Context-Aware Depth and Pose Estimation for Bronchoscopic Navigation
Mali Shen\textsuperscript{1*}, Yun Gu\textsuperscript{1*}, Ning Liu\textsuperscript{1}, and Guang-Zhong Yang\textsuperscript{1}

Abstract—Endobronchial intervention is increasingly used as a minimal invasive means of lung intervention. Vision-based localization approaches are often sensitive to image artifacts in bronchoscopic videos. In this paper, a robust navigation system based on a context-aware depth recovery approach for monocular video images is presented. To handle the artifacts, a conditional generative adversarial learning framework is proposed for reliable depth recovery. The accuracy of depth estimation and camera localization is validated on an in vivo dataset. Both quantitative and qualitative results demonstrate that the depth recovered with the proposed method preserves better structural information of airway lumens in the presence of image artifacts, and the improved camera localization accuracy demonstrates its clinical potential for bronchoscopic navigation.

Index Terms—Computer Vision for Medical Robotics; Deep Learning in Robotics and Automation; Visual Learning

I. INTRODUCTION

In recent years, bronchoscopic biopsy technique (Fig. 1) is emerging as a more favourable approach compared to the traditional transthoracic needle aspiration (TTNA) for lung cancer staging [1]. Common post-operative complications of TTNA include pneumothorax, hemorrhage and air embolism [2]. Bronchoscopic biopsy has shown a reduced complication rate than that of TTNA for the diagnosis of solitary pulmonary nodules [1]. However, with the restricted field of view, it remains a challenging task to orient and steer the flexible bronchoscope inside the complex airway networks. Therefore, the development of effective navigation systems is important to assist endobronchial interventions [1]. The key to automatic bronchoscopic camera localization system is estimation of intra-operative six degree-of-freedom (6 DoF) position and orientation of the scope with respect to the airway model reconstructed from pre-operative computed tomography (CT). It can assist not only pulmonologists but also surgical robotic platforms [3], [4] to maintain device orientation towards target airway locations.

Manuscript received: September 9, 2018; Revised November 29, 2018; Accepted December 27, 2018.

This paper was recommended for publication by the Associate Editor and Reviewers’ comments. This work was supported by Engineering and Physical Sciences Research Council (EPSRC), United Kingdom (EP/N019318/1).

\textsuperscript{*}The authors are with the Hamlyn Centre for Robotic Surgery, Imperial College London, SW7 2AZ, London, UK \{mali.shen09, y.gu16, ning.liu13, g.z.yang\}@imperial.ac.uk

\textsuperscript{*}Mali Shen and Yun Gu contributed equally to the paper

Mali Shen, Yun Gu and Ning Liu are supported by Chinese Scholarship Council (CSC).

Fig. 1. Bronchoscopic biopsy technique for lung cancer staging.

Navigation systems based on video-CT registration\textsuperscript{1} and electromagnetic (EM) tracking have been extensively studied to assist bronchoscopic camera localization [5]. The reliability of vision-based techniques mostly rely on the quality of video and CT airway model, while EM-based approaches are sensitive to sensor calibration and field distortion. Compared to EM-based systems, vision-based frameworks also experience smaller influence from tissue deformations [6]. Furthermore, the diameter of endobronchial devices for peripheral airway intervention ideally needs to be less than 2 mm [4], incorporating tracking devices therefore places extra burden on device minimization. The purpose of this paper is focused on the development of a vision-based localization framework. Several 2D/3D registration methods based on different similarity measures [6]–[9] have been developed for bronchoscopic navigation. Among them, the method based on depth maps [9] is shown to be more robust to illumination variations compared to other methods based on intensity or gradient features. However, the depth estimation from bronchoscopic images is still challenging due to the following issues: 1) conventional Shape from Motion (SFM) technique is difficult to properly recover the depth with the small field of view and the lack of salient features; 2) constrained by the small airway dimensions, only monocular bronchoscopes are widely available, making the depth recovery approaches using stereo camera settings inapplicable; 3) matched in vivo video frames and CT virtual views are difficult to obtain because of the time-consuming and error-prone manual alignment, while supervised depth learning algorithms often require a large amount of labelled data; 4) in real clinical scenarios, image artifacts such

\textsuperscript{1}In this paper, we use video-CT registration and 2D/3D registration interchangeably.
as blurring and water bubbles as shown in Fig. 3 are likely to be introduced by patients’ intensive coughing, fast scope steering, and bronchial washout, while most of the existing methods could be very sensitive to those image artifacts.

In this paper, we propose a bronchoscopic navigation framework based on a context-aware depth estimation approach that generates the depth maps from in vivo video frames without paired video and CT depth images. Inspired by the recent development of generative adversarial networks (GAN) [10], [11], view-attribute conditions and neighbourhood warping are introduced to handle the partial occlusion from image artifacts. The estimated depth maps of video frames are then registered to the virtual depth maps from CT data for camera pose estimation. The proposed navigation framework is validated on in vivo datasets. The evaluation on both the depth estimation from video images and the camera localization accuracy in bronchoscopic navigation demonstrate the promising performance of the proposed method.

The remainder of this paper is organized as follows: the related works are discussed in Section II. In Section III, the details of the proposed method are presented. Section IV shows the experiments, followed by the conclusion in Section V.

II. RELATED WORKS

Current bronchoscopic navigation systems can be broadly classified into three main categories including sensor-based [12], vision-based [6], [8], [9] and the hybrid systems [13], [14]. External tracking devices such as EM sensors can be integrated into flexible endoscopes to report real-time tip pose. However, calibration and complicated instrument setup affect normal clinical workflow and its sensitivity to metal interference and tissue deformation is likely to cause tracking failure [6].

In contrast to sensor-based navigation systems, the vision-based approaches rely only on the bronchoscopic videos to perform 2D/3D registration between the video frames and virtual CT views with a specific similarity metric. The image-based registration approaches are more robust to respiratory motion because the scope moves with the airway simultaneously [6]. The similarity metric can be based on image intensity [6], [7], gradient [8], [14], depth [9] or airway lumen features [15]. The similarity in terms of image intensity proposed in [6], [7] requires the generation of realistic renderings from the CT airway model. However, the surface textures of the endobronchial walls and the illumination conditions are difficult to recover. Instead of using image intensity, several attempts transform the registration into pq space [8], [14] or depth maps [9] which are more texture-independent while better preserving the morphology of airway bifurcations. Since pixel-based image registration is computationally expensive, more efficient localization methods based on airway lumen detection [15], [16] are developed for navigation at distal airways where multiple bifurcations can be observed. Among these methods, 2D/3D registration based on depth is invariant to surface texture and it preserves the rich morphological information of the scene.

Recently, depth estimation from monocular images has been tackled by several supervised [17]–[19] or unsupervised [20], [21] deep learning methods. Supervised depth recovery methods such as [17] require pair-wise matching between images and depth data which is hardly available for bronchoscopic video images. To address this problem, [18], [19] applied domain adaption to transfer the video frames to textureless images where the mapping from rendered views to depth maps can be learned with a large number of synthetic data. The model proposed in [18] provides good depth estimation result on bronchial phantom data. Mahmood et al. [19] demonstrates satisfying structure-preserving depth recovery performance on pig colon endoscopic images using the model trained on synthetic data.

Furthermore, unsupervised methods [20], [21] are also proposed for monocular depth estimation without using synthetic data. However, [20] uses stereo images for training which is not applicable for bronchoscopic images. A self-supervised approach proposed in [21] estimates both depth maps and camera pose using view synthesis and SFM. While the method requires smooth camera motion and rich visual information. The lack of salient features on endobronchial surface and fast scope motion may restrict its use for bronchoscopic images.

Our work is also related to recent developments on GAN models [10], [11], [22]. Isola et al. [22] adapted the GAN to a general image-to-image translation framework which requires exact correspondence between multiple domains. To relax this constraint, Zhu et al. [11] proposed an unpaired image-to-image translation framework based on cyclic consistency (CycleGAN). This unpaired framework is appreciated for the depth map generation from bronchoscopic videos. However, the cycle consistency cannot always be preserved with the existence of artifacts. For a specific video frame, the corresponding depth map is fixed while the inverse transformation is not unique. A depth map could be matched with video frames with clear view or artifacts at the same airway location. To address this problem, we construct the complementary conditions to model the artifacts in bronchoscopic images.

III. METHOD OVERVIEW

The workflow of the proposed camera localization method for bronchoscopic navigation is illustrated in Fig. 2(a), which includes three main steps: the extraction of depth maps from multiple virtual camera projections of the CT airway model (Section III-A); the reconstruction of a depth map from a single bronchoscopic video frame based on context-aware generative adversarial learning (Section III-B); and the estimation of the camera pose that maximizes the similarity between the video depth map and the corresponding virtual depth maps (Section III-C).

A. Virtual depth generation from CT data

Chest CT scans are taken pre-operatively from which a 3D bronchial tree is segmented using a 3D active contour segmentation approach provided in ITK-SNAP [23]. The bronchi segmentation is then used to compute the 3D airway mesh. Given a specific camera pose \( p_t \) which is a 6DoF vector at time stamp \( t \), the virtual depth map \( z_C(p_t) \in Z_C \) is generated by modelling the virtual camera with the same
intrinsic parameters as the real video bronchoscope camera where $Z_c$ is the set of CT depth map. To mimic all possible views, the virtual camera pose $p$ is set to follow the airway centerline, computed by fast marching algorithm [24], with the controlled amount of variation. The generated virtual depth maps from CT airway model is used for training the proposed depth estimation network.

### B. Context-aware depth estimation from video frames

Given a specific bronchoscopic video frame $x \in X$, the aim of depth estimation is to learn a mapping $G : X \rightarrow Z$ to generate the corresponding depth map. For a proper depth map $z_t = G(x_t)$ estimated from the video frame at time step $t$, it requires the following properties: 1) The local characteristics of $z_t$ should be similar to the virtual depth map $z_C \in Z_C$ generated by CT airway mesh. 2) The depth map $z_t$ should be informative to indicate the bifurcations of $x_t$, e.g., the lumens are normally with larger depth. 3) The depth map $z_t$ is not easily affected by artifacts. Therefore, a depth map may correspond to the same airway location, however, with different types of artifacts. 4) The visual consistency should be preserved between the depth maps $z_t$ and $z_{t-1}$ from the consecutive video frames.

The properties listed above inspire us to build the context-aware depth estimation as shown in Fig. 2(b). The optimal depth estimation function is obtained via solving the learning framework which consists of the following loss terms:

**Adversarial Loss**: Instead of learning the mapping $G_{\text{Depth}} : X \rightarrow Z$ based on manually paired data, we adopt an unsupervised framework by introducing the inverse mapping $G_{\text{Video}} : Z \rightarrow X$ as well as the discriminators $D_{\text{Depth}}$ and $D_{\text{Video}}$ to identify the real depth and bronchoscopic frames respectively. Following the general settings in adversarial learning [10], adversarial losses are applied to both mapping functions. For $G_{\text{Depth}}$ and $D_{\text{Depth}}$, we express the objective as the cross entropy for classifying the real and synthetic depth data as:

$$
G_{\text{Depth}}, D_{\text{Depth}} = \arg \min_{G_{\text{Depth}}} \arg \max_{D_{\text{Depth}}} \mathcal{L}_{\text{adv}}(G_{\text{Depth}}, D_{\text{Depth}})
= \arg \min_{G_{\text{Depth}}} \arg \max_{D_{\text{Depth}}} \mathbb{E}_{z \sim Z} \log D_{\text{Depth}}(z) + 
\mathbb{E}_{x \sim X}(1 - \log D_{\text{Depth}}(G_{\text{Depth}}(x)))
$$

(1)

where $G_{\text{Depth}}$ tries to generate depth $G_{\text{Depth}}(x)$ that looks similar to real depth from CT data, while $D_{\text{Depth}}$ aims to distinguish between synthetic samples $G_{\text{Depth}}(x)$ and real samples $z \in Z$.

Similar to Eq. (1), the adversarial loss term for video frames $\mathcal{L}_{\text{adv}}(G_{\text{Video}}, D_{\text{Video}})$ is also defined to distinguish the real and synthetic data.

**Cyclic Consistency Loss**: Unlike the supervised methods, the corresponding depth maps of bronchoscopic frames are assumed not available in this work. Considering the similar morphology shared by bronchoscopic frames and depth maps, a good depth estimation should be informative to recover the original bronchoscopic frames which is formulated as a cycle consistency in [11]:

$$
G_{\text{Depth}} = \arg \min_{G_{\text{Depth}}} \mathcal{L}_{\text{cycle}}(G_{\text{Depth}})
= \arg \min_{G_{\text{Depth}}} \mathbb{E}_{x \sim X} \| x - G_{\text{Video}}(G_{\text{Depth}}(x)) \|
$$

(2)

where the error between the original and recovered bronchoscopic frames is minimized.
In order to avoid the bias, the cyclic consistency is also preserved for the estimated video frames via:

\[ G_{\text{Video}} = \arg \min_{G_{\text{Video}}} L_{\text{cycle}}(G_{\text{Video}}) \]
\[ = \arg \min_{G_{\text{Video}}} \mathbb{E}_{z \sim z} \| z - G_{\text{Depth}}(G_{\text{Video}}(z)) \| \]  

(3)

**Conditional Context Loss:** Although the cyclic loss enables the recovery of video frames from depth maps, it cannot handle the artifacts in bronchoscopy images. A depth map may correspond to the same airway location, however, with different types of artifacts. Therefore, we label the video frames with specific artifacts and clear view are illustrated in Fig. 3. By accepting the artifacts and clear view are illustrated in Fig. 3.

The cyclic consistency on \( G_{\text{Video}} \) is also extended as follows:

\[ G_{\text{Depth}} = \arg \min_{G_{\text{Depth}}} L_{\text{condition}}(G_{\text{Depth}}) \]
\[ = \arg \min_{G_{\text{Depth}}} \mathbb{E}_{x \sim X,a} \| x - G_{\text{Video}}(G_{\text{Depth}}(x,a)) \| \]  

(4)

The cyclic consistency on \( G_{\text{Video}} \) is extended as follows:

\[ G_{\text{Video}} = \arg \min_{G_{\text{Video}}} L_{\text{condition}}(G_{\text{Video}}) \]
\[ = \arg \min_{G_{\text{Video}}} \mathbb{E}_{z \sim z,a} \| z - G_{\text{Depth}}(G_{\text{Video}}(z,a)) \| \]  

(5)

In addition, we also extend the discriminator \( D_{\text{Video}} \) for view attributes. Besides distinguishing the real and synthetic data, \( D_{\text{Video}} \) is also designed to identify the view-attributes of the video frame as follows:

\[ D_{\text{Video}} = \arg \min_{D_{\text{Video}}} L_{\text{att}}(D_{\text{Video}}) \]
\[ = \arg \min_{D_{\text{Video}}} -\mathbb{E}_{x \sim X,a} \log D_{\text{Video}}(a|x) \]
\[ -\mathbb{E}_{z \sim Z,a} \log D_{\text{Video}}(a|G_{\text{Video}}(z,a)) \]  

(6)

This is implemented by adding another classification output on top of the original discriminator.

**Consecutive Warping Loss:** Considering the consecutive frames \( x_t \) and \( x_{t-1} \), their corresponding depth maps \( z_t \) and \( z_{t-1} \) should also be similar. Let \( F(x_{t-1}, x_t) \) denote the flowmap between \( x_{t-1} \) and \( x_t \), the warping similarity [25] can be presented as follows:

\[ G_{\text{Video}} = \arg \min_{G_{\text{Video}}} L_{\text{warp}}(x_{t-1}, x_t) \]
\[ = \arg \min_{G_{\text{Video}}} \| x_t - S(x_{t-1}, F(x_{t-1}, x_t)) \| + \| G_{\text{Video}}(x_t) - S(G_{\text{Video}}(x_{t-1}), F(x_{t-1}, x_t)) \| \]  

(7)

where \( S(\cdot, \cdot) \) is the spatial transformation function.

However, due to the presence of image artifacts, the image transformation between the consecutive video frames cannot be always adapted to the depth maps. As shown in Fig. 5, sequential frames with clear views are easy to warp. For images with different types of artifact status, the arbitrary constraints on depth map may lead to singular results. For adjacent frames with the same type of artifacts, the warping error is minimized with smaller weights due to the large variation of the artifact regions. Therefore, the final warping error is formulated as follows:

\[ L_{\text{warp}} = \sum \lambda(x_{t-1}, x_t) L_{\text{warp}}(x_{t-1}, x_t) \]
\[ \lambda(x_{t-1}, x_t) = \begin{cases} 
1 & \text{if } a_t = 1 \text{ and } a_{t-1} = 1, \\
0 & \text{if } a_t \neq a_{t-1}, \\
\gamma & \text{otherwise.}
\end{cases} \]  

(8)

where \( \gamma = 0.01 \) in this paper to achieve best performance.

**Overall Loss:** By combining the loss terms into the unified learning framework, the final optimization problem is formulated as follows:

\[ D_{\text{Depth}} = \arg \max_{D_{\text{Depth}}} L_{\text{adv}}(D_{\text{Depth}}) \]
\[ D_{\text{Video}} = \arg \max_{D_{\text{Video}}} L_{\text{adv}}(D_{\text{Video}}) - L_{\text{att}}(D_{\text{Video}}) \]
\[ G_{\text{Depth}}, G_{\text{Video}} = \arg \min_{G_{\text{Video}}, G_{\text{Depth}}} L_{\text{condition}}(G_{\text{Depth}}) \]
\[ + L_{\text{condition}}(G_{\text{Video}}) \]
\[ + L_{\text{adv}}(G_{\text{Video}}) + L_{\text{adv}}(G_{\text{Depth}}) \]
\[ + L_{\text{warp}}(G_{\text{Video}}) \]

The model is trained in an alternative manner. In each iteration, the generators are firstly fixed to obtain the optimal \( D_{\text{Video}} \) and \( D_{\text{Depth}} \). Then the discriminators are fixed to obtained optimal \( G_{\text{Video}} \) and \( G_{\text{Depth}} \).

**C. Camera pose estimation via 2D/3D registration**

For continuous camera localization, after generating the depth map \( z_t \) of the input video frame \( x_t \), the pose of the camera \( p_t \) is estimated by maximizing the similarity between the video depth map \( z_t \) and the virtual CT depth maps. Given the camera pose \( p_{t-1} \) of previous video frame \( x_{t-1} \) as the initialization, the optimal pose \( p_t \) is determined following the similar strategy in [9] as described in Eq. (9).

\[ p_t = p_{t-1} + \Delta p_{t|t-1} \]
\[ \Delta p_{t|t-1} = \arg \max_{\Delta p_{t|t-1}} \{ SIM(z_t, z_C(p_{t-1} + \Delta p_{t|t-1})) \} \]  

(9)

where \( SIM(\cdot, \cdot) \) is the similarity metric between depth maps.

In this paper, we use Normalized Cross Correlation (NCC) as the similarity measure [9]. Each value of the camera pose change \( \Delta p_{t|t-1} \) corresponds to a unique CT depth map \( z_C \). The optimal pose change \( \Delta p_{t|t-1} \) is estimated by maximizing the NCC between \( z_t \) and \( z_C \). Since the objective function is not fully differentiable, the Powell method [26] was applied as the optimization strategy. The maximum magnification allowed for a parabolic-fit step is set to 100. The default ratio by which successive intervals are magnified is set to 1.62.
A. Dataset and Baselines
The proposed framework is validated on in vivo data. CT scans with voxel spacing of $[0.4, 0.4, 0.5] \text{mm}$ are acquired by Siemens SOMATOM Definition Edge CT scanner. Videos with dimensions $307 \times 313$ and frame rate of $30 \text{fps}$ are captured using Olympus BF-1T260 bronchoscope. The data is acquired in standard clinical routine. The dataset belongs to the Hamlyn lung and bronchoscopy database (Hamlyn Centre, Imperial College), which is under an ethical approval (REC reference: 16/LO/0170, IRAS project ID: 198081) by the research ethics committee, London, UK. We used the paired CT scans and bronchoscopy videos from two subjects for training and data from another subject for testing.

For comparison, Shape from Shading (SFS) [27], supervised FCN [28], SFM [21] and CycleGAN [11] are employed as baselines. The supervised FCN is based on the residual blocks and trained with $350$ manually paired CT depth maps and video images. Due to the deformation and artifacts in the in vivo data, the number of available paired data with precise correspondence is limited. To ensure smooth motion between consecutive frames for training the SFM network, $1520$ video frames are selected using a threshold based on optical flow from the training video sequences. $2951$ video frames and $7200$ CT depth maps are used for training the CycleGAN and the proposed method. For testing, two video sequences consisting of $180$ video frames in total with the corresponding registered camera poses and CT depth maps are used.

B. Implementations
In this paper, the generator is formulated as a fully convolutional neural network based on U-Net [29] architecture. We replace the original bottleneck blocks with a six layer ResNet block [30] to encode rich information as illustrated in Fig. 4. The discriminator is a multi-layer convolutional network based on PatchGAN [22] that identifies the real data on multiple scales (Fig. 4).

We use the Adam solver [31] with a batch size of $1$. The Pytorch\(^\dagger\) framework is adopted to implement the deep convolution neural networks and the experiment platform is a workstation with Xeon E5-2630 and NVIDIA GeForce Titan Xp.

\(^\dagger\)https://github.com/pytorch/pytorch

IV. EXPERIMENTS

C. Validation and Discussion
Both the quality of depth estimation and the accuracy of camera localization are compared between the proposed method and the baselines.

The depth maps recovered from three sample video frames and the registered CT depth maps estimated by the baselines and the proposed method are illustrated in Fig. 6. For clear views such as the one in the first row in Fig. 6, all the methods are able to recover the overall depth map reasonably well with the correct size and location of the bifurcation except SFM. The SFM approach performs poorly on bronchoscopic video images because of the fast and unstable maneuvering of the scope. Unlike the outdoor dataset in which the camera moves at a relatively constant speed with small inter-frame translational and rotational variation in one direction, the bronchoscope moves with large translational and rotational

---

Fig. 3. Sample clear view and frames with image artifacts. (a) colour blurring, no occlusion of lumens; (b) bubbles, no occlusion of lumens; (c) bubbles with occlusion; (d) colour stripes; (e) foggy blurring; (f) bubbles causing specular reflection; (g) overlapping shadow and (h) clear view.

Fig. 4. The neural network architecture for (a) the generator, (b) residual block and (c) discriminator. $M \times M \text{conv (N}_a, N}_b)$ denotes a 2D convolution layer with $M \times M$ kernel. $N_a$ input channels and $N_b$ output channels. $M \times M \text{deconv (N}_a, N}_b)$ denotes a 2D deconvolution layer. ‘IN’ denotes the instance normalization layer. ‘RELU’ denotes the ReLU function. ‘BN’ denotes the batch normalization layer.
movement in all directions. SFS is based on a Lambertian surface model which estimates the location of the lumens correctly but with inaccurate lumen boundary. The performance of the FCN method is limited by the small number of paired data and the large tissue deformation in video images. For frames with artifacts, all the other baseline methods are jeopardized while the depth estimated by the proposed method still preserves good lumen structure.

The similarity measure including Peak-Signal-Noise-Ratio (PSNR), Structural SIMilarity (SSIM), NCC and mutual in-
corresponding ground-truth CT depth. This is consistent with the most accurate registered CT depth compared with the methods as shown in Table I. All similarity measures indicate the proposed depth estimation approach produces the smallest ATE and the highest number of successfully registered video frames. Overall, the proposed context-aware depth estimation approach recovers clear structural characteristics from images without image artifacts greatly improves the robustness of the navigation framework for clinical use.

5A video demonstrating the registration results for the whole video sequence can be found in the supplementary materials.

camera localization. SFS, FCN and CycleGAN approaches can estimate the structural characteristics from images without image artifacts but any presence of bubbles and blur can deteriorate their performance for continuous camera tracking. Our proposed context-aware depth estimation approach takes into account the effect from image artifacts, making the camera localization framework much more robust. Apart from SFS, the depth estimation methods based on deep neural networks all have promising time efficiency for real-time application. In the future, GPU acceleration will be incorporated to reduce the processing time required for video-CT registration.

V. CONCLUSION

In summary, we have proposed a context-aware depth estimation approach for bronchoscopic navigation with satisfying tolerance to image artifacts. The conditional generative adversarial networks with cycle-consistency are capable of learning the specific features of real depth maps from the unpaired data. The validation results on the in vivo dataset collected from the real bronchoscopy routine demonstrate the superior performance of our proposed approach in comparison with the baseline approaches. Our method is capable of generating depth maps from video frames that encodes the structural characteristics of the size and presence of bifurcations even under partial occlusion by image artifacts. Its resilience to image artifacts greatly improves the robustness of the navigation framework for clinical use.

ACKNOWLEDGMENT

The authors would like to thank Prof Pallav Shah at Royal Brompton Hospital, UK for the collaboration in developing the Hamlyn lung and bronchoscopy database.

REFERENCES

TABLE I
QUANTITATIVE EVALUATIONS ON DEPTH ESTIMATION, NAVIGATION AND TIME EFFICIENCY. ‘V’ DENOTES VIDEO FRAMES. ‘D’ DENOTES THE DEPTH MAPS. ‘C’ DENOTES THE EXACT CORRESPONDENCE BETWEEN VIDEO FRAMES AND DEPTH MAPS. ‘SR-N’ DENOTES PERCENTAGE OF VIDEO FRAMES REGISTERED WITH ATE LESS THAN N MM. ‘FPS’ DENOTES FRAMES PER SECOND.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Supervision</th>
<th>Depth Map</th>
<th></th>
<th></th>
<th></th>
<th>Navigation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
<td>NCC ↑</td>
<td>MI ↑</td>
<td>fps</td>
</tr>
<tr>
<td>SFS</td>
<td>-</td>
<td>29.83±3.12</td>
<td>0.89±0.08</td>
<td>0.86±0.09</td>
<td>0.92±0.33</td>
<td>0.12</td>
</tr>
<tr>
<td>SFM</td>
<td>V</td>
<td>17.59±7.08</td>
<td>0.62±0.22</td>
<td>0.12±0.29</td>
<td>0.57±0.28</td>
<td>75.88</td>
</tr>
<tr>
<td>FCN</td>
<td>V+D+C</td>
<td>29.51±1.69</td>
<td>0.92±0.04</td>
<td>0.84±0.07</td>
<td>0.89±0.28</td>
<td>50.00</td>
</tr>
<tr>
<td>CycleGAN</td>
<td>V+D</td>
<td>29.03±1.88</td>
<td>0.91±0.05</td>
<td>0.82±0.08</td>
<td>0.85±0.29</td>
<td>50.00</td>
</tr>
<tr>
<td>Ours</td>
<td>V+D</td>
<td>34.15±4.05</td>
<td>0.92±0.09</td>
<td>0.93±0.07</td>
<td>1.22±0.46</td>
<td>50.00</td>
</tr>
</tbody>
</table>

Fig. 8. The registered virtual CT views for one of the testing video sequence by the proposed method and baselines in comparison with ground truth.