DG-3DGS: Dynamic Growing 3D Gaussian Splatting for Sequential Scene Reconstruction of Monocular Endoscope Video

Ziang Zhang[a] , Hong Song[b,*], Jingfan Fan[c,*], Long Shao[b], Tianyu Fu[a], Danni Ai[c], Deqaing Xiao[c], Yuanyuan Wang[c], Yucong Lin[c], and Jian Yang[c,*]

***Abstract*—Scene reconstruction of monocular endoscope video is essential for the enhancement of Surgical Endoscope Image Analysis and Application. However, restricted by the narrow space of endoscopic movement and the obstruction of vision within cavities, it's difficult for most conventional methods to conduct high-quality reconstruction. To overcome these challenges, a novel dynamic growing 3D gaussian splatting architecture is proposed to construct the 3D model of endoscopic scene without pre-compute camera poses or Structure from Motion. Firstly, to establish spatial feature associations between interframe, a 2D-3D displacement field is designed based on dense feature matches and depth prediction. On this basis, a novel displacement field variational optimization is developed to acquire relative poses by minimizing the energy functional of field transformation. Secondly, to address the constraint of endoscopic view, by gaussian dynamic transformation and differential gradient field optimization, a novel dynamic gaussian growing strategy is proposed to sequentially grow the local gaussian model. Finally, a novel Forward-Reconstruction&Backward-Optimization architecture is proposed to generate the global gaussian model. The evaluation is conducted on two public endoscopic datasets: Scared and C3VD. The experimental results show the proposed method outperforms state-of-the-art methods in quantitative (PSRN, SSIM and LIPIS) and qualitative comparisons. The project page is https://iheckzza.github.io/DG-3DGS/.**

***Index Terms*—Monocular Endoscopic Video, 3D Gaussian Splatting, Dynamic Gaussian Growing, Displacement Field Optimization, Scene Reconstruction**

I. Introduction

RECONSTRUCTING surgical scenes from endoscopic video stream is essential for Robotic Assisted Minimally Invasive Surgery [1] by recovering the 3D model of the observed anatomy. This technique can assist in simulating the surgical environment for pre-operative planning and Augmented-Reality/Virtual-Reality surgical navigation [2].

With the emergence of Neural Radiance Field (NeRF) [3] and Neural Implicit Surfaces (NeuS) [4], The field of scene reconstruction and view synthesis has been largely advanced during recent years [5]. Moreover, recently more efforts have been focused on representing the surgical scene as the radiance field [6]–[8], which can learn continuous functions that implicitly represent the 3D scene trained from 2D images and paired camera poses. Neural-Field-Based methods exploit deep neural networks to implicit model complex geometry and appearance, and outperforms Discrete-Representation-Based methods [9]. While Neural-Field-Based methods have achieved reasonably good results, during the rendering process for each image, an enormous number of points and rays will be queried repeatedly from the radiance fields. This method requires significantly constrains the rendering speed and poses substantial obstacles for practical intraoperative applications [10].

Aiming to address these limitations, the 3D Gaussian Splatting (3DGS) [11] approach extends the volumetric rendering in NeRF to accommodate high-quality point clouds. Recently researches have shifted their attentions to 3DGS in the field of computer and surgical vision [12]–[14]. The 3DGS can describe the scene as anisotropic representation and render images with efficient Tile-Based rasterizer, allowing real-time rendering with superior reconstruction quality [15]. Nonetheless, adopting 3DGS for surgical scenes still faces challenges. Generally, the endoscopic scenes are more complex and it's cumbersome to acquire motion through tracking devices. What's more, the endoscope motion is usually constrained by the instrument characteristics or surgeon's capacity, which makes it difficult for endoscope to observe the full view of cavity [16] [17]. The typical 3DGS mainly relies on the Structure-from-Motion (SfM) [18] algorithm like Colmap [19] to initialize the gaussian position. It is a time-consuming pipeline with multiple stages and requires stringent requirements for the motion trajectory of endoscope and the range of view field. Furthermore, it is challenging for surgeons to acquire panoramic views by maneuvering the handheld endoscope around the patient's organs [20] [21]. As a result, the view field of endoscope is typically confined to a limited range of motion, making it difficult to recover sparse point clouds and estimate absolute camera poses with SfM pipeline [13].

To perform gaussian model reconstruction without camera pose prior or SfM preprocessing, in this paper, a novel dynamic growing 3D gaussian splatting (DG-3DGS) architecture is established by sequentially "growing" one frame at a time as endoscope moves to process scene reconstruction. In this architecture, there are two-step Interframe Pose Estimating and Dynamic Gaussian Growing procedures during each growing stage: (1) To address the challenge of estimating tiny relative

Corresponding Authors: Hong Song (songhong@bit.edu.cn), Jingfan Fan (fjf@bit.edu.cn) and Jian Yang (jyang@bit.edu.cn)

[a] The School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

[b] The School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China

[c] The School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China

poses between neighboring frames, a novel sequential 2D-3D displacement field variational optimization is designed to achieve continuous tiny relative camera pose. By utilizing the dense key-points matching from Feature Matching network (FM-Net) and relative depth from Dense Prediction Transformer network (DPT-Net), the 2D-3D displacement field is established with 3D coordinates from former frame and 2D corresponding locations from next frame. The optimization goal is to learn the relative pose transformation by minimising the gaussian rendering variational gradients of displacement field. (2) To conduct gaussian reconstruction within constrained endoscopic view, by gaussian dynamic transforming with relative poses, a sequential dynamic gaussian growing module is developed to conduct gaussian growth progressively. During each rendering iteration, the whole gaussian model is transformed to current camera coordinate and the constrained area is fitted with new gaussian ellipsoids, and a novel differential gradient field optimization is proposed to refine the local gaussian model coordinates under the condition of lacking multi-view geometric constraints. Freeing from reliance on SfM preprocessing or other prior information, based on the relative poses between neighboring frames and the gaussian growing module, a novel reconstruction architecture with Forward-Reconstruction & Backward-Optimization strategy is proposed to train and generate the new gaussian ellipsoids of new viewpoint frame-by-frame. Therefore, it's possible for the proposed architecture to receive unlimited frames continuously and perform gaussian rastering consistently.

In brief, the main contributions are as follows:

• To perform accurate tiny relative pose estimation between interframe, a novel Interframe Pose Estimation (IPE) module is designed to learn the transform association from the proposed 2D-3D displacement field, by optimizing the variational gradients of field transformation.

• By gaussian dynamic transformation and differential gradient field optimization, a novel Dynamic Gaussian Growing (DGG) module is proposed to perform new gaussian synthesis and refine the gaussian coordinates within constrained endoscopic view.

• Based on the proposed IPE and DGG modules, a novel Forward-Reconstruction & Backward-Optimization 3D gaussian splatting architecture is proposed to conduct high-quality monocular endoscopic scene reconstruction without SfM preprocessing.

II. RELATED WORKS

Recently, to overcome the shortcomings of various preprocessing priors in the field of 3DGS-Based reconstructions, great interests have been increasing for researchers to estimate accurate camera poses or modify SfM pipeline with alternative methods.

Pose Estimation in Scene Reconstruction: Conventional pose estimation methods in SfM typically rely on feature extracting and matching (such as SIFT [22], SURF [23], or ORB [24]), which leverage epipolar geometry to estimate the essential matrix and decompose it to obtain the relative rotation and translation between interframes. NeRF-Based methods usually require initializing the camera poses to achieve photo-realistic scene reconstruction [25]. Recently several studies [26]–[28] have focused on minimizing the dependence on SfM by integrating pose estimation directly within the NeRF or 3DGS framework. Rosinol *et al*. [29] leveraged recent advances in dense monocular Simultaneous Localization and Mapping (SLAM) and proposed a novel 3D mapping pipeline for accurate pose estimation and scene reconstruction from monocular images taken casually. Sucar *et al*. [30] utilized a multi-layer perceptron as the only scene representation in a real-time SLAM system with handheld RGB-D cameras, and achieved real-time system via continual training of a neural network against live image streams. Chen *et al*. [31] presented a novel framework called iNeRF that performs mesh-free pose estimation by inverting radiance field, showing that poses for novel view images can be estimated with a reconstructed NeRF model. Wang *et al*. [32] proposed an efficient RGB-D SLAM method called EndoGSLAM, leveraging differentiable rasterization to enable Gradient-Based optimizing the camera poses in each incoming frame. Chng *et al*. [33] presented a new positional embedding-free neural radiance field architecture called Gaussian Activated neural Radiance Fields (GARF), which outperforms the current state-of-the-art in terms of high-fidelity reconstruction and pose estimation. However, limitation still exists for accurate tiny sequential pose estimating, most of the current methods lack effective utilization of interframe feature association between neighboring frames. Therefore, to overcome this limitation, a novel relative pose estimation module is designed by leveraging feature correspondence between interframes effectively to estimate and optimize the tiny pose transform matrix.

3DGS without SfM Prepossessing: Recently, great interest has been growing in eliminating the requirements of preprocessing steps for NeRF-Based or 3DGS-Based methods. Several approaches [26], [28], [31], [33]–[37] are mainly focusing on incorporating undistorted monocular or binocular depth priors to progressively train and optimize the local radiance fields. Huang *et al*. [12] exploited a depth predict network to generate pseudo depth maps as prior and presented Endo-4DGS with lightweight MLPs to capture temporal dynamics with gaussian deformation fields. Bonilla *et al*. [20] proposed Gaussian Pancakes that synergizes a geometrically-regularized 3D gaussian splatting with the SLAM system [38] to enhance texture rendering and anatomical accuracy while boosting training speeds. By integrating the efficient gaussian representation and highly-optimized rendering engine, Liu *et al*. [13] proposed a real-time endoscopic scene reconstruction framework named EndoGaussian, which is built on 3DGS and significantly boosts the rendering speed to a real-time level. Batlle *et al*. [6] exploited the relation between pixel brightness and depth map as the supervised information, they modified the NeuS [4] architecture to explicitly account for it and introduce a calibrated photometric model of the endoscope and light source. Zha *et al*. [1] modeled surface dynamics, shape, and texture with three neural fields to propose a novel Neural-Field-Based method called EndoSurf, which can effectively learn to represent a deforming surface from RGBD videos. Nonetheless, recent works are mainly focusing on promoting

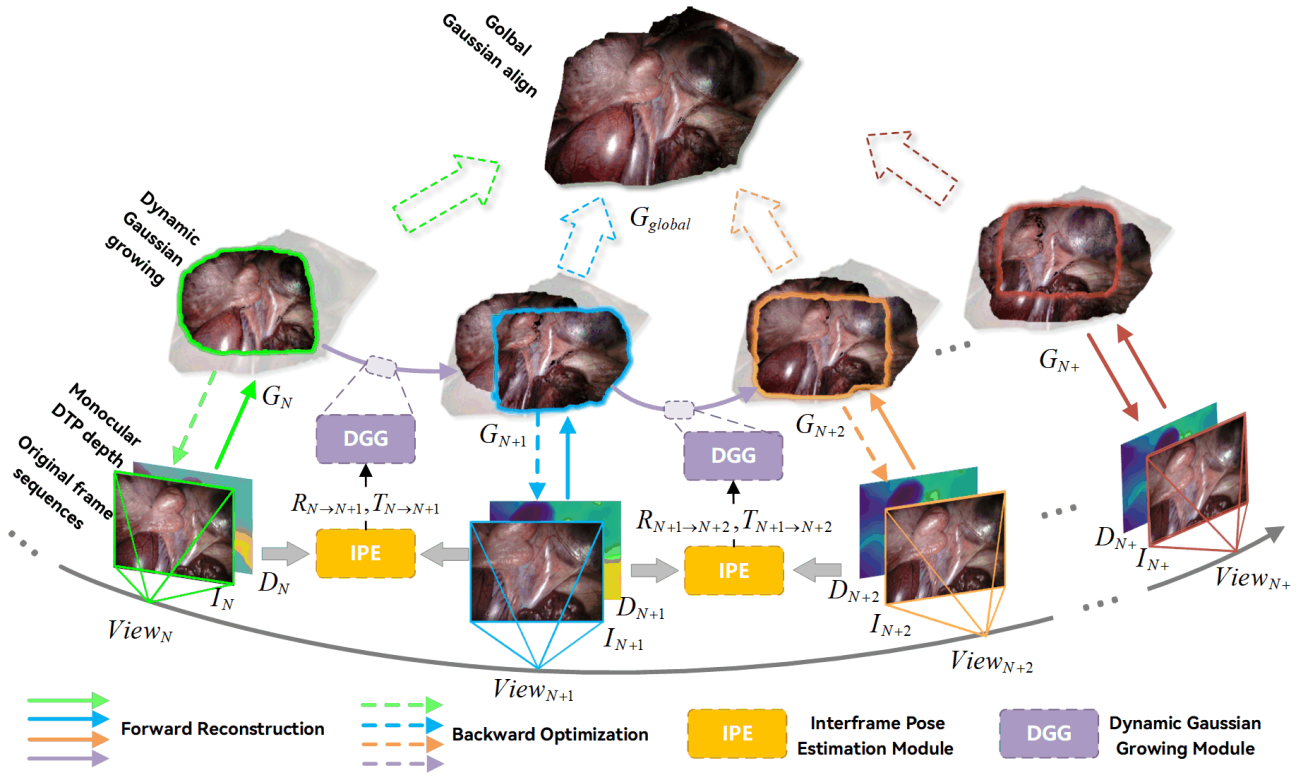

Fig. 1. The architecture of DG-3DGS From $View_N$ to $View_{N+1}$, $View_{N+1}$ to $View_{N+2}$, and to $View_{N+}$. The gaussian evolution path is presented from G_N to G_{N+1}, G_{N+1} to G_{N+2}, and finally to G_{N+}.

the gaussian performance and optimizing render process with depth prior information of discrete frames, which lacks the geometric feature between consecutive sequence frames within the limited field of views. Hence, to maintain the consistency of the geometric perspective, by fully utilizing the continuous features of the sequence interframes and depth prior, a dynamic sequential gaussian growing strategy is proposed to gradually expand the gaussian model with interframe features and depth prior information, and the optimization goal is to refine the coordinates within limited geometry constraint to realize high-quality endoscopic scene reconstruction.

III. METHODS

The overall architecture of the proposed method is shown in Fig. 1. The whole process of gaussian reconstruction is based on the premise of sequential monocular endoscopic video. The proposed DG-3DGS mainly consists of two modules during each training stage: the IPE module and DGG module. In the beginning of proceed, the first frame with predicted depth map is considered as the gaussian initialization step, and the *XYZ* location of gaussian model is calculated from the depth map with intrinsic matrix. When the next frame flows in (e.g. from $View_N$ to $View_{N+1}$), the relative transform pose matrix (e.g. $R_{N\to N+1}$,$T_{N\to N+1}$) is estimated and optimized with the IPE module. By setting $R_{N\to N+1}$,$T_{N\to N+1}$ as the transformation matrix, the gaussian model is transformed from the local cameras coordinate system (*LCCS*) of $View_N$ to that of $View_{N+1}$, the DGG module is proceeded with current frame I_{N+1} and predicted depth map D_{N+1} to grow the local gaussian set appearing in new region under current viewpoint from gaussian model G_N to G_{N+1}. With sequential frames flow to the end, the G_N can finally evolve to G_{golbal} and finish scene reconstruction.

A. *Interframe Pose Estimation Module*

Given interframes $View_N$ and $View_{N+1}$, the main target is aiming to estimate the relative pose transformation $R_{N\to N+1}$,$T_{N\to N+1}$ between right-handed camera coordinates system. Therefore, an Interframe Pose Estimation module is proposed and the main processing of the module is shown in Fig. 2.

An end-to-end pretrained monocular deep predict network DPT[35] is used to estimate the local depth map D_N, and the corresponding point cloud Pt_N is converted on local camera coordinate system (*LCCS*) with intrinsic matrix K_N and distortion parameters $Dist_N$:

$$Pt_N = D_N * K_N^{-1} \otimes F_u(Dist_N, Girdmap_(h, w, 1)) \quad (1)$$

Where $Girdmap_(h, w, 1)$ represents 3D normalized plane with *h* height and *w* weight, F_u represents distortion correction function, and $Pt_N \mid \{x_n, y_n, z_n \in R^{n\times 2}\}$.

In order to obtain robust correlated feature information for local camera pose transformation, a monocular endoscopic feature matching network [40] is leveraged to fulfill dense key-points pairs matching between interframes. The matching distribution is located in a $(h/8, w/8)$ sized gird-map, in each grid a maximum matching score key-points pairs is extracted and be formulated as $kp_N \mid \{u_n, v_n \in \mathbb{R}^{n\times 2}\}$ in $View_N$ and corresponding $kp_{N+1} \mid \{u_{n+1}, v_{n+1} \in \mathbb{R}^{n\times 2}\}$ in $View_{N+1}$.

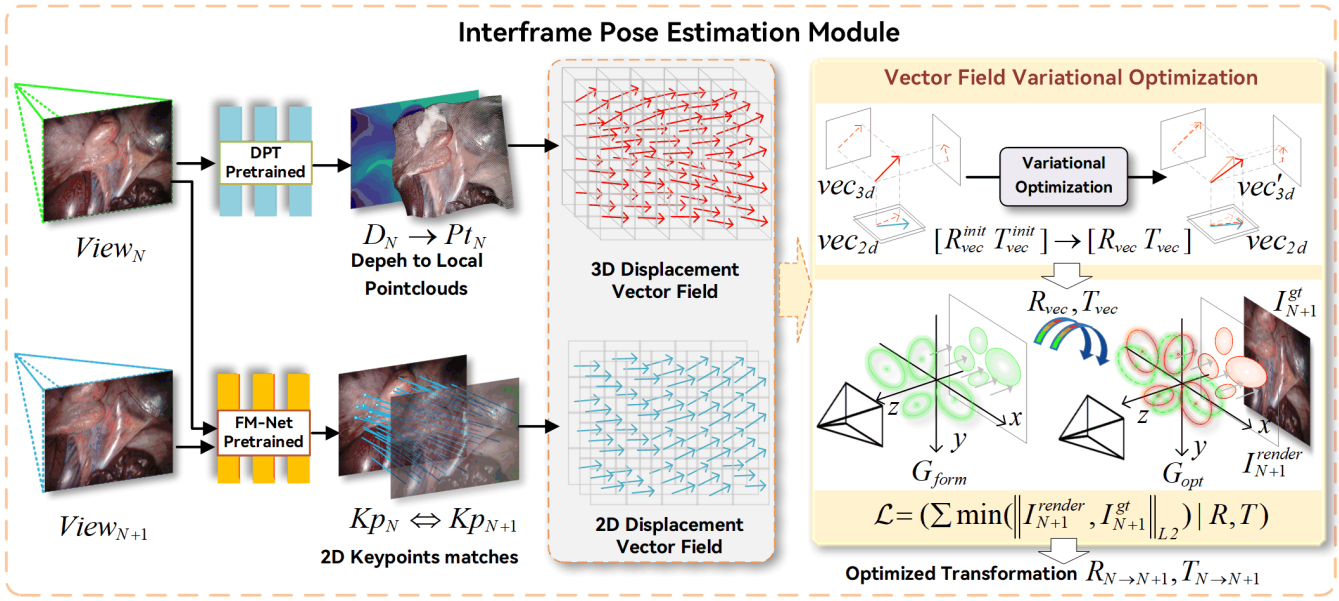

Fig. 2. The architecture of proposed Interframe Pose Estimation module based on sequential 2D-3D displacement field variational optimization. The DTP pretrained module denotes the Dense Prediction Transformer of depth map pretrained network [39], and the FM-Net Pretrained denotes the Feature Matching Pretrained network [40].

Based on the kp_N and kp_{N+1}, the 2D displacement field can be constructed as the follow:

$$DVF_{2D}(\vec{u}, \vec{v}) = \Gamma_{n=0}^{N}([u_{n+1} - u_n, v_{n+1} - v_n]) \quad (2)$$

The $\Gamma_{n=0}^{N}$ denotes the displacement field formation with N corresponding tensors. In the DVF_{2D} each vector element represents the matched local 2D feature offset between interframe, which also symbolizes the projection location of 3D key-points onto the local camera coordinate plane. Thus, according to the aforementioned plane the 3D displacement field denotes the 3D key-points offset between interframes in the *LCCS* of $View_{N+1}$, and the formula can be denoted as:

$$DVF_{3D}(\vec{x}, \vec{y}, \vec{z}) = \Gamma_{n=0}^{N}([R_{vec}^{init}\ T_{vec}^{init}] * Pt_N - Pt_H) \quad (3)$$

Where the $[R_{vec}^{init}\ T_{vec}^{init}]$ denotes the initial RT matrix between interframes, and the Pt_H denotes homogeneous coordinate location of kp_{N+1} in camera $View_{N+1}$. As shown in Fig. 2, the $[R_{vec}^{init}\ T_{vec}^{init}]$ is initiated by homogeneous matrix estimating between Pt_N and kp_{N+1} with RANSAC algorithm and optimized with variational functional. Given initial $\mathcal{T} \in [R_{vec}^{init}\ T_{vec}^{init}]$, considering an energy functional:

$$E[\mathcal{T}] = \int_{\Omega} ||\mathcal{T} * DVF_{3D}(x) - DVF_{2D}(x)||_2 dx \quad (4)$$

The Ω defines the domain of the independent variable and it's related to the definition of x. Here x is defined as the tensor of the displacement field with respect to the mapping of the coordinates, while ∇x can be denoted as the magnitude of the displacement field. Here the formula $||\mathcal{T} * DVF_{3D}(x) - DVF_{2D}(x)||_2$ is denoted as $F(\mathcal{T})$. To achieve the extremum of the functional $E[\mathcal{T}]$, the function $F(\mathcal{T})$ must satisfy the Euler-Lagrange equation under the following condition.

$$\partial F/\partial y - d/dx(\partial F/\partial y') = 0 \quad (5)$$

Here the gradient descent method is used to solve the equation, by calculating the gradient of the energy functional $E[\mathcal{T}]$ with respect to y. It can be indicated convergence when the change of the energy functional $E[\mathcal{T}]$ is less than the threshold, and the corresponding ∂F is the minimum function yielding from the optimization.

To improve computational efficiency, during training step the original 3D displacement field of size $(h, w, 1)$ is mapped and obtained by stacking the channels to high dimensional feature space with size $(h/8, w/8, 64)$, while for further intuitive understanding, it is mapped to $(h, w, 1)$ during presentation. The example of 3D and 2D displacement field and corresponding displacement field variational optimization in Scared [41] dataset and C3VD [42] dataset are shown in Fig. 3. Specifically, in Fig. 3 the corresponding displacement fields are presented with different transformation changes in camera poses: the planar translation on XY-plane of *LCCS* shown in row D1K1; the z-axis longitudinal translation of *LCCS* shown in row desc_t4_a; the z-axis rotation of *LCCS* shown in row D2K3. The column (a) represents the display of optimized 3D displacement field $DVF_{3d}(\vec{x}, \vec{y}, \vec{z})$ and its corresponding optimized 2D vector projection on XY-plane, the x axis and y axis represent the corresponding *LCCS* x_3d and y_3d, respectively.The value of z axis is taken as a negative number for better visual results. And the basic unit of length is millimeters(mm). The column (b) represents $DVF_{3d}^{pred}(\vec{u}, \vec{v})$ from $DVF_{3d}(\vec{x}, \vec{y}, \vec{z})$ with optimized pose transformation, which demonstrates the planar movement vector of the same feature point between interframes, and the color distribution of each two-dimensional vector represents its corresponding scalar value from lime to indigo. The column (c) displays the 3D-2D displacement field variational optimization. The set of red vector arrows represents the ground truth $DVF_{3d}^{gt}(\vec{u}, \vec{v})$ and the set of green vector arrows represents the $DVF_{3d}^{pred}(\vec{u}, \vec{v})$, and the basic unit of length is pixel. To specifically illustrate the vector filed distribution results, some figure details are presented with a locally magnified manner in (b) and (c).

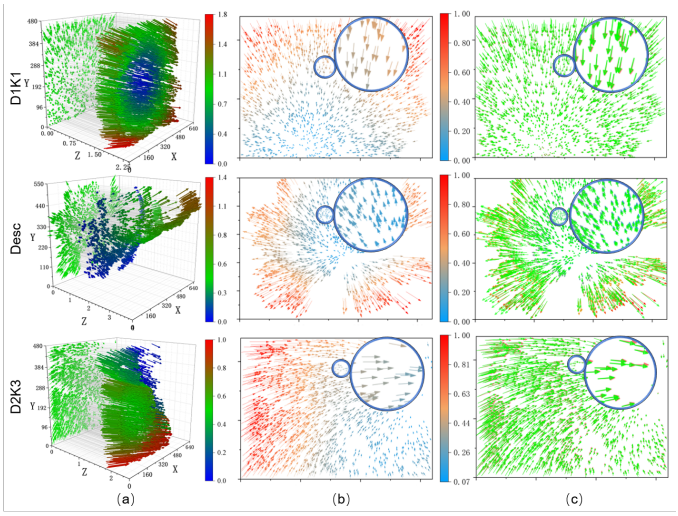

Fig. 3. The visualization of three transformation changes in camera poses under *LCCS* (D1K1: XY-planar translation, desc_t4_a: z-axis longitudinal translation, and D2K2: z-axis rotation). (a) shows the optimized 3D displacement field and the corresponding 2D vector projection on XY-plane; (b) represents the 2D displacement field demonstrates the planar movement vector of the same feature point between interframes, and the color distribution of each two-dimensional vector represents its corresponding scalar value from lime to indigo; (c) indicates the display of 3D-2D displacement field optimization. The red arrow represents the ground truth of 2D displacement field in (b) and the green one represents the 3D displacement field with optimized pose transformation in (a).

With 2D-3D displacement field variational optimization step, the initial transformation between interframes $View_N$ and $View_{N+1}$ is obtained. Different with the original gaussian splatting architecture[11], the global camera poses are replaced with the relative camera poses under the transformation in *LCCS* of current gaussian model. Given former gaussian model G_{Form} in $View_N$, assuming the transformation of $G_{form} \in View_N$ is identity rotation matrix with zero translation, and the local camera pose transformation of $View_N \rightarrow View_{N+1}$ is $[R_{vec}\ T_{vec}]$. Thus, as the displacement field variational optimization step shown in Fig. 2, the gaussian model transformation of $G_{form} \rightarrow G_{opt}$ can be formulated as:

$$[x, y, z] = [R_{vec}^T\ - R_{vec}^T * T_{vec}] * [x, y, z]_{form} \tag{6}$$

Where $[x, y, z]_{form}$ denotes the location gaussian model in *LCCS*, and in Fig. 2 the green ellipsoid represents the original gaussian model in $View_N$, the red ellipsoid represents the transformed gaussian model in $View_{N+1}$. With the function of gaussian raster rending f_{raster}, the rendered frame can be formulated as

$$\begin{aligned} I_{N+1}^{render} &= f_{raster}(G_{opt}) \\ &= f_{raster}([R_{vec}^T\ - R_{vec}^T * T_{vec}] \times G_{opt}) \end{aligned} \tag{7}$$

The next frame I_{N+1}^{gt} with L2-norm is set as supervision function in $View_{N+1}$, and the final loss function of Interframe Pose Estimation can be formulated as:

$$\mathcal{L} = \sum (min(\|I_{N+1}^{render}, I_{N+1}^{gt}\|_{L2}) | R, T) \tag{8}$$

The optimized matrix R_{opt} and T_{opt} represent the local gaussian model transformation between G_{form} and G_{opt} in *LCCS* of $View_{N+1}$, which can fulfill continuous transform optimization with sequential frames flow.

B. *Dynamic Gaussian Growing Module*

The original gaussian splatting method [11] mainly focuses on the initial point cloud extracted from the SfM preprocessing, and the camera pose matrices are based on the coordinate system of the initial point cloud. Therefore, the process of training the gaussian model can be seen as the simulation of object capturing in real-world, which also means the new render frame is obtained and gaussian model is trained by transforming the viewpoint position while remaining model static. In [12]–[14], the work of endoscopic reconstruction based on gaussian splatting is also based on the approach of static gaussian model & dynamic camera pose, which heavily relies on the initial point cloud as well as camera pose data.

In contrast to the aforementioned approach, a dynamic gaussian growing strategy with differential gradient field optimization is proposed to release the reliance on the initial point cloud and generate new views of gaussian within constrained field of camera view. The overall architecture of the module is shown in Fig. 4.

During each view a DPT pretrained network is used to predict the depth map as prior, and set it as initial local supervision for current view of gaussian model. Therefore, the gaussian model can be grown through corresponding depth map in constrained area of current view. The gaussian model is defined based on *LCCS* and each gaussian location will be transformed with camera pose moving, which means the pose matrix of current gaussian model is always the identity matrix in current camera coordinate system. When gaussian model is transformed from $View_N$ to $View_{N+1}$, there should be regions of model deficiency on the raster view due to the incompleteness of the initial gaussian model. As shown in Fig. 4, on the rendered image the corresponding representation would be the appearance of empty regions. Thus, the growth of the gaussian model can be achieved through the synthesis of the new viewpoint. For new blank region, at first a DTP pretrained network is used to output local depth map $Depth_{N+1}$ from I_{N+1} as supervision, then the $Depth_{N+1}$ is normalized to 0-1 and converted into local pointcloud Pts_{N+1}, based on the previous mask of the blank regions, the initial *xyz* coordinates of the new gaussian point set can be obtained. Since the previous gaussian model has already been transformed to the current camera viewpoint, in other words, the corresponding pose transformation matrix is the identity matrix, and the new initialized gaussian positions can be automatically aligned with the previous gaussian coordinates. The new gaussian initialization is shown with green color in Fig. 5, the α, θ and γ denote the initial radius of the gaussian distribution, and the rest parameters can also be initialed alone with the corresponding spherical harmonics, opacity and rotation. Different with the original gaussian model in [11], the proposed gaussian model designed in this paper not only retains the original feature parameters, but also added a novel parameter called the gradient optimization direction $vec_{optimize}$ for differential gradient field optimization.

Regarding the refinement of the generated gaussian model, since the new growth local gaussian set in the current viewpoint $View_{N+1}$ are generated based on 2D planar points. The

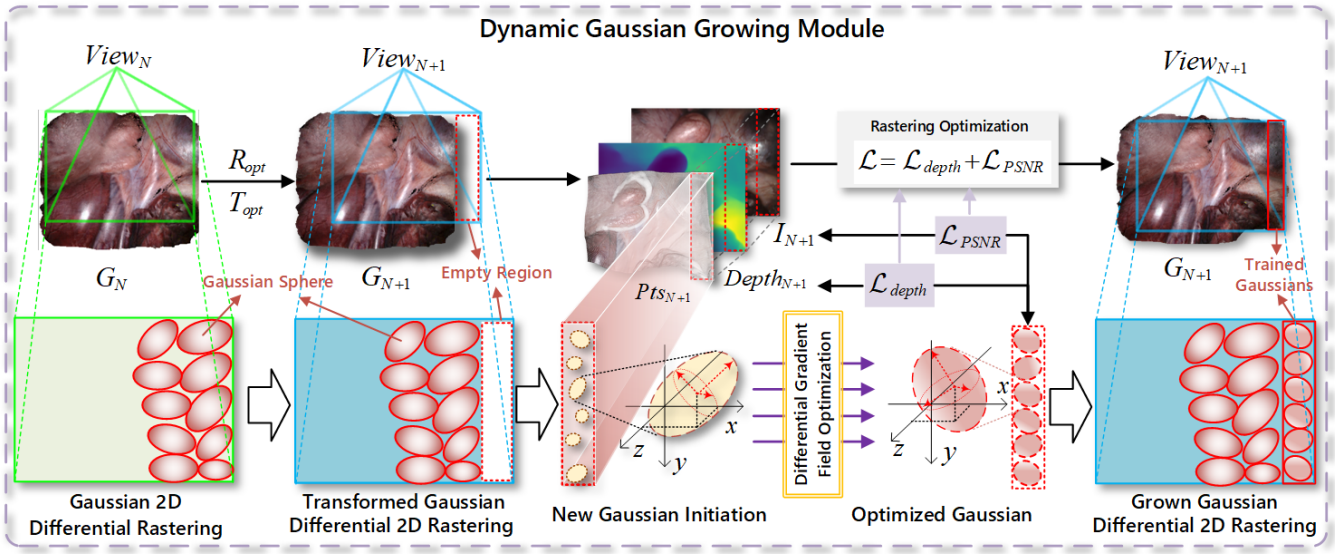

Fig. 4. The architecture of proposed dynamic gaussian growing module between interframe $View_N$ and $View_{N+1}$. The G_N and G_{N+1} represent the gaussian model under the *LCCS* of $View_N$ and $View_{N+1}$, respectively.

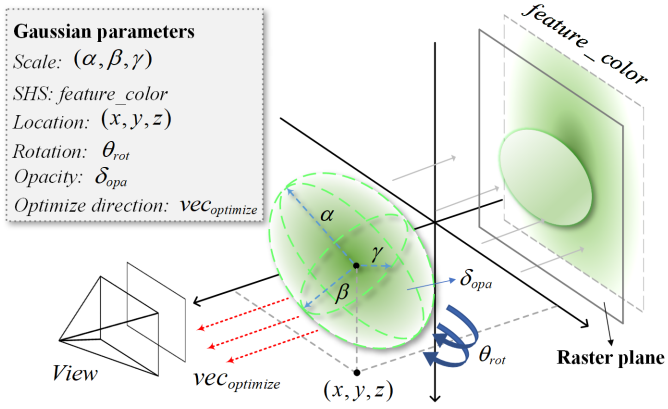

Fig. 5. The schematic diagram of the gaussian model shows the single gaussian model with parameters defined under view and rendered on raster plane.

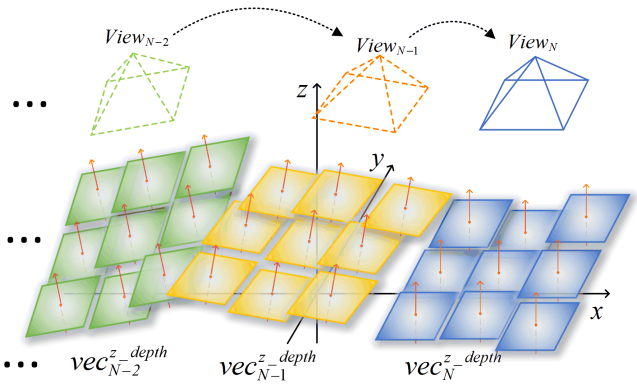

Fig. 6. Example of different gradient directions of local gaussian set to be optimized under camera viewpoints $View_N$, $View_{N-1}$ and$View_{N-2}$. The red arrows represent the descending direction of the gradients. Each rectangle represents the normal plane which is perpendicular to the gradient of the voxel.

calculation is show below:

$$Pts_{N+1} = \sum([u_{2d}, v_{2d}, z_{depth}] \times K * coords_{h \times w}) \quad (9)$$

Where u_{2d}, v_{2d}, z_{depth} represent the local depth points, K represents the camera intrinsic matrix, and $coords_{hw}$ represents homogeneous coordinate map, which only relies on the dimensions of image. The projected coordinates u_{2D} and v_{2D} on the plane are absolutely accurate, therefore the parameter that needs to be optimized can be optimized as a single z_{depth} value during the training process. And the optimized z_{depth} is the local depth value under the current viewpoint, whose optimized direction is kept consistent with current camera viewpoint direction (identity matrix).

When the gaussian model undergoes the next stage of growth, due to the occurrence of gaussian changes, the optimized direction of the z_{depth} in the previously grown gaussian model will be transformed to a new direction, and its transformation matrix is associated with the new gaussian model transformation matrix. Therefore, based on gaussian differential rastering, a differential gradient field is established for optimizing the *XYZ* of gaussian along with each gradient vector direction. The example of differential gradient field to be optimized in different camera viewpoints is shown in Fig. 6. Since the gradient of the field is derived from a finite number of gaussians, in essence it's only differentiable locally with discrete representation. The formula for the new optimized field direction is as follows:

$$vec_N^{z_{depth}} = vec_{N-n}^{z_{depth}} * \prod_{i<n}^{i=0} [R_{optimized}^{N-i} \; T_{optimized}^{N-i}]^T \quad (10)$$

The example visualization of gaussian growth with differential gradient field optimization is shown in Fig. 7. (a) demonstrates the comparison of *xyz* surface differential gradient field optimization in local gaussian set growth between $View_N$, $View_{N-1}$, and named G_{xyz}^N, G_{xyz}^{N-1}. (b) shows the difference

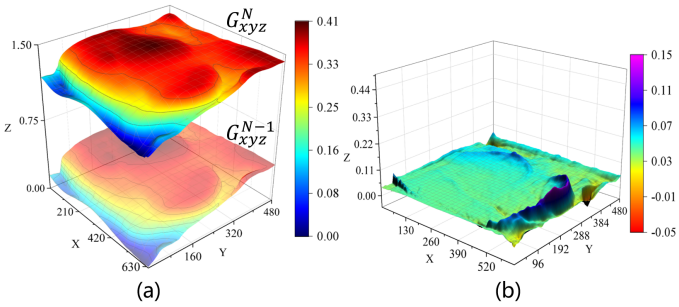

Fig. 7. The visualization of gaussian growth with differential gradient field optimization between interframes. (a) presents the dynamic changes of the gaussian point set; (b) presents the 3D graph of the difference differential gradient map of each voxel point in the *XY* plane of the *LCCS*.

differential gradient map of local gaussian growth where larger values indicate a higher gradient rate.

With sequence frame flows in, the proposed module can traverse and optimize all of the local gaussian set that appear within the field of view as much as possible. Meanwhile, dense property would appear due to the generated local gaussian set is based on pixel points, a novel scale activation function is designed to replace the original gaussian scale activation function to accommodate the dense property. Aiming to achieve lower radius threshold through slower learning rate, thereby reducing the radius of the gaussian ellipsoid to generate more refined results. The proposed activation function formula is as follows:

$$f(x) = 1/4 \times (e + ln(1 - x)) \tag{11}$$

The loss function is designed as a combination of the raster supervision and depth supervision function. The raster supervision loss function is set as the PSNR similarity to generate outputs that maximize the image similarity measure. The formula is as follow:

$$\mathcal{L}_{\mathcal{PSNR}} = 10 \times log_{10}(Max(I_N)/MSE(I_N)) \tag{12}$$

Due to the occurrence of gaussian changes, the depth supervision loss function is designed with different optimization directions of gaussian models under each viewpoint. Additionally, the training directions of the same gaussian point set under different viewpoints shall change according to the gaussian transformation during the training process, and the transformation method has already been described in Equ 10.

For a subset of the gaussian model corresponding to one single growth iteration, firstly the point clouds optimization function is as follows:

$$f^N_{Pts\_opt} = \sum([u_{2D}, v_{2D}, z_{depth}] \times K * coords_{h \times w}) \tag{13}$$

Then the loss function of the subset estimation depth map can be represented as:

$$\begin{aligned} l^{sub\_N}_{Depth} =& f_{Raster\_depth}(G_N(xyz \in f^N_{Pts\_opt})) * I^{Mask}_N \\ & - Depth_N * I^{Mask}_N \end{aligned} \tag{14}$$

Where I^{Mask}_N represents the bool mask map under current viewpoint, which filters out the growth gaussian models consisting with $vec^{z_{depth}}_N$, and f_{Raster_depth} represents the differential gaussian rasterization of depth map. Therefore, the total modified depth supervision loss function can be denoted as:

$$\mathcal{L}^{total}_{depth} = \sum_{n<N}^{n=0} l^{subset\_n}_{Depth} \tag{15}$$

And the final loss function can be denoted as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{PSNR}} + \mathcal{L}^{total}_{depth} \tag{16}$$

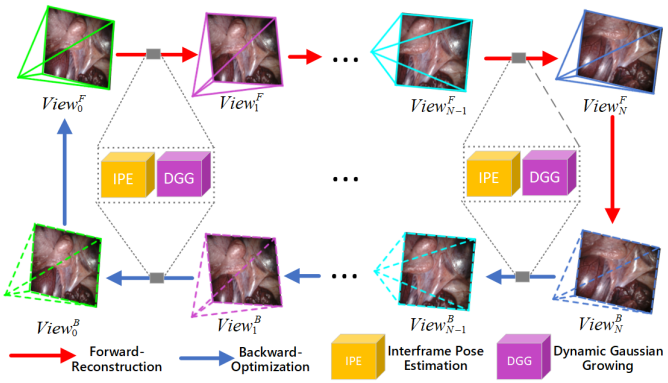

Fig. 8. The architecture of sequential scene reconstruction with "forward-reconstruction & backward-optimization".

C. Sequential Scene Reconstruction Architecture

In order to acquire gaussian reconstruction frame-by-frame without SfM prior, based on the two proposed modules above, a novel unconstrained sequential gaussian reconstruction framework is constructed with "Forward-Reconstruction & Backward-Optimization" (FR&BO) strategy. Different with the original 3D gaussian splatting, the proposed architecture can support continuous frames input with scalability, the overall architecture of sequential scene reconstruction is shown in Fig. 8. In the architecture, one epoch contains the complete forward reconstruction and inverse optimization process by iterating the sequence frames twice.

First Frame Initialization. For the initial frame, the viewpoint is initialized into the *LCCS* of the first frame, and the corresponding point cloud Pts_{init} is calculated from depth map with Formula (1). The *xyz* of gaussian model is initialed with the point cloud Pts_{init}, and the spherical harmonics *SHS* parameter is initialized with the RGB value of the frame. The *SHS* contains the main feature color $feature\_dc$ and the rest feature color $feature\_rest$, and the formula is as follow:

$$SHS = feature\_dc \oplus feature\_rest \tag{17}$$

Forward-Reconstruction. During the forward process of sequential order video frame, with the help of IPE and DGG, the interframe poses are estimated and new local gaussian sets are grown for $View_0 \rightarrow View_1$, $View_1 \rightarrow View_2$ ... and $View_{N-1} \rightarrow View_N$. The forward reconstruction process aims to maintain the integrity of the reconstruction process as much as possible.

Backward-Optimization. During the backward process of sequential order video frame, since the estimated poses represent the transformation matrices between every double interframes, and the current gaussian model G_N is in the *LCCS*

of $View_N$, it's not feasible to perform repeated operations directly from viewpoint $View_0$. The gradients of the gaussian model need to be guaranteed continuous during the training process. Therefore, it's a feasible approach to achieve cyclical training by reversing the order of the video frames, and the backward training orders are: $View_N \rightarrow View_{N-1}$, $View_{N-1} \rightarrow View_{N-2}$, ... , and $View_1 \rightarrow View_0$. The backward optimization process aims to perform optimization on the estimated gaussian parameters during the forward reconstruction process and back-propagate the gradients.

IV. EXPERIMENT AND RESULTS

A. *Experiment Settings*

*1) **Datasets Detail**:* The proposed network is evaluated on two public available datasets: the stereo correspondence and reconstruction of endoscopic dataset (Scared) [41] and the colonoscopy 3D video dataset (C3VD) [42].

Stereo Correspondence and Reconstruction of Endoscopic Dataset: For Scared dataset, the original frames and depth maps were obtained using the da Vinci Xi surgical robot. The dataset consists of 7 training datasets and 2 testing datasets with a resolution of 1024×1280 pixels. Each dataset corresponds to a single porcine subject and contains between 4 and 5 keyframes. And one keyframe is a single unique view of the scene where the structured light pattern was projected into the field of view of the camera and a dense stereo reconstruction was computed [41]. For all of the keyframes in each dataset, where 80 percent of random picked keyframes are used for training, and the remaining 20 percent are used for validation. The resolution of the original frame is downsampled into 480×640. In this paper, four distinct sequences ("dataset_1, keyframe_1", "dataset_2, keyframe_3", "dataset_4, keyframe_2" and "dataset_8, keyframe_1") are selected for example visualize demonstration.

Colonoscopy 3D Video Dataset: For C3VD dataset, the original frames are acquired with a high definition clinical colonoscope and high-fidelity colon models for benchmarking computer vision methods in colonoscopy. In total, 22 short video sequences are registered to generate 10,015 total frames with paired ground truth depth, surface normal maps, optical flow, occlusion, six degree-of-freedom pose, coverage maps, and 3D models. The dataset also includes screening videos acquired by a gastroenterologist with paired ground truth pose and 3D surface models [42]. The ratio of the training set as well as the validation set is consistent with the aforementioned Scared. The resolution of the original frame is down-sampled into 480×675. In this paper, four distinct sequences ("cecum, cecum_t2_a", "desc, desc_t4_a", "sigmoid, sigmoid_t3_b" and "trans, trans_t1_a") are selected for example visualize demonstration.

*2) **Evaluation Metrics**:* For a comprehensive quality assessment of comparative view synthesis, the proposed method is evaluated in four aspects several common metrics. For novel view synthesis, following the previous methods [11]–[13], [32] Peak Signal-To-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS), specifically the VGG loss [43] and Structural Similarity Index Measure (SSIM) [44] are selected as evaluation metrics. For relative pose evaluation, since current methods do not include the relative pose calculation step, in this paper only the proposed relative pose estimation metrics is exhibited and ground truth. The standard visual odometry metrics is used, which including the Absolute Trajectory Error (ATE) and Relative Pose Error (RPE). The estimated trajectory is aligned with the ground truth with 6 degrees of freedom.

*3) **Compared Methods**:* The DG-3DGS is evaluated with four relevant methods: the modified 3D gaussian splatting (3DGS*) approach [11]; the monocular endoscopic scene reconstruction with 4D gaussian splatting (Endo4DGS) [12]; the realtime dense reconstruction and tracking in endoscopic surgeries nursing gaussian splatting (EndoGSLAM) [32] and the realtime gaussian splatting for dynamic endoscopic scene reconstruction (EndoGaussian) [13]. The different training strategies of all the compared methods are briefly introduced as follows. The 3DGS* is based on the standard reconstruction baseline and contains the classical gaussian render and raster approach. Since the Colmap pipeline from the original 3DGS cannot complete the initialization to generate initial point cloud. Therefore, the input method of the 3DGS approach is modified by directly using the initial point cloud and pose ground truth as the initialized data. Theoretically, this approach has better advantages, and it will be marked as 3DGS* in the following discussions. The Endo-4DGS utilizes lightweight MLPs to capture temporal dynamics with gaussian deformation fields and obtain a satisfactory gaussian initialization by generating pseudo-depth maps as a geometry prior with Depth-Anything [45], the provided depth map from original datasets is used for the ground truth input. The EndoGSLAM integrates streamlined gaussian representation and differentiable rasterization to facilitate high rendering speed during online camera tracking and tissue reconstructing. The EndoGaussian contains holistic gaussian initialization (HGI) and spatio-temporal gaussian tracking (SGT), to handle the non-trivial gaussian initialization and tissue deformation problems, respectively. To make a fair comparison to methods which are ready to be tested, for each dataset the networks are used with the same training and validation data. During each test sequence, use the default hyperparameters in the original implementations for all of the baselines.

*4) **Implementation Details**:* The entire proposed network is trained using AdamW [46] optimizer with initial learning rate of 1×10^{-6}, the AdamW decay parameter is set 0.1 and adopt linear learning rate warming-up strategy after 2000 steps. The learning rate scheduler is set multistepLR with gamma 0.5. The batch size is set 1 on RTX 3080Ti GPU with 10 epochs. The original endoscopic image size is set to $(480 \times 640 \times 3)$ with RGB format. For depth map period, the DPT pretrained network [47] is used to output the relative local depth estimation results, and the dual-correlate optimized coarse-fine feature matching network [40] is used for interframe keypoints matching and output distributed matches uniformly.

B. *Comparison with Related Methods*

In this subsection, the DG-3DGS is compared with several baselines mentioned above on new view synthesis and relative

TABLE I
THE COMPARISON RESULTS OF PSNR, SSIM AND LPIPS ON SCARED DATASET FROM DATASET_1 TO DATASET_9. EACH DATA SET CONTAINS KEYFRAMES FROM 1 TO 4, AND THE EVALUATION METRICS UNDER 20% EVAL FRAME VIEWPOINTS ARE COLLECTED.

scenes	Endo4DGS [12]			3DGS* [11]			EndoGSLAM [32]			EndoGaussian [13]			DG-3DGS		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Dataset_1	16.56	0.5096	0.4299	24.13	0.6893	0.2093	15.98	0.4222	0.5213	21.29	0.6463	0.4247	**24.35**	**0.9747**	**0.1556**
Dataset_2	12.19	0.4802	0.6062	**24.81**	0.6624	0.3702	12.39	0.4100	0.5905	13.28	0.4664	0.6157	22.27	**0.9617**	**0.2554**
Dataset_3	11.87	0.4253	0.5812	**22.66**	0.6075	0.3196	11.61	0.3398	0.5906	16.19	0.4894	0.5631	22.20	**0.9550**	**0.2216**
Dataset_4	15.74	0.6061	0.4535	24.75	0.7701	0.2754	14.81	0.5662	0.5172	18.20	0.6618	0.5045	**25.79**	**0.9788**	**0.1519**
Dataset_5	16.50	0.5565	0.4455	20.29	0.6694	0.3990	14.64	0.5204	0.5147	18.43	0.6194	0.4788	**21.42**	**0.9472**	**0.2419**
Dataset_6	16.39	0.6265	0.4741	**24.29**	0.7900	0.3665	15.31	0.6071	0.5271	18.40	0.6968	0.5203	23.65	**0.9713**	**0.2514**
Dataset_7	13.71	0.5433	0.5244	**24.40**	0.7557	0.3469	14.31	0.5054	0.5462	15.36	0.5272	0.5621	23.73	**0.9487**	**0.2734**
Dataset_8	16.35	0.5691	0.4598	25.37	0.7691	0.2971	16.58	0.5665	0.5003	20.09	0.6784	0.4995	**25.61**	**0.9730**	**0.1684**
Dataset_9	13.58	0.3940	0.6507	22.49	0.5579	0.3372	13.06	0.2854	0.6160	14.47	0.3186	0.6529	**22.63**	**0.9551**	**0.2442**
Mean	14.81	0.5279	0.5105	**23.68**	0.7011	0.3259	14.33	0.4761	0.5450	17.37	0.5736	0.5337	23.55	**0.9628**	**0.2182**

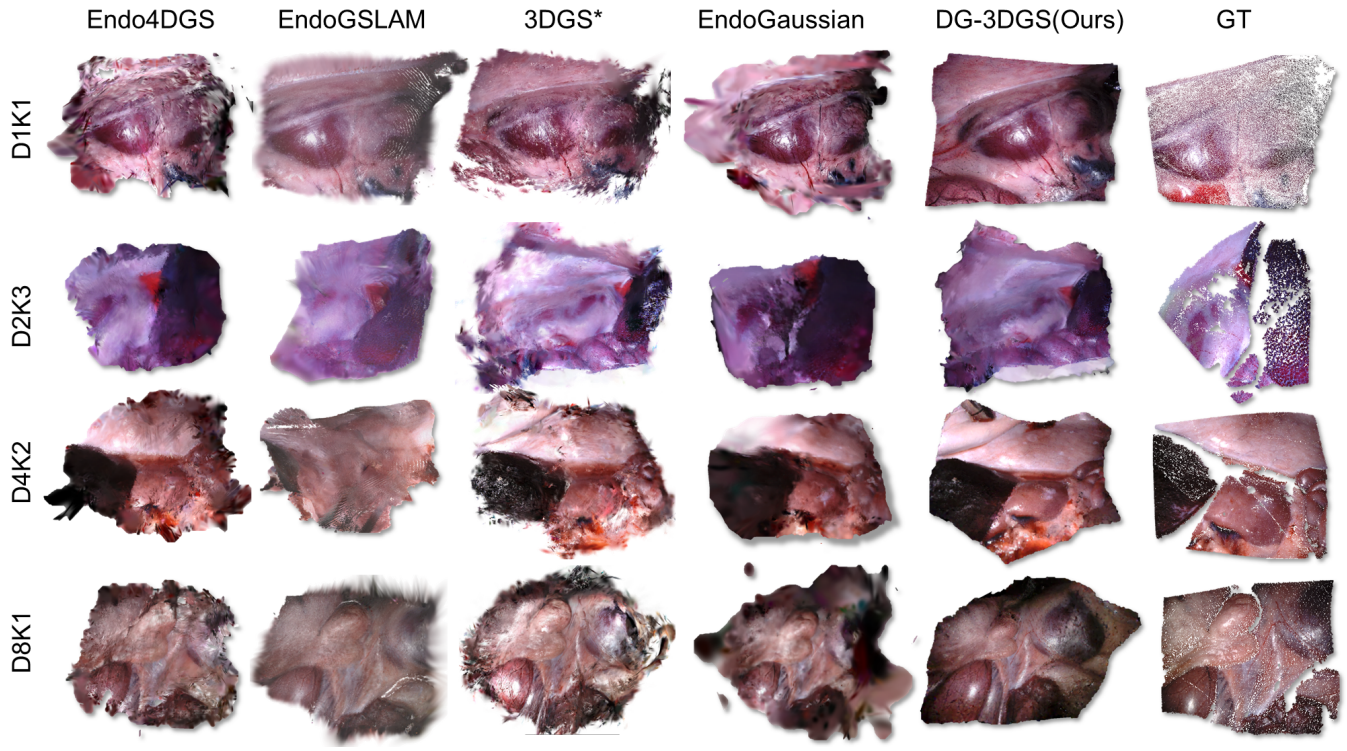

Fig. 9. The visualization of qualitative 3D reconstruction results of Endo4DGS, EndoGSLAM, 3DGS*, EndoGaussian and DG-3DGS (ours) from Scared dataset, the GT represents the sparse point clouds extracted from original sense data with voxel rendering. The example results are reported including dataset_1 keyframe_1(D1K1), dataset_2 keyframe_1(D1K1), dataset_4 keyframe_2(D4K2), dataset_8 keyframe_1(D8K1).

pose estimation.

*1) **Results on Scared dataset**:* A comprehensive comparison of the DG-3DGS is conducted with state-of-the-art approaches for abdominal cavity scene reconstruction. The comparison results on Scared from dataset_1 to dataset_9 are reported in Table I. Upon analysis, the DG-3DGS outperforms the others across most of metrics consistently. For Endo4DGS, EndoGSLAM and EndoGaussian, due to limited viewpoints or low accuracy of initial point cloud, the view reconstruction after model training may exhibit uneven gaussian distributions, significant distortion, or excessively large radii, resulting in poor image quality. In terms of 3DGS*, since the provided initial point cloud and pose matrices data are both derived from the ground truth without Colmap initialization, the reconstruction results largely depend on the ground truth data. Therefore, the value of PSNR under several viewpoints does not outperform those of 3DGS*.

Analyzing from specific detail prospectively, the Scared dataset contains more limited camera views, whose motion trajectories are mainly located on the *XY*-plane of *LCCS*. The shortage of multi-view supervision results in lacking effective bundle adjustment (BA) as well as richer multi-view representations of the gaussian model. In contrast, with the help of DGG proposed above, the DG-3DGS can achieve interframe bundle adjustment reprojecting optimization and supervise the coordinates of the generated point cloud of gaussian model with monocular depth map priors. As is shown in Table. I, the DG-3DGS can achieve higher SSIM scores

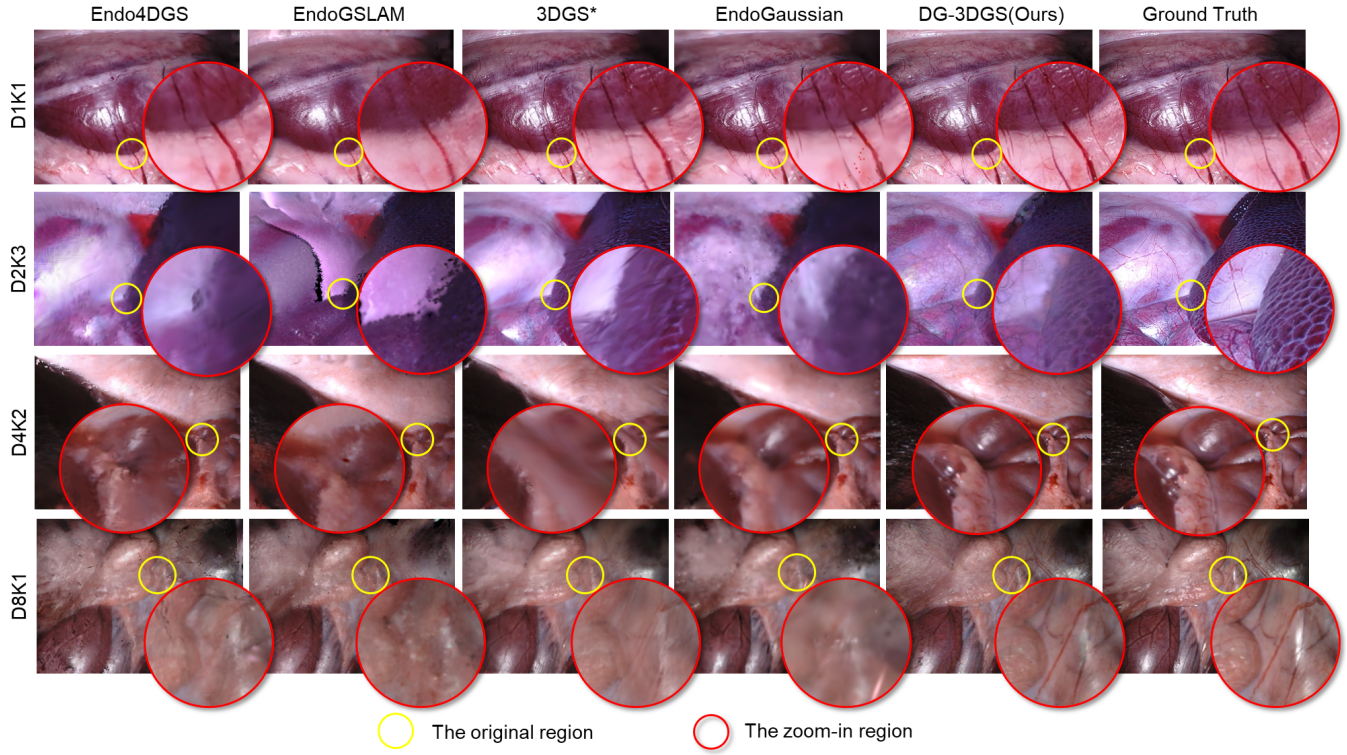

Fig. 10. The visualization of qualitative 2D rendering results of Endo4DGS, EndoGSLAM, 3DGS*, EndoGaussian and DG-3DGS from Scared dataset. The example results are reported including dataset_1 keyframe_1(D1K1), dataset_2 keyframe_1(D1K1), dataset_4 keyframe_2(D4K2), dataset_8 keyframe_1(D8K1).

together with lower LPIPS scores.

The qualitative 3D reconstruction results are presented in Fig. 9. From the figure it can be observed that compared with the reconstruction results of other methods against the GT sparse point clouds, the DG-3DGS can demonstrate more refined reconstruction outcomes with complete gaussian model. Followed with the original 3DGS*, the Endo3DGS, EndoGSLAM and EndoGaussian mainly maintain the gaussian training pipeline with Pruning and Split strategy, which mainly relies on enough view points to establish multi-view geometry constraint. Due to the constrained field of views, it's difficult for this strategy to conduct gaussian *XYZ* location constraint, lacking geometry supervision under multi-view during training. Compared with the related work, with the help of differential gradient field optimization in DGG, the gaussian gradient tensor can be optimized to a clear direction without random gradient descending. By establishing the gaussian differential gradient field, it's possible for optimizer to search for the lowest descending locations under geometry constrains. In fact, due to the narrow space of endoscope scene, during training the directions of gaussian gradients are mainly located on the *XY* plane, gaussian tends to be optimized along with the tensor perpendicular to z-axis. By adopting the pixel-level gaussian dynamic grow strategy, the occurrence of holes is reduced on the surface of reconstruction results. The example rendering results of each method are reported in Fig. 10. In this figure the render images with zoom-in detail are presented under eval frame viewpoints, and it can be shown that the DG-3DGS can generate clearer images and sufficient texture details compared to other methods. With the function of DGG, because of the consistency between optimized direction and current camera viewpoint direction, the gaussian ellipsoid is constrained within a local range while ensuring the gaussian ellipsoids exhibit different spherical harmonics from different viewpoints. Therefore, it's possible for DG-3DGS to raster more detailed and accurate render results compared with the others.

*2) **Results on C3VD dataset**:* A comprehensive comparison of the DG-3DGS is conducted with state-of-the-art approaches for rectum scene reconstruction and report the comparison results on C3VD in Table II. The C3VD dataset contains 4 sub-data labeled as cecum, desc, sigmoid and trans. Upon analysis, the DG-3DGS still outperforms the others across most of metrics consistently. For EndoGSLAM, 3DGS*, EndoGSLAM and EndoGaussian, the PSNR and SSIM evaluation metric shows relatively low performance, and the LPIPS score perform higher compared with the DG-3DGS, which means lower render quality. Similar with the experiment on Scared dataset, during evaluation of the 3DGS*, the provided initial point cloud and pose matrices data are both derived from the ground truth without Colmap initialization, and the quantitative analysis shows that the DG-3DGS still outperforms the 3DGS* across most of metrics consistently.

Different with the Scared dataset, the clinical scenes of C3VD dataset come from inside the rectum. The camera motion trajectories primarily follow the internal motion along the rectum, Therefore, its movement directions are mainly restricted to the z-axis of *LCCS*. Similarly, it's challenge to process BA under such limited viewpoints without accurate pose period. The qualitative 3D reconstruction results of

TABLE II
THE COMPARISON RESULTS OF PSNR, SSIM AND LPIPS ON C3VD DATASET. EACH DATATSET CONTAINS VARIOUS KEYFRAMES AND THE EVALUATION METRICS UNDER 20% EVAL FRAME VIEWPOINTS ARE COLLECTED.

scenes	Endo4DGS [12]			3DGS* [11]			EndoGSLAM [32]			EndoGaussian [13]			DG-3DGS		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Cecum	19.32	0.7595	0.4675	20.05	0.8217	0.3615	17.52	0.6902	0.4797	17.58	0.7468	0.5262	**21.03**	**0.9091**	**0.3475**
Desc	16.38	0.6988	0.4264	**19.10**	0.8135	**0.3667**	15.11	0.6247	0.4693	16.14	0.7054	0.4642	16.81	**0.8769**	0.3831
Sigmoid	19.34	0.7567	0.4065	21.06	0.8331	**0.2957**	17.51	0.6652	0.4750	17.76	0.7522	0.4947	**22.149**	**0.9118**	0.3064
Trans	18.10	0.7784	0.3622	19.91	0.8340	0.3078	18.50	0.7084	0.4230	16.63	0.7579	0.4878	**22.12**	**0.9651**	**0.2550**
Mean	18.63	0.7615	0.4121	20.12	0.8282	0.3329	17.71	0.6874	0.4562	17.14	0.7489	0.5006	**21.34**	**0.9286**	**0.3068**

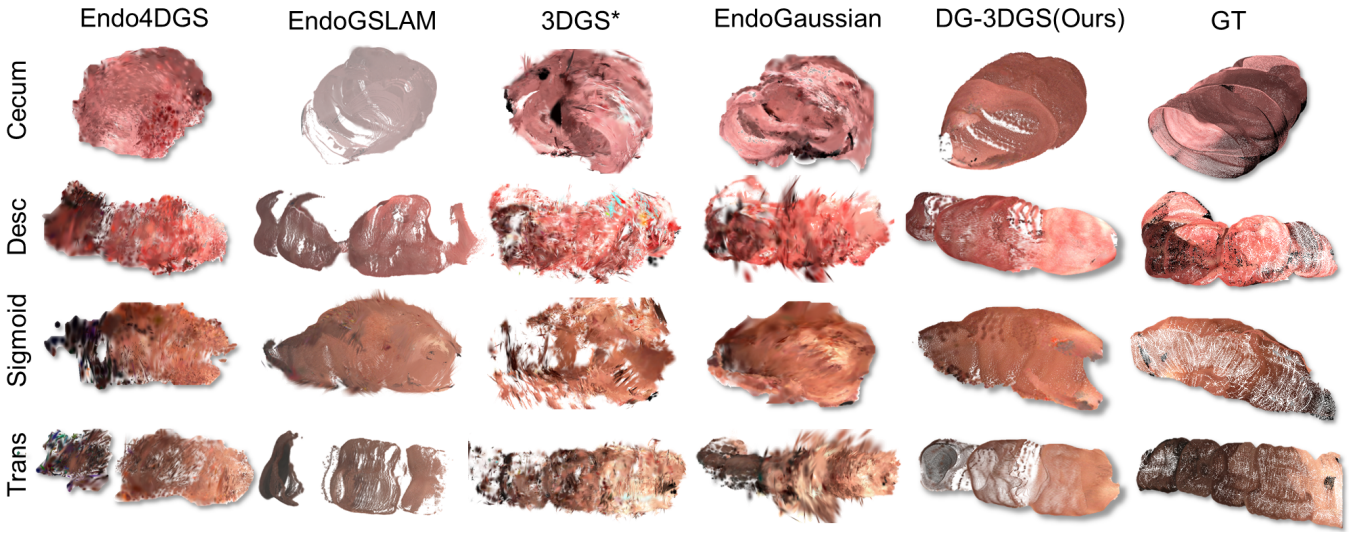

Fig. 11. The visualization of qualitative 3D reconstruction results of Endo4DGS, EndoGSLAM, 3DGS*, EndoGaussian and the DG-3DGS from C3VD dataset, and the GT presents the sparse point cloud provided from original data with voxel rendering. The example from four sample cases are reported: cecum_t2_a(Cecum), desc_t4_a(Desc), sigmoid_t3_b(Sigmoid), trans_t1_a(Trans).

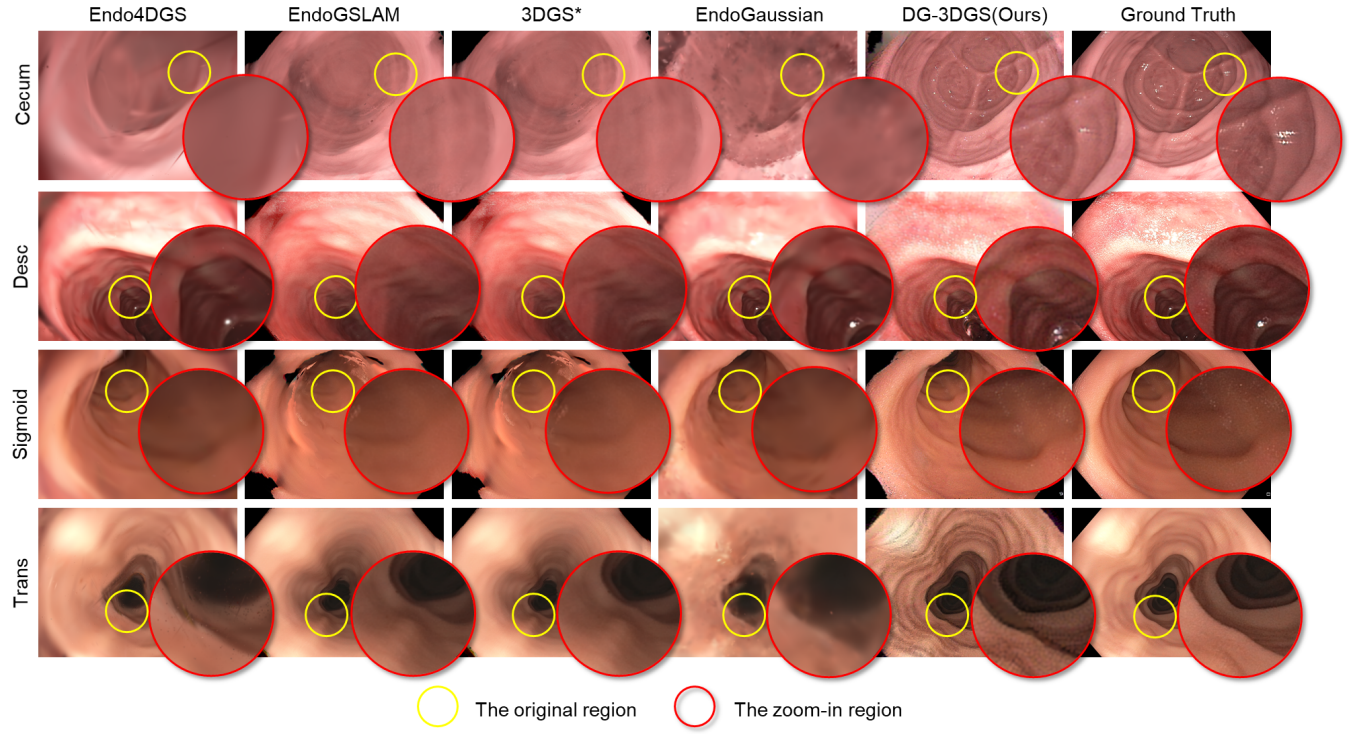

Fig. 12. The visualization of qualitative 2D rendering results of Endo4DGS, EndoGSLAM, 3DGS*, EndoGaussian and the DG-3DGS from C3VD dataset, and the GT presents the sparse point cloud provided from original data with voxel rendering. The example from four sample cases are reported: cecum_t2_a(Cecum), desc_t4_a(Desc), sigmoid_t3_b(Sigmoid), trans_t1_a(Trans).

C3VD are reported in Fig. 11. Compared with the other results, the DG-3DGS can generate more detailed and high-quality reconstruction gaussian models. Similar with the results on the Scared dataset, the other methods tend to optimized the gaussian ellipsoids with random gradient directions, resulting in unreasonable radius of gaussian ellipsoids and lacking geometry supervision under multi-view. In DG-3DGS, the gaussian gradient tensor can be optimized to a clear direction without random gradient descending with differential gradient field optimization, and more reasonable gaussian ellipsoids can be optimized to guarantee the detailed gaussian model. The example rendering results of each method are also reported in Fig. 12. In this figure the render images with detail under eval frame viewpoints are presented, and it can be shown that the DG-3DGS can generate clearer images and sufficient texture details compared to other methods. The zoom-in view of the local details shown in the image indicates that the DG-3DGS is more beneficial for generating texture details of the cecum wall. Compared with the Scared dataset, the reconstruction process is mainly focused on the motion along with z-axis in *LCCS* of C3VD dataset. Therefore, the requirements for high-accurate camera pose are the primary influencing factors. In DG-3DGS, the IPE module can generate more accurate relative camera pose compared with the others, and transform the gaussian model continuously to align with the current *LCCS* to avoid absolute pose errors. Therefore, there are more detailed and accurate gaussian rendering results in DG-3DGS compared with the others.

*3) **Relative pose estimation**:* The relative pose estimation accuracy against ground truth poses of Scared and C3VD are reported in Table 3. RPEt represents the root mean square error of the translation part in pose estimation, which measures the difference between estimated translation and actual translation over a series of time steps. RPEt is commonly used to evaluate the translation accuracy between consecutive frames. PREr represents the root mean square error of the rotating part in pose estimation. Similar with RPEt, RPEr measures the difference between estimated rotation and actual rotation over a series of time steps, typically used to evaluate the rotation accuracy between consecutive frames. ATE represents absolute trajectory error, which measures the difference between the estimated trajectory and the actual trajectory. ATE is evaluated by calculating the point-to-point distance between the estimated trajectory and the actual trajectory, typically used to assess the accuracy of overall pose estimation. Since the DG-3DGS implies the relative pose estimation for interframe transformation, each pair of pose matrix is calculated based on the *LCCS* of previous frame as baseline. Compared to the Colmap-Based pipeline, the proposed IPE pipeline demonstrates superior performance across various evaluation metrics, which relies entirely on the estimated absolute poses and point clouds without using Colmap or ground truth poses for assistance during training.

We further present the visualization of local camera poses estimation results extracted from example cases in Fig. 13. The horizontal color bar represents the rotation matrix errors corresponding to different frames, obtained by converting them to rotational Euler angles to derive the angular errors relative to the ground truth rotation angles. From this figure it can be seen that the variations of the relative pose are very close to those of the ground truth, which provides significant assistance for the subsequent viewpoint localization and reconstruction. The rotation matrix errors primarily concentrated in some scattered frames, and there are little impact on the accuracy of overall gaussian reconstruction.

TABLE III
THE EVALUATION OF POSE ACCURACY ON SCARED AND C3VD BETWEEN IPE AND COLMAP-BASED PIPELINES. THE UNIT OF RPER IS IN DEGREES, ATE IS IN THE GROUND TRUTH SCALE AND THE UNIT OF RPET IS PIXEL WITH 100 SCALING.

dataset	scenes	IPE Pipeline			Colmap-Based Pipeline		
		RPEt↓	RPEr↓	ATE↓	RPEt↓	RPEr↓	ATE↓
Scared	Dataset_1	**0.1777**	**0.0020**	**0.1774**	0.9855	1.9952	0.9859
	Dataset_2	**0.3719**	**0.0066**	**0.3715**	1.2008	2.0113	1.0323
	Dataset_3	**0.2125**	**0.0044**	**0.2125**	1.0552	2.0603	1.0397
	Dataset_4	**0.2441**	**0.0029**	**0.2439**	1.0435	2.0584	1.0465
	Dataset_5	**0.2959**	**0.0057**	**0.2958**	1.0440	2.1246	1.0537
	Dataset_6	**0.3611**	**0.0040**	**0.3613**	1.0998	1.9563	1.0627
	Dataset_7	**0.2899**	**0.0051**	**0.2898**	1.0585	2.0213	1.0744
	Dataset_8	**0.2395**	**0.0021**	**0.2393**	1.0582	2.0050	1.1045
	Dataset_9	**0.2896**	**0.0041**	**0.2896**	1.0628	2.0434	1.2022
	Mean	**0.2758**	**0.0041**	**0.2757**	1.0665	2.0316	1.0664
C3VD	Cecum	**0.1638**	**0.0136**	**0.1637**	0.9855	2.0959	0.9856
	Desc	**0.2433**	**0.1728**	**0.2426**	0.9082	0.9082	0.9082
	Sigmoid	**0.0984**	**0.0136**	**0.0984**	1.0153	2.1062	1.0152
	Trans	**0.2322**	**0.0042**	**0.2316**	1.0764	2.0487	1.0766
	Mean	**0.1844**	**0.0511**	**0.1841**	1.0202	2.0349	1.0204

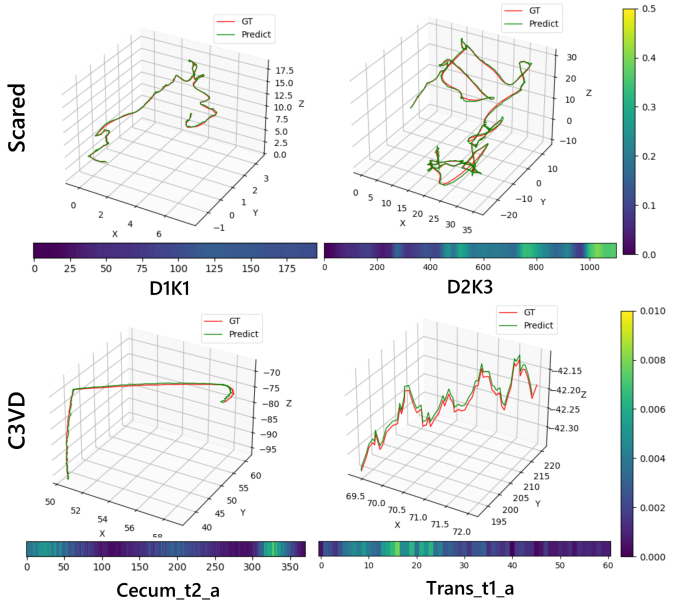

Fig. 13. The visualization of local camera poses estimation extracted from Scared and C3VD with different example cases. The unit of translation is millimeter (mm) in *LCCS*, and the unit of rotation is degree.

*4) **Ablation Study**:* To comprehensively verify the effects of the core architecture design on the DG-3DGS, we conduct the ablation experiments under specific conditions. The detail conditions are as follows, which keep the same setting and training strategies and test on the Scared dataset. The com-

parison results is presented in Tab. IV, and the qualitative 3D reconstruction results of Scared are reported in Fig. 14.

TABLE IV
THE COMPARISON RESULTS OF PSNR, SSIM AND LPIPS BETWEEN ABLATION EXPERIMENTS ON SCARED AND C3VD DATASETS.

dataset	eval metrics	w/o IPE	w/o DGG	w/o SSR	DG-3DGS
Scared	PSNR↑	20.80	20.61	21.24	**23.55**
	SSIM↑	0.9380	0.9186	0.8588	**0.9628**
	LPIPS↓	0.2644	0.3524	0.3264	**0.2182**
C3VD	PSNR↑	21.03	20.18	20.42	**21.34**
	SSIM↑	0.9040	0.9126	0.8978	**0.9286**
	LPIPS↓	0.3312	0.3123	0.3272	**0.3068**

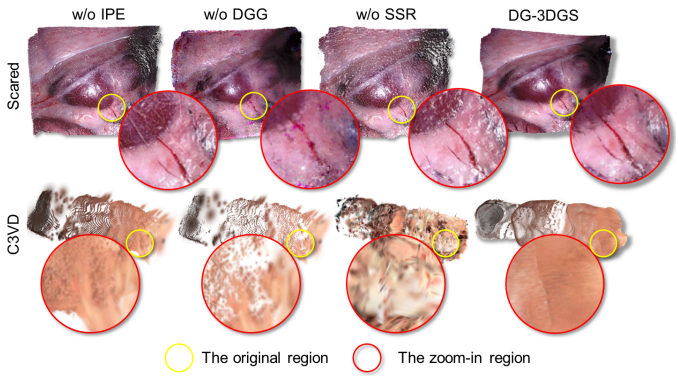

Fig. 14. The visualization of qualitative 3D reconstruction results of w.o. IPE, w.o. DGG, w.o. SSR and DG-3DGS on Scared D1K1 and C3VD trans_t1_a senses.

(a) W/o Interframe Pose Estimation.

The Interframe Pose Estimation module is replaced by the classical local pose transform estimation method based on SIFT/RANSAC. Each transformation matrix is based on the previous frame image as the reference coordinate system with consistent global scale maintaining. The result shows that the PSNR and SSIM decreases due to the lack of pose estimation accuracy, resulting in the rendering sense shifts.

(b) W/o Dynamic Gaussian Growing.

The dynamic gaussian growing module is replaced by the original density and prune function proposed from the original 3DGS [11], this method achieves gaussian filling by separating existing large-scale gaussian ellipsoids and generating new gaussian models. The result shows that due to the lack of multi-view supervision in training process, the density and prune function can't fully perceive the 3D structural information of the scene, and the LPIPS increases compared with DG-3DGS.

(c) W/o Sequential Scene Reconstruction.

The standard gaussian training strategy with random sense view sampling are used during training process, and the gaussian model is frozen with transforming relative poses into absolute poses. The result shows that the PSNR, SSIM and LPIPS still fails to exceed the metric of DG-3DGS. Without sequential scene reconstruction architecture, the estimation of relative pose will become quite challenging and the gaussian model growth for new view will also face more difficulties.

V. CONCLUSION

In this work, to overcome the challenges of tiny pose variation, constrained field of views and non-prior SfM in monocular endoscopic scene reconstruction, an effective dynamic growing 3D gaussian splatting architecture called DG-3DGS is proposed for high-quality reconstruction without any pre-computed poses or SfM priors. A novel Interframe Pose Estimation module is designed by learning and optimizing the variational gradients of gaussian rastering to perform accurate tiny pose estimation. Based on differential gradient field optimization, a novel Dynamic Gaussian Growing strategy is designed to perform new gaussian synthesis within constrained endoscopic view. Combining these two modules above, a novel 3D gaussian splatting architecture with Forward-Reconstruction&Backward-Optimization strategy is proposed to generate the gaussian model sequentially and resolve the limitation of SfM prior. The improved relative pose estimation and novel gaussian growing strategy lead to better novel view synthesis quality and model geometry reconstruction. The proposed DG-3DGS outperforms state-of-the-art methods in quantitative (PSRN, SSIM and LIPIS) and qualitative comparisons. It's believed that the proposed method will be an important step towards applying the unknown-pose or non-SfM 3D gaussian splatting methods to more complex clinical scenes in the future.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation Program of China (62025104, 62171039, 62422102), the National Key R&D Program of China (2023YFC2415300), and the Haidian Original Innovation Joint Fund of Beijing Natural Science Foundation (L222149).

REFERENCES

[1] R. Zha, X. Cheng, H. Li, M. Harandi, and Z. Ge, "Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, Conference Proceedings, pp. 13–23.

[2] R. Tang, L.-F. Ma, Z.-X. Rong, M.-D. Li, J.-P. Zeng, X.-D. Wang, H.-E. Liao, and J.-H. Dong, "Augmented reality technology for preoperative planning and intraoperative navigation during hepatobiliary surgery: A review of current methods," *Hepatobiliary & Pancreatic Diseases International*, vol. 17, no. 2, pp. 101–112, 2018.

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[4] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.

[5] R. Yunus, J. E. Lenssen, M. Niemeyer, Y. Liao, C. Rupprecht, C. Theobalt, G. Pons-Moll, J. Huang, V. Golyanik, and E. Ilg, "Recent trends in 3d reconstruction of general non-rigid scenes," in *Computer Graphics Forum*. Wiley Online Library, 2024, Conference Proceedings, p. e15062.

[6] V. M. Batlle, J. M. Montiel, P. Fua, and J. D. Tardós, "Lightneus: Neural surface reconstruction in endoscopy using illumination decline," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, Conference Proceedings, pp. 502–512.

[7] Y. Wang, Y. Long, S. H. Fan, and Q. Dou, "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, Conference Proceedings, pp. 431–441.

[8] C. Yang, K. Wang, Y. Wang, X. Yang, and W. Shen, "Neural lerplane representations for fast 4d reconstruction of deformable tissues," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, Conference Proceedings, pp. 46–56.

[9] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, Conference Proceedings, pp. 3504–3515.

[10] G. Chen and W. Wang, "A survey on 3d gaussian splatting," *arXiv preprint arXiv:2401.03890*, 2024.

[11] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[12] Y. Huang, B. Cui, L. Bai, Z. Guo, M. Xu, and H. Ren, "Endo-4dgs: Distilling depth ranking for endoscopic monocular scene reconstruction with 4d gaussian splatting," *arXiv preprint arXiv:2401.16416*, 2024.

[13] Y. Liu, C. Li, C. Yang, and Y. Yuan, "Endogaussian: Gaussian splatting for deformable surgical scene reconstruction," *arXiv preprint arXiv:2401.12561*, 2024.

[14] L. Zhu, Z. Wang, J. Cui, Z. Jin, G. Lin, and L. Yu, "Endogs: Deformable endoscopic tissues reconstruction with gaussian splatting."

[15] M. Yu, T. Lu, L. Xu, L. Jiang, Y. Xiangli, and B. Dai, "Gsdf: 3dgs meets sdf for improved rendering and reconstruction," *arXiv preprint arXiv:2403.16964*, 2024.

[16] X. Gu and H. Ren, "A survey of transoral robotic mechanisms: Distal dexterity, variable stiffness, and triangulation," *Cyborg and Bionic Systems*, vol. 4, p. 0007, 2023.

[17] L. Li, X. Li, B. Ouyang, H. Mo, H. Ren, and S. Yang, "Three-dimensional collision avoidance method for robot-assisted minimally invasive surgery," *Cyborg and Bionic Systems*, vol. 4, p. 0042, 2023.

[18] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, Conference Proceedings, pp. 4104–4113.

[19] A. Fisher, R. Cannizzaro, M. Cochrane, C. Nagahawatte, and J. L. Palmer, "Colmap: A memory-efficient occupancy grid mapping framework," *Robotics and Autonomous Systems*, vol. 142, p. 103755, 2021.

[20] S. Bonilla, S. Zhang, D. Psychogyios, D. Stoyanov, F. Vasconcelos, and S. Bano, "Gaussian pancakes: Geometrically-regularized 3d gaussian splatting for realistic endoscopic reconstruction," *arXiv preprint arXiv:2404.06128*, 2024.

[21] X. Duan, D. Xie, R. Zhang, X. Li, J. Sun, C. Qian, X. Song, and C. Li, "A novel robotic bronchoscope system for navigation and biopsy of pulmonary lesions," *Cyborg and Bionic Systems*, vol. 4, p. 0013, 2023.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[23] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, pp. 404–417, 2006.

[24] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, Conference Proceedings, pp. 2564–2571.

[25] A. Fisher, R. Cannizzaro, M. Cochrane, C. Nagahawatte, and J. L. Palmer, "Colmap: A memory-efficient occupancy grid mapping framework," *Robotics and Autonomous Systems*, vol. 142, p. 103755, 2021.

[26] W. Bian, Z. Wang, K. Li, J.-W. Bian, and V. A. Prisacariu, "Nope-nerf: Optimising neural radiance field with no pose prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, Conference Proceedings, pp. 4160–4169.

[27] M. Kim, S. Seo, and B. Han, "Infonerf: Ray entropy minimization for few-shot neural volume rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, Conference Proceedings, pp. 12 912–12 921.

[28] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "Nerf–: Neural radiance fields without known camera parameters," *arXiv preprint arXiv:2102.07064*, 2021.

[29] A. Rosinol, J. J. Leonard, and L. Carlone, "Nerf-slam: Real-time dense monocular slam with neural radiance fields," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, Conference Proceedings, pp. 3437–3444.

[30] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, Conference Proceedings, pp. 6229–6238.

[31] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, "inerf: Inverting neural radiance fields for pose estimation. in 2021 ieee," in *RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, Conference Proceedings, pp. 1323–1330.

[32] K. Wang, C. Yang, Y. Wang, S. Li, Y. Wang, Q. Dou, X. Yang, and W. Shen, "Endogslam: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting," *arXiv preprint arXiv:2403.15124*, 2024.

[33] S.-F. Chng, S. Ramasinghe, J. Sherrah, and S. Lucey, "Garf: Gaussian activated radiance fields for high fidelity reconstruction and pose estimation," *arXiv e-prints*, p. arXiv: 2204.05735, 2022.

[34] Y. Xia, H. Tang, R. Timofte, and L. Van Gool, "Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction," *arXiv preprint arXiv:2210.04553*, 2022.

[35] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, Conference Proceedings, pp. 5741–5751.

[36] Z. Cheng, C. Esteves, V. Jampani, A. Kar, S. Maji, and A. Makadia, "Lu-nerf: Scene and pose estimation by synchronizing local unposed nerfs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, Conference Proceedings, pp. 18 312–18 321.

[37] A. Meuleman, Y.-L. Liu, C. Gao, J.-B. Huang, C. Kim, M. H. Kim, and J. Kopf, "Progressively optimized local radiance fields for robust view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, Conference Proceedings, pp. 16 539–16 548.

[38] R. Ma, R. Wang, Y. Zhang, S. Pizer, S. K. McGill, J. Rosenman, and J.-M. Frahm, "Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy," *Medical image analysis*, vol. 72, p. 102100, 2021.

[39] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.

[40] Z. Zhang, H. Song, J. Fan, T. Fu, Q. Li, D. Ai, D. Xiao, and J. Yang, "Dual-correlate optimized coarse-fine strategy for monocular laparoscopic videos feature matching via multilevel sequential coupling feature descriptor," *Computers in Biology and Medicine*, vol. 169, p. 107890, 2024.

[41] M. Allan, J. Mcleod, C. Wang, J. C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K. X. Fu, T. Zeffiro, and W. Xia, "Stereo correspondence and reconstruction of endoscopic data challenge," *arXiv preprint arXiv:2101.01133*, 2021.

[42] T. L. Bobrow, M. Golhar, R. Vijayan, V. S. Akshintala, J. R. Garcia, and N. J. Durr, "Colonoscopy 3d video dataset with paired depth from 2d-3d registration," *Medical image analysis*, vol. 90, p. 102956, 2023.

[43] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, Conference Proceedings, pp. 586–595.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[45] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, Conference Proceedings, pp. 10 371–10 381.

[46] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[47] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

Ziang Zhang received the B.E. degree in automation from Beijing University of Chemical Technology in 2017, and received the M.A. degree in robotics science and engineering from Northeastern University in 2020. He is currently pursuing the Ph.D. degree with the School of Medical Technology, Beijing Institute of Technology. His research interests include medical image analysis and computer vision.

Hong Song received the Ph.D. degree in computer science from the Beijing Institute of Technology in 2004. She is currently a Professor at the School of Computer Science and Technology, Beijing Institute of Technology, China. She was a Visiting Professor at the College of Computing and Informatics, Drexel University, and the Department of Computer Science, Columbia University, USA. Her current research interests include medical image analysis, augmented reality and computer vision.

Jingfan Fan received the Ph.D. degree in optical engineering from the Beijing Institute of Technology in 2016. He worked as a Postdoctoral Researcher with the University of North Carolina at Chapel Hill, from 2017 to 2019. He is currently an Associate Professor with the School of Optics and Photonics, Beijing Institute of Technology. He focuses his research interests on medical image processing, computer vision, and augmented reality.

Long Shao received the B.E. degree in measurement and control technology and instrumentation from Xi 'an Technological University in 2015, and the Ph.D. degree in optical engineering from Beijing Institute of Technology in 2022. He is a postdoctoral fellow in the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His research interests include binocular stereo vision, intelligent image sensing, and computer vision.

Tianyu Fu received the Ph.D. degree from the Beijing Institute of Technology, Beijing, China, in 2019. He is currently an Assistant Professor with the School of Medical Technology, Beijing Institute of Technology. His research interests include medical image registration and reconstruction.

Danni Ai received the M.S. degree in Information Engineering from Xi'an Jiaotong University, China, in 2008, and the Ph.D. degree from Ritsumeikan University, Japan, in 2011. She is currently an Associate Professor in the School of Optics and Photonics, Beijing Institute of Technology. Her research interests include medical image analysis, surgical navigation, virtual reality and augmented reality.

Deqiao Xiao received the B.S. degree in computer science from Henan University, Kaifeng, China, in 2011, and the Ph.D. degree from the Shenzhen Institutes of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, in 2017. He is currently an Assistant Professor with the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China. His research interests include medical image analysis, image-guided intervention/surgery, and computer vision.

Yuanyuan Wang received the B.S. degree in biomedical engineering from Capital Medical University, Beijing, China, in 2016, and the Ph.D. degree in biomedical engineering from Tsinghua University, Beijing, in 2022. She is currently a Postdoctoral Research Scientist with the School of Optics and Photonics, Beijing Institute of Technology University. She was supported by the General Fund and Special Fund of Chinese Post-Doctoral Science Foundation in 2023 and 2024, respectively. She received the Young Scientists Fund from the National Natural Science Foundation of China (NSFC) in 2023. Her research interests include ultrasound imaging, photoacoustic imaging, and medical image analysis.

Yucong Lin received the Ph.D. degree from the Tsinghua University, in 2021. He is currently an Assistant Professor with the School of Optics and Photonics, Beijing Institute of Technology. His research interests include multimodal medical image analysis, multimodal large models, and medical informatics.

Jian Yang received the Ph.D. degree in optical engineering from the Beijing Institute of Technology, Beijing, China, in 2007. He is currently a full Professor with the School of Optics and Photonics, Beijing Institute of Technology. His research interests focus on medical image processing, surgical navigation, augmented reality, and computer vision.