
Article:
Zhao, C., Shen, M., Sun, L. orcid.org/0000-0002-0393-8665 et al. (1 more author) (2020) Generative localisation with uncertainty estimation through video-CT data for bronchoscopic biopsy. IEEE Robotics and Automation Letters, 5 (1). pp. 258-265.

https://doi.org/10.1109/LRA.2019.2955941

Generative Localisation with Uncertainty Estimation through Video-CT data for Bronchoscopic Biopsy

Cheng Zhao[1*], Mali Shen[1], Li Sun[2], *Member, IEEE,*
and Guang-Zhong Yang[1,3], *Fellow, IEEE*

***Abstract*—Robot-assisted endobronchial intervention requires accurate localisation based on both intra- and pre-operative data. Most existing methods achieve this by registering 2D videos with 3D CT models according to a defined similarity metric with local features. Instead, we formulate the bronchoscopic localisation as a learning-based global localisation using deep neural networks. The proposed network consists of two generative architectures and one auxiliary learning component. The cycle generative architecture bridges the domain variance between the real bronchoscopic videos and virtual views derived from pre-operative CT data so that the proposed approach can be trained through a large number of generated virtual images but deployed through real images. The auxiliary learning architecture leverages complementary relative pose regression to constrain the search space, ensuring consistent global pose predictions. Most importantly, the uncertainty of each global pose is obtained through variational inference by sampling within the learned underlying probability distribution. Detailed validation results demonstrate the localisation accuracy with reasonable uncertainty achieved and its potential clinical value.**

***Index Terms*—Computer Vision for Medical Robotics, Medical Robots and Systems, Localization, Visual Learning, Deep Learning in Robotics and Automation**

I. Motivation

LUNG cancer is now the leading cause of cancer-related death world-wide, and an efficient early diagnosis is in high-demand. A more favourable approach, the bronchoscopic biopsy is an emerging technology for lung cancer diagnosis staging. During the bronchoscopic biopsy, the physician needs to estimate the intra-operative position and orientation of the scope through the intra-operative 2D image from camera in the coordinates of the preoperative 3D computed tomography (CT) scan. The conventional bronchoscopic localisation approaches [1][2][3][4] are mainly based on video-CT registration or electromagnetic (EM) tracking. The localisation error gradually accumulates when using a continuous video-CT registration approach whose current camera pose is initialised on the basis of a prior guess given by previous image observation. Due to the incremental nature of video-CT registration approaches, continuous tracking is likely to be suspended when one registration fails due to paucity of airway features. For the EM-based localisation, the EM field can be distorted by any presence of surrounding ferromagnetic objects. Therefore, special setup in the operating room is required for deploying any EM-based localisation system for bronchoscopic examination. Moreover, the existing conventional bronchoscopic localisation approaches only provide the position prediction but without the corresponding uncertainty estimation.

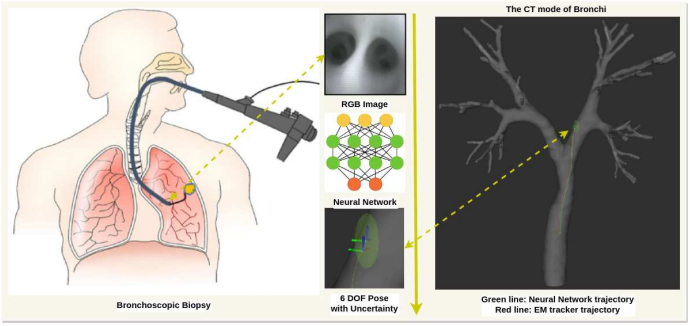

Fig. 1: The pipeline of the proposed system.

Recently, the deep learning significantly improved the visual-based localisation [5][6] and mapping [7][8] in the computer vision and robotics community. Also, the deep learning-based approaches [9][10][11][12] have achieved remarkable performance in a variety of clinical applications. Inspired by these researches, this letter investigates the ability of deep neural network (DNN) to advance the vision-based bronchoscopic localisation, as shown in Figure 1. For the bronchoscopic localisation, we formulate the conventional 2D/3D registration problem as a data-driven learning problem. To be specific, we resolve the bronchoscopic localisation as a continuous global pose estimation problem through deep regression. Comparing with conventional methods, it can significantly eliminate the localisation drift caused by continuous video-CT registration.

Different from the conventional global localisation methods [13][14] that utilise the same sensor for both observation and mapping, bronchoscopic localisation uses the camera to obtain intra-operative observation while relies on preoperative CT scans to generate the global 3D map. Therefore, there is a domain variance between the current observation reading i.e. RGB image and the global map i.e. CT scans. Moreover, it is very challenging to generate a huge number of training data with ground truth for the DNN training in the clinical

Manuscript received: September, 10, 2019; Revised November, 22, 2019; Accepted November, 10, 2019.

This paper was recommended for publication by Editor Pietro Valdastri upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Engineering and Physical Sciences Research Council (EPSRC), U.K. under Grant EP/N019318/1. *(*Corresponding author: Cheng Zhao.)*

[1]Cheng Zhao, Mali Shen and Guang-Zhong Yang are with Hamlyn Centre, Imperial College London, UK cheng.zhao@imperial.ac.uk; mali.shen09@imperial.ac.uk; g.z.yang@imperial.ac.uk

[2]Li Sun is with Oxford Robotics Institute, University of Oxford, UK kevin@robots.ox.ac.uk

[3]Guang-Zhong Yang is with Institute of Medical Robotics, Shanghai Jiao Tong University, China gzyang@sjtu.edu.cn

Digital Object Identifier (DOI): see top of this page.

application. In order to solve the above two problems, we employed a domain transfer learning network to transfer both RGB image and CT depth to the same virtual RGB domain. In this case, the proposed deep network not only can solve the domain variance between the different sensor readings but also can be trained through the virtual RGB images while deployed on the real RGB images.

Another challenge of bronchoscopic localisation is non-unique image readings, and in other words, there are locations with similar appearances in the endobronchial environment. It is very challenging to eliminate the ambiguity through a single image measurement. Therefore, leveraging the previous motion information can provide a geometric consistency with current measurement for pose estimation. Inspired by [14], an auxiliary learning approach that jointly optimizes relative and global pose is adopted to constrain the search space for the global pose regression.

Lastly and most importantly, different from the general global localisation problem, the uncertainty estimation has critical importance for the bronchoscopic localisation when being applied in the real clinical scenario. The uncertainty estimation indicates the confidence level of the scope position predicted by the neural network for the physician. The uncertainty of the DNN model, i.e. $epistemic$ uncertainty usually can be attributed to distinguish two different sites with similar appearances or to extrapolate unknown locations with insufficient training data for training a discriminative model. Our intuition is to leverage variational inference to approximate the localisation posteriori using a variation distribution in the latent space. Benefiting from the sampling mechanism over the latent variable distribution, the $espistemic$ uncertainty of the prediction can be naturally modelled.

In summarise, our novel contributions for learning-based bronchoscopic localisation are: 1) as the domain variance is bridged by the domain transfer learning, the proposed localisation network can be trained through virtual data but deployed on the real data, 2) an auxiliary learning architecture can constrain the search space of global pose prediction to guarantee consistent predictions, 3) the deep variational regression can estimate the predictive uncertainties through sampling within the learned underlying probability distribution. To the best of our knowledge, this is the first learning-based localisation with uncertainty estimation for the bronchoscopic localisation. A video demo is available on the website[1].

II. Related work

The most recent research effort in bronchoscopic localisation mainly depends on 2D/3D image registration or EM tracking. The visual bronchoscopic localisation methods can be broadly grouped into two main categories including geometry-based localisation and learning-based localisation.

A. Geometry-based localisation

The conventional approaches of visual bronchoscopic localisation reply on 2D/3D registration between the intra-operative video frames and the pre-operative CT airway model. Various similarity measures based on image intensity [1], gradient [15][2], depth [3] or airway lumen features [16][17][4] have been investigated to improve the registration accuracy.

Similarity metrics based on image intensity [1] require generating realistic virtual bronchoscopic views using CT airway model. However, the illumination effects caused by the moving light source and endoluminal surface texture are difficult to recover in the rendered views. Since intensity-based measures often suffer from the paucity of image features and illumination artefacts, texture invariant gradient-based similarity metrics [15][2] and a depth-based similarity metric [3] are applied for comparing the structural similarity between video images and CT virtual views. Furthermore, rather than using the time consuming pixel-wise similarity measures, more efficient registration approaches based on airway lumen feature matching have been proposed in [17][16][4] to perform real-time bronchoscopic localisation. In addition, a feature-based visual SLAM [18] is also investigated for localisation in the endobronchial environment. However, the visual odometry approach is prone to insufficient visual features such as SIFT or ORB for tracking.

B. Learning-based localisation

Recently, the deep learning-based approaches significantly improve the depth estimation from monocular images. Therefore, some research [9][10][11] employ the DNN to estimate the depth, which further can be registered to the 3D CT scan for localisation. However, generating large annotated *in vivo* datasets for DNN training is difficult due to ethical issues and the labour-intensive labelling process, so some research [9][10][12] try to train the DNN through generated synthetic images.

Marco *et al.* [9] take advantage of two CNNs to transfer the real image to rendering image and then map the generated rendering image to depth image. Faisal *et al.* [10] propose a reverse domain adaptation to make the real image look like the synthetic image through adversarial training. With the domain adaptation process, a large number of synthetic images can be used to train the depth estimation network. However, those methods only focus on the depth estimation but the camera localisation is not investigated.

Our previous work [11] proposes a context-aware depth recovery approach through a CycleGan-like network trained using unpaired videos and CT depth maps. The camera pose is estimated through maximising the similarity between predicted video depth and CT depth. However, the localisation part is still dependent on the geometry registration rather than neural network.

OffsetNet [12] employs DNN to regress the 6 DOF relative pose between two adjacent real and rendering images, and further accumulate them to generate the whole trajectory of camera. The performance of proposed network is improved by augmenting the training data with rendering images. The domain gap between the real and rendering RGB images is bridged by a generative adversarial network. However, the localisation error will accumulate gradually when estimating the global pose from a sequence of predicted relative poses along

[1]https://www.dropbox.com/home/RAL2020_demo?preview=RAL2020_demo.mp4

a long trajectory. The localisation accuracy will be severely jeopardised for the subsequent frames after a tracking failure occurs due to such incremental pose estimation approach.

III. Methodology

A. Overview

The overview of the proposed neural network for bronchoscopic localisation is shown in Figure 2. It consists of two generative architectures and one auxiliary learning architecture. The first generative architecture is a cycle generative network for the domain transfer learning between the real and virtual RGB image. The second generative architecture is a variational inference for the uncertainty estimation of predicted global pose. An auxiliary learning architecture is a dual-stream network to jointly optimise the relative and global pose during training.

B. Domain Transfer Learning

The pipeline of the proposed domain transfer between the real RGB image and the CT depth is illustrated in Figure 3.

1) CT Depth and Virtual RGB Generation: Using the 3D active contour segmentation approach provided in ITK-SNAP [19], the 3D bronchial tree is segmented from the pre-operative chest CT scans. The 3D airway mesh is further generated as a global map from this bronchial segmentation. Given a specific 6 DOF camera pose, a corresponding depth image can be generated by modelling a virtual camera with the same intrinsic parameters as the real bronchoscope camera. Finally, a virtual RGB image is rendered from the depth image using VTK[2]. A large amount of virtual RGB images can be generated for the localisation network training.

2) Cycle Generative Architecture: In order to make the deep model trained on virtual images and deployable on real images, a CycleGAN-like architecture is employed to bridge the domain variance between the real RGB image $x \in X$ and the virtual RGB image $y \in Y$. LSGAN [20] loss $\mathcal{L}_{LSGAN}$ is adopted to achieve a stable adversarial training to generate high-quality images. To empower the network to learn the transformation of an individual input to the desired output domain from unpaired data, the cycle consistency loss $\mathcal{L}_{cyc}$ in the original CycleGAN [21] is adopted to reduce the space of possible mapping functions, which makes the learned mapping function cycle-consistent. Moreover, a consecutive warping loss $\mathcal{L}_{warp}$ in our previous work [11] is incorporated to provide a spatial transformation constrain between two adjacent frames of both real and virtual images. The overall loss $\mathcal{L}$ of the domain transfer learning is formulated as,

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{LSGAN}((G_{X\to Y}, D_Y, G_{Y\to X}, D_X)(x_t, y_t)) \\ &+\mathcal{L}_{cyc}((G_{X\to Y}, G_{Y\to X})(x_t, y_t)) \\ &+\mathcal{L}_{warp}(G_{X\to Y}(x_{t-1}, x_t)),\end{aligned} \tag{1}$$

where G is the generator and D is the discriminator. The training data of the cycle generative architecture are unpaired real RGB video and rendering virtual RGB video. The generator and discriminator have the same architectures as those in the CycleGAN [21].

[2]https://vtk.org/doc/nightly/html/classvtkRenderWindow.html

C. Auxiliary Learning

Auxiliary learning which leverages the complementary relative pose regression to constrain the search space of global pose regression, is employed to alleviate the non-unique image reading problem. To be specific, the propose network is a dual-stream architecture including a global pose regression stream and a relative pose regression stream.

During training, the network can predict the global pose $Pose_g = [p_t, q_t]$ and the relative pose $Pose_r = [p_{t,t-n}, q_{t,t-n}], n \in [1, 10]$ through a pair of images, i.e. current image I_t and one image randomly selected from the previous ten images $I_{t-n}, n \in [1, 10]$. In the deployment, the network can predict the global pose $Pose_g = [p_t, q_t]$ with the corresponding uncertainty u_t only given the current image I_t. The position $p \in \mathbb{R}^3$ is described by a 3D position (x, y, z) and the orientation $q \in \mathbb{R}^4$ is described by a quaternion (q_w, q_x, q_y, q_z).

1) Global Pose Learning: In order to estimate both translational and rotational pose components, the loss function of global pose regression is defined as

$$\mathcal{L}_p(I_t) = \| \, p_t - \widehat{p}_t \, \|_2, \tag{2}$$

$$\mathcal{L}_q(I_t) = 1 - |q_t \cdot \widehat{q}_t|, \tag{3}$$

$$\mathcal{L}_G(I_t) = \mathcal{L}_p(I_t) + \lambda_1 \mathcal{L}_q(I_t). \tag{4}$$

$\widehat{p}_t$ and $\widehat{q}_t$ denote the predicted translational and rotational global pose through current image I_t, p_t and q_t denote the corresponding ground-truth translational and rotational global pose. λ_1 is the scale factor to balance the weights of translation and orientation. ℓ_2 loss is adopted for the position regression and the inner product loss is applied for the orientation regression to mitigate the gimbal problem.

2) Relative Pose Learning: The relative pose is regressed using a pair of images directly, and it is also optimised by the predicted global poses from the global pose network. Integrating the predicted global poses for the relative pose regression can provide a strong geometric consistency to the network training. The loss function of relative pose regression is defined as,

$$\begin{aligned}\mathcal{L}_p(I_t, I_{t-n}) = &\| \, p_{t,t-n} - \widehat{p}_{t,t-n} \, \|_2 \\ &+ \| \, p_{t,t-n} - (\widehat{p}_t - \widehat{p}_{t-n}) \, \|_2,\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}_q(I_t, I_{t-n}) = &1 - |q_{t,t-n} \cdot \widehat{q}_{t,t-n}| \\ &+1 - |q_{t,t-n} \cdot (\widehat{q}_{t-n}^{-1} \cdot \widehat{q}_t)|,\end{aligned} \tag{6}$$

$$\mathcal{L}_R(I_t, I_{t-n}) = \mathcal{L}_p(I_t, I_{t-n}) + \lambda_2 \mathcal{L}_q(I_t, I_{t-n}), \tag{7}$$

where $n \in [1, 10]$. $\widehat{p}_{t,t-n}$ and $\widehat{q}_{t,t-n}$ denote the predicted translational and rotational relative pose between the current image I_t and one image randomly selected from previous ten images I_{t-n}. $p_{t,t-n}$ and $q_{t,t-n}$ denote the corresponding ground-truth translational and rotational relative pose. $\widehat{p}_t$, $\widehat{q}_t$ and $\widehat{p}_{t-n}$, $\widehat{q}_{t-n}$ denote the predicted global position and orientation through the current image I_t and previous image I_{t-n}. λ_2 is also the scale factor to balance the weights of translation and orientation.

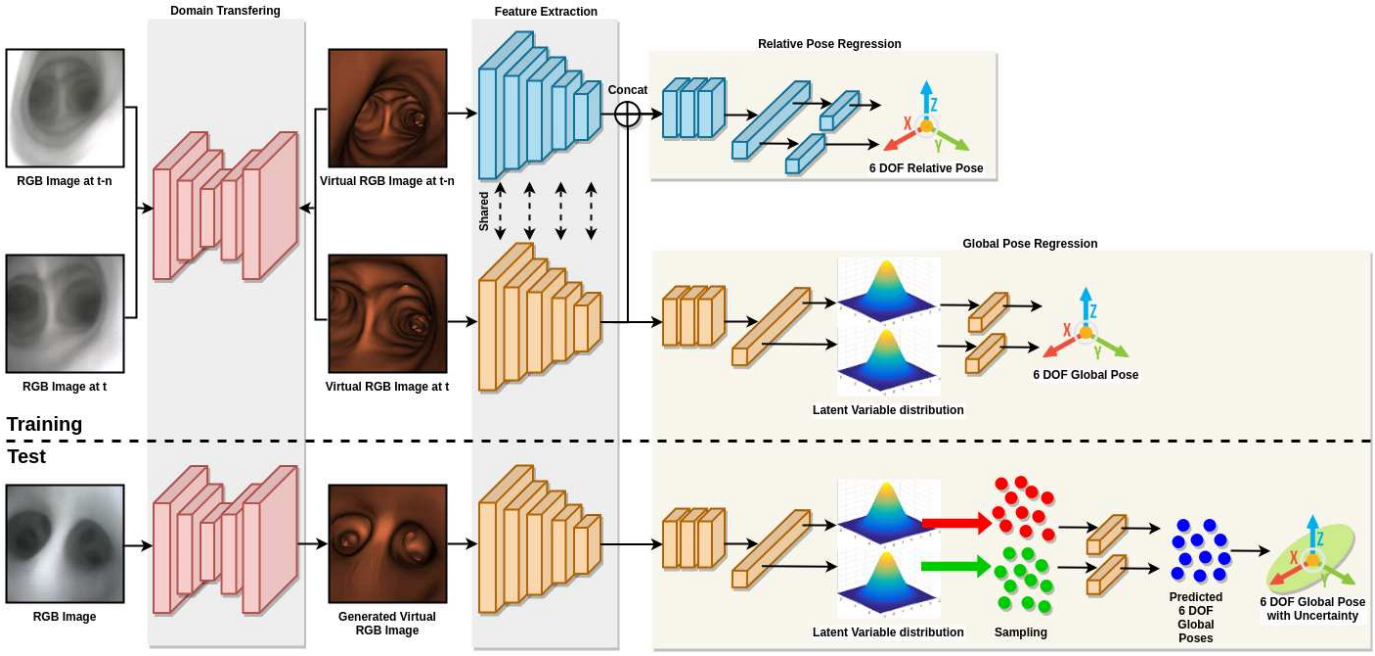

Fig. 2: The architecture of proposed neural network.

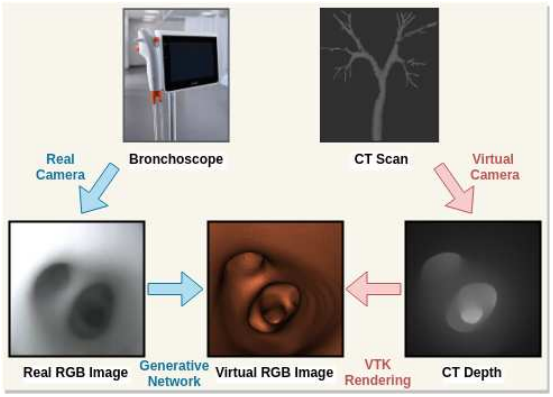

Fig. 3: The pipeline of proposed domain transfer between the real RGB image and the CT depth.

Jointly performing the global and relative pose regression enables collaborative learning between sub-networks during backpropagation to minimise the loss function. The neural network can learn spatially consistent features to constrain the search space of global pose estimation, which can guarantee consistent global pose predictions. During the test, the localisation network can be deployed jointly using both global and relative pose sub-networks or individually using only global relative pose sub-network, because there is no inter-network dependency between the global and relative pose sub-networks. In order to boost the runtime performance and lighten the size of the trained model for real-time bronchoscopic operation, only the global pose sub-network is employed for the bronchoscopic localisation. Therefore, only the current image is required instead of a pair of images during the test. By only using the current image as input, the model can also automatically rescue the "kidnapped robot" (lost and recover) for more robust localisation.

D. Uncertainty Modelling

The variational inference rooted in Bayesian inference is employed to estimate the uncertainty of each global pose prediction. It can model the probability distribution of the observation as a variational distribution in the latent variable space so that the prediction can be made by marginalising the estimated posteriori.

Our objective is to predict the 6 DOF global pose $pose = [p, q]$ with uncertainty u given the observation I. This can be achieved by introducing a normally distributed latent variables ξ whose mean and variance are approximated by neural network. Then the predictive probability can be obtained by marginalising over ξ,

$$P(pose|I) = \int P(pose|\xi)P(\xi|I)d\xi. \tag{8}$$

However, this integral is analytically intractable. This can be addressed by Monte-Carlo sampling with reparameterisation trick i.e. variational inference. The problem is to minimise the $\mathcal{KL}$ divergence between the true distribution $P(\xi|I)$ and variational distribution $Q(\xi|I)$. By the Bayesian rule, the objective function of variational inference is:

$$\mathcal{KL}(Q(\xi|I)||P(\xi|I)) \geq \mathcal{KL}(Q(\xi|I)||P(\xi)) - \mathbb{E}[logP(pose|\xi)], \tag{9}$$

where the first term is the $\mathcal{KL}$ divergence between the variational distribution and the prior distribution of ξ, and the second term is the negative log likelihood of the prediction. A simple normal distribution $\mathcal{N}(0, 1)$ is used as the prior distribution of ξ, and the variational distribution $Q(\xi|I)$ is also a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, where μ, Σ are obtained from logits of the neural network ($encoder$). Hence, the first term can be expressed as,

$$\mathcal{KL}(Q(\xi|I)||P(\xi)) = \mathcal{KL}(\mathcal{N}(\mu, \Sigma)||\mathcal{N}(0, 1)). \tag{10}$$

This $\mathcal{KL}$ divergence can be resolved by maximisation of Evidence of Lower Bound (ELBO):

$$\frac{1}{2}\sum_{k}(\Sigma+\mu^2-1-log\Sigma)-\mathbb{E}[logP(pose|\xi)]. \quad (11)$$

Above, k is the dimension of the Gaussian distribution. In practice, we use two $\mathcal{KL}$ distances $p_{\mathcal{KL}}$, $q_{\mathcal{KL}}$ for the global position p and orientation q regression separately in the loss function,

$$\mathcal{L}_{G_{\mathcal{KL}}}(I_t)=\mathcal{L}_{p_{\mathcal{KL}}}(I_t)+\lambda_3\mathcal{L}_{q_{\mathcal{KL}}}(I_t). \quad (12)$$

λ_3 is also the scale factor to balance the weights of translation and orientation. λ_1, λ_2 and λ_3 in Equations 4, 7 and 12 are chosen according to the experience. Considering the loss functions mentioned in Section III-C, the overall loss function of localisation network is defined as,

$$\mathcal{L}=\mathcal{L}_G(I_t)+\mathcal{L}_{G_{\mathcal{KL}}}(I_t)+\mathcal{L}_R(I_t,I_{t-n}), \quad (13)$$

where $n\in[1,10]$. During inference, the predicted global poses are obtained through sampling N times within the variational distribution $Q(\xi|I)$ of latent variables. The final global pose $pose$ with the corresponding uncertainty u can be obtained through,

$$\xi_1,\xi_2,...,\xi_N\sim\mathcal{N}(\mu,\Sigma), \quad (14)$$

$$pose=\frac{1}{N}\sum_{i=1}^{N}decoder(\xi_i), \quad (15)$$

$$u=\frac{1}{N}\sum_{i=1}^{N}(decoder(\xi_i)-pose)^2. \quad (16)$$

E. Network Architecture

As shown in Figure 2, the sub-networks of global and relative pose regression have similar architecture, including a fast feature extraction, a convolution stack and two dense connected stacks. The differences are that two variational units are inserted within the two dense connected stacks respectively in the sub-network of global pose regression.

During the bronchoscopic localisation, the real-time operation is required to enable closed-loop control. Therefore, the MobileNet V2 [22] truncated before the last pooling layer is employed for the feature extraction. Its architecture is comprised of a sequence of inverted residual blocks. MobileNet V2, which is tailored for the mobile computational resource, can retain a satisfying accuracy and meanwhile significantly decrease the memory requirement and the number of operations. The sub-networks of global and relative pose regression share the same weights respect to the visual feature extraction.

The convolution stack consists of a sequence of convolution layers with 3 $\times$ 3 filter and the stride of 1. The numbers of their channels are 512, 256 and 64. Two four-layer dense connected stacks are used for position and orientation regression separately. The numbers of their hidden dimensions are 256, 128, 64 and 32. The sub-networks of global and relative pose regression have the same architectures of convolution and dense connected stacks but without sharing the weights.

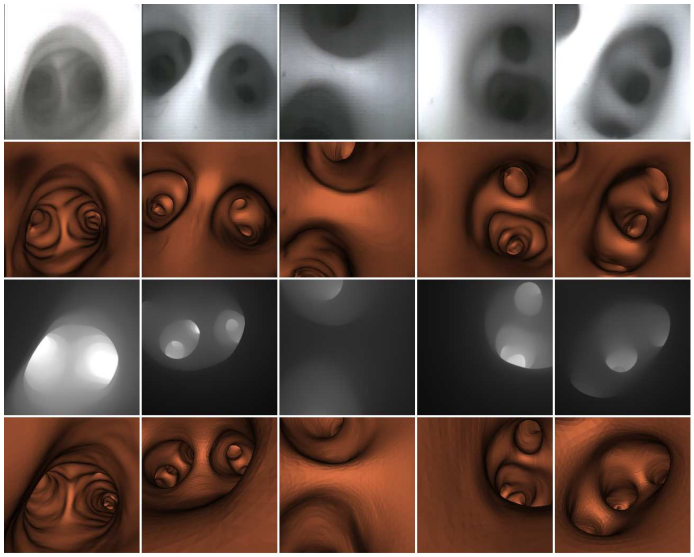

Fig. 4: Examples of the generated virtual RGB images through domain transfer learning. First row: real RGB images from Bronchoscope, second row: generated virtual RGB images by neural network, third row: CT depths from virtual camera, fourth row: virtual RGB images rendered by VTK. The network of domain transfer learning is trained by the unpaired images.

IV. EXPERIMENTS

A. Sensors and Calibration

In this letter, an airway phantom is used to collect data for the evaluation. The real RGB videos with dimensions of 307×313 and frame rate of 30 fps are collected using an Ambu[3] bronchoscope. High resolution CT scans with voxel spacing of [0.4, 0.4, 0.5] mm is acquired for the airway phantom by the Siemens SOMATOM Definition Edge[4]. The trajectories of bronchoscopic camera are captured using a 6DOF NDI Aurora EM sensor[5] as a reference for the comparison. The precision of the EM sensor is around $0.80mm$ for position and 0.70° for orientation, which depends on the tool design and the presence of metal.

The transformation between the CT scan and EM tracker is calibrated using a Matlab toolbox ABSOR[6]. 30 CT markers are placed on the airway phantom to acquire their 3D positions in those two coordinate systems. A $7mm\times6mm$ checkerboard with the size of $17.5mm\times15mm$ is used for the calibration between the EM tracker and camera. The Matlab Camera Calibration toolbox[7] is adopted to compute the intrinsic parameters of the camera and the transformation matrix between the EM tracker to the camera.

[3]https://www.ambu.com/products/flexible-endoscopes/bronchoscopes/product/ambu-ascope-3-large

[4]https://www.siemens-healthineers.com/en-uk/computed-tomography/single-source-ct/somatom-definition-edge

[5]https://www.ndigital.com/medical/products/aurora/

[6]https://uk.mathworks.com/matlabcentral/fileexchange/26186-absolute-orientation-horn-s-method

[7]http://www.vision.caltech.edu/bouguetj/calib_doc/

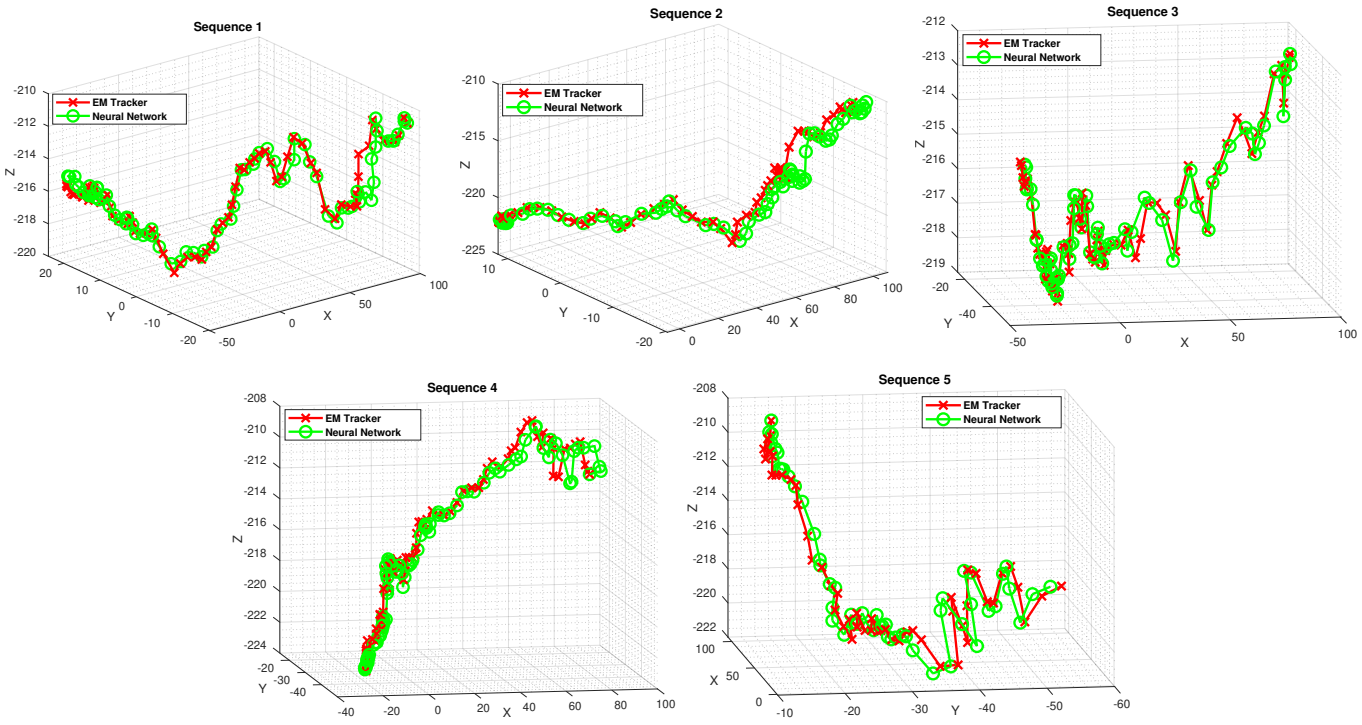

Fig. 5: The predicted 3D trajectories of proposed method on sequence 1-5.

Method	Sequence 1		Sequence 2		Sequence 3		Sequence 4		Sequence 5		Average	
	$t(mm)$	$r(^\circ)$	$t(mm)$	$r(^\circ)$	$t(mm)$	$r(^\circ)$	$t(mm)$	$r(^\circ)$	$t(mm)$	$r(^\circ)$	$t(mm)$	$r(^\circ)$
Geometry Registration[1][3]	9.67	16.74	5.72	12.34	3.23	19.81	3.99	**8.10**	5.25	14.98	5.57	14.39
Learning localisation[2]	**1.83**	**12.28**	**2.07**	**10.68**	**1.27**	**9.25**	**1.23**	10.15	**1.79**	**11.10**	**1.64**	**10.69**
Learning localisation[3]	1.65	10.92	1.32	9.87	0.83	9.01	0.97	8.50	1.06	10.23	1.17	9.71

$t(mm)$ and $r(^\circ)$ are median translational position offset and median rotational angle offset. [1] is tested using real images, [2] is trained using virtual images and tested using real images, [3] is both trained and tested using virtual images.

TABLE I: The performance comparison between the proposed method and geometry registration baseline on sequence 1-5.

TABLE II: The ablation analysis for the overall performance of localisation. B: CNN feature extraction with dense regressor, A: auxiliary learning, V: variational inference.

Architecture	$t(mm)$	$r(^\circ)$
B	3.27	11.83
B+A	1.77	10.96
B+A+V	**1.49**	**10.81**

$t(mm)$ and $r(^\circ)$ are median translational position offset and median rotational angle offset. The network is trained using virtual images and tested using real images.

B. Dataset Generation and Network Training

Using the virtual camera as mentioned in the Section III-B1, 500 video sequences of CT depths with different lengths, totally 43743 frames, are captured from the 3D CT model. Then those CT depth images are transformed to the virtual RGB images by VTK rendering. 150 videos of real RGB images with different lengths, totally 14259 frames, are captured by the bronchoscope from the phantom. Those unpaired real-virtual RGB videos are employed to train the network of domain transfer learning. Both the geometric augmentation (translation, rotation, scaling) and image augmentation (colour, brightness, gamma) are adopted for data augmentation.

The network of domain transfer learning is trained for 200 epochs with a batch size of 1. The input images are cropped to 256×256. The linear learning policy is adopted during training. The initial learning rate is 2e-4 and the momentum is fixed to 0.5. The size of the image buffer is set to 50, which stores previously generated images. Some selected generated virtual RGB images are shown in Figure 4. It can be seen that the network of domain transfer learning achieves the satisfying results for the style transfer between the real and virtual RGB images.

For the data generation for training the localisation network, 45 long videos of CT depths with the virtual 6 DOF ground truth trajectories are generated by the virtual camera through the 3D CT model. Similarly, those CT depth images are also converted to virtual RGB images by VTK rendering. In order to simulate the real trajectories generated by the physician's operation, noises are added to the virtual trajectories. For the test data of the localisation network, 5 long videos of real RGB images are captured using the bronchoscope from the phantom. The corresponding 6 DOF trajectories (after calibration) are obtained by EM tracker as a reference for the evaluation.

The network of localisation is trained for 900 epochs with a batch size of 64. The input images are cropped to 256×256. The step learning policy is employed and the learning rate de-

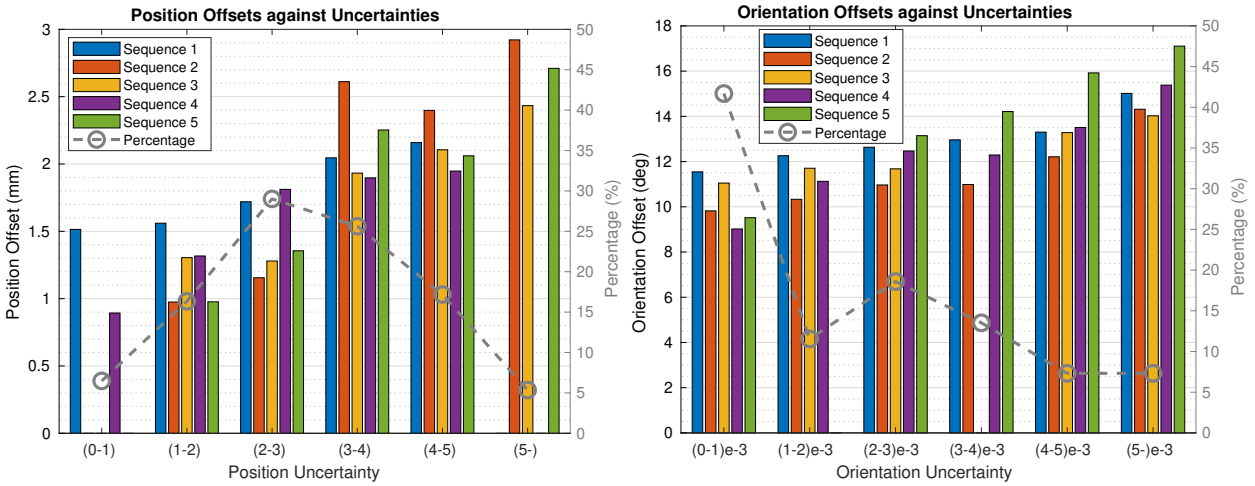

Fig. 6: The statistic results of the positional offsets against positional uncertainties (left) and the angular offsets against rotational uncertainties (right).

cay is fixed to 0.95. The initial learning rate is 1e-3 and the momentum is fixed to 0.9. The pre-trained truncated MobileNet V2 model is integrated into the localisation network for the feature extraction. In order to increase the training robustness, the gradient clipping is adopted for training. During the test, the number of sampling times from the latent distribution is set to 10000. The evaluation and analysis of localisation network are provided in the following two subsections.

All the networks are implemented through Pytorch[8], which are trained on an NVIDIA Titan GPU accelerated by CUDA and CUDNN. Using the captured training data, it takes around 24 hours and 6 hours in training the cycle generative model and localisation model, respectively. The whole system is implemented under ROS[9] framework using C++ and Python mixed coding. The runtime performance can be boosted to 11-13Hz for the whole system.

C. *localisation Evaluation*

Because we formulate a bronchoscopic localisation as a global localisation problem, the standard evaluation metrics for global localisation i.e. the positional error and angular error in PoseNet [13] are employed to evaluate the localisation performance. Specifically, the median error of $\ell_2(p-\widehat{p})$ distance offset between the ground truth position and predicted position is used for the translational evaluation. And the median error of $2\arccos(q\cdot\widehat{q})$ angular offset between the ground truth quaternion and predicted quaternion is used for the rotational evaluation.

The predicted 3D trajectories for the five testing sequences are plotted in Figure 5. The quantitative results of the localisation accuracy are shown in Table I for each testing sequence and Table II for the ablation analysis. It can be observed from Figure 5, the trajectories predicted by the localisation network closely follow the ground-truth trajectories calibrated from the readings of EM tracker for each sequence. The average positional error and angular error of sequences 1-5 are $1.64mm$ and 10.69° respectively. The overall positional error and angular error of all the sequences are $1.49mm$ and 10.81° respectively. According to Table I, the performance of the proposed localisation network is superior to the conventional geometry registration method which suffers from sudden tracking failures. Taking Sequence 1 as an example, the tracking loss of one frame can result in an extremely large offset for the localisation of the subsequent frames. In contrast, our method can significantly alleviate this problem because we solve it as a global localisation problem instead of an incremental registration problem. We also provide the results of localisation network which is tested on the virtual RGB images instead of the real RGB images. Those virtual RGB images are generated by making the virtual camera move along the calibrated EM trajectories in the 3D CT model. The localisation network achieves inferior performance using the real RGB images than the virtual RGB images due to the error from the domain transferring. Furthermore, it can be observed from Table II that the auxiliary learning makes the main contribution to the performance improvement. The variational inference also slightly improves the performance as well as estimates the uncertainty of each prediction.

D. *Uncertainty Evaluation*

How to evaluate the uncertainty predicted by the localisation network remains an open problem. Intuitively, we can propose an assumption that the uncertainty is proportional to the localisation error, i.e. the predicted global pose with higher uncertainty should have a larger localisation discrepancy, and vice versa.

To verify this hypothesis, we first calculate the ℓ_2 norm of the position (3 dimensions) uncertainty and orientation (4 dimensions) uncertainty. The overall mean uncertainties of position and orientation are 2.388 and 0.002 respectively. We further divide those translational and rotational uncertainties into different intervals separately. Lastly, the mean (not median) positional and angular errors of the predictions whose uncertainties lie within the corresponding intervals are calculated. We also provide the percentage of the predictions in each interval. The statistical results of the positional offsets against positional uncertainties and the angular offsets against

[8]https://pytorch.org/
[9]https://www.ros.org/

rotational uncertainties are shown in Figure 6. It can be observed that as the positional/angular uncertainties increase, the positional/angular geometry errors also generally increase in each sequence, which can verify the proposed assumption at the beginning. Comparing with the existing research, the proposed method is the first work to provide the predictive uncertainty of each estimation during the bronchoscopic localisation.

V. Conclusion

In this letter, we present a novel generative localisation approach with uncertainty estimation using only video and CT data for bronchoscopic biopsy. Comparing with the conventional 2D/3D geometry registration methods, we formulate the bronchoscopic localisation task as a global localisation problem solved by a data-driven deep neural network. More importantly, the proposed method can not only predict the global pose, but also estimate the corresponding uncertainty of each prediction via variational pose generation. Finally, the experiment results verify the resultant improvement of the localisation accuracy and the rationality of uncertainty estimation.

The future work will investigate the proposed method in the real clinical workflow. The basic procedure can be summarised into three steps. Firstly, a patient-specific airway model is reconstructed from the pre-operative CT scans. Secondly, a large number of virtual data are captured from the CT airway model through the virtual camera for training the neural networks. Finally, the trained model will be deployed to localise the camera in patient's lung with the corresponding uncertainty during the bronchoscopy.

VI. Acknowledgement

The authors would like to thank Wei Li for the data collection.

References

[1] X. Luo and K. Mori, "A discriminative structural similarity measure and its application to video-volume registration for endoscope three-dimensional motion tracking," *IEEE transactions on medical imaging*, vol. 33, no. 6, pp. 1248–1261, 2014.

[2] T. D. Soper, D. R. Haynor, R. W. Glenny, and E. J. Seibel, "In vivo validation of a hybrid tracking system for navigation of an ultrathin bronchoscope within peripheral airways," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 736–745, 2009.

[3] M. Shen, S. Giannarou, and G.-Z. Yang, "Robust camera localisation with depth reconstruction for bronchoscopic navigation," *International journal of computer assisted radiology and surgery*, vol. 10, no. 6, pp. 801–813, 2015.

[4] C. Sánchez, A. Esteban-Lansaque, A. Borrás, M. Diez-Ferrer, A. Rosell, and D. Gil, "Towards a videobronchoscopy localization system from airway centre tracking." in *VISIGRAPP (4: VISAPP)*, 2017, pp. 352–359.

[5] C. Zhao, L. Sun, P. Purkait, T. Duckett, and R. Stolkin, "Learning monocular visual odometry with dense 3d mapping from dense 3d flow," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 6864–6871.

[6] C. Zhao, L. Sun, Z. Yan, G. Neumann, T. Duckett, and R. Stolkin, "Learning kalman network: A deep monocular visual odometry for on-road driving," *Robotics and Autonomous Systems*, vol. 121, p. 103234, 2019.

[7] C. Zhao, L. Sun, P. Purkait, T. Duckett, and R. Stolkin, "Dense rgb-d semantic mapping with pixel-voxel neural network," *Sensors*, vol. 18, no. 9, p. 3099, 2018.

[8] C. Zhao, L. Sun, and R. Stolkin, "A fully end-to-end deep learning approach for real-time simultaneous 3d reconstruction and material recognition," in *2017 18th International Conference on Advanced Robotics (ICAR)*. IEEE, 2017, pp. 75–82.

[9] M. Visentini-Scarzanella, T. Sugiura, T. Kaneko, and S. Koto, "Deep monocular 3d reconstruction for assisted navigation in bronchoscopy," *International journal of computer assisted radiology and surgery*, vol. 12, no. 7, pp. 1089–1099, 2017.

[10] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2572–2581, 2018.

[11] M. Shen, Y. Gu, N. Liu, and G.-Z. Yang, "Context-aware depth and pose estimation for bronchoscopic navigation," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 732–739, 2019.

[12] J. Sganga, D. Eng, C. Graetzel, and D. Camarillo, "Offsetnet: Deep learning for localization in the lung using rendered images," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5046–5052.

[13] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[14] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," in *2018 IEEE International Conference on Robotics and Automation*. IEEE, 2018, pp. 6939–6946.

[15] F. Deligianni, A. Chung, and G.-Z. Yang, "Patient-specific bronchoscope simulation with pq-space-based 2d/3d registration," *Computer Aided Surgery*, vol. 9, no. 5, pp. 215–226, 2004.

[16] M. Shen, S. Giannarou, P. L. Shah, and G.-Z. Yang, "Branch: Bifurcation recognition for airway navigation based on structural characteristics," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 182–189.

[17] C. Sánchez, J. Bernal, D. Gil, and F. J. Sánchez, "On-line lumen centre detection in gastrointestinal and respiratory endoscopy," in *Workshop on Clinical Image-Based Procedures*. Springer, 2013, pp. 31–38.

[18] P. D. Byrnes and W. E. Higgins, "Construction of a multimodal ct-video chest model," in *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 9036. International Society for Optics and Photonics, 2014, p. 903607.

[19] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[20] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.