3D Reconstruction of Subglottic Stenosis using Neural Radiance Fields and Robotics

Jinjie Sun and Julia Wiercgroch

Abstract—Subglottic stenosis (SGS) is a rare disease that is difficult for clinicians to measure during routine endoscopic examination. We propose the use of continuum robotics and computer vision techniques to reconstruct the stenosis from collected endoscopic videos to improve the decision-making of the clinician. A virtual SGS model was constructed with textures taken from endoscopic videos. A continuum robot was controlled in Unity to capture different viewpoints of the SGS from above the vocal chords. The collected images were used to train NeRF and MonoSDF models. The best performing model was the vanilla-NeRF model with the full dataset that was able to reconstruct the stenosis with minimal artefacts. This project was a stepping stone to determine which models could be used as a backbone for the 3D reconstruction of subglottic stenosis.

Index Terms—3D Reconstruction, Neural Radiance Fields, Laryngeal Stenosis, Endoscopic Video

1 Introduction

Subglottic stenosis (SGS) is a rare recurrent disease characterized by the gradual narrowing of the airway between the vocal cords and trachea. The current standard of care is routine laryngeal examinations with a flexible monoscopic endoscope to monitor the length and width of the stenosis and to determine when surgical intervention will be required. However, conventional exams provide clinicians only with a top-down view of the stenosis, making it difficult to visualize and quantify the extent of the disease progression. Additionally, patient intolerance and the narrow airway heavily restrict the laryngoscope movements and accessible camera viewpoints to create robust 3D reconstructions.

Computer-assisted diagnosis with continuum robotics and endoscopic 3D reconstruction of the affected area has the potential to provide clinicians with an appropriate measurement tool of SGS. Continuum robots are flexible, jointless structures that can perform complex bending motions. Continuum robots have a low diameter to length ratio and have been developed to navigate through confined spaces and reach sites of interests through complex trajectories. These characteristics make continuum robots ideal candidates to help increase camera viewpoints in SGS examination while minimizing patient discomfort when navigating an obstructed airway.

The task of 3D reconstruction from multiple RGB images has been a fundamental problem in computer vision and has been particularly challenging in medical applications. With the recent emergence of coordinate-based neural networks, compact, memory-efficient multi-layer perceptrons have been used to parameterize implicit shape representations such as occupancy or signed distance functions (SDF) and reconstruct scenes. Neural radiance fields (NeRFs) have achieved improved reconstruction results by expressing volume density as a function of the underlying 3D surface. Although NeRFs were initially designed for simple scenes with dense viewpoint sampling, recent improvements have allowed this reconstruction technique to perform well in the presence of limited input views (at least 3 images) and for scenes with large textureless regions. This makes NeRF an exciting technique to explore for 3D reconstruction with medical endoscopic images, which typically have been limited by their restricted camera viewpoints and large textureless regions.

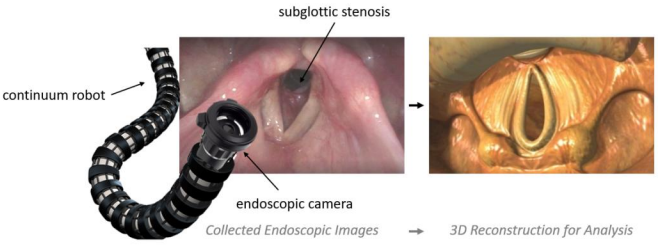

Fig. 1. Anticipated Examination Procedure

In this paper, we present the development of a method for 3D reconstruction of the subglottic region with sparse viewpoints captured by a continuum robot in a virtual environment (Fig. 1). This paper aims to compare the performance of the original NeRF algorithm to a neural radiance field capable of processing depth and surface normal cues (MonoSDF) on simulated endoscopic images. Additionally, by replacing the traditional laryngoscope with a continuum robot, we examine how the number of camera poses affects the quality of 3D reconstruction.

In summary, we make the following contributions:

- We create a dataset for SGS with virtual models of the larynx and stenosis, accounting for surface specularities and tissue texture. We develop a Unity scene with a continuum robot and 3D models of the larynx with varying severity of subglottic stenosis.

- *J. Sun and J. Wiercigroch are with the Department of Computer Science, University of Toronto, Ontario in the Medical Computer Vision and Robotics (MEDCVR) Lab.*

- We analyse and compare the novel view synthesis and 3D reconstruction results using vanilla-NeRF and MonoSDF on our custom dataset.

2 Related Work

2.1 Structure from Motion

Structure from Motion (SfM) is a reconstruction method that detects features on collected images, finds correspondences between different frames, and uses triangulation to create a surface for visualization. However, its application in endoscopic videos has been challenging due to the deformability of tissues, smooth and textureless surfaces, and imaging specularities [1].

The Medical and Computer Vision Lab (MEDCVR) has been working on developing a reconstruction technique for the laryngeal region and monoscopic endoscope videos based on the SfM method. The reconstruction algorithm includes a pre-filtering step from endoscopic videos to ensure that the frames used for feature detection and matching have a similar anterior glottic angle of the vocal cords, creating a pseudo-rigid problem. A learning-based approach with Correspondence Transformer for Matching Across Images (COTR) was used for feature matching to account for the low-textured regions [2]. The current pipeline is capable of reconstructing virtual models of the larynx, but fails to reconstruct the stenosis beyond the vocal cords. The method is extremely sensitive to lighting conditions and specularities, and creates a misaligned model when applied to real endoscopic videos.

2.2 Neural Radiance Fields

In contrast to SfM which only may use deep learning for parts of the reconstruction pipeline, recent neural approaches use multilayered perceptrons (MLP) to parameterize the surfaces of objects. Neural field representations are compact, differentiable and easy to optimize. Neural radiance fields (NeRF) represents scenes as an MLP that outputs volume density and view-dependent emitted radiance as a function of 3D location and 2D viewing direction [3]. Given a viewing angle, novel scenes can be created by casting simulated camera rays and using the volume rendering equation to determine the resulting view-dependent colour. Positional encoding is used to map the 3D location and 2D viewing angles into higher dimensional space to allow the model to learn higher frequency details. NeRF drastically outperformed other methods for novel scene synthesis and 3D reconstruction. The method drastically reduces the amount of space required to represent a single scene in 3D, but requires a lot of time to train. Furthermore, neural radiance fields need diverse camera poses, a large image dataset, and were originally not adapted to specular, dynamic or deformable scenes.

2.3 Monocular Geometric Cues for Neural Implicit Surface Reconstruction

In MonoSDF, monocular geometric priors were incorporated in neural implicit surface reconstruction methods to improve the reconstruction quality [4]. This framework uses a pretrained Omnidata model [5], [6], a neural network trained to estimate surface normals and depth from monocular images. This method was developed to address the poor reconstruction quality in NeRF for textureless regions and to decreases the number of photos required for accurate reconstructions. The method also explored different choices for neural implicit surface representations including dense signed distance function grids, single MLPs, single-resolution feature grid with MLP decoder, and multi-resolution feature grips with MLP decoders. MonoSDF was able to handle complex 3D scenes with many different surfaces. For sparse inputs, depth and normal cues improved the reconstruction quality and was more robust to less-observed regions than other methods. However, the results of the model were extremely dependent on the quality of the extracted monocular cues.

3 Methods

3.1 NeRF

3.1.1 Scene Representation

Continuous scenes can be parameterized as a function taking 3D position $\mathbf{x} = (x, y, z)$ and 2D viewing direction (θ, ϕ) as inputs and outputs an emitted colour $\mathbf{c} = (r, g, b)$ and volume density σ. The MLP architecture used to estimate the function $F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ is shown in Figure 3.1.1. Inputs are mapped to higher-dimensional space using positional encoding to improve its performance on data with high frequency variations in colour and geometry.

Let $p = (x, y, z)$. Then the positional encoding used for NeRF is given by equation 1.

$$\gamma(p) = \begin{bmatrix} \cos(2^0\pi p) & \sin(2^0\pi p) \\ \cos(2^1\pi p) & \sin(2^1\pi p) \\ \cos(2^2\pi p) & \sin(2^2\pi p) \\ \vdots & \vdots \\ \cos(2^{L-1}\pi p) & \sin(2^{L-1}\pi p) \end{bmatrix} \tag{1}$$

The MLP processes the input with 8 full-connected layers with ReLU activations and 256 channels per layer, and outputs σ and a 256-dimensional feature vector. The feature vector is concatenated with the viewing direction and passed through one more layer with 128 channels and ReLU activation and outputs the view-dependent RGB.

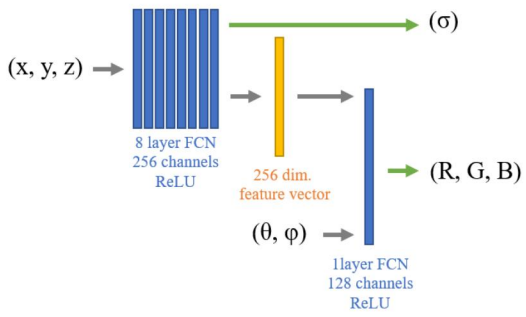

Fig. 2. NeRF architecture

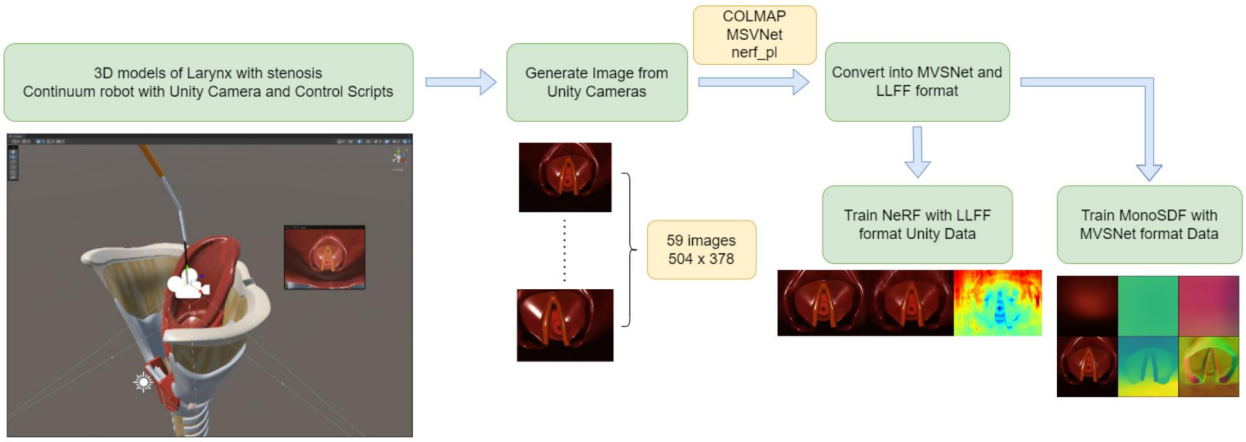

Fig. 3. Virtual SGS NeRF Reconstruction Workflow

3.1.2 *Volumetric Rendering of Network Output*

Let $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ represent the parameterized camera ray. The expected colour of the camera ray $\mathbf{r}(t)$ is given by:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} \sigma(\mathbf{r}(t))T(t)\mathbf{c}(\mathbf{r}(t), \mathbf{d})\, dt \tag{2}$$

where: $T(t) = \exp^{-\int_{t_n}^{t} \sigma(\mathbf{r}(s))\, ds}$ is the accumulated transmittance, $\sigma(\mathbf{r}(t))$ is the absorption coefficient, and $\mathbf{c}(\mathbf{r}(t))$ is emissive radiance. This integral is discretized by sampling points along the ray and summing them according to equation (3):

$$C(\mathbf{r}) \approx \sum_{i=1}^{N}(1 - \exp(-\sigma_i\delta_i)) \exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j)\mathbf{c}_i \tag{3}$$

Here $\delta_i = t_{i+1} - t_i$ is the distance between adjacent samples.

To optimize the network, NeRF randomly samples a batch of camera rays from the set of all pixels from the dataset to render a volume as outlined above. The loss that the MLP aims to minimize is the total squared error between the predicted and ground truth R, G, B values as shown in equation (4).

$$\mathcal{L}_{rgb} = \sum_{\mathbf{r}\in\mathcal{R}} \|C_{pred}(\mathbf{r}) - C_{gt}(\mathbf{r})\|_2^2 \tag{4}$$

3.2 MonoSDF

MonoSDF explored different design choices for representation neural implicit surfces such as a dense signed distance function (SDF) grid, a single MLP, as well as hybrid options such as a single-resolution feaure grid with MLP decoder and a multi-resolution feature grid with MLP decoder. Similar to NeRF, to render a pixel, a camera ray is cast using the parameterized equation $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$. N points are sampled along this ray to predict the SDF $\hat{s}^i_{\mathbf{r}}$ and colour values $\hat{\mathbf{c}}^i_{\mathbf{r}}$. Next, density values are learned by transforming SDF values with a learnable parameter β as follows:

$$\sigma_\beta(s) = \begin{cases} \frac{1}{2\beta}\exp\frac{s}{\beta} & s \le 0 \\ \frac{1}{\beta}(1 - \frac{1}{2}\exp(-\frac{s}{\beta})) & s > 0 \end{cases} \tag{5}$$

The colour for the current ray $C_{pred}(\mathbf{r})$ is calculated similarly to NeRF in equation (3). Depth $D_{pred}(\mathbf{r})$ and normal $N_{pred}(\mathbf{r})$ of the surface of the ray is calculated as:

$$D_{pred}(\mathbf{r}) = \sum_{i=1}^{N} T^i_{\mathbf{r}}\alpha^i_{\mathbf{r}}t^i_{\mathbf{r}} \tag{6}$$

$$N_{pred}(\mathbf{r}) = \sum_{i=1}^{N} T^i_{\mathbf{r}}\alpha^i_{\mathbf{r}}\hat{\mathbf{n}}^i_{\mathbf{r}} \tag{7}$$

where:

$$T^i_{\mathbf{r}} = \prod_{j=1}^{i-1}(1 - \alpha^j_{\mathbf{r}}) \tag{8}$$

$$\alpha^i_{\mathbf{r}} = 1 - \exp -\sigma^i_{\mathbf{r}}\delta^i_{\mathbf{r}} \tag{9}$$

and where $\hat{\mathbf{n}}$ is the 3D unit normal vector or the analytical gradient of the SDF function.

To optimize the network, in addition to the RBG reconstruction loss as in equation (4), MonoSDF also aims to minimize the Eikonal loss (10), depth consistency loss (11), and normal consistency loss (12).

$$\mathcal{L}_{\text{eikonal}} = \sum_{\mathbf{x}\in\mathcal{X}} (\|\nabla F_\Theta(\mathbf{x})\|_2 - 1)^2 \tag{10}$$

$$\mathcal{L}_{\text{depth}} = \sum_{\mathbf{r}\in\mathcal{R}} (\|\omega D_{pred}(\mathbf{r}) + q) - D_{gt}(\mathbf{r})\|^2 \tag{11}$$

$$\mathcal{L}_{\text{normal}} = \sum_{\mathbf{r}\in\mathcal{R}} (\|N_{pred}(\mathbf{r}) - N_{gt}(\mathbf{r})\|_1 + \|1 - N_{pred}(\mathbf{r})^\intercal N_{gt}(\mathbf{r})\|_1 \tag{12}$$

4 Experiments

4.1 Data Generation and Prepossessing

As Figure 3.1.1 shows, to create the Unity dataset, we generate a 3D model of the subglottic region and larynx with surface specularity and tissue texture by using MAYA and import the 3D model into Unity. We further create a continuum robot-like 3D object in Unity, with a controller script that can change the robot's position and orientation. At last, we attach a Unity camera at the robot's tip as an end-effector. We design a controller to save the camera view

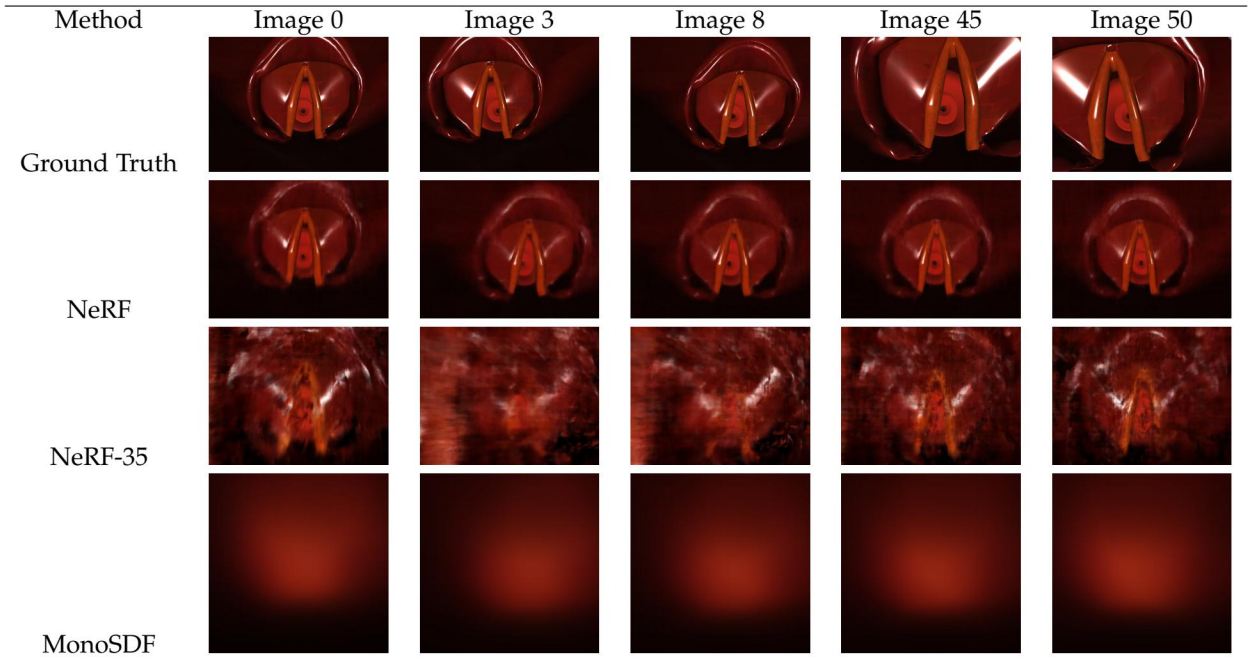

Fig. 4. Rendering Results of Different Reconstruction Models on Test Images

as a PNG image with a width and height of 504 and 378 pixels, respectively. In our dataset, we generate 59 images using the Unity camera to capture the subglottic region with a different position and orientation, all from the top view of the model with the larynx model placed vertically in the Unity axis.

We convert the dataset into two formats: Local Light Field Fusion(LLFF) format [3], [7] and Replica format [4] to train the vanilla NeRF model and MonoSDF model. For LLFF format, we use COLMAP [8], [9] to recover the camera poses of the input images, as suggested by the LLFF repository. [7]. We then use the output from COLMAP and the converter scripts from the LLFF repository to convert our custom data into LLFF format. The LLFF format data is further used to train the PyTorch-version of NeRF model, named nerf_pl [4].

For Replica format data, we use the output from COLMAP and the converter from the MVSNet repo to convert the custom data into MVSNet format [10]. We then extract the monocular depth and normal information of each image by running a pre-trained Omnidata model following the instructions from the MonoSDF repository. We will train the MonoSDF model with the Replica format custom data.

4.2 Experiment Design

We train 3 NeRF-based models: two are Vanilla NeRF models based on a PyTorch re-implementation of NeRF, named nerf_pl [11], and a MonoSDF model. One NeRF model is trained with the full dataset, and another with a subset of the dataset that contains 35 images from the Unity dataset in LLFF format. Both models are trained with default hyperparameters suggested in nerf_pl, and both models are trained with 20 epochs, the top 5 epochs with the highest PSNR values are saved, and the models with the highest PSNR values are used in the reconstruction comparison. For MonoSDF, we train the model with the default hyperparameters used in the MonoSDF repository. We use the model trained with 500 epochs for the reconstruction evaluation and comparison with the other two NeRF models.

5 Results

We test the reconstruction quality of the 3 trained models on 5 select test images and report the mean, and standard deviation of the PSNR values among the test results, as well as the PSNR value for each test image, as Table 1 shows. We also include the comparison of the rendering results from three models with the ground truth of the test images, demonstrates in Figure 4.

TABLE 1
PSNR Values of Different Reconstruct Methods on Test Images

Method	NeRF	NeRF-35	MonoSDF
Training Epoch	1	1	500
Mean	25.32	17.28	18.89
Standard Deviation	1.30	1.20	2.27
Test Image 0	24.25	16.69	20.97
Test Image 3	26.87	15.80	20.79
Test Image 8	26.82	16.50	20.45
Test Image 45	24.9	18.56	16.25

Figure 5 demonstrates the rendering results from 12 different simulated viewpoints, generated by our best performing model (the NeRF model trained with a full dataset).

6 Discussion

We chose to evaluate the NeRF model with a full dataset and with half the dataset to see how the reconstruction quality changes with the number of input images. As shown, more

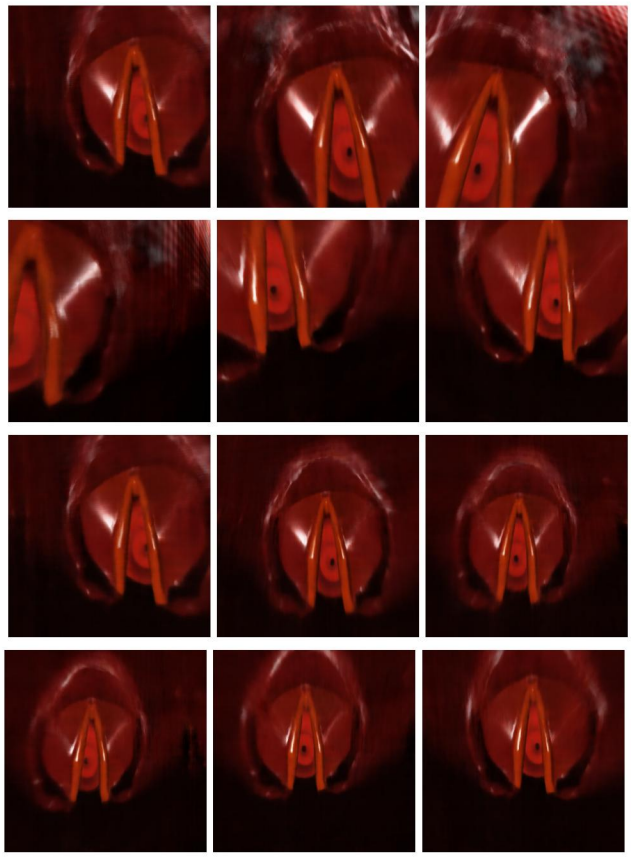

Fig. 5. Rendering Results of NeRF on Simulated Camera Positions

input images improves the overall reconstruction quality- as highlighted by the higher PSNR. Although NeRF-35 is able to render images where the subglottic region is identifiable, the artefacts make it difficult to visualize the stenosis correctly. Additionally, the specular reflections and textureless regions makes the subglottic region difficult to capture with a vanilla-NeRF model.

We also chose to compare the vanilla-NeRF model with the MonoSDF model, to see whether the MonoSDF model would be able to use the normal and depth cues to create a better reconstruction. As outlined in the paper, MonoSDF was intended to help decrease the number of images required in the dataset as well as improve the rendering of textureless regions. Our model was not able to properly learn information about the depth and surface normals. Despite our RGB and Eikonal loss decreasing, the depth consistency and normal consistency losses did not converge during training.

We believe that MonoSDF's poor performance could be attributed to the quality of monocular cues derived from the pretrained Omnidata model or the noise in the camera positions and directions, as determined by COLMAP. The paper mentions high dependency on the monocular cues for model performance. The poor reconstruction results along with the high depth and normal losses, suggest that this may our model's shortcomings. We predict that the depth and normal cues can be better estimated using an Omnidata model, or other monocular cue model, that is trained on endoscopic images, or simulated endoscopic images. The Omnidata model in the MonoSDF repository was trained on regular inputs such as chairs, tables, etc. and therefore may be inadequate for textureless, endoscopic simulations.

6.1 Future Work

To improve the scene reconstruction, we propose testing our dataset on VolSDF, a model that parameterizes the density in neural volume rendering [12]. Since the MonoSDF model is derived from this model, we want to assess whether the rendering issues arise from the noise in the camera positions. The only losses this model considers is the RGB loss and the SDF loss, therefore, it will help us understand what a good starting point will be for model improvement.

Furthermore, MonoSDF is extremely dependent on the quality of extracted depth and normal cues. By retraining the Omnidata model on endoscopic images or simulated endoscopic images, we could anticipate that the model would perform better with this type of information. Therefore, retrying our dataset with an improved Omnidata model could help the model create better reconstructions.

Finally, the subglottic region is dynamic and deformable. During an actual endoscopic examination, the vocal cords are constantly opening and closing, reducing the frames where the stenosis is visible. We should experiment with reg-NeRF, a model that has been specialized to deal with an extremely low number of input images.

7 Conclusion

Clinicians treating subglottic stenosis (SGS) have a difficult time visualizing and measuring the length and width of the disease in the narrow airway. A virtual SGS model with specular reflections and low-textured regions was modeled in Unity with a continuum robot controlling an endoscope. A dataset was constructed to simulate endoscopic images collected during a SGS examination. NeRF was tested on the full and half the dataset, whereas MonoSDF was implemented for the full dataset. NeRF was able to create the best reconstruction results of SGS, whereas MonoSDF was unable to learn the depth and normal information correctly. This project identified that improvements need to be made in existing neural radiance fields methods to accurately capture the textureless and shiny surfaces found in endoscopic videos.

Acknowledgments

We would like to thank Dr. David Lindell for an engaging Computational Imaging course that will be helpful for our graduate research. We would also like to thank Dr. Lueder Kahrs for being our mentor throughout our graduate research and guiding us with our project concept.

References

[1] A. F. Vélez, J. M. Marcinczak, and R. R. Grigat, "Structure from motion based approaches to 3d reconstruction in minimal invasive laparoscopy," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7325 LNCS, pp. 296–303, 2012. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-31298-4_35

[2] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi, "COTR: Correspondence Transformer for Matching Across Images," in *ICCV*, 2021.

[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," in *ECCV*, 2020.

[4] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[5] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 786–10 796.

[6] O. F. Kar, T. Yeo, A. Atanov, and A. Zamir, "3d common corruptions and data augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 963–18 974.

[7] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (TOG)*, 2019.

[8] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[9] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," *European Conference on Computer Vision (ECCV)*, 2018.

[11] C. Quei-An, "Nerf_pl: a pytorch-lightning implementation of nerf," 2020. [Online]. Available: https://github.com/kwea123/nerf_pl/

[12] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.