superior accuracy and diversity. Our code is available at: https://github.com/gzq17/VesselDiffusion.**

***Index Terms*— Vascular Structure Generation, Diffusion Model, Dual-Stream Feature Extraction**

I. Introduction

Vascular structure models are integral to various applications, such as clinical diagnosis, surgical planning, virtual interventional vascular surgery, and medical education [1], [2]. The manual extraction of numerous vascular structures for analysis is labor-intensive, resulting in a scarcity of vascular data. Recently, many vascular structure generation methods have been proposed [2]–[5]. Nevertheless, the inherent variability in the shape, size, and structure of the vascular system continues to pose substantial challenges in generating accurate vascular models.

Traditional vascular generation methods predominantly rely on the relationship between vessel diameter and blood flow [3], [6], utilizing a set of fixed rules and vascular dynamics constraints to model blood vessels [7]–[9]. While these approaches introduce some variability in vascular structure generation, they often fail to capture the full diversity of vascular architectures. Moreover, they typically represent blood vessels using simplified geometric shapes such as cylinders or truncated cones (as shown in Fig. 1(a)), which inadequately reflect the actual complex morphology of blood vessels [6]. Recently, the development of deep learning algorithms has significantly advanced the generation of natural images, prompting research into their application in vessel generation via Generative Adversarial Networks (GANs) [4] and Variational Autoencoders (VAEs) [5]. However, these efforts primarily focus on the generation of local or single vessel (Fig. 1(b) and (c)), lacking the capability to produce detailed vascular structures. This limitation is likely due to the inherent difficulty in capturing the detail and global connectivity of the vascular network.

In addition to specialized vascular synthesis techniques, methods for generating natural images and 3D objects can also be applied to vascular generation studies. Notably, the recent surge in popularity of diffusion probabilistic models [10], [11] has led to significant advancements in the generation of natural images, videos, and 3D objects [12], [13]. The efficacy of these models is contingent upon access to vast amounts of training data. For instance, the classic Stable Diffusion[1] was trained on a dataset comprising 5 billion images. The training of 3D generative models also necessitates extensive data support [14], [15]. However, the acquisition of medical imaging data is considerably constrained by ethical and privacy concerns, making it challenging to amass large datasets. Consequently, training generative models with the limited available 3D data poses significant difficulties in ensuring the diversity and accuracy of the generated vascular structure models.

To address these challenges, we propose a novel two-stage framework, termed VesselDifussion, for comprehensive 3D vascular network generation (Fig. 1(d)). We represent vascular data as surface point clouds to mitigate the complexities inherent 3D space. In comparison to 3D vascular data, some pub-

Manuscript received November 1, 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62321005. This study got ethical approval of Wuhan Union Hospital of China and Xuanwu Hospital of Capital Medical University (2020009) for using the clinically collected dataset. *(Corresponding author: Jianjiang Feng.)*

Zhanqiang Guo, Zimeng Tan, Jianjiang Feng, and Jie Zhou are with the Department of Automation, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China (e-mail: guozq21@mails.tsinghua.edu.cn; tzm19@mails.tsinghua.edu.cn; jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

licly accessible 2D datasets contain larger sample sizes (e.g., X-ray dataset [16] consists of over 1,000 images), likely due to the greater ease of storage and annotation. Furthermore, 3D vascular structures can be projected from multiple viewpoints to generate additional 2D representations, further expanding the accessible dataset size. Given the relative abundance of 2D vascular data, which captures diverse structural and topological variations, coupled with the comparative ease of 2D data generation, we initially train an improved diffusion model [11] using a large corpus of 2D images, including Maximum Intensity Projection (MIP) images from 3D volumes and other publicly accessible datasets. Subsequently, special 3D vascular data and their corresponding 2D MIP images are employed to train a 3D point cloud conditional generation model. The model incorporates a Dual-Stream Feature Extraction (DSFE) module, consisting of a Vision Transformer (ViT) and a Graph Convolutional Network (GCN), to effectively extract structural features of local vascular details and visual features of global connection rationality from the 2D images and the corresponding graph. The structural features enhance the precision of critical parameters such as the radius and curvature of local blood vessels and ensure the consistency of the generated 3D results with the input 2D images. Simultaneously, the visual features uphold the global integrity and anatomical rationality of the vascular constructs, collectively facilitating a thorough and accurate representation of the vascular system. Furthermore, combining GCN with ViT effectively simulates the arboriform structural characteristics of blood vessels, thereby enhancing the network's capacity to learn comprehensive and intricate vascular features. Experimental results indicate that our method surpasses existing models in both qualitative and quantitative assessments.

The main contributions of our study are as follows:

- We introduce VesselDiffusion for 3D vascular structure generation, which, to the best of our knowledge, is the first study to generate detailed and anatomically realistic vascular networks based on diffusion model.
- To mitigate the limitations posed by the scarcity of 3D vascular data, we propose a novel training paradigm that disentangles the process of 3D vascular generation into the creation of 2D MIP masks and the reconstruction of 3D vascular systems. The usage of extensive 2D data enhances the diversity of generated MIP images, and the subsequent transition from 2D to 3D generation maintains the heterogeneity and detail of vascular structures.
- In proposed point cloud conditional generation model, we combine a ViT and a GCN to extract both visual and structural features, thereby ensuring the rationality and diversity of the generated vascular structures.

II. Related Work

A. Medical Image Generation

Obtaining large-scale medical image datasets poses significant ethical and privacy challenges, making medical image generation a persistent area of interest. Traditional approaches have predominantly employed Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) to capture the underlying data distributions [17]. With the recent success of diffusion models across various domains, their application to medical image generation has gained traction, yielding promising outcomes [18], [19]. While the generation of 2D images has been well explored, efforts to extend these techniques to 3D images remain limited. The inherent complexity and high dimensionality of 3D data often result in a loss of fine-grained detail [20], [21]. This issue is particularly pronounced in the synthesis of 3D vascular structures, where the sparse nature of the data and the need for high-fidelity detail present additional challenges. In our study, we address this by modeling sparse blood vessels as surface point clouds and leveraging a 3D diffusion model to synthesize detailed 3D vascular structures.

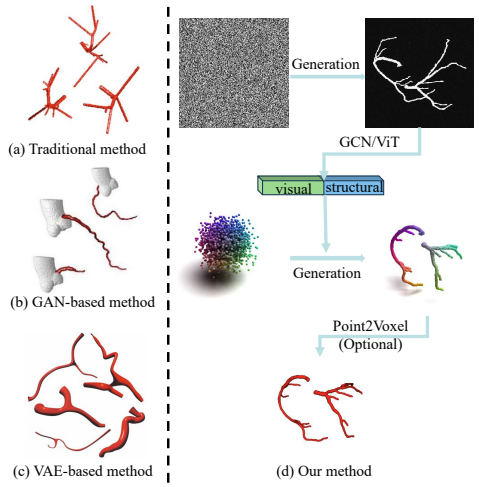


Fig. 1. Vascular structures generated by different methods. (a) the generated vessels of [3]. (b) the results of [4]. (c) the synthetic vessels of [5]. (d) the process of our method and the synthetic data.

B. Vessel Generation

Research on vascular generation can be categorized into traditional algorithms and learning-based methods. Traditional techniques primarily involve the iterative growth of blood vessels based on hemodynamics and predefined rules [2], [3], [6], [22], [23]. For instance, Hamarneh et al. [3] iteratively developed vascular structures using a user-defined oxygen demand map, considering bifurcation locations, branching properties, and tree hierarchy. However, these methods often produce blood vessels with limited diversity and overly idealized structures. Learning-based methods utilize natural image generation techniques, such as GANs [24] and VAEs [25], to learn blood vessel distributions from real data [4], [5]. Wolterink et al. [4] proposed a GAN-based model to generate single coronary vessel, and Feldman et al. [5] employed a recursive variational neural network to generate the local vascular structures.

Recent advancements in vascular network modeling have seen the exploration of implicit neural representations, with Sinha et al. [26] leveraging diffusion models for vascular generation. Their study introduced a novel representation method designed to mitigate storage demands, demonstrating

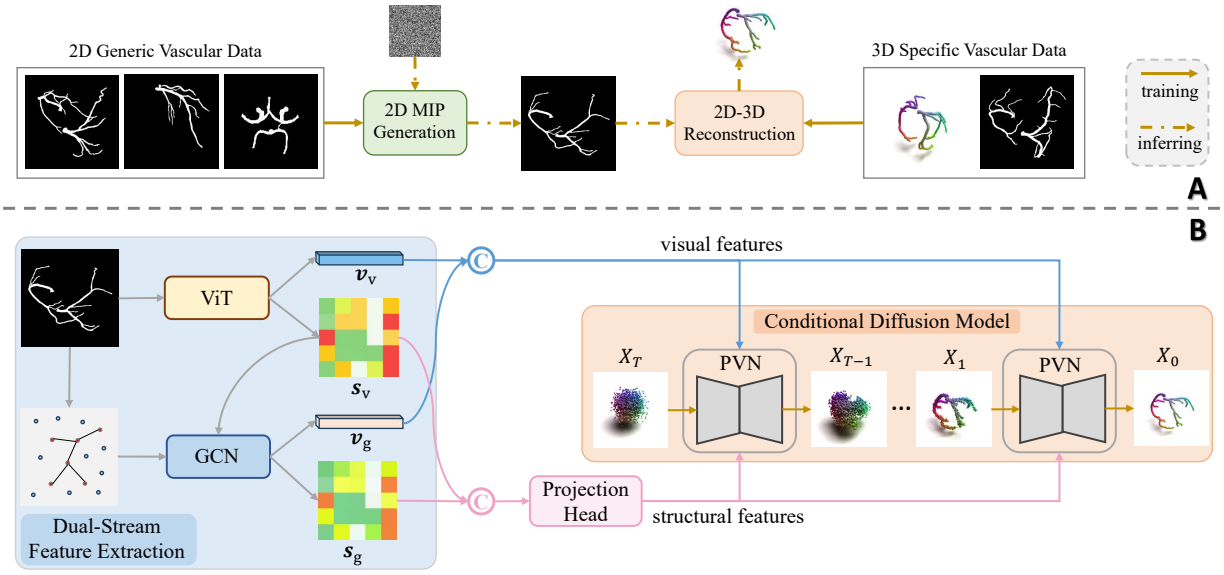


Fig. 2. The proposed VesselDiffusion framework for generating 3D vascular systems. (A) depicts the training and inference process of two stages within our method. (B) provides a detailed overview of the 3D point cloud generation network conditioned on 2D MIP images, where Point-Voxel Network (PVN) [39] is used as the denoising function of diffusion model.

its efficacy through localized vascular generation tasks. However, potential representation errors, coupled with the absence of structural constraints and insufficient utilization of vascular topology information, may limit its ability to generate anatomically complex vascular networks. The success of discrete diffusion model-based graph generation [27], [28] in various domains, such as molecular generation [29] and protein design [30], has motivated further investigations into its applicability for vascular structure synthesis. For example, Prabhakar et al. [31] proposed a simplified graph-based representation of blood vessels, synthesizing vascular networks by generating nodes and edges. While effective in constructing connectivity patterns, the abstraction of vascular structures into discrete points and lines compromises anatomical realism, potentially limiting its applicability in biomedical contexts. Despite these advancements, current learning-based methods are limited to generating local blood vessel or over-simplified vascular network. To the best of our knowledge, few studies have explored the synthesis of comprehensive and anatomically realistic vascular networks using learning-based techniques.

C. Diffusion Model and 3D Point Generation

The diffusion probabilistic model, inspired by non-equilibrium thermodynamics, employs Markov chains to transform Gaussian distributions into target data distributions [10], [11], [32]. This approach facilitates image generation through learning the perturbation removal, achieving notable success. The diffusion models have outperformed GANs in text-to-image generation [33], image editing [34], and video synthesis [35]. In the realm of 3D point cloud generation, the diffusion model has also demonstrated impressive results [14], [36], [37]. For instance, Luo et al. [12] and Zhou et al. [38] adapted the diffusion model of 2D image generation to 3D point cloud generation and achieved good results. Subsequently, text-to-3D [14] and 2D-to-3D generation [15], [36] has also made significant strides. However, the success of the diffusion models is heavily reliant on the availability of extensive datasets. In contrast, obtaining large datasets for 3D vascular data is challenging due to ethical and confidentiality constraints, making it difficult to ensure the structural diversity and detailed accuracy of results when directly training 3D vascular surface point cloud generation models. In this paper, we propose to leverage the abundant available 2D vascular images to generate diverse MIP images, which are used to reconstruct corresponding 3D point clouds.

III. Method

A. Overview

As depicted in Fig. 2(A), our proposed framework decomposes the generation of 3D vascular systems into two distinct stages. In the initial phase, an improved 2D diffusion model [11] is trained using MIP images derived from 3D volumes alongside other publicly available 2D vascular datasets, which encompass abundant vascular topologies and structural variations, ensuring the diversity of the generated 2D data.

In the subsequent stage, we leverage specific 3D vessel data in conjunction with the corresponding MIP masks to train a 3D point cloud generation network for reconstructing 3D vascular structures from 2D images, where the 2D vascular structures serve as guidance and facilitate the model's acquisition of structural information. Meanwhile, the 2D-to-3D conversion also ensures the diversity of the 3D generation by enforcing the consistency with the 2D vascular structures. During the inference phase, random Gaussian noise is fed into the 2D generative model to produce the vascular MIP mask, which is then utilized by the trained 2D-to-3D reconstruction network to derive the comprehensive 3D vascular structures. In the initial stage, our framework is compatible with most existing 2D generation models. Our primary focus, however, lies in the second stage, as detailed in Fig. 2(B), where we elaborate on

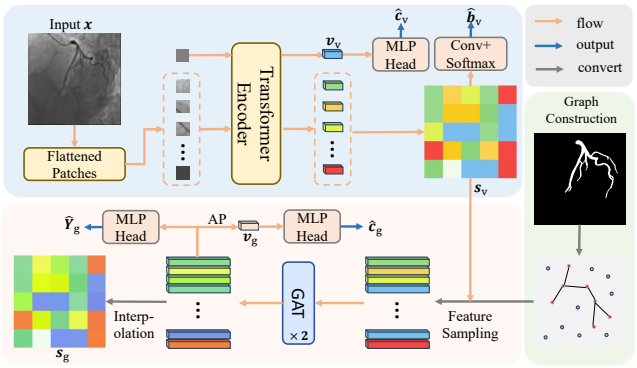


Fig. 3. The Dual-Stream Feature Extraction (DSFE) module. During pre-training, the input image $\boldsymbol{x}$ is processed by a Transformer Encoder to extract features, which are used to predict vessel segmentation map ($\hat{\boldsymbol{c}}_{\mathbf{v}}$) and image category ($\hat{\boldsymbol{b}}_{\mathbf{v}}$). Subsequently, the vessel graph integrated with features extracted by ViT is processed by a two-layer Graph Attention Network (GAT) to predict node classification result ($\hat{\boldsymbol{Y}}_{\mathbf{g}}$) and graph-level category ($\hat{\boldsymbol{c}}_{\mathbf{g}}$).

the generation of 3D vascular surface point clouds from the generated 2D MIP images.

B. Problem Formulation

1) Point Cloud and Corresponding MIP Image.: In the initial stage, using the trained 2D generation model, we generate a variety of MIP images, where some noise is inevitably present. Subsequently, we train a 3D generation network conditioned on these 2D images. For a 3D vascular volume $V \in \mathbb{R}^{H\times W\times D}$, represented as a binary image with vessel voxels valued at 1 and background voxels at 0, we extract the surface point cloud $X \in \mathbb{R}^{N\times 3}$ and employ farthest point sampling (FPS) to obtain a fixed number of points. If the sampled points are insufficient, additional points are generated through interpolation. Concurrently, the corresponding 2D MIP image $I_{\mathrm{m}} \in \mathbb{R}^{H\times W}$ is generated, mathematically expressed as $I_{\mathrm{m}}(h,w) = \max_{d=1,2..,D} V(h,w,d)$. To align it with the generated 2D data in the first stage, random noise is introduced to produce $\tilde{I}_{\mathrm{m}}$. This process yields paired 3D point cloud X and 2D MIP image $\tilde{I}_{\mathrm{m}}$, facilitating the training of the 2D-to-3D generation network.

2) Point Cloud Generation with Conditional Diffusion Model.: A point cloud $X_0 \in \mathbb{R}^{N\times 3}$ is randomly sampled from the distribution $q(X_0)$, and noise is gradually introduced to X_0 following the forward process described in [10] until it conforms to a Gaussian distribution. The noise addition process adheres to the Markov chain assumption:

$$q\left(X_{0:T}\right) = q\left(X_0\right)\prod_{t=1}^{T} q\left(X_t \mid X_{t-1}\right), \tag{1}$$

$$q\left(X_t \mid X_{t-1}\right) = \mathcal{N}\left(\sqrt{1-\beta_t}X_{t-1}, \beta_t\mathbf{I}\right), \tag{2}$$

where $\{\beta_t\}_{t=0}^{T}$ are variance schedule hyper-parameters. The objective is to generate the corresponding 3D point cloud conditioned on $\tilde{I}_{\mathrm{m}}$. We consider the reverse diffusion process, which starts from the randomly sampled Gaussian distribution X_T and generates X_0 step-by-step conditioned on $\tilde{I}_{\mathrm{m}}$. A neural network parameterized by θ is employed to predict each step of the reverse diffusion process $p_\theta(X_{t-1} \mid X_t, \tilde{I}_{\mathrm{m}})$. The transition from the Gaussian distribution X_T to X_0 can be formulated as:

$$p_\theta\left(X_{0:T}\right) = p\left(X_T\right)\prod_{t=1}^{T} p_\theta\left(X_{t-1} \mid X_t, \tilde{I}_{\mathrm{m}}\right), \tag{3}$$

$$p_\theta\left(X_{t-1} \mid X_t, \tilde{I}_{\mathrm{m}}\right) = \mathcal{N}\left(\mu_\theta\left(X_t, t, \tilde{I}_{\mathrm{m}}\right), \sigma_t^2\mathbf{I}\right). \tag{4}$$

For effective network training, it is crucial that the noise predicted in the reverse process aligns with the noise introduced during the forward process. This alignment ensures that the desired 3D point cloud can be accurately generated from the standard Gaussian distribution. The loss function can be derived as:

$$\mathcal{L}_t = \mathbb{E}_{X_0\sim q(X_0)}\mathbb{E}_{\epsilon_t\sim\mathcal{N}(0,\mathbf{I})}\left\|\epsilon_t - \mu_\theta\left(X_t, t, \tilde{I}_{\mathrm{m}}\right)\right\|^2. \tag{5}$$

C. DSFE Module and Pretraining

The DSFE structure is shown in Fig. 3. We employ the ViT architecture to process the input image x and obtain the image classification result ($\hat{c}_{\mathrm{v}}$) and vessel segmentation prediction ($\hat{b}_{\mathrm{v}}$). Similar to the approach described in [40], we utilize GCN to further enhance the extraction of vessel features. Specifically, we construct a graph from the vessel label and utilize the intermediate features from ViT as node features to predict both the graph category ($\hat{c}_{\mathrm{g}}$) and the vessel membership of each node ($\hat{Y}_{\mathrm{g}}$). Next, we will provide a detailed introduction to the pretraining module.

1) The Flow of ViT.: The DSFE module is pre-trained using a diverse collection of 2D vascular data, including fundus vessels, cerebral vascular projection images, and X-ray coronary angiography images. From this collection, we randomly select a set of data (x, b, c), where $x \in \mathbb{R}^{H\times W}$ represents the grayscale image, $b \in \mathbb{R}^{H\times W}$ the corresponding binary vascular image, and $c \in \{0, 1, 2\}$ the category label, indicating the dataset of origin. Subsequently, ViT is employed to derive structural features $s_{\mathrm{v}} \in \mathbb{R}^{H\times W\times C}$ and visual features $v_{\mathrm{v}} \in \mathbb{R}^{C}$, the latter corresponding to the class token feature. The structural and visual features are then processed through convolution and Multi-Layer Perceptron (MLP) to yield the segmentation prediction ($\hat{b}_{\mathrm{v}}$) and classification result ($\hat{c}_{\mathrm{v}}$). The loss function is formulated as a linear combination of a Dice loss and a cross entropy loss:

$$\mathcal{L}_{\mathrm{ViT}} = \mathrm{Dice}(\hat{b}_{\mathrm{v}}, b) + \alpha_1\mathrm{CE}(\hat{c}_{\mathrm{v}}, c). \tag{6}$$

2) Graph Construction and GCN.: The vascular structure is distinctive in that it occupies a minimal foreground area and exhibits an arboriform configuration. Previous studies have demonstrated that graph structures based on images can more accurately represent blood vessels [40], [41]. Therefore, we employ GCN in conjunction with ViT to achieve a more comprehensive representation of vascular structures. Initially, we randomly select a set of data (x, b, c) from collected datasets. To maintain consistency with the generated MIP image, random noise is added to obtain $\tilde{b}$. Inspired by [40], [42], we construct a graph from the vascular image $\tilde{b}$ to

model the vascular structure and intuitively characterize vessel connectivity. Specifically, we partition $\tilde{b}$ into non-overlapping regions of size $h \times w$ (where each region may contain vessels, bifurcations, or background; in our experiments, $h = w = 6$). The center points of these regions serve as graph nodes $V = \{v_i\}_{i=1}^{N_n} = \{(h_i, w_i)\}_{i=1}^{N_n}$, with the corresponding node labels $Y = \{y_i\}_{i=1}^{N_n} = \{b(h_i, w_i)\}_{i=1}^{N_n}$, where $N_n = \lceil \frac{H}{h} \rceil \times \lceil \frac{W}{w} \rceil$. Subsequently, we compute the geodesic distance $d(i, j)$ between each point p_i and its surrounding point p_j based on $\tilde{b}$. A connection is established between two points (p_i, p_j) if $d(i, j)$ is below a specified threshold d_{deo} ($d_{\text{deo}} = 32$ in our experiments). The adjacency matrix A records the connection weights between nodes, and is calculated based on the geodesic distance, mathematically expressed as:

$$A_{ij} = \begin{cases} \frac{d_{\text{deo}}}{d_{\text{deo}} + d_{ij}} & d(i,j) \le d_{\text{deo}}, \\ 0 & d(i,j) > d_{\text{deo}}. \end{cases} \tag{7}$$

The initial features of each node comprise its coordinate values (h_i, w_i) and the corresponding structural features s_{v} extracted by ViT: $fea_i = \text{Concat}([h_i, w_i], s_{\text{v}}(h_i, w_i))$, $fea_i \in \mathbb{R}^{C+2}$.

The resulting graph structure is processed through two layers of Graph Attention Network (GAT), extracting features for each node. These node features are then organized into image feature based on their coordinates and upsampled to produce the structural feature map $s_{\text{g}} \in \mathbb{R}^{H \times W \times C}$. Each node's features are passed through a MLP head to generate the classification result for each node $\hat{Y}_{\text{g}} = \{\hat{y}_i\}_{i=1}^{N_n}$. By performing Average Pooling (AP) on node features, global visual features v_{g} and corresponding graph classification result $\hat{c}_{\text{g}}$ are obtained. The loss function of GCN is formulated as:

$$\mathcal{L}_{\text{GCN}} = \frac{1}{N_n} \sum_{i=1}^{N_n} \text{CE}(\hat{y}_i, y_i) + \alpha_2 \text{CE}(\hat{c}_{\text{g}}, c). \tag{8}$$

The total loss during pre-training is represented as

$$\mathcal{L}_{\text{pre-train}} = \alpha_3 \mathcal{L}_{\text{ViT}} + \mathcal{L}_{\text{GCN}}. \tag{9}$$

3) Pretraining of DSFE.: During the pre-training phase, we employ $\mathcal{L}_{\text{pre-train}}$ as the loss function to jointly train the ViT and GCN, enabling the DSFE module to effectively capture structural information of blood vessels. We utilize a diverse collection of 2D vascular data for training, which includes fundus vascular data, cerebral vascular projection data, and X-ray coronary angiography data. Each dataset is accompanied by corresponding segmentation annotations and category labels (indicating the source dataset). The extensive datasets allow the network to perceive a wide range of topological and structural variations, preserving both visual and structural cues of blood vessels in the feature space.

D. 3D Point Cloud Generation Based on MIP Image

In the field of medical image generation, the limited availability of 3D data often results in limited diversity and insufficient detail in the generated results. However, it is feasible to construct a comprehensive generic 2D vascular dataset by leveraging existing 2D vascular data along with multi-angle projections from 3D vascular data. The inherent vascular topologies and structural variations contained within these data can be effectively utilized to generate diverse MIP images. With strong 2D-to-3D consistency, the diversity inherent in 2D generation results can be effectively transferred to the synthetic 3D data, thus compensating for the lack of diversity in directly generating 3D images. Moreover, conditioning the 3D generation on 2D inputs simplifies the complexity of 3D vascular modeling, ensuring the anatomical plausibility of the generated details.

In most previous studies [15], [43], [44], the approach of extracting conditional image features is analogous to the method used when text serves as a condition. The global visual features of the input 2D image are directly extracted via a network and used as conditions for the denoising function (Eq. 3). It ensures the rationality and diversity of the generated images when there is a sufficiently large amount of training data. However, the generated 3D data often exhibit weak consistency with the input 2D images, limiting the diversity and details of the results when training data is scarce. Local structural features were used in the work of [36]; however, directly applying this approach to vascular structure generation presents challenges due to significant differences in the acquisition of 2D images and the potential inadequacy of the extracted features in capturing the unique structural characteristics of blood vessels.

To ensure the correspondence between the generated 3D data and the 2D MIP images, we extract both visual features of global rationality (v_{v}, v_{g}) and structural features of local details (s_{v}, s_{g}) from the input image $\tilde{I}_{\text{m}}$ in the DSFE module. The visual features ensure the global integrity and rationality of the generated image, while the structural features, consistent with the image size, ensure the correctness of the generated point cloud in local details and the alignment with the input 2D image. At the t-th step of the diffusion process, the 3D point cloud is denoted as X_t, which is derived from the input point cloud X_0 according to Eq. 2, with the added noise represented as ϵ_t. For each point p_i in X_t, the coordinate feature f_i is (h_i, w_i, d_i). To fully utilize the extracted image information, we extend the visual and structural features to f_i. The coordinates and local features are processed through the projection head (shown in Fig. 2) to obtain the corresponding features. Specifically, the coordinates of point p_i are first mapped to the original 3D image coordinate system $(h_i, w_i, d_i) \rightarrow (\tilde{h}_i, \tilde{w}_i, \tilde{d}_i)$, and the corresponding local features are extracted based on the projection coordinates $(\tilde{h}_i, \tilde{w}_i)$. The final features $f_i' \in \mathbb{R}^{4C+3}$ for each point are represented as:

$$f_i' = \text{Concat}(f_i, v_{\text{v}}, v_{\text{g}}, s_{\text{v}}(\tilde{h}_i, \tilde{w}_i), s_{\text{g}}(\tilde{h}_i, \tilde{w}_i)). \tag{10}$$

Thus, the denoising function μ_θ takes X_t along with the corresponding feature F' as input to predict the added noise:

$$\hat{\epsilon}_t = \mu_\theta(X_t, F'), \mathbb{R}^{N \times (4C+3)} \rightarrow \mathbb{R}^{N \times 3}, \tag{11}$$

followed by computing the loss with Eq. 5.

IV. Experiments

A. Experimental Setup

1) 3D Coronary Dataset.: We acquired 3D coronary artery data from three distinct datasets: 40 CTA volumes from

ASOCA [45], 72 CTA volumes from orCaScore [46], and 224 CTA volumes from Wuhan Union Hospital of China. ASOCA datasets include provided annotations, whereas the orCaScore and in-house datasets were semi-automatically annotated by a collaborating physician. For our study, we randomly selected 300 volumes for training and 36 for testing.

2) 3D Cerebral Dataset.: A total of 274 3D cerebrovascular volumes are assembled for further validation of our method, comprising 109 volumes sourced from the publicly available dataset TubeTK[2] and 165 volumes collected from Xuanwu Hospital of Capital Medical University, China. Two collaborating doctors annotated the Willis vessels for the in-house data. 250 data were randomly selected for training, while the remaining 24 data for testing.

3) Dataset for Pretraining.: During pre-training, we utilized vascular data from three categories: 1) fundus vascular data [47], [48]; 2) vascular projection images from 3D Cerebral Dataset; and 3) X-ray coronary data [16]. They comprised 2,046 images to train the dual-stream feature extraction network. Notably, projection images from the 3D Coronary Dataset were excluded from this analysis due to significant occlusions present in the coronary data when projected [49]. During pre-training, all images are resized to 336×336 pixels, and random noise is added. For classification tasks, category labels 0, 1, and 2 are assigned to represent fundus vascular, cerebrovascular, and X-ray coronary data, respectively.

4) 2D Vessel Dataset for MIP Image Generation.: Ensuring the diversity and adequacy of the training set is paramount for generating varied 2D MIP images. Notably, the training of MIP vascular image generation network requires segmented vascular datasets. Our training data are sourced from three datasets: 1) MIP images obtained by projecting 3D coronary volumes, augmented through random rotation within a specified 3D space range, resulting in 1,008 images; 2) coronary angiography data [16], contributing 1,156 2D coronary images; and 3) 274 cerebral vascular projection images. The inclusion of these data ensures that the training set encompasses a wide range of topological and structural variations, thereby positively contributing to the learning of MIP image vascular structures. During the training of 2D generative network, we initially pretrain it with all the data, followed by fine-tuning exclusively on specific MIP images. The pretrained model, trained on all the datasets, acquires robust representations of various vascular patterns, including coronary projections, coronary angiography, and cerebral vascular projections. To adapt the model for specific vessel generation task, we subsequently fine-tune it using targeted MIP images. This strategic approach maintains the model's capacity to generate diverse vascular structures while specializing its output for particular vascular types. For instance, in 3D coronary vessel generation tasks, we specifically fine-tune the model using 2D coronary MIP data, ensuring optimal performance for the target application.

[2] https://public.kitware.com/Wiki/TubeTK/Data

5) Training Strategy.: As outlined in Sec. III, our proposed framework composes two distinct stages. In the first stage, we focus on learning the generation of 2D MIP images. This is achieved through a comprehensive training approach where the 2D generative network is initially trained on a diverse collection of generic vascular data, followed by task-specific fine-tuning using targeted MIP images (tailored to different generation tasks). The second stage involves learning the transformation from 2D MIP images to 3D vascular structures. Initially, the DSFE module is pretrained using the dataset described in Sec. IV-A.3. Subsequently, we integrate the pretrained module with our conditional generation network to learn the mapping from 2D representations to their corresponding 3D vascular structures.

6) Implementation.: The network is implemented using PyTorch on four NVIDIA GeForce GTX 3090 GPUs. In the pre-training loss function, parameters α_1, α_2, and α_3 are set to 1.0, 0.2, and 0.2, respectively. The extracted feature dimension C is 32. The settings in stage-I of 2D MIP image generation are consistent with [11], taking $\approx$ 2 days for training. In the 2D-to-3D diffusion model, the learnable denoising function μ_θ is implemented using a Point-Voxel Network [39]. This architecture employs a dual-branch design: a point-based branch that processes point cloud and a voxel-based branch that hierarchically aggregates local point features. During the training phase, the point cloud consists of $N = 4096$ points, with the diffusion steps (T) of 1000. The size of input MIP images is 336×336 pixels. The AdamW optimizer [50] is employed, with $\beta = (0.95, 0.999)$ and the initial learning rate of 1×10^{-3}. The batch size is set to 64. The training process for the second stage requires approximately 22.5 hours, resulting in a cumulative training time $\approx$ 3 days when combined with the first stage.

TABLE I

COMPARISON WITH OTHER METHODS IN TERMS OF DISTRIBUTION STATISTICS (COSINE DISTANCE) AND SCORING INDICATORS ON CORONARY DATASET, WITH THE BEST PERFORMANCE HIGHLIGHTED IN BOLD. IN THE USER SCORE, THE AVERAGE RESULTS OF ALL RATINGS AND THE BEST 50% ARE CALCULATED RESPECTIVELY.

Category	Method	RCA			LAD			LCX			Ave	∠(Dis)	∠(Err)	DIV	User Score	
		L	R	T	L	R	T	L	R	T					All	50%
One-stage	PDM [12]	0.084	0.963	0.302	0.078	1.001	0.374	0.064	0.999	0.219	0.454	0.176	12.60	0.089	3.97	4.63
	PVD [38]	0.044	0.421	0.062	0.106	0.612	0.144	0.077	0.677	0.093	0.248	0.191	13.60	0.125	6.32	6.86
	TIGER [37]	0.047	0.542	0.148	0.092	0.638	0.070	**0.051**	0.792	0.091	0.275	0.223	14.06	0.113	6.71	7.08
	TrIND [26]	0.039	0.329	0.066	0.124	0.322	0.085	**0.051**	0.515	0.097	0.181	0.140	10.70	**0.160**	8.21	8.73
Two-stage	PSGN [43]	0.911	0.457	0.794	0.754	0.731	0.146	0.782	0.589	0.107	0.586	0.150	8.13	0.075	4.64	6.35
	Point-e [15]	0.052	0.125	0.117	0.095	0.216	0.136	0.089	0.206	0.122	0.129	0.142	8.56	0.122	7.13	7.78
	PC2 [36]	0.031	0.088	0.045	0.087	0.098	**0.052**	0.078	0.137	0.056	0.075	0.129	2.64	0.153	8.01	8.31
	Ours	**0.018**	**0.033**	**0.043**	**0.077**	**0.087**	0.081	0.073	**0.044**	**0.027**	**0.054**	**0.097**	**2.61**	0.152	**8.48**	**9.16**

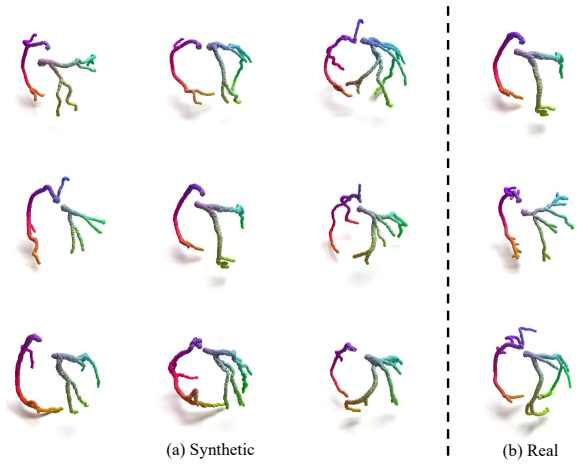


Fig. 4. Examples of coronary systems generated by our model and real data for training.

7) Baselines.: For more complete comparisons, we compare the proposed VesselDiffusion with the two categories of baseline models: (1) 3D point clouds generation directly from noise, including PDM [12], PVD [38], TIGER [37], VGGD [31], and TrIND [26]. Among these, VGGD and TrIND are specifically designed for vascular generation, while PDM, PVD, and TIGER are general point cloud generation methods. (2) models that generate 3D point clouds from 2D images using the same 2D generation method as ours, specifically models PSGN [43], Point-e [15], and PC2 [36]. All reproduced baseline methods are trained to convergence following the procedures outlined in their original publications and publicly available codes, with hyperparameter settings kept consistent with the reported implementations. Notably, the VGGD method requires annotation of vascular key points. Following the original study [31], we conduct comparative validation exclusively on cerebral dataset.

8) Evaluation for Generation Results.: Given the unique structure of blood vessels, certain evaluation metrics used for natural images are not suitable. Consistent with [5], we select three critical indicators of blood vessels for evaluation: length (L), average radius (R), and tortuosity (T) [51], [52]. These metrics are specifically applied to evaluate the three main coronary arteries: the right coronary artery (RCA), the left anterior descending artery (LAD) and the left circumflex artery (LCX), as well as the basilar artery and posterior cerebellar artery within the cerebrovascular context. For vascular analysis, we employ an automated approach based on prior anatomical rules to detect the endpoints, followed by manual refinement. The path length L measures the vascular path length between the identified endpoints, while the radius R is defined as the average vascular radius along this path. Tortuosity T, a widely adopted metric in vascular analysis, is computed as the ratio between the actual path length of a convoluted vessel and the linear distance between its endpoints [52]. To assess structural similarity, we examine the statistical distribution histograms of these metrics in both real and generated vessels (denoted as $\mathbf{p}, \mathbf{q}$) and compute the cosine distances between their distributions:

$$d_{\cos}(\mathbf{p}, \mathbf{q}) = 1 - \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \sqrt{\sum_{i=1}^{n} q_i^2}}. \tag{12}$$

Similarly, we compute the bifurcation angle between LAD and LCX arteries, following the definition in [53]. We statistically analyze the distribution of bifurcation angles in both generated and real data, quantifying their similarity using cosine distance ($\angle$(Dis)). Also, we report the average error ($\angle$(Err)) of the bifurcation angle between generated and real data (in degrees). To further assess the diversity of the generated 3D point clouds, we compute a diversity metric (DIV) by measuring pairwise distances within the generated data using the Chamfer Distance, calculating its average value, which is a metric commonly employed to evaluate statistical diversity [54]–[56]. Additionally, to assess the anatomical rationality and detailed accuracy of the synthetic vascular systems, a cooperating physician scores the generated results and the real data on a scale of 1 to 10, with higher scores indicating better results. To ensure fairness, we shuffle the generated data from various methods with the real data before scoring, and then calculate the average score for each method as well as the average score of the top 50% results. The overall and the top 50% average score on real coronary data are 9.98 and 10.0, respectively, underscoring the reliability of the scores.

9) Evaluation For 2D-3D Reconstruction Results.: To quantitatively evaluate the effectiveness of our proposed second-stage 2D-to-3D transformation, we measure the similarity between 3D point clouds generated from 2D MIP images and their corresponding real 3D point clouds in the test set. Specifically, given a generated point cloud $\hat{X}$ and its corresponding real data X, we compute the following metrics: a) Chamfer Distance (CD), b) Earth Mover's Distance (EMD), c) F1 score:

$$\text{Precision} = \frac{\left|\left\{\hat{x} \in \hat{X} \mid \min_{x \in X} \|\hat{x} - x\| < \tau\right\}\right|}{|\hat{X}|}, \tag{13}$$

$$\text{Recall} = \frac{\left|\left\{x \in X \mid \min_{\hat{x} \in \hat{X}} \|x - \hat{x}\| < \tau\right\}\right|}{|X|}, \tag{14}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{15}$$

where τ is the set threshold, which is 0.15 in our experiment (for normalized point cloud), and d) 95% Hausdorff Distance

TABLE II
COMPARISON WITH OTHER METHODS IN TERMS OF DISTRIBUTION STATISTICS AND SCORING INDICATORS ON CEREBRAL DATASET, WITH THE BEST PERFORMANCE HIGHLIGHTED IN BOLD.

Category	Method	Ave	DIV	User Score	
				All	50%
One-stage	PDM (2021) [12]	0.227	0.028	4.02	5.13
	PVD (2021) [38]	0.152	0.033	6.59	7.18
	TIGER (2024) [37]	0.096	0.034	6.89	7.31
	VGGD (2024) [31]	0.390	**0.067**	6.12	6.89
	TrIND (2024) [26]	0.157	0.047	8.36	9.04
Two-stage	PSGN (2017) [43]	0.428	0.015	4.33	6.05
	Point-e (2021) [15]	0.412	0.038	6.30	6.89
	PC2 (2023) [36]	0.138	0.044	8.33	8.79
	Ours	**0.053**	0.045	**8.64**	**9.32**

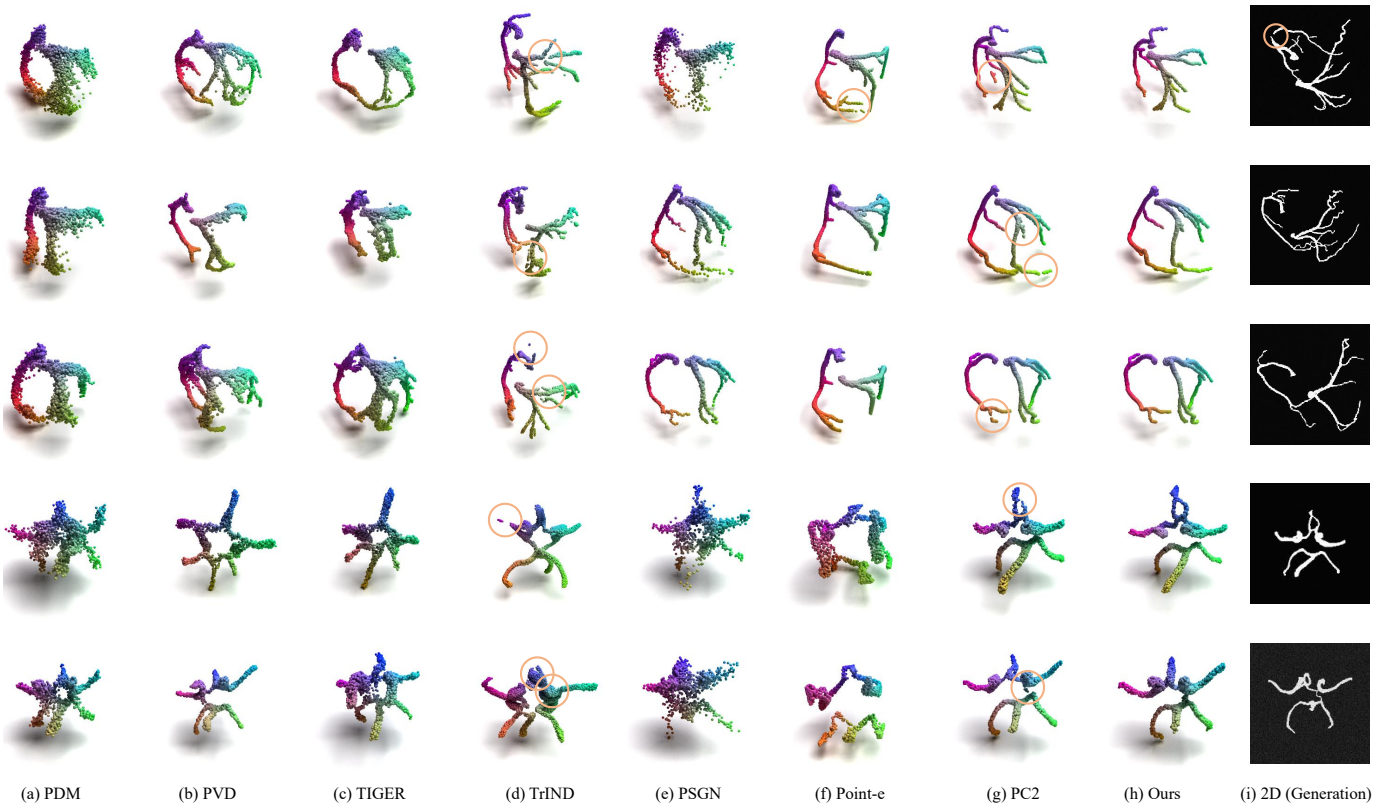
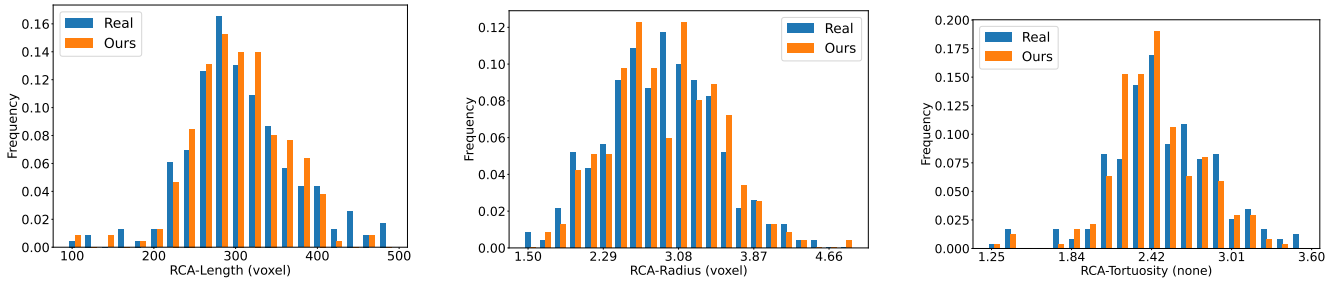


Fig. 5. Visual comparison of the results generated by different methods. For the two-stage methods (PSGN, Point-e, PC2, and Ours), the results obtained using the same 2D input ((i), generated in the first stage) are presented. The orange circles in the figure highlight certain unreasonable generated details.



(a) Cosine Distance = 0.018 (b) Cosine Distance = 0.033 (c) Cosine Distance = 0.043

Fig. 6. The statistical distribution histograms of three vascular properties of the generated and real data on RCA of coronary dataset. L and R are the voxel distances, and T is a dimensionless number. The cosine distances between the distributions are displayed.

(HD95):

$$d_{\mathrm{HD95}} = \mathrm{percentile}_{95}(d_X \cup d_{\hat{X}}), \tag{16}$$

where

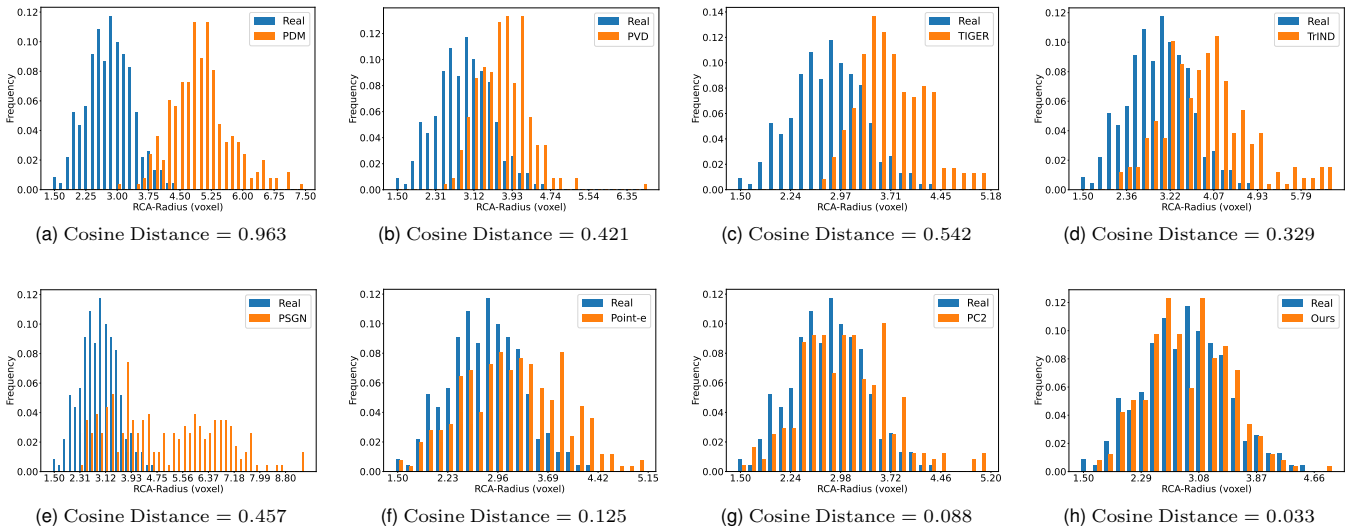
$$d_X = \left\{ \min_{\hat{x}\in\hat{X}} \|x - \hat{x}\| \mid x \in X \right\}, \tag{17}$$

$$d_{\hat{X}} = \left\{ \min_{x\in X} \|\hat{x} - x\| \mid \hat{x} \in \hat{X} \right\}. \tag{18}$$

B. Generation Result

1) Quantitative Comparisons: Table I and Table II present the quantitative indicators of results obtained by various methods applied to two datasets, detailing the distribution distance between the generated and real data, as well as the physician's scoring outcomes. In Table II, only the average distance performance is reported for the cerebrovascular data.

Overall, our method outperforms the compared methods, indicating its ability to generate diverse, plausible, and realistic data. VGGD exhibits greater diversity in generated results (Table II), likely attributable to the inherent randomness of discrete node generation. However, it performs significantly worse than our approach in terms of statistical similarity to the real distribution and expert evaluations. This limitation is expected, as VGGD models blood vessels using point and line representations, leading to substantial deviations from real data. Furthermore, the TrIND method employs neural networks to represent vascular structures. Due to the potential reconstruction errors and insufficient utilization of structural



(a) Cosine Distance = 0.963 (b) Cosine Distance = 0.421 (c) Cosine Distance = 0.542 (d) Cosine Distance = 0.329

(e) Cosine Distance = 0.457 (f) Cosine Distance = 0.125 (g) Cosine Distance = 0.088 (h) Cosine Distance = 0.033

Fig. 7. Distribution histograms of different methods on the average radius (R) of RCA.

information, it may lead to anatomical inconsistencies, such as discrete noise (potentially increasing the DIV metric) and vascular discontinuities, resulting in suboptimal performance in both distribution-based statistical evaluations and expert assessments. Notably, our approach achieves results that are close to real data in terms of both expert physician ratings and DIV (DIV results are 0.159 for coronary dataset and 0.050 for cerebral dataset in real data). The quantitative evidence in DIV, combined with the expert evaluation scores, strongly indicates that our generated results exhibit comparable levels of anatomical fidelity and structural diversity to real data.

In contrast to direct 3D point cloud generation methods (PDM, PVD, and TIGER), the two-stage methods based on the diffusion model (Point-e, PC2, and Ours) achieve results more closely aligned with the real distribution. This finding underscores that the complexity of vascular structures and the limited amount of training data render direct 3D generation methods less authentic and diverse. Furthermore, among the two-stage methods (employing the same 2D image generation approach), our proposed framework exhibits superior performance in both distribution distance and scoring, highlighting its advantages in 2D feature extraction and structural feature utilization.

2) Qualitative Comparisons: Fig. 4 presents the generated coronary systems of our method alongside several real data. Visually, our method produces blood vessel surface point clouds that closely resemble the real point clouds. To facilitate comparison with other methods, all data are displayed as point clouds. Note that the color of the point is used solely for visualization purposes.

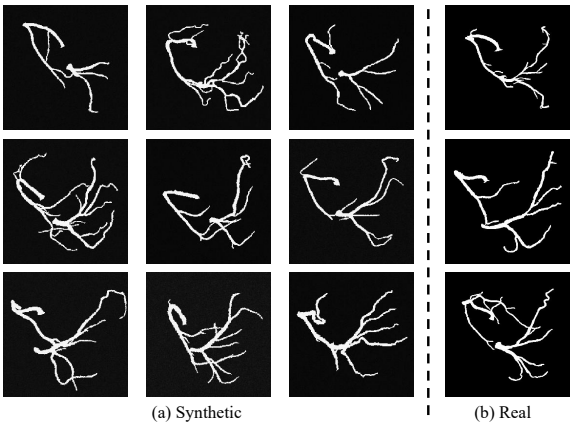
Fig. 5 presents a comparative analysis of the results obtained by our method against those from other methods. Consistent with the conclusions drawn from Table I and Table II, our method demonstrates superior performance in terms of both overall plausibility and details. Moreover, the 2D data produced in the initial phase inevitably exhibit certain inaccuracies, such as vascular discontinuities (as illustrated in the representative cases shown in the top-right corner of Fig. 5 and the top-left corner of Fig. 10). Our approach to generating 3D data maintains fidelity to the 2D inputs while effectively correcting the unreasonable aspects present in the 2D conditions, as evidenced by the corresponding 3D reconstruction results, which demonstrate anatomically plausible vascular networks with enhanced structural continuity. This indicates that our method ensures the 3D results remain consistent with the 2D input while capturing the overall structural rationality.

C. Distribution Histogram

Fig. 6 illustrates the distribution of our generated data compared to the real data on RCA. The high similarity between the histograms indicates that the generated blood vessels closely resemble the real ones. Fig. 7 presents the statistical distribution histograms of the average radius (R) on RCA across real data and the results obtained from various methods. In contrast, compared methods tend to overestimate the vessel radius compared to the ground truth. This discrepancy arises from the sensitivity of radius computation to vascular surface position deviations in generated results, with most biases resulting in overestimated values. Our method addresses this limitation through effective extraction and utilization of local structural information during the 2D-to-3D generation process. This enables better preservation of consistency between the 3D output and 2D input, resulting in minimal position deviations and accurate radius measurements that closely follow the distribution of real data. This finding underscores the capability of our method to accurately capture the statistical properties of the training distribution, ultimately leading to more realistic and diverse 3D vessel data.

D. Visualization of 2D MIP Generation Results

In our proposed method, the generation of diverse 3D vascular structures is predicated on producing 2D MIP images with rich topological and structural variations in the initial stage. Since the proposed VesselDiffusion preserves consistency between the resulting 3D data and the input 2D images



(a) Synthetic (b) Real

Fig. 8. The generated 2D coronary MIP images in the first stage.

during 2D-to-3D generation, the diversity observed in the 2D outputs is effectively transferred to the resulting 3D structures. Fig. 8 presents the coronary synthesis outcomes of the first stage alongside several real images. As demonstrated, the 2D generation model produces realistic and diverse MIP images, which are critical for preserving anatomical rationality and structural diversity in the subsequent 3D generation.

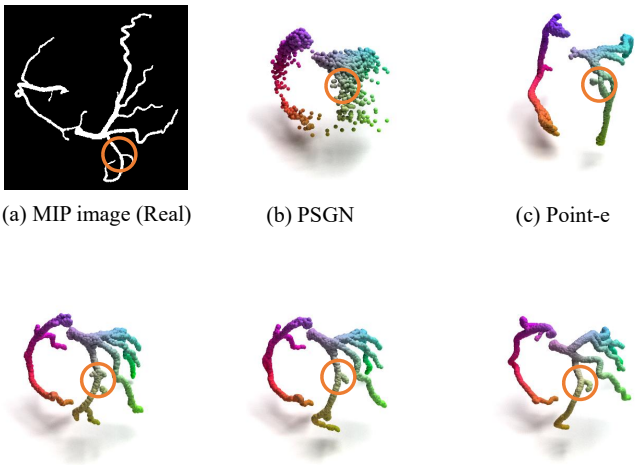
E. 2D-To-3D Comparison

TABLE III

QUANTITATIVE RESULTS BETWEEN 3D DATA GENERATED FROM 2D MIP IMAGES AND THE CORRESPONDING REAL POINT CLOUDS. THE METRICS OF HD95, CD, AND EMD ARE COMPUTED USING NORMALIZED COORDINATES WITHIN THE [-1,1] RANGE AND REPORTED IN NORMALIZED VOXEL UNITS.

Method	HD95(↓)	F1(↑)	CD(↓)	EMD(↓)
PSGN (2017)	0.365	0.635	0.296	1.093
Point-e (2021)	0.394	0.694	0.251	0.394
PC2 (2023)	0.348	0.705	0.256	0.319
Ours	**0.331**	**0.781**	**0.210**	**0.275**

To quantitatively assess the consistency between the 2D conditions and the results of 2D-to-3D generation methods, we compare the differences between the 2D MIP based generated 3D point clouds and their corresponding real 3D point clouds across test data on Coronary Dataset, as presented in Table III. Our method leverages a combination of ViT and GCN to more effectively extract vascular features, fully utilizing the synergy between visual and structural features during the generation process. This enables our method to produce results that are closer to the real 3D data and better maintain consistency between the generated 3D data and the 2D inputs. This advantage helps preserve the diversity of the 2D data when generating 3D structures. Fig. 9 illustrates a representative example of the generated results. Consistent with the findings discussed earlier, the results reaffirm that our method surpasses other two-stage approaches in terms of detail rationality and alignment with the 2D input.



(a) MIP image (Real) (b) PSGN (c) Point-e

(d) PC2 (e) Ours (f) Ground truth

Fig. 9. The input MIP image, results generated by different 2D-to-3D methods, and the corresponding real 3D data. The orange circles in the figure highlight the same location of generated results and ground truth.

F. Ablation Study

1) The Effects of Different Modules.: In Table IV, we assess the impact of various experimental configurations on the coronary generation outcomes. As anticipated, directly generating 3D vascular structures (one-stage) poses significant challenges, leading to poor performance in generating vascular details and insufficient diversity in results. When only global visual features are employed (visual only), there is a marked improvement in structural rationality and detailed accuracy (as reflected in user scores), indicating that the incorporation of 2D images alleviates the challenge of learning complex 3D vascular structures. However, the absence of structural features hinders the generated 3D results from maintaining consistency with the 2D input, leading to an increased distance between distributions. The subsequent inclusion of structural features leads to improvements across all performance indicators, particularly in the similarity between the distributions of the generated and real data. The integration of both visual and structural features thus ensures the rationality of details and the diversity of generated structures. Additionally, a comparison

TABLE IV

THE DISTRIBUTION DISTANCES BETWEEN THE GENERATED AND REAL DATA ON CORONARY DATASET UNDER DIFFERENT EXPERIMENTAL SETTINGS.

Method	Ablation			RCA			LAD			LCX			Ave	∠(Dis)	∠(Err)	DIV	User score	
	V	S	GCN	L	R	T	L	R	T	L	R	T					All	50%
one-stage	✗	✗	✗	0.065	0.064	0.040	0.169	0.125	0.136	0.115	0.121	0.092	0.103	0.182	14.48	0.137	7.64	8.33
visual only	✓	✗	✗	0.271	0.204	0.262	0.359	0.122	0.212	0.301	0.113	0.347	0.243	0.111	8.22	0.143	8.18	8.90
w/o GCN	✓	✓	✗	0.056	0.051	**0.032**	**0.068**	0.118	0.090	0.080	0.098	0.030	0.069	0.098	8.57	0.151	8.21	8.98
w/o pretrain	✓	✓	✓	0.045	0.055	0.052	0.151	0.088	**0.039**	0.140	0.091	0.036	0.077	0.110	2.73	0.139	8.14	8.67
Ours	✓	✓	✓	**0.018**	**0.033**	0.043	0.077	**0.087**	0.081	**0.073**	**0.044**	**0.027**	**0.054**	**0.097**	**2.61**	**0.152**	**8.48**	**9.16**

between the results obtained without GCN (w/o GCN) and final results (Ours) demonstrates that the combination of GCN and ViT allows for the extraction of richer and more comprehensive image features, thereby enhancing the quality of the generated results for the complex structure of vessels.

TABLE V

THE EVALUATION RESULTS BETWEEN THE GENERATED POINT CLOUD AND THE CORRESPONDING REAL ONE ACROSS THE TESTING DATA IN THE ABLATION EXPERIMENT. THE METRICS OF HD95, CD, AND EMD ARE COMPUTED USING NORMALIZED COORDINATES WITHIN THE [-1,1] RANGE AND REPORTED IN NORMALIZED VOXEL UNITS.

Method	HD95(↓)	F1(↑)	CD(↓)	EMD(↓)
visual only	0.865	0.239	0.683	0.619
w/o GCN	0.334	0.771	0.221	0.282
w/o pretrain	0.335	0.760	0.223	**0.264**
Ours	**0.331**	**0.781**	**0.210**	0.275

Furthermore, we evaluate the impact of pre-training on the performance of our results. Pre-training the DSFE module on a large dataset allows the network to effectively capture the vascular structure features of interest, thereby improving the fidelity of the generated results. These observations are further supported by the findings in Fig. 10. Table V provides the evaluation results comparing the outcomes generated under different experimental settings with the corresponding real data. The results further substantiate and reinforce the conclusions drawn from Table IV and Fig. 10.

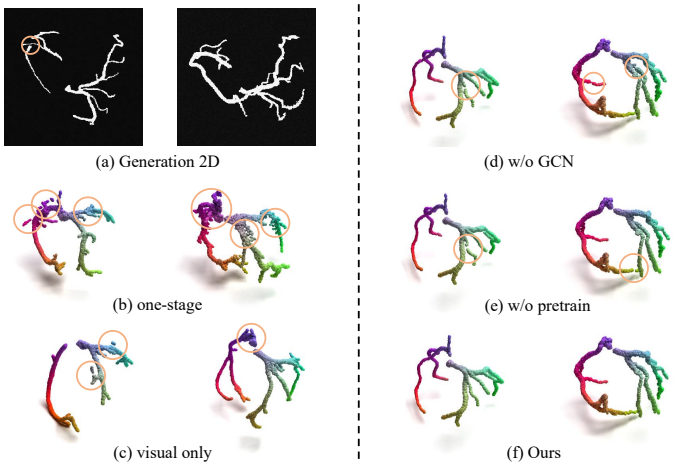


Fig. 10. Two examples of ablation study on Coronary Dataset.

2) The Pre-Training Strategy of DSFE Module.: We investigate the impact of various pre-training strategies for the DSFE module (on coronary dataset). Our comparative analysis encompasses two distinct approaches: (1) training exclusively on coronary projection data (Only MIP), and (2) implementing self-supervised learning (Self-Sup) through a Masked AutoEncoder framework [57] (Table VI). The removal of coronary X-ray angiography and fundus vascular data (Only MIP) leads to a decline in both the diversity and statistical similarity of the generated data. The performance gap occurs because incorporating additional datasets enhances vascular diversity and structural variation during training, enabling the DSFE module to more effectively capture visual and morphological characteristics across different vessel types. When trained with self-supervision, the overall performance shows a slight decrease (compared to supervised model) due to the absence of categorical and vessel-specific label information and the removal of GCN, which aligns with expectations. Nonetheless, self-supervised training offers greater adaptability for extending our method to other datasets, mitigating the need for manual 2D annotations and facilitating the incorporation of larger-scale data into the pre-training process.

TABLE VI

ABLATION EXPERIMENT ON PRE-TRAINING STRATEGY OF DSFE MODULE, WITH THE BEST PERFORMANCE HIGHLIGHTED IN BOLD.

Method	Ave	∠(Dis)	∠(Err)	DIV	User Score	
					All	50%
w/o pretrain	0.077	0.110	2.73	0.139	8.14	8.67
Only MIP	0.071	**0.095**	2.64	0.139	8.36	9.09
Self-Sup	0.065	0.101	**2.52**	**0.156**	8.29	9.02
Full-Sup	**0.054**	0.097	2.61	0.152	**8.48**	**9.16**

3) The Number of Points.: In our framework, we employ farthest point sampling to extract a fixed number (N) of representative points from the vascular surface point cloud, with interpolation implemented for cases existing fewer points. To investigate the effect of different point cloud sizes on performance, we conduct an analysis summarized in Table VII. When using a relatively small number of points ($N = 2048$), the representation of the vascular network exhibits more geometric approximation errors, leading to a decline in generation performance. While denser point clouds intuitively provide more accurate reconstructions, they also impose greater storage and computational demands. Therefore, selecting an appropriate point cloud resolution is essential to achieving an optimal balance between resource efficiency and generation quality.

TABLE VII

ABLATION EXPERIMENT ON THE NUMBER POINTS OF 3D VESSELS.

Method	Ave	∠(Dis)	∠(Err)	DIV	User Score	
					All	50%
$N = 2048$	0.107	0.122	2.59	0.150	7.96	8.83
$N = 4096$	**0.054**	0.097	2.61	0.152	8.48	9.16
$N = 6144$	0.056	**0.088**	**2.43**	**0.157**	**8.50**	**9.28**

G. Voxel Visualization and Point2Voxel

In addition to the point cloud visualization of the vascular system, we provide voxel visualization results on Coronary Dataset (Fig. 11). These results are derived from the generated 3D point clouds. The process involves several key steps: (1) transforming the generated point cloud coordinates into the original 3D coordinate system; (2) filling each point with a specified radius and refining to obtain the centerline structure; (3) applying the minimum spanning tree algorithm to extract all vascular paths; (4) for each point along the path, identifying the corresponding point on the 2D MIP image to calculate the radius; and (5) smoothing the radii along the centerline, and reconstructing the blood vessels based on the centerline and associated radii.

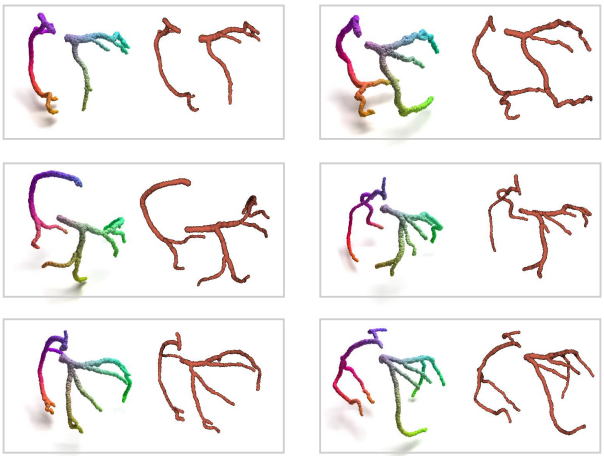


Fig. 11. Synthetic point clouds and corresponding voxel visualization.

V. Limitation and Future Works

A. Generation Result

A limitation of our approach lies in its reduced applicability to densely structured vascular networks, such as cerebrovascular systems, where excessive vessel crossings and overlaps in 2D MIP images introduce challenges. Consequently, our validation is restricted to the vessels of the Circle of Willis on Cerebral Dataset. While our method demonstrates improved capability in generating detailed and realistic vascular structures compared to existing approaches [3]–[5], [26], [31], the vascular structures synthesized by VesselDiffusion are currently limited to specific anatomical regions and scales. Extending its applicability to larger and more complex vascular networks remains an important avenue for future research.

B. Framework Design

Our method follows a two-stage process, meaning the quality of the generated 2D images directly influences the fidelity of the final 3D vascular reconstructions. The first stage requires segmented vascular data to train the MIP image generation. Although incorporating a more extensive set of segmented 2D vascular images could enhance generation performance, the manual segmentation process incurs additional labor costs, imposing constraints on the model's scalability. While our proposed VesselDiffusion (trained on 3D point clouds and MIP images) mitigates some of the deficiencies present in the generated 2D images (as outlined in Sec. IV-B.2), this dependency nonetheless constrains the broader utility of the method. Exploring techniques to reconstruct accurate 3D vascular structures from real 2D X-ray images remains a compelling direction for future research.

VI. Conclusion

In this study, we introduce VesselDiffusion, a novel method for generating 3D vascular systems based on diffusion model. Unlike natural image generation, which benefits from vast datasets comprising thousands to millions of images, vascular synthesis is constrained by the relatively limited availability of 3D training data. To address this challenge and effectively capture the complexity and diversity of vascular structures, we propose a two-stage generation framework. In the first stage, we utilize a large corpus of available generic 2D vascular data to generate a variety of MIP images. Conditioned on these 2D images, a 3D point cloud generation network is employed, where the MIP image guides the formation of vascular structures, thereby reducing the complexity of 3D generation. The 3D generation model combines ViT and GCN to extract comprehensive 2D image features, integrating both visual features of global rationality and structural features of local vascular details within the denoising function of the diffusion model to ensure the consistency of the generated 3D structures with the 2D images. Our proposed method produces 3D vascular systems characterized by rich diversity and anatomical accuracy. Extensive experiments on different datasets demonstrate the effectiveness of our approach.

References

[1] M. Piccinelli, A. Veneziani, D.A. Steinman, A. Remuzzi, and L. Antiga, "A framework for geometric analysis of vascular structures: application to cerebral aneurysms," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1141–1155, 2009.

[2] J. Wu, Q. Hu, and X. Ma, "Comparative study of surface modeling methods for vascular structures," *Comput. Med. Imag. Graph.*, vol. 37, no. 1, pp. 4–14, 2013.

[3] G. Hamarneh and P. Jassi, "Vascusynth: Simulating vascular trees for generating volumetric image data with ground-truth segmentation and tree analysis," *Comput. Med. Imag. Graph.*, vol. 34, no. 8, pp. 605–616, 2010.

[4] J.M. Wolterink, T. Leiner, and I. Isgum, "Blood vessel geometry synthesis using generative adversarial networks," *arXiv preprint arXiv:1804.04381*, 2018.

[5] P. Feldman, M. Fainstein, V. Siless, C. Delrieux, and E. Iarussi, "Vessel-vae: Recursive variational autoencoders for 3D blood vessel synthesis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2023, pp. 67–76.

[6] M.A. Galarreta-Valverde, M.M. Macedo, C. Mekkaoui, and M.P. Jackowski, "Three-dimensional synthetic blood vessel generation using stochastic l-systems," in *Med. Imag. 2013: Imag. Process.*, 2013, pp. 414–419.

[7] N. Rauch and M. Harders, "Interactive synthesis of 3D geometries of blood vessels.," in *Eurographics*, 2021, pp. 13–16.

[8] M. Schneider, J. Reichold, B. Weber, G. Székely, and S. Hirsch, "Tissue metabolism driven arterial tree generation," *Med. Image Anal.*, vol. 16, no. 7, pp. 1397–1414, 2012.

[9] G.D.M. Talou, S. Safaei, P.J. Hunter, and P.J. Blanco, "Adaptive constrained constructive optimisation for complex vascularisation processes," *Sci. Rep.*, vol. 11, no. 1, pp. 6180, 2021.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

[11] A.Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.

[12] S. Luo and W. Hu, "Diffusion probabilistic models for 3D point cloud generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2837–2845.

[13] H. Yuan *et al.*, "Instructvideo: Instructing video diffusion models with human feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 6463–6474.

[14] Z. Wu, Y. Wang, M. Feng, H. Xie, and A. Mian, "Sketch and text guided diffusion model for colored point cloud generation," in *Int. Conf. Comput. Vis.*, 2023, pp. 8929–8939.

[15] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3D point clouds from complex prompts," *arXiv preprint arXiv:2212.08751*, 2022.

[16] Y. Ma *et al.*, "Self-supervised vessel segmentation via adversarial learning," in *Int. Conf. Comput. Vis.*, 2021, pp. 7536–7545.

[17] C. Han *et al.*, "Gan-based synthetic brain MR image generation," in Proc. IEEE 15th Int. Symp. Biomed. Imag., IEEE, 2018, pp. 734–738.

[18] N. Konz, Y. Chen, H. Dong, and M.A. Mazurowski, "Anatomically-controllable medical image generation with segmentation-guided diffusion models," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Springer, 2024, pp. 88–98.

[19] M. Özbey *et al.*, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3524–3539, 2023.

[20] P. Subramaniam *et al.*, "Generating 3D TOF-MRA volumes and segmentation labels using generative adversarial networks," *Med. Image Anal.*, vol. 78, pp. 102396, 2022.

[21] K. Han *et al.*, "Medgen3D: A deep generative framework for paired 3D image and mask generation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Springer, 2023, pp. 759–769.

[22] A. Merrem, S. Bartzsch, J. Laissue, and U. Oelfke, "Computational modelling of the cerebral cortical microvasculature: effect of X-ray microbeams versus broad beam irradiation," *Phys. Med. Biol.*, vol. 62, no. 10, pp. 3902, 2017.

[23] S. Di Gregorio *et al.*, "A computational model applied to myocardial perfusion in the human heart: from large coronaries to microvasculature," *J. Comput. Phys.*, vol. 424, pp. 109836, 2021.

[24] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[25] D.P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[26] A. Sinha and G. Hamarneh, "TrIND: Representing anatomical trees by denoising diffusion of implicit neural fields," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Springer, 2024, pp. 344–354.

[27] J. Jo, S. Lee, and S.J. Hwang, "Score-based generative modeling of graphs via the system of stochastic differential equations," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 10362–10383.

[28] C. Vignac *et al.*, "Digress: Discrete denoising diffusion for graph generation," *arXiv preprint arXiv:2209.14734*, 2022.

[29] X. Peng, J. Guan, Q. Liu and J. Ma, "Moldiff: Addressing the atom-bond inconsistency problem in 3D molecule diffusion generation," *arXiv preprint arXiv:2305.07508*, 2023.

[30] K. Yi, B. Zhou, Y. Shen, P. Liò and Y. Wang, "Graph denoising diffusion for inverse protein folding," *Adv. Neural Inform. Process. Syst.*, vol. 36, pp. 10238–10257, 2023.

[31] C. Prabhakar *et al.*, "3D vessel graph generation using denoising diffusion," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Springer, 2024, pp. 3–13.

[32] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[33] A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[34] Z. Yu, H. Li, F. Fu, X. Miao, and B. Cui, "Accelerating text-to-image editing via cache-enabled sparse diffusion inference," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 16605–16613.

[35] D. Danier, F. Zhang, and D. Bull, "Ldmvfi: Video frame interpolation with latent diffusion models," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 1472–1480.

[36] L. Melas-Kyriazi, C. Rupprecht, and A. Vedaldi, "Pc2: Projection-conditioned point cloud diffusion for single-image 3D reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12923–12932.

[37] Z. Ren, M. Kim, F. Liu, and X. Liu, "Tiger: Time-varying denoising model for 3D point cloud generation with diffusion process," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9462–9471.

[38] L. Zhou, Y. Du, and J. Wu, "3D shape generation and completion through point-voxel diffusion," in *Int. Conf. Comput. Vis.*, 2021, pp. 5826–5835.

[39] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel cnn for efficient 3D deep learning," *Adv. Neural Inform. Process. Syst.*, vol. 32, pp. 963–973, 2019.

[40] R. Li *et al.*, "3D graph-connectivity constrained network for hepatic vessel segmentation," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 3, pp. 1251–1262, 2021.

[41] G. Zhao *et al.*, "Graph convolution based cross-network multiscale feature fusion for deep vessel segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 183–195, 2022.

[42] S. Shin, S. Lee, I. Yun, and K. Lee, "Deep vessel segmentation by learning graphical connectivity," *Med. Image Anal.*, vol. 58, pp. 101556, 2019.

[43] H. Fan, H. Su, and L.J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 605–613.

[44] H. Xie, H. Yao, X. Sun, S. Zhou, and S. Zhang, "Pix2vox: Context-aware 3D reconstruction from single and multi-view images," in *Int. Conf. Comput. Vis.*, 2019, pp. 2690–2698.

[45] R. Gharleghi *et al.*, "Automated segmentation of normal and diseased coronary arteries–the asoca challenge," *Comput. Med. Imag. Graph.*, vol. 97, pp. 102049, 2022.

[46] J.M. Wolterink *et al.*, "An evaluation of automatic coronary artery calcium scoring methods with cardiac ct using the orcascore framework," *Med. Phys.*, vol. 43, no. 5, pp. 2361–2373, 2016.

[47] J. Staal, M.D. Abràmoff, M. Niemeijer, M.A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, 2004.

[48] C.G. Owen *et al.*, "Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program," *Invest. Ophthalmol. Vis. Sci.*, vol. 50, no. 5, pp. 2004–2010, 2009.

[49] Z. Guo, Z. Tan, J. Feng, and J. Zhou, "3D vascular segmentation supervised by 2D annotation of maximum intensity projection," *IEEE Trans. Med. Imag.*, vol. 43, no. 6, pp. 2241–2253, 2024.

[50] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[51] S. Lang *et al.*, "Three-dimensional quantification of capillary networks in healthy and cancerous tissues of two mice," *Microvasc. Res.*, vol. 84, no. 3, pp. 314–322, 2012.

[52] E. Bullitt, G. Gerig, S.M. Pizer, W. Lin, and S.R. Aylward, "Measuring tortuosity of the intracerebral vasculature from MRA images," *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1163–1171, 2003.

[53] P. Medrano-Gracia *et al.*, "A computational atlas of normal coronary artery anatomy," in *EuroIntervention*, vol. 12, no. 7, pp. 845–854, 2016.

[54] M. Mironov and L. Prokhorenkova, "Measuring diversity: Axioms and challenges," *arXiv preprint arXiv:2410.14556*, 2014.

[55] F. Velikonivtsev, M. Mironov, and L. Prokhorenkova, "Challenges of generating structurally diverse graphs," *Adv. Neural Inform. Process. Syst.*, 2024.

[56] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22500–22510.

[57] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.