Article | Received 6 August 2026; Revised 21 October 2026; Accepted 30 October 2026; Published XX XXX 2026
https://doi.org/10.55092/rl20260029

BiSST: bidirectional structure-guided style translation for robotic bronchoscopy navigation

Fengxinyun Fang[1], Dongsheng Xie[2], Xiangyun Xian[1], Changsheng Li[1,3,4,*] and Xingguang Duan[1,3,4,*]

[1] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, China

[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China

[3] Key Laboratory of Biomimetic Robots and Systems, Ministry of Education, Beijing Institute of Technology, Beijing, China

[4] School of Medical Engineering, Beijing Institute of Technology, Zhuhai, China

* Correspondence authors; E-mails: lics@bit.edu.cn (C.L.); duanstar@bit.edu.cn (X.D.).

Highlights:

- We propose BiSST, a unified bidirectional sim-to-real framework for autonomous bronchoscopic navigation via structure-preserving translation between simulated and real images.
- We introduce a latent semantic consistency module and a mask-guided translation module to preserve task-relevant airway anatomy across visual domains.
- BiSST enables realistic synthetic data generation and effective deployment of imitation- and reinforcement-learning navigation policies in real bronchoscopic environments.

Abstract: Robotic bronchoscopy relies on visual perception for navigation, but limited image quality and the visual domain gap hinder the deployment of simulation-trained navigation policies in real environments. To address these challenges, we propose a bidirectional structure-guided style translation (BiSST) framework that learns structure-consistent mappings between simulated and real bronchoscopic images. Without relying on conventional electromagnetic (EM)-to-computed tomography (CT) registration to render virtual endoscopic views, BiSST incorporates a latent semantic consistency module and a mask-guided translation module to preserve task-relevant airway anatomy across visual domains. It translates simulated images into realistic images to enrich policy training and converts real observations into a simulation-like style to support reliable deployment of learned navigation policies. Experiments in simulation and physical airway phantoms demonstrate the effectiveness of BiSST, which improves the average path completion rate by 15.7 percentage points over baseline methods and enhances the robustness of sim-to-real bronchoscopic navigation.

Keywords: robotic bronchoscopy; sim-to-real transfer; surgical navigation; learning-based control

Copyright©2026 by the authors. Published by ELSP. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

1. Introduction

Endoscopic imaging plays a critical role in minimally invasive procedures such as bronchoscopy and gastroscopy, where reliable visual guidance is essential for safe and effective navigation. The decrease in image quality may lead to navigation errors and missed detection of lesions, thereby increasing surgical risks. In robot-assisted surgery, visual perception further serves as the foundation for autonomous decision-making and motion control. However, the miniaturization of surgical robots required to access deep and narrow anatomical regions inevitably constrains sensor size and imaging capability. Therefore, accurate interpretation of endoscopic images is crucial for reliable surgical navigation, particularly when image quality and the field of view are limited in deep and narrow anatomical regions.

In bronchoscopic navigation, virtual bronchoscopy navigation (VBN) is widely used to improve navigation accuracy (Figure 1). In a typical system, electromagnetic (EM)sensors installed at the end of the robot estimate the real-time pose of the robot through an external magnetic field [1,2]. This pose is registered with the airway model reconstructed by Computed Tomography (CT) to provide navigation guidance and generate corresponding virtual endoscopic images. Although this method has certain effectiveness, VBN relies on electromagnetic sensing and is easily affected by respiratory motion, thereby reducing registration accuracy. Therefore, visual navigation methods have gradually attracted the attention of researchers and are seen as an alternative solution. However, the robustness of such methods in complex endoscopic environments is still limited, and accurately understanding the visual semantics in endoscopic images remains an important challenge for achieving reliable navigation and clinical applications [3,4]. In addition to providing navigation guidance, virtual bronchoscopy also provides an effective platform for large-scale synthetic data generation. Recent research has begun to explore the synthesis of realistic endoscopic datasets [5,6], which involves rendering virtual images based on reconstructed anatomical models and transferring them to the real camera domain through style transfer. This process improves the realism of image texture while preserving anatomical structures, thereby generating realistic endoscopic images suitable for different clinical and research applications [7,8].

With the rapid development of reinforcement learning (RL), autonomous navigation has become an important research direction in the field of bronchoscopy robots [9,10]. The artificial intelligence (AI) assisted bronchoscopy robot system demonstrates the feasibility of learning based navigation methods [11]. The typical technical process is to first use limited expert demonstration data for pre training, then train through large-scale reinforcement learning in a simulation environment, and finally deploy to real scenarios [12]. Despite these advances, there are still two main issues: first, there are visual differences between simulated and real endoscopic images, which affect the migration of strategies from simulated environments to real environments; The second issue is that reinforcement learning still has shortcomings in training efficiency, and its generalization ability is limited in situations where anatomical structures change and data is limited. These problems constrain its reliable application in clinical scenarios [13–15].

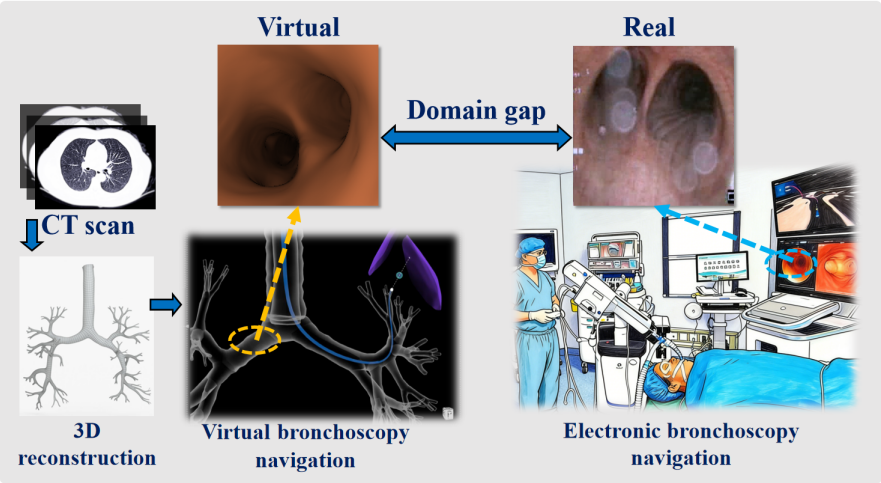


Figure 1. Virtual bronchoscopy and electronic bronchoscopy are commonly used for navigation, but there are significant differences in their visual appearance, resulting in significant visual domain differences.

In response to the above issues, this article proposes bidirectional structure-guided style translation (BiSST), a bidirectional structure guided style transfer framework, as the visual front-end for bronchoscope robot navigation. BiSST is capable of converting endoscopic images between the real domain and the simulation domain while preserving anatomical structures related to navigation tasks. This framework provides structurally consistent visual input for manual navigation and perceptual information for learning based autonomous navigation. In addition, its bidirectional conversion capability can also be used to generate endoscopic images with realism and preserved structural information for constructing datasets. The main contributions of this article are summarized as follows:

(1) We propose BiSST, a bidirectional structure guided style transfer framework for bidirectional conversion of bronchoscopy images between the real domain and the simulation domain, providing a unified visual front-end for manual operation and learning based autonomous navigation.

(2) We introduced a latent semantic consistency constraint and mask guided image conversion method to preserve task related anatomical structures between different domains, thereby improving the structural fidelity and semantic consistency of converted bronchoscopic images.

(3) We will apply BiSST to the generation of a realistic endoscopic dataset and incorporate it as a visual front-end into navigation strategies trained through imitation learning (IL) and reinforcement learning. The effectiveness of the overall framework is validated through comprehensive comparative experiments across different evaluation settings.

The remainder of this article is organized as follows. Section 2 reviews related work on sim-to-real style translation and learning-based robotic navigation. Section 3 describes the proposed BiSST framework, including its overall architecture, bidirectional structure-guided translation process, key consistency and mask-guided translation modules, and integration with downstream navigation policy learning. Section 4 describes the experimental setup. Section 5 reports and discusses the experimental results. Finally, Section 6 discusses the limitations of this work and presents the conclusion.

2. Related work

In this section, we review recent advances in sim-to-real transfer and learning-based robotic control. We first discuss general approaches for bridging discrepancies between simulated and real environments, and then summarize imitation learning and reinforcement learning in robotic control. We further consider their applications in autonomous and semi-autonomous surgical robotics, and finally highlight the key challenges in robotic bronchoscopy that motivate this work.

2.1. Sim-to-real style translation

Existing sim-to-real transfer methods aim to reduce discrepancies between simulated environments and the real world, particularly those caused by differences across style domains. Common strategies include domain randomization, adversarial domain adaptation, progressive transfer learning, among others [16–19]. Domain randomization increases appearance diversity in simulation [20], while progressive transfer learning strategies progressively adapt models trained in simulation to real-world conditions, gradually narrowing the sim-to-real gap [7]. Image-level adversarial domain adaptation often uses generative adversarial networks (GANs) or variational autoencoders (VAEs) to translate images from one domain to another [5,21,22]. The translated images retain the original content while adopting the visual style of the target domain, thereby narrowing the domain gap [23–25]. More recently, high-quality image generation models, such as diffusion-based approaches, have also been explored for sim-to-real transfer [26], though they often introduce higher computational cost.

Some representative studies have further explored the simulation to reality transfer between different style domains. Lin *et al.* [27] developed a simulation to reality reinforcement learning framework for visual dexterity operations, achieving robust policy transfer in complex dual arm tasks. Wu *et al.* [24] proposed RL GSBridge, which integrates three-dimensional (3D) Gaussian splashing technology into the reinforcement learning process to construct realistic scene representations, thereby achieving vision based transfer. Samak *et al.* [26] proposed a conditional latent diffusion framework to improve perceptual alignment between simulation and real driving environments. At present, the main challenges include limited generation quality, high training difficulty, and insufficient semantic understanding ability.

2.2. Imitation and reinforcement learning

IL and RL are two fundamental paradigms in learning based robot control, often combined and applied in autonomous tasks of robots [28,29]. IL derives policies from expert demonstrations through approaches such as behavioral cloning, adversarial imitation learning, and inverse reinforcement learning [30], while recent diffusion-based models further improve the modeling of multi-modal continuous action distributions [26,31]. RL instead learns policies through interaction with the environment under designed reward functions, and simulation-based training is widely adopted due to the high cost of real robotic data [3,32]. Recently, there has been an increasing exploration of hybrid frameworks combining imitation learning and reinforcement learning. Among them, imitation learning provides expert priors for policy initialization, while reinforcement learning further optimizes policies through interaction to cope with complex control tasks [25,30,33,34].

In recent years, the field of surgical robots has increasingly explored autonomous and semi autonomous capabilities to improve operational efficiency and reduce dependence on operators [13,35]. Zhang *et al.* [11] developed an artificial intelligence assisted bronchoscopy robot that adopts a human-machine shared control strategy based on expert demonstration training, enabling junior physicians to safely navigate in bronchial structures. Peloso *et al.* [30] proposed a learning based robot mitral valve repair path planning framework that utilizes imitation learning techniques to generate safe catheter trajectories in dynamic cardiac anatomical structures. Yao *et al.* [36] proposed a simulation based reinforcement learning method for autonomous navigation of guide wires and verified the successful transfer of the method from simulation environment to real robot system. Compared with other robotics fields, autonomous surgical robots require more accurate scene understanding ability and higher basic stability to ensure their safety in complex clinical environments [37,38].

3. Method

This section introduces the proposed simulation to reality visual transfer method and robot autonomous navigation method. Firstly, introduce the overall architecture and bidirectional style transfer process, followed by Mask Guided Translation and Latent Semantic Consistency modules and their corresponding optimization objectives. Next, describe autonomous navigation strategies, including reinforcement learning and imitation learning strategies. Finally, introduce the overall training process and the configuration of various optimization objectives and loss functions.

3.1. Framework and bidirectional style translation

We propose BiSST, an end-to-end framework for robotic bronchoscopy navigation that integrates BiSST, simulation-based policy learning, and real-world deployment, as illustrated in Figure 2. BiSST is designed to bridge the visual gap between virtual and real endoscopic domains while learning structure-consistent visual representations for navigation. The framework contains a bidirectional image translation branch for visual adaptation and a shared structural feature extraction branch that provides visual observations for downstream control.

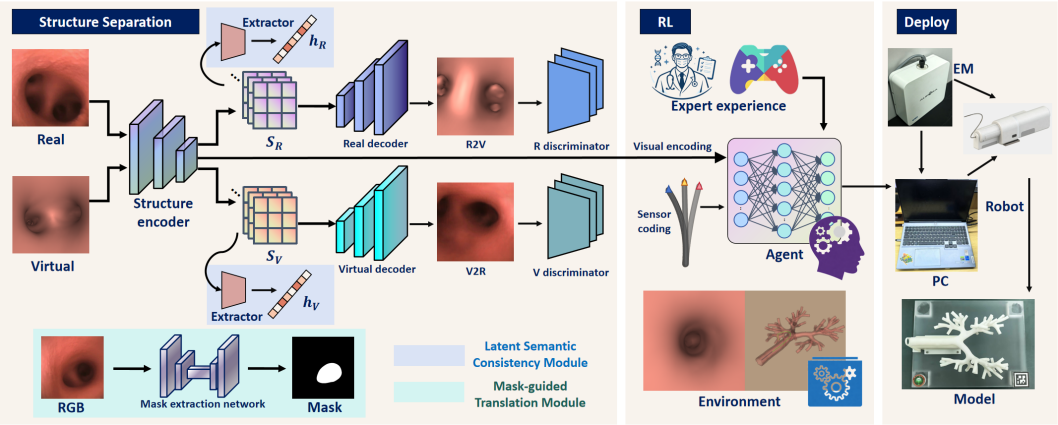


Figure 2. This framework covers style transfer, reinforcement learning, and end-to-end deployment, ultimately deployed on physical robots through expert imitation of pre training and simulation training strategies.

Let V and R denote the virtual and real endoscopic image domains, respectively. BiSST learns two appearance translation mappings, namely $V \rightarrow R$ and $R \rightarrow V$, from unpaired samples in the two domains. As shown in Figure 3a, an image from either domain is processed by its corresponding translation branch to generate an image with the appearance characteristics of the opposite domain. In this manner, virtual images can be translated into visually realistic endoscopic images for simulation-based learning, while real images can be translated into the virtual-domain style. This bidirectional architecture allows BiSST to leverage observations from both domains and model their distinct appearance characteristics.

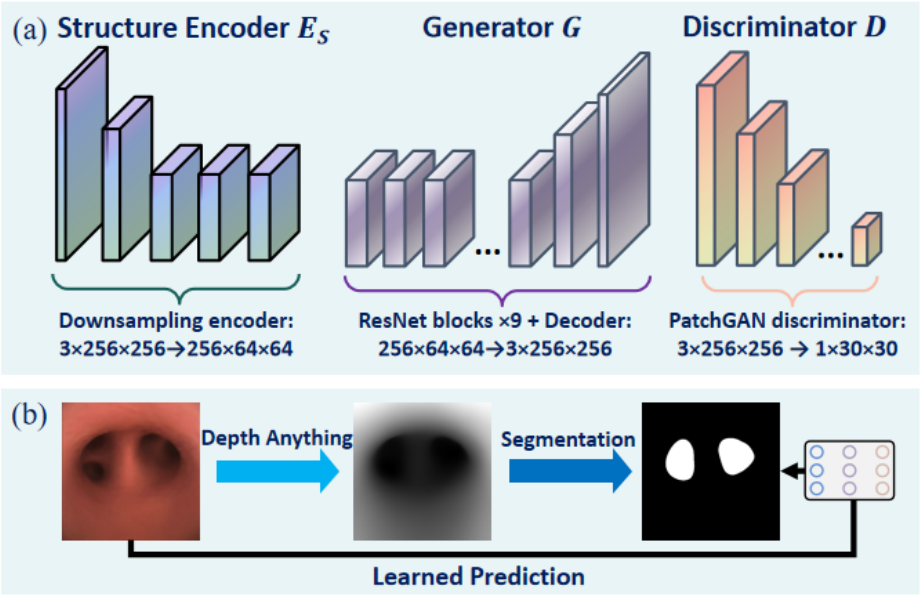


Figure 3. Network details. **(a)** Separating structural feature maps from the generator; **(b)** The lightweight mask prediction network F_{mask} estimates lumen masks from RGB endoscopic images using depth-based segmentation targets.

As illustrated in Figure 3a, the two translation branches share a common structure encoder E_s, while each branch has its own generator G and discriminator D. The structure encoder is implemented as a downsampling encoder that maps an input image of size $3 \times 256 \times 256$ to structural feature maps of size $256 \times 64 \times 64$. These feature maps are passed to the generator, which consists of nine ResNet blocks followed by an upsampling decoder to synthesize an output image of size $3 \times 256 \times 256$. The discriminator adopts a PatchGAN architecture and produces a spatial prediction map of size $1 \times 30 \times 30$ for assessing the realism of translated images in the target domain. The two translation branches share the same architectural design and perform the mappings $V \rightarrow R$ and $R \rightarrow V$, respectively.

The structure encoder E_s serves as the shared visual backbone of BiSST. Its feature maps retain structural information that is less sensitive to the appearance differences between the virtual and real domains. As shown in Figure 2, the Latent Semantic Consistency module processes the structural feature maps using a lightweight embedding extractor E_h to produce a normalized 128-dimensional latent representation, encouraging corresponding anatomical structures from the two domains to remain consistent in the latent space. The Mask-Guided Translation module is designed to preserve lumen-related geometric information during style translation. As illustrated in Figure 3b, a lightweight mask prediction network F_{mask} estimates lumen masks from RGB endoscopic images to guide the translation process.

3.2. Mask-guided translation

Image translation may inadvertently alter anatomically important structures, particularly the lumen region and its boundary. Preserving the underlying airway geometry is therefore essential during translation. A conventional solution is to estimate depth and then obtain the lumen mask through morphological segmentation. However, this pipeline is non-differentiable and cannot provide effective gradient supervision for the translation network. To address this issue, we introduce a Mask-Guided Translation module with a lightweight mask predictor F_{msk}. The predictor adopts a standard convolutional neural network (CNN) architecture and is pretrained to estimate differentiable soft lumen masks directly from RGB endoscopic images.

For a virtual image x^V, its real-style translation is denoted as $\hat{x}^{V\to R}$. The corresponding masks are obtained by

$$M^V = F_{\text{msk}}(x^V), \quad \hat{M}^{V\to R} = F_{\text{msk}}(\hat{x}^{V\to R}). \tag{1}$$

Since style translation should modify image appearance without changing the underlying airway geometry, the predicted masks are encouraged to remain consistent. We therefore use the Dice loss to measure structural agreement:

$$\mathscr{L}_{\text{mask}}^{V\to R} = 1 - \frac{2\langle M^V, \hat{M}^{V\to R}\rangle + \varepsilon}{\|M^V\|_1 + \|\hat{M}^{V\to R}\|_1 + \varepsilon}. \tag{2}$$

By directly constraining the luminal masks before and after migration, the proposed module can preserve anatomically significant regions and suppress structural translation caused by domain adaptation. This enables the generator to adjust texture, lighting, and color statistical features while maintaining the airway structure required for subsequent navigation. The reverse migration direction $R \to V$ is defined similarly.

3.3. Latent semantic consistency

To promote domain invariant structural understanding, we further introduced the latent semantic consistency module. As shown in the Figure 2, the lightweight embedding extractor E_h maps the structural feature maps generated by the shared encoder E_s to a compact 128 dimensional latent space. This design aims to provide similar latent representations for images depicting the same anatomical structure, without being affected by its visual domain.

For a virtual image x^V and its real-style translation $\hat{x}^{V\to R}$, their latent representations are computed as

$$h^V = E_h(E_s(x^V)), \quad \hat{h}^{V\to R} = E_h(E_s(\hat{x}^{V\to R})). \tag{3}$$

Since the translation changes only the image appearance while preserving the underlying anatomy, the two representations are encouraged to remain close:

$$\mathscr{L}_{\text{sem}}^{V\to R,\,pos} = \left\| h^V - \hat{h}^{V\to R} \right\|_2^2. \tag{4}$$

However, enforcing similarity only between positive pairs may reduce the discriminability of the latent space. We therefore use another randomly sampled virtual image $x^{V'}$ as a negative sample and

encourage its embedding $h^{V'}$ to be separated from h^{V}. Since randomly sampled images may still exhibit similar lumen structures, the separation strength is determined by their anatomical difference. Specifically, we reuse the pretrained mask predictor F_{msk} introduced in Section 3.2 to obtain M^{V} and $M^{V'}$. Their mask difference is measured by

$$w(x^{V},x^{V'}) = 1 - \frac{2\langle M^{V}, M^{V'}\rangle + \varepsilon}{\|M^{V}\|_1 + \|M^{V'}\|_1 + \varepsilon}. \tag{5}$$

A larger $w(x^{V},x^{V'})$ indicates a greater difference between the two lumen configurations and therefore imposes stronger separation in the latent space. The negative separation term is defined as

$$\mathscr{L}_{\text{sem}}^{V\to R,neg} = -\mathbb{E}_{x^{V'}}\left[w(x^{V},x^{V'})\,\|h^{V} - h^{V'}\|_1\right]. \tag{6}$$

The final semantic consistency loss $\mathscr{L}_{\text{sem}}$ is obtained as a weighted sum of the positive consistency term and the mask-aware negative separation term, preserving cross-domain semantic consistency while maintaining the distinction between different anatomical observations. The reverse direction is defined analogously.

3.4. Navigation policy learning

To enable autonomous navigation of the robotic bronchoscope, the downstream policy learning in this work adopts a two-stage strategy that combines imitation learning and reinforcement learning [9]. Firstly, expert demonstration data is collected through remote manipulation of the joystick in a simulated airway environment with curved and branching structures. Subsequently, these data were used to initialize the policy through behavior cloning (BC), and further optimized through RL and autonomous interaction with the simulation environment.

Given an expert dataset $\mathscr{D}_E$ containing observation–action pairs (o,a^{E}), BC trains the policy by minimizing the discrepancy between the predicted and expert actions:

$$\mathscr{L}_{\text{BC}}(\theta) = \mathbb{E}_{(o,a^{E})\sim\mathscr{D}_E}\left[\ell\big(\pi_\theta(o),a^{E}\big)\right], \tag{7}$$

where π_θ is the navigation policy and $\ell(\cdot,\cdot)$ is the imitation loss. This initialization allows the policy to acquire basic navigation behavior before RL-based refinement.

The navigation task is then formulated as a Markov decision process, with the objective of maximizing the expected discounted return

$$J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^{t}R_t\right], \tag{8}$$

where R_t is the reward at time step t and γ is the discount factor. For continuous control, we employ proximal policy optimization (PPO) with the clipped surrogate objective

$$\mathscr{L}_{\text{PPO}}^{\text{clip}}(\theta) = \mathbb{E}_t\left[\min\left(q_t(\theta)\hat{A}_t,\,\text{clip}\left(q_t(\theta),1-\varepsilon,1+\varepsilon\right)\hat{A}_t\right)\right], \tag{9}$$

where $q_t(\theta) = \pi_\theta(a_t \mid o_t)/\pi_{\theta_{\text{old}}}(a_t \mid o_t)$ denotes the probability ratio between the current and previous policies, $\hat{A}_t$ is the estimated advantage, and ε is the clipping threshold.

The continuum robotic bronchoscope has three controllable degrees of freedom: insertion, bending, and rotation. Accordingly, the continuous action is represented as $a_t = [d_t, r_{x,t}, r_{y,t}]$, where d_t denotes insertion, while $r_{x,t}$ and $r_{y,t}$ represent the two joystick-controlled motion variables.

As shown in Figure 4, BiSST style translation relies solely on endoscopic images and requires neither EM information nor EM-to-CT registration. In the autonomous navigation module, however, EM localization is used online, and the EM measurements are registered to the CT-derived airway coordinate system to obtain the pose and construct auxiliary navigation states. Specifically, the endoscopic image is processed by the BiSST structure encoder E_s and embedding extractor E_h to obtain a structure-consistent visual representation. This representation is fused with an eight-dimensional navigation-state vector and an EM-derived 6-DoF pose encoded by an multilayer perceptron (MLP). The eight-dimensional vector comprises the three-dimensional displacement to the target, the three-dimensional displacement to the nearest centerline point, the current path completion rate, and a scalar description of the current collision state. The resulting multimodal feature z_t is passed through a long short-term memory (LSTM) to capture temporal dependencies and produce a 512-dimensional hidden representation h_t, which is used by the policy to predict continuous control actions. Here, h_t denotes the recurrent representation produced by the LSTM.

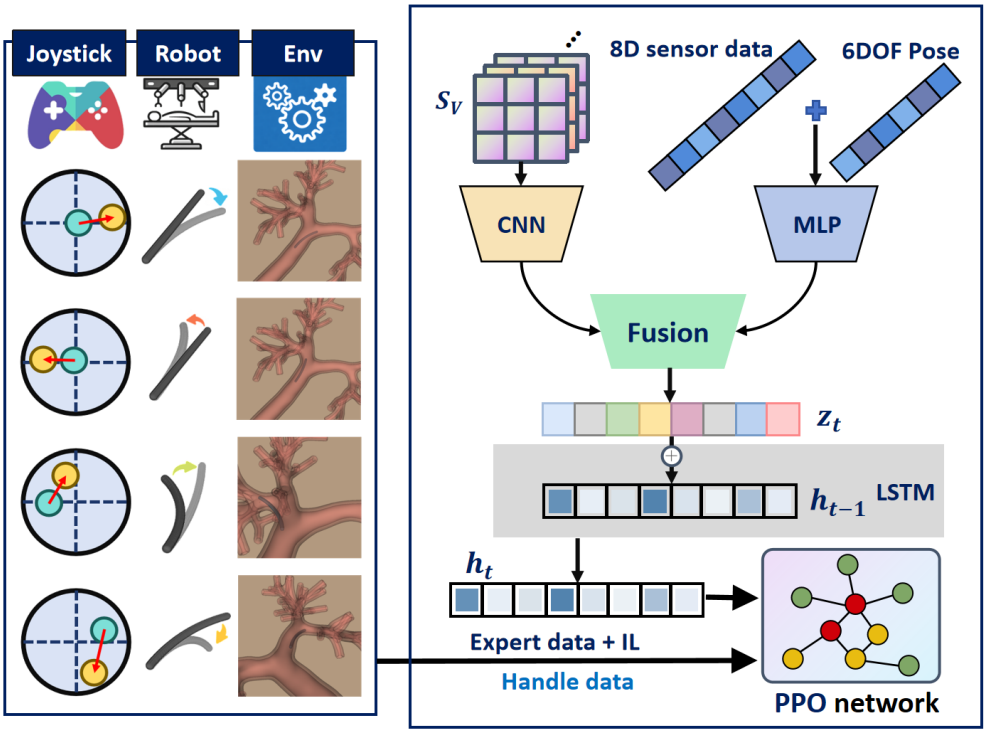


Figure 4. Two-stage learning pipeline for autonomous bronchoscope navigation. Joystick demonstrations initialize the policy through imitation learning, followed by PPO refinement with multimodal observations.

3.5. Training procedure and loss configuration

The training procedure consists of two stages: BiSST training for structure-preserving style translation, followed by navigation policy learning. For BiSST, we adopt the adversarial, contrastive, and identity objectives from CUT as the basic translation objectives, whose contrastive formulation was found to be

effective for image translation in our preliminary experiments [39]. However, CUT does not explicitly provide bidirectional translation, anatomical structure preservation, or task-relevant structural semantics. Accordingly, we further incorporate the mask-based structural loss and latent semantic consistency loss proposed in this work, together with grayscale structural and edge-smoothing regularization, to enable bidirectional and structure-preserving translation.

(1) Adversarial loss. For translation from the virtual domain V to the real domain R, the discriminator is trained using the hinge loss

$$\begin{aligned}\mathscr{L}_{\text{adv}}^{D}(D_R) = \mathbb{E}_{x^R\sim p_R}\left[\max\left(0,1-D_R(x^R)\right)\right] \\ +\mathbb{E}_{x^V\sim p_V}\left[\max\left(0,1+D_R(\hat{x}^{V\to R})\right)\right],\end{aligned} \tag{10}$$

while the corresponding generator objective is

$$\mathscr{L}_{\text{adv}}^{G}(G_R) = -\mathbb{E}_{x^V\sim p_V}\left[D_R(\hat{x}^{V\to R})\right]. \tag{11}$$

The adversarial losses for the reverse direction $R\to V$ are defined analogously.

(2) Contrastive loss. The PatchNCE objective preserves patch-level content correspondence between the input image and its translated result:

$$\mathscr{L}_{\text{NCE}} = \sum_{l\in\mathscr{L}_\phi}\frac{1}{N}\sum_{i=1}^{N}-\log\frac{\exp\left(\text{sim}(q_i^l,k_i^l)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(q_i^l,k_j^l)/\tau\right)}, \tag{12}$$

where q_i^l denotes the query embedding of the translated patch at spatial location i in layer l, and k_i^l denotes the corresponding positive key extracted from the input image. The remaining sampled patches serve as negative keys. Here, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature parameter.

(3) Identity regularization. For images already sampled from the target domain, an identity-style PatchNCE loss is imposed to avoid unnecessary appearance changes:

$$\mathscr{L}_{\text{id}}^{R} = \mathscr{L}_{\text{NCE}}\left(x^R,\hat{x}^{R\to R}\right), \quad \hat{x}^{R\to R} = G_R(E_s(x^R)). \tag{13}$$

The identity loss $\mathscr{L}_{\text{id}}^{V}$ is defined analogously for the virtual domain.

(4) Basic structural regularization. To further preserve low-level structural information while suppressing undesirable artifacts in the generated images, we introduce a basic regularization term composed of a grayscale structural loss and an edge-smoothing loss:

$$\mathscr{L}_{\text{basic}}^{d\to d'} = \mathscr{L}_{\text{gray}}^{d\to d'} + \lambda_{\text{edge}}\mathscr{L}_{\text{smooth}}^{d\to d'}, \quad (d,d')\in\{(V,R),(R,V)\}. \tag{14}$$

For an input image x^d and its translated image $\hat{x}^{d\to d'}$, the grayscale structural loss is defined as

$$\mathscr{L}_{\text{gray}}^{d\to d'} = \frac{1}{N}\left\|\text{Gray}\left(x^d\right)-\text{Gray}\left(\hat{x}^{d\to d'}\right)\right\|_1, \tag{15}$$

where $\text{Gray}(\cdot)$ denotes grayscale conversion and N is the number of pixels. This loss encourages preservation of the input's coarse intensity structure while allowing domain-specific appearance changes.

The edge-smoothing loss is defined as

$$\mathscr{L}_{\text{smooth}}^{d\to d'} = \frac{1}{N}\left(\left\|\nabla_x \hat{x}^{d\to d'}\right\|_1 + \left\|\nabla_y \hat{x}^{d\to d'}\right\|_1\right), \tag{16}$$

where ∇_x and ∇_y denote the horizontal and vertical image-gradient operators, respectively. The coefficient λ_{edge} balances the edge-smoothing loss against the grayscale structural loss within the basic regularization term.

Together with the mask-based structural loss $\mathscr{L}_{\text{msk}}$ and the latent semantic consistency loss $\mathscr{L}_{\text{sem}}$ introduced above, the overall BiSST training objective is formulated as

$$\begin{aligned}\mathscr{L}_{\text{total}} = {} & \lambda_{\text{adv}}\left(\mathscr{L}_{\text{adv}}^{V\to R} + \mathscr{L}_{\text{adv}}^{R\to V}\right) + \lambda_{\text{NCE}}\left(\mathscr{L}_{\text{NCE}}^{V\to R} + \mathscr{L}_{\text{NCE}}^{R\to V}\right) \\ & + \lambda_{\text{NCE-id}}\left(\mathscr{L}_{\text{id}}^{R} + \mathscr{L}_{\text{id}}^{V}\right) + \lambda_{\text{basic}}\left(\mathscr{L}_{\text{basic}}^{V\to R} + \mathscr{L}_{\text{basic}}^{R\to V}\right) \\ & + \lambda_{\text{msk}}\left(\mathscr{L}_{\text{msk}}^{V\to R} + \mathscr{L}_{\text{msk}}^{R\to V}\right) + \lambda_{\text{sem}}\left(\mathscr{L}_{\text{sem}}^{V\to R} + \mathscr{L}_{\text{sem}}^{R\to V}\right).\end{aligned} \tag{17}$$

Here, λ_{adv}, λ_{NCE}, $\lambda_{\text{NCE-id}}$, λ_{basic}, λ_{msk}, and λ_{sem} are balancing coefficients for the corresponding loss terms. After BiSST training, the parameters of the structure encoder E_s and embedding extractor E_h are frozen and used as the visual backbone for navigation policy learning. The navigation policy is first initialized through behavior cloning using joystick demonstrations and is subsequently refined with PPO through interaction with the simulated airway environment.

The navigation reward is formulated as

$$R_t = r_{\text{terminal},t} + \lambda_{\text{p}} r_{\text{progress},t} + \lambda_{\text{c}} r_{\text{centerline},t} + \lambda_{\text{s}} r_{\text{step},t}, \tag{18}$$

where $r_{\text{terminal},t}$ is assigned only when an episode terminates, providing a positive reward for successfully reaching the target and a penalty for failure. The progress reward $r_{\text{progress},t}$ encourages the bronchoscope tip to move toward the selected target, whereas the centerline reward $r_{\text{centerline},t}$ encourages the tip to remain close to the corresponding reference airway centerline, thereby promoting stable and accurate navigation. The step reward $r_{\text{step},t}$ is a reward applied at each time step to discourage unnecessarily long trajectories and encourage efficient task completion. The coefficients λ_{p}, λ_{c}, and λ_{s} balance the contributions of the progress, centerline, and step reward terms, respectively.

4. Experiments

This section introduces the experimental validation of the proposed framework. Firstly, train the BiSST bidirectional style transfer network on the GPU server to achieve alignment between the appearance of simulated and real bronchoscopy images. Each domain collects approximately 3000 basic images and divides them into a training set and an evaluation set before model training. After the model converges, the transferred data is used to train reinforcement learning strategies in a self-developed bronchoscope robot simulator.

For the mask prediction network, pseudo-depth maps are first generated using Depth Anything V2-Large [40] and then processed by the segmentation algorithm to obtain airway masks. The masks are generally reliable in most cases, with a small number of inaccurate masks will be manually corrected.

The experimental setup is shown in the Figure 5, including a physical bronchoscope robot, airway model, and electromagnetic navigation component. In the physical robot experiment, the trained algorithm is deployed on a computer equipped with NVIDIA RTX 4060 GPU. The self-developed bronchoscope robot has three degrees of freedom, namely insertion, bending, and rotation. The robot is equipped with an internal endoscope channel and an electromagnetic navigation sensor installed at the end. The robot is controlled through a joystick interface, where the left joystick controls bending and rotation, and the right joystick controls insertion. The airway model is adapted from a publicly available open-source model [11]. In the experiment, three representative and commonly used airway models were selected for simulation and physical experimental verification. The main training and navigation parameters are summarized in the Table 1.

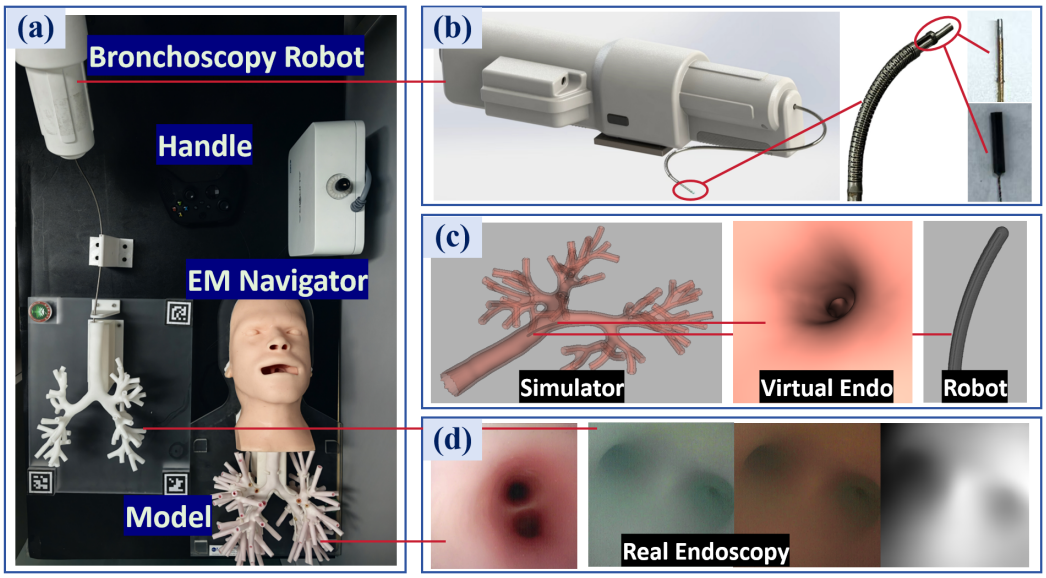


Figure 5. Experimental setup and airway models used in this study. **(a)** An experimental platform consisting of a bronchoscope robot, airway model, and electromagnetic navigation system; **(b)** Bronchoscopy robot equipped with an embedded electromagnetic sensor and a 1 mm endoscope; **(c)** Reinforcement learning simulator that generates virtual bronchoscopic images using a realistic robot model; **(d)** Endoscopic images captured from the phantom and 3D-printed airway models.

Table 1. Main training and reinforcement learning parameters.

Module	Parameter	Value	Parameter	Value
BiSST	Learning rate	1×10^{-4}	Batch size	4
Behavior cloning	Training steps	100,000		
Reinforcement learning	Learning rate	2×10^{-4}	Batch size	16
	PPO clipping parameter	0.15	λ_p	0.5
	λ_c	-0.1	λ_s	-0.1
	Terminal reward	$+20$ / -20	Control frequency	20 Hz

For visual domain adaptation, the BiSST network and all baseline style translation methods are repeatedly trained under different parameter configurations for 100 epochs. For each method, the model

with the best performance is selected for subsequent experiments. The real-domain images used for style translation are collected from a publicly available bronchoscopy visual odometry dataset [41,42]. After style translation training, the generated data are used for navigation policy learning. All reinforcement learning policies are trained for 500 k steps, with behavior cloning used for initialization and PPO used for subsequent policy optimization.

The evaluation consists of four parts. First, we evaluate the visual quality and structural fidelity of BiSST through qualitative and quantitative comparisons with representative image translation methods, including CycleGAN, CUT, UVCGAN, and diffusion-based approaches [19,39,43,44]. Second, simulation experiments are conducted to evaluate the benefit of BiSST for sim-to-real autonomous navigation learning, with representative trajectories and quantitative navigation results reported. Third, we evaluate the capability of BiSST to generate high-fidelity virtual-to-real bronchoscopic datasets through human evaluations of image and video realism. Fourth, we deploy the learned policy on the physical bronchoscopy robot to evaluate sim-to-real autonomous navigation and its robustness under simulated surgical conditions, including visual occlusion and blood-like artifacts. Human-operated navigation experiments are also conducted to assess the effects of different translated visual inputs using metrics such as average trajectory error and task completion time.

5. Results

5.1. Bidirectional style translation

Firstly, a qualitative evaluation of the bidirectional style transfer module is conducted by generating visual comparisons of the images. The experiment covers two transfer directions: virtual to real and real to virtual, to examine the ability of each method to transfer specific domain appearance features while preserving the underlying airway structure. For navigation tasks, image clarity and structural consistency are crucial, as inaccurate representation of airway lumen and bifurcation structures may directly affect spatial perception and control decisions. Although various image transfer methods have been used for endoscopic image synthesis and domain adaptation, there is still a relative lack of research on bidirectional structure guided transfer methods specifically designed for bronchoscopy navigation. Therefore, we selected representative general image transfer methods, including CycleGAN [45], CUT [39], and UVCGAN [19], as well as BronchoGAN for bronchoscopy and a diffusion model-based transfer method, as comparison baselines. The diffusion model baseline adopts a cyclic consistency architecture similar to CycleGAN, but replaces the traditional generator with a conditional diffusion model. For simplicity, this method will be referred to as Diffusion in the following discussion [43,44].

As shown in the Figure 6, Diffusion is usually able to generate images with high perceptual quality and realistic textures. However, its cyclic consistency training is relatively unstable and may not converge. CycleGAN and CUT have high computational efficiency, but their overall generation quality is limited. Specifically, both methods may produce erroneous textures in the later stages of training, which may be due to the instability of the adversarial optimization process caused by overly strong discriminators. In addition, they tend to exaggerate areas related to depth and generate unnatural folds at the bifurcation of the airway, which may introduce misleading visual cues. Although UVCGAN improves visual quality

through a stronger generator architecture, its universal image transfer objective does not explicitly constrain anatomical structures. In contrast, BiSST can achieve more stable image transfer while better preserving lumen boundaries, branch continuity, and branching geometry.

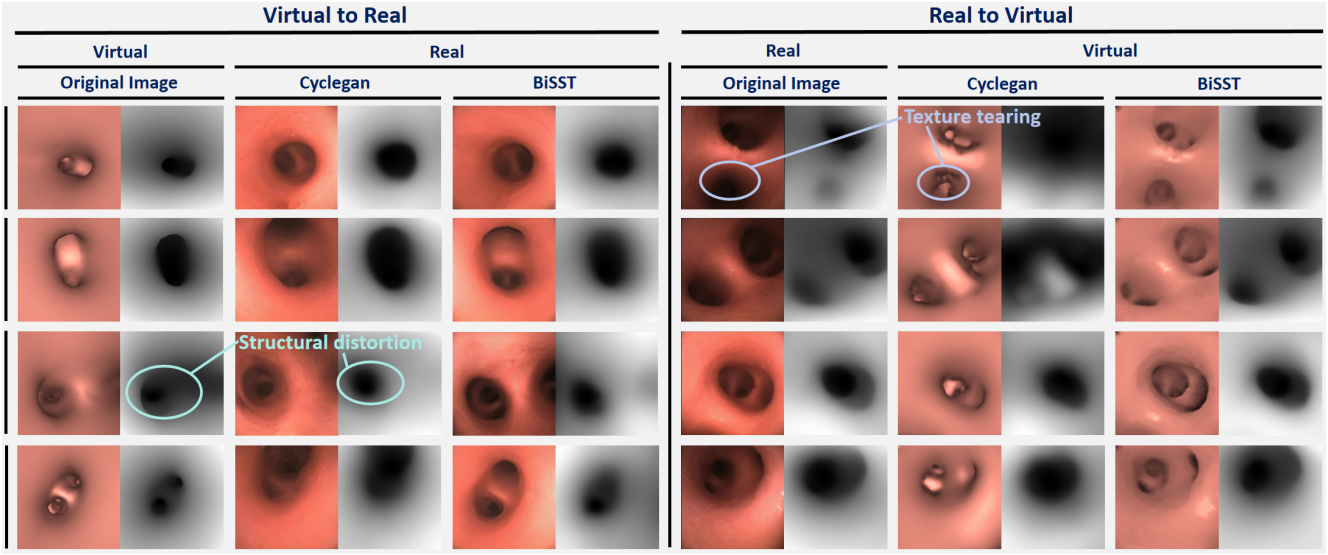


Figure 6. Visual comparison of style transfer results. BiSST can effectively maintain structural consistency without introducing artifacts. In contrast, the transfer images generated by baseline methods often suffer from structural distortion and texture tearing.

The results of quantitative and ablation experiments are shown in the Table 2, and the bold values in all tables indicate the best performance. Use Dice coefficient to evaluate the preservation of airway structure. During BiSST training, the mask predictor provides Dice based structural supervision; When evaluating quantitatively, the airway mask is independently obtained by combining depth estimation with traditional morphological segmentation, avoiding the repeated use of mask predictors in the training stage. SSIM, LPIPS, and FID are used to evaluate image similarity, perceived distance, and distribution quality, respectively.

Table 2. Quantitative comparison and ablation study of virtual-to-real and real-to-virtual image translation.

Methods	**Sim-to-Real**				**Real-to-Sim**			
	SSIM ↑	DICE ↑	LPIPS ↓	FID ↓	SSIM ↑	DICE ↑	LPIPS ↓	FID ↓
CycleGAN	0.612	0.862	0.302	303.46	0.648	0.889	0.281	253.17
UVCGAN	0.652	0.878	0.276	226.55	0.709	0.912	0.252	197.37
CUT	0.635	0.861	0.321	279.31	0.666	0.895	0.305	199.37
Diffusion	0.709	0.917	0.217	222.75	0.729	0.935	0.202	137.37
BronchoGAN	0.701	0.919	0.224	188.64	0.718	0.927	0.211	156.82
BiSST (w/o L_{sem})	0.693	0.925	0.229	178.94	0.732	0.952	0.220	140.41
BiSST (w/o L_{mask})	0.720	0.906	0.206	159.69	0.749	0.936	0.196	152.58
BiSST (w/o $L_{\text{sem}} + L_{\text{mask}}$)	0.639	0.879	0.275	242.72	0.679	0.902	0.246	193.73
BiSST (unidirectional)	0.741	0.929	0.188	128.57	0.775	0.948	0.184	106.35
BiSST (Ours)	**0.756**	**0.945**	**0.175**	**114.83**	**0.779**	**0.966**	**0.162**	**97.79**

The test set consists of paired virtual and real bronchoscopy images collected through registered virtual navigation, with each pair of images corresponding to the same airway structure. Compared with the baseline method, BiSST achieved the best overall performance in both Sim to Real and Real to Sim directions, with the highest SSIM and Dice scores and the lowest LPIPS and FID values, indicating that it can better balance visual transfer effects and structural preservation. The ablation experiment showed that removing semantic consistency loss or mask loss would lead to performance degradation, and removing both would result in a more significant decline, indicating that both contribute to improving the anatomical reliability of image transfer. The performance of BiSST trained unidirectionally is close to that of a complete bidirectional model, verifying its transferability. However, considering that bidirectional training can provide more effective visual semantic representation for downstream navigation, bidirectional training is better.

Mask loss reduces geometric distortion and maintains branching continuity between different domains by constraining alignment within anatomical regions, which is particularly important for maintaining consistency in luminal morphology and branching structures. In contrast, the loss of structural semantic consistency does not directly constrain high-dimensional feature maps, but rather constrains the consistency between the structural representations extracted by the embedding head E_h in a compact embedding space. This design enables the encoder to learn domain invariant structural representations while avoiding degradation collapse of feature representations, thereby enhancing the robustness of cross domain correspondence and improving stability under domain offset.

The combined effect of these two constraints makes the migration results more stable, the geometric structure more coordinated, and more in line with anatomical features, especially in complex areas such as airway bifurcations and depth changes. Overall, compared to baseline methods, BiSST performs better in terms of structural preservation and cross domain consistency. Its ability to align structures in the learning embedding space has also enhanced the transferability from simulation to reality, and the related effects will be further evaluated in subsequent experiments.

5.2. Simulation experiments

Conduct simulation experiments in a self-developed bronchoscope simulator to evaluate the effectiveness of BiSST in downstream reinforcement learning navigation tasks. The experiment compared four methods: traditional reinforcement learning (RL), reinforcement learning based on imitation learning initialization (RL+IL), BiSST without imitation learning (BiSST w/o IL), and the complete BiSST framework. The training process and generalization experiment results are shown in the Figure 7, and the average navigation performance is summarized in the Table 3.

The completion rate represents the percentage of the navigation trajectory completed by the intelligent agent to the expected total trajectory length; Reward is the accumulated reward during the navigation process. In generalization evaluation, completion rate refers to the proportion of experiments that complete at least 90% of the expected trajectory. The number of steps required to achieve a success rate of 90% (Steps to 90% SR) represents the additional training steps required to achieve this success rate after the initial 500 k steps of training.

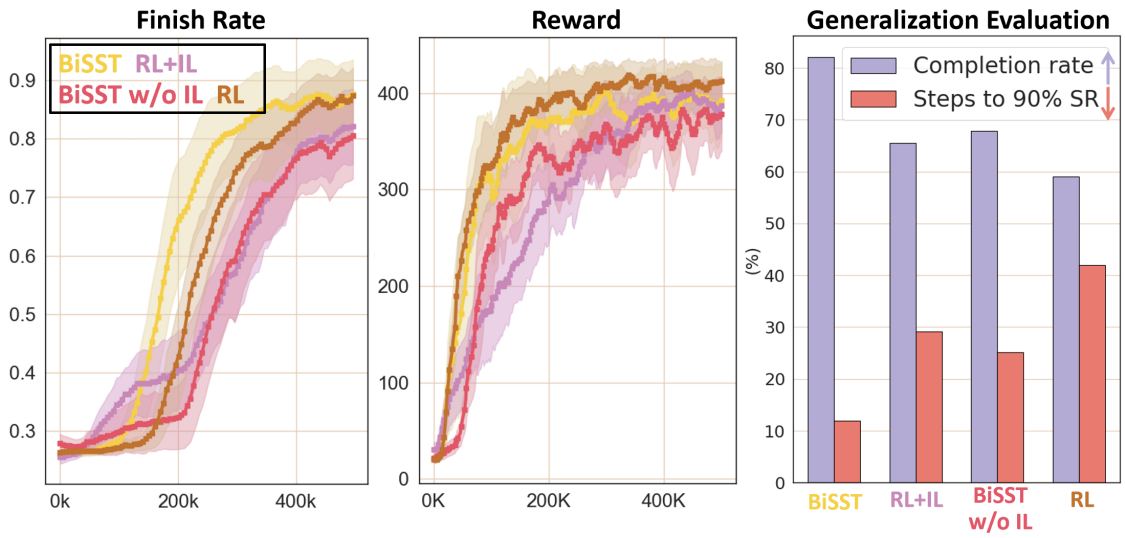


Figure 7. Training and generalization performance of different navigation methods in simulation environments.

Figure 7 compares the training and generalization performance of four methods. BiSST achieves faster improvement and better final finish-rate and reward performance than the baseline methods. BiSST w/o IL also consistently outperforms RL and RL+IL, demonstrating the benefit of structure-consistent visual representations for downstream navigation. RL+IL shows improved early-stage learning over RL, indicating that imitation learning provides a better policy initialization. In the generalization evaluation, BiSST achieves the highest completion rate on unseen airway environments and requires the fewest additional training steps to reach 90% SR. This result indicates that BiSST improves navigation performance and adaptability to unseen airway geometries.

Table 3 shows the average reward and average completion rate in the simulation experiments. BiSST achieves the best overall performance, with an average reward of 419.7 and an average completion rate of 94.3%. BiSST (w/o IL) obtains the second-best results, reaching an average reward of 407.3 and an average completion rate of 90.9%. RL+IL achieves an average reward of 389.5 and an average completion rate of 89.1%, while RL obtains lower values of 384.5 and 86.4%, respectively. These results indicate that the proposed BiSST framework can improve cumulative rewards and further advance the robot along the planned navigation path.

Overall, the results indicate that the proposed BiSST framework benefits reinforcement learning-based bronchoscope navigation in three complementary ways. First, the structure-consistent representation improves the policy's ability to preserve task-relevant anatomical information, resulting in higher navigation completion rates and cumulative rewards. Second, imitation learning further improves the initial policy and increases the completion rate compared with conventional RL. Third, the combination of BiSST and imitation learning achieves the strongest overall performance, suggesting improved learning effectiveness and generalization to unseen airway structures.

Table 3. Simulation navigation performance of different methods.

Method	Average Reward ↑	Average Completion Rate (%) ↑
RL	384.5 ± 12.6	86.4 ± 2.8
RL+IL	389.5 ± 11.8	89.1 ± 2.6
BiSST (w/o IL)	407.3 ± 10.4	90.9 ± 2.1
BiSST (Ours)	$\mathbf{419.7 \pm 11.1}$	$\mathbf{94.3 \pm 1.8}$

5.3. Dataset generation experiments

To evaluate the perceptual quality of the generated bronchoscopy datasets, we conducted a human evaluation study comparing BiSST with two visually competitive image translation methods, UVCGAN and Diffusion. The evaluation included both independently generated images and generated navigation videos. Five participants, all of whom had a robotics background but no medical background, took part in the study. Before the evaluation, they were familiarized with real bronchoscopic images to establish a basic understanding of the target-domain visual appearance. For each method, every participant evaluated 5–10 generated images. During testing, the outputs from the different methods were presented in a randomized order. The evaluation included four criteria: realism, quality, structure preservation, and preference. The results are shown in Figure 8.

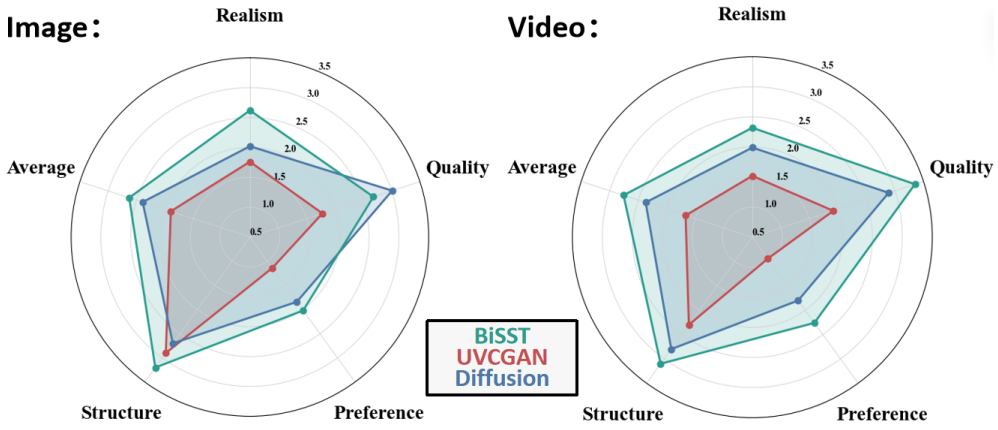


Figure 8. Generate manual evaluation results for images and videos. The higher the score, the better the performance.

In realism assessment, participants are presented with both real target domain images and generated images, and are required to distinguish between the two. A higher misclassification rate indicates that generated images are more difficult to distinguish from real target-domain images and are therefore more realistic. For quality, participants assigned subjective scores according to the visual quality of the generated results. For structure preservation, participants assigned subjective scores to the structural consistency between each source image and its corresponding style-transferred image, evaluating how well the generated image preserved the structure of the original image. For preference, participants compared outputs generated from the same source image and selected the most favorable result. The overall average score was calculated as a weighted combination of realism, quality, structure preservation, and preference, with weights of 0.4, 0.2, 0.2, and 0.2, respectively.

As shown in the Figure 8, BiSST achieved the best overall performance in both image and video generation. Specifically, BiSST performs the best in terms of realism, structural preservation, and preference scores, indicating that its generated results are more difficult to distinguish from real bronchoscopic images, can better preserve the structural information of the source image, and are more favored by evaluators. The Diffusion method achieved relatively good quality scores, especially in video generation, indicating that well-trained diffusion models can generate local details with good visual effects, but their overall performance is still lower than BiSST. UVCGAN scores lower on various evaluation indicators, which may be due to its sensitivity to the quality of pre trained representations and their correlation with the target domain. The above results indicate that BiSST can provide more realistic, structurally consistent and preferred generated images and videos for dataset construction.

5.4. Physical robot navigation and robustness evaluation

We conducted physical experiments on the bronchoscope robot platform to evaluate the practical feasibility of BiSST and its ability to simulate and transfer to reality. The physical robot shared the same kinematic configuration as the simulation model used for policy training. During autonomous navigation, the actions predicted by the policy were directly converted into motor commands for insertion, bending, and rotation. Human operators without prior joystick-control experience were recruited as the manual-control baseline.

The airway phantom was reconstructed from a segmented CT airway tree and fabricated by 3D printing. Because the 3D-printed phantom was not painted, its surface was rendered in a dark-red endoscopic style to simulate the color appearance of real bronchoscopic scenes and improve the visual realism of the experimental environment. As shown in Figure 9, experiments were performed on both the 3D-printed airway model and a pulmonary phantom model.

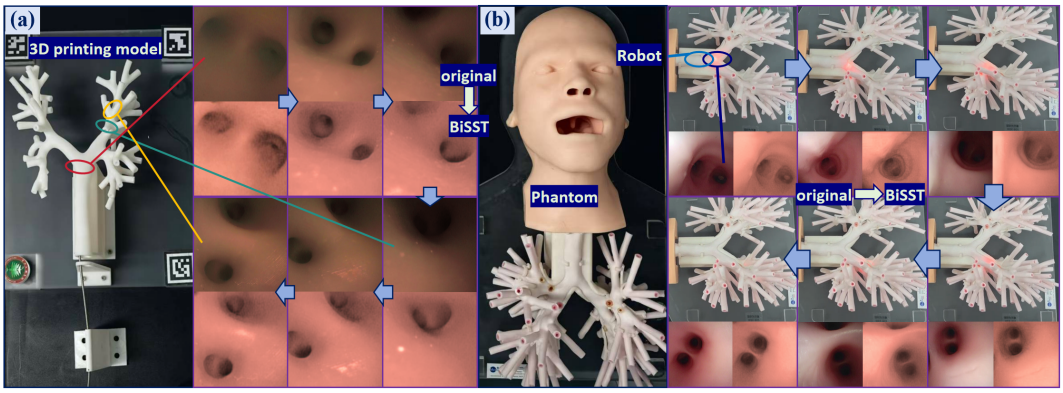


Figure 9. Physical bronchoscopy robot experiments on a 3D-printed airway model and a pulmonary phantom.

Approximately 30 target points were selected from airway branches spanning up to five levels. For each target, a reference centerline from the airway entrance to the target was extracted in advance. At the beginning of each navigation task, a random target point is selected as the navigation target, and the task is defined as controlling the robot to move from the entrance to the vicinity of the target point. A shared policy was trained across all target-specific navigation paths rather than training a separate policy for each target. The eight-dimensional sensor vector provided the policy with information related to the selected

target, the corresponding airway centerline, the current path completion rate, and the current collision state. Together with the EM-derived 6-DoF pose and the visual representation extracted by BiSST, these auxiliary states were used as multimodal observations for autonomous navigation.

Each target was evaluated in five independent trials, and the robot trajectory was recorded using electromagnetic tracking. Autonomous navigation performance was evaluated using average trajectory error (ATE), success rate, and path completion rate. ATE measures the average Euclidean distance between the executed trajectory and the target-specific reference centerline. The success rate (SR) SR-5, SR-15, and SR-25 denote the percentages of trials that traverse at least 95%, 85%, and 75% of the corresponding reference centerline path, respectively. The path completion rate measures the proportion of the corresponding reference path reached by the robot.

Figure 10a compares the physical navigation behavior of BiSST, conventional RL, manual operation, and manual operation with centerline guidance. The proposed BiSST method can generate more stable navigation trajectories with smaller centerline deviations and fewer pipe wall collisions. In contrast, manual operations have greater trajectory fluctuations in bifurcation and high curvature areas, mainly due to the need for operators to repeatedly adjust control inputs based on delayed visual feedback. Although traditional reinforcement learning improves navigation consistency, it still exhibits significant trajectory deviation and oscillation when there are visual differences between simulation and physical environments. The local trajectory comparison in Figure 10b further shows that BiSST recovers earlier from potential deviations and maintains more stable motion in challenging airway regions.

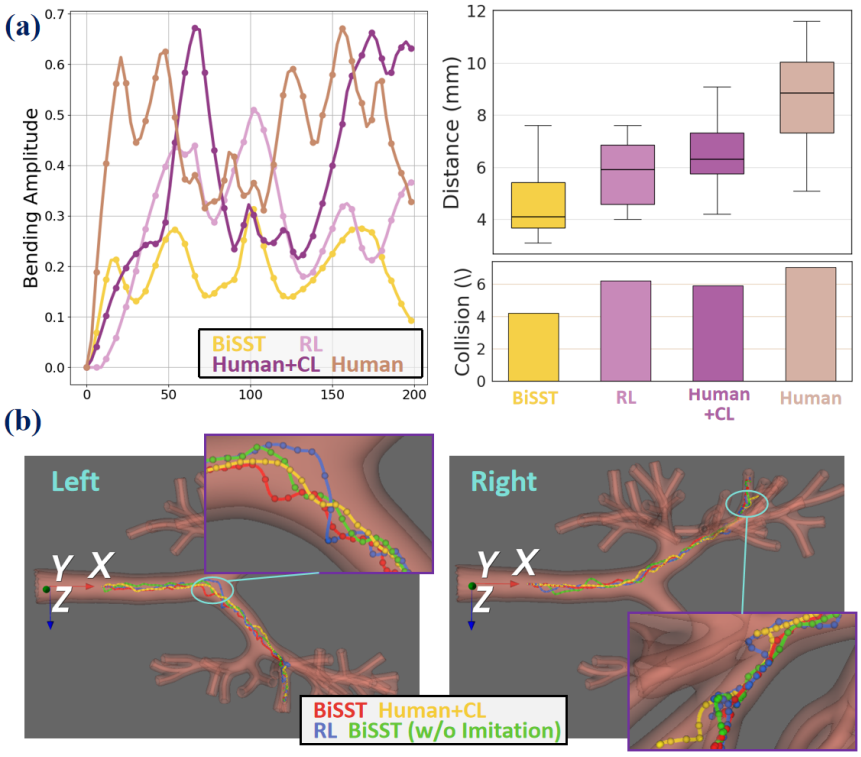


Figure 10. Physical autonomous navigation comparison. **(a)** Quantitative comparison of four methods in terms of bending amplitude, centerline distance, and wall-collision count; **(b)** Local trajectory comparison reconstructed from experimental position samples and visualized in the simulator.

The quantitative comparison in Figure 11 provides a comprehensive evaluation of autonomous navigation performance. The path completion rate of BiSST is the highest, reaching 89.1%, which is better than manual operation (88.2%), UVCGAN based navigation (79.0%), diffusion model-based navigation (74.3%), and traditional reinforcement learning (73.4%). In addition, BiSST maintained the highest success rate on the SR-25, SR-15, and SR-5 indicators. Especially when the success criteria are stricter, the performance degradation is smoother, indicating that this method can maintain accurate centerline tracking ability in longer and more challenging trajectories. These results validate that BiSST can effectively reduce the visual differences between simulation and reality, and improve the stability of bronchoscopy autonomous navigation.

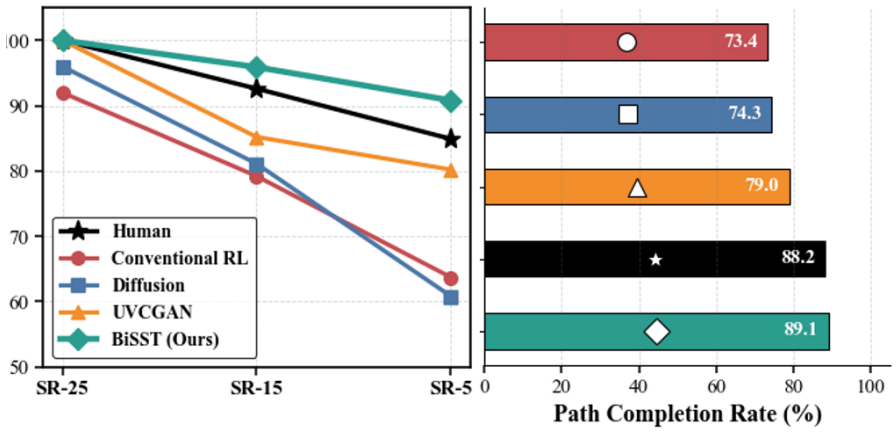


Figure 11. Autonomous navigation performance in physical experiments. The higher the success rate, the greater the proportion of experiments that reach the corresponding path completion threshold, reflecting that the robot can navigate more reliably to the target position.

We further evaluated the robustness of autonomous navigation under visual interference in simulating complex endoscopic environments. As shown in the Figure 12, the input image was subjected to lighting variation processing, including locally over-illuminated areas and simulated blood like artifacts that caused partial occlusion. In addition to evaluating these interferences separately, the combination of changes in lighting and blood like artifacts was also considered. These interferences can alter image brightness, reduce local contrast, and obscure anatomical structures, thereby increasing the difficulty of visual navigation.

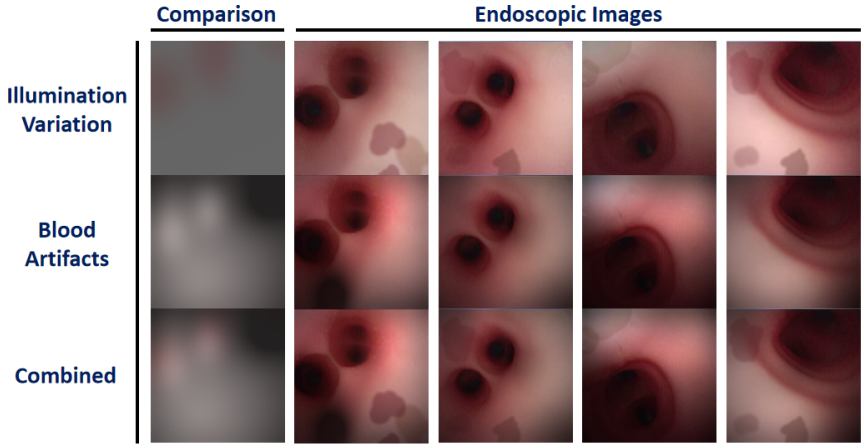


Figure 12. Visual pollution simulation.

Table 4 summarizes autonomous navigation performance under illumination variation, blood artifacts, and their combined condition. All learning-based methods exhibit varying degrees of performance degradation under these visual disturbances, demonstrating the difficulty of maintaining robust navigation under degraded visual conditions. Under illumination variation, conventional RL achieves SR-5 and SR-15 values of 54.6% and 60.3%, respectively, while Diffusion+RL improves these values to 62.5% and 73.1%. UVCGAN+RL achieves 74.8% and 87.4%, whereas BiSST achieves the highest SR-5 among the learning-based methods at 79.2% and a comparable SR-15 of 87.1%. Under blood artifacts, conventional RL decreases to 41.8% and 58.1%, respectively, indicating substantial sensitivity to partial visual occlusion. Diffusion+RL and UVCGAN+RL achieve 44.3%/73.5% and 68.6%/74.2%, respectively, while BiSST achieves the highest performance among the learning-based methods, with SR-5 and SR-15 values of 72.6% and 79.3%.

Table 4. Autonomous navigation robustness under visual degradation.

Method	Illumination		Blood Artifacts		Combined	
	SR-5 ↑	SR-15 ↑	SR-5 ↑	SR-15 ↑	SR-5 ↑	SR-15 ↑
Human	84.9% ± 1.2%	92.0% ± 0.9%	83.8% ± 1.4%	90.7% ± 1.0%	84.5% ± 2.5%	87.2% ± 2.2%
Conventional RL	54.6% ± 2.8%	60.3% ± 2.5%	41.8% ± 3.1%	58.1% ± 2.6%	52.7% ± 3.0%	57.6% ± 2.8%
Diffusion+RL	62.5% ± 2.4%	73.1% ± 2.0%	44.3% ± 2.9%	73.5% ± 2.1%	59.8% ± 2.7%	61.4% ± 2.2%
UVCGAN+RL	74.8% ± 1.8%	**87.4 ± 1.3%**	68.6% ± 2.0%	74.2% ± 1.7%	63.4% ± 2.1%	68.1% ± 1.8%
BiSST	**79.2% ± 1.6%**	87.1% ± 1.2%	**72.6% ± 1.8%**	**79.3% ± 1.4%**	**68.8% ± 2.0%**	**73.5% ± 1.6%**

Under the combined condition, which simultaneously introduces illumination variation and blood-like artifacts, BiSST maintains the highest performance among the learning-based methods, reaching SR-5 and SR-15 values of 68.8% and 73.5%, respectively, followed by UVCGAN+RL with 63.4% and 68.1%. Overall, these results indicate that the structure-guided design of BiSST helps preserve task-relevant anatomical information and maintain navigation performance under different forms of visual degradation. Nevertheless, human operators consistently outperform all evaluated learning-based methods across the three disturbance conditions, indicating that further improvement is still required in the robustness and generalization of autonomous navigation systems.

Finally, we evaluated BiSST in a master-slave navigation mode with manual operation, where visual input was simultaneously affected by changes in lighting and blood artifacts. Table 5 compares the performance of artificial navigation using original images and enhanced with Diffusion, UVCGAN, BronchoGAN, and BiSST. Among them, navigation time refers to the time required to complete a navigation task and reach the vicinity of the target point. The operating frequency of the controller is 10 frames per second, which is selected based on the response characteristics of the robot mechanism. Under the experimental hardware configuration, both the style transfer module and reinforcement learning strategy inference run at a stable speed of over 30 frames per second, meeting the real-time requirements of physical experiments.

Table 5. Human navigation performance under visual degradation with visual enhancement.

Method	ATE (mm) ↓	Navigation Time (NT, s) ↓
Human	3.60 ± 0.18	35.8
Human + Diffusion	3.36 ± 0.16	32.6
Human + UVCGAN	3.12 ± 0.13	32.7
Human + BronchoGAN	3.23 ± 0.14	34.9
Human + BiSST	$\mathbf{3.10 \pm 0.13}$	**30.3**

Compared to manual operations without assistance, Human+Diffusion reduces the average trajectory error from 3.60 ± 0.18 mm to 3.36 ± 0.16 mm. Human+UVCGAN and Human+BiSST further reduced ATE to 3.12 ± 0.13 mm and 3.10 ± 0.13 mm, respectively. In addition, the navigation time of Human+BiSST is the shortest, at 30.3 seconds, while the manual operation without assistance is 35.8 seconds. The results indicate that visual enhancement can provide effective assistance for manual navigation, but the difference in ATE between UVCGAN and BiSST is relatively small. The proposed structural guidance module and loss function help preserve task related anatomical information for autonomous navigation and manual navigation.

6. Limitations and conclusion

Despite extensive experiments conducted in simulation and physical airway model environments, there are still some limitations. Firstly, the current image transfer framework adopts a CNN based architecture to meet the real-time requirements of bronchoscopy navigation. Comparative experiments have shown that methods based on diffusion models have advantages in image quality. Therefore, in the future, we will consider introducing diffusion model generation to further improve visual fidelity and structural consistency, but we need to balance its inference speed and training cost. Secondly, rapid camera motion and vibration in physical experiments may lead to motion blur, thereby reducing image transfer quality and affecting navigation performance. Thirdly, the style transfer module independently processes each frame of the image, which may result in flickering and structural discontinuity between consecutive transfer frames. In the future, we will explore the loss of temporal consistency in motion perception and sequence architectures such as RNN or spatiotemporal Transformer to improve temporal coherence. Fourthly, the current validation is mainly conducted in simulation and physical airway model environments, and further evaluation of its robustness under more realistic conditions such as tissue deformation, liquid interference, lighting changes, and complex anatomical structures is still needed through animal experiments.

In summary, this article proposes a bidirectional structure guided style transfer framework for aligning structural representations between simulated and real bronchoscopy domains, providing consistent cross domain perception capabilities for navigation. In addition, combining imitation learning with reinforcement learning in the training process has improved control stability and training efficiency, achieving effective simulation to reality transfer of autonomous navigation for continuum robots. The experimental results in simulation and physical airway model environments show that compared with traditional methods, this method has improved navigation accuracy, robustness, and stability. In addition,

the proposed framework improves the performance of autonomous navigation and manual master-slave navigation by providing visual feedback that is more in line with task requirements and has higher structural consistency.

Future work will focus on promoting the development of visual perception frameworks towards more realistic and clinically valuable navigation scenarios. We will explore more expressive generative architectures and spatiotemporal modeling strategies to improve the quality and consistency of visual representations during navigation. ~~Meanwhile~~Furthermore, the generalization ability and practical application potential of the proposed framework will be further evaluated through animal experiments.

Data availability statement

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, we used OpenAI GPT for language editing assistance. We have reviewed and revised the generated content, and take full responsibility for the accuracy and completeness of the manuscript and figures.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62573058) and the Special Project for Basic Research Cooperation in the Beijing–Tianjin–Hebei Region (Grant No. F2024202102).

Authors' contribution

Conceptualization, F.F., C.L. and X.D.; methodology, F.F. and D.X.; data curation, F.F. and X.X.; writing—original draft preparation, F.F.; writing—review and editing, F.F., D.X., C.L. and X.D.; funding acquisition, C.L. and X.D. All authors have read and agreed to publish the final version of this manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Giri M, Dai H, Puri A, Liao J, Guo S. Advancements in navigational bronchoscopy for peripheral pulmonary lesions: a review with special focus on virtual bronchoscopic navigation. *Front. Med.* 2022, 9:989184.

[2] Duan X, Xie D, Zhang R, Li X, Sun J, *et al.* A novel robotic bronchoscope system for navigation and biopsy of pulmonary lesions. *Cyborg Bionic Syst.* 2023, 4:0013.

[3] Gu Y, Gu C, Yang J, Sun J, Yang G. Vision–kinematics interaction for robotic-assisted bronchoscopy navigation. *IEEE Trans. Med. Imag.* 2022, 41(12):3600–3610.

[4] Xu S, Li W, Du Z, Jin H, Fang Y, *et al.* Visual localization for robotic bronchoscopy with transferred domain: a fluoroscopy validation. *IEEE Trans. Instrum. Meas.* 2025, 74:4021711.

[5] Deng B, Zheng X, Chen X, Zhang M. A swin transformer encoder-based StyleGAN for unbalanced endoscopic image enhancement. *Comput. Biol. Med.* 2024, 175:108472.

[6] Soliman A, Keuth R, Himstedt M. BronchoGAN: anatomically consistent and domain-agnostic image-to-image translation for video bronchoscopy. *Int. J. Comput. Assist. Radiol. Surg.* 2025, 20(8):1741–1748.

[7] Shu H, Soberanis-Mukul RD, Xu J, Ding H, Ringel M, *et al.* BronchOpt: vision-based pose optimization with fine-tuned foundation models for accurate bronchoscopy navigation. *arXiv* 2025, arXiv:2511.09443.

[8] Martyniak S, Kaleta J, Dall'Alba D, Naskręt M, Płotka S, *et al.* SimuScope: realistic endoscopic synthetic dataset generation through surgical simulation and diffusion models. In *Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Tucson, USA, February 26–March 6, 2025, pp. 4268–4278.

[9] Zhao J, Chen H, Tian Q, Chen J, Yang B, *et al.* BronchoCopilot: towards autonomous robotic bronchoscopy via multimodal reinforcement learning. In *Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, UAE, October 14–18, 2024, pp. 6923–6930.

[10] Chen Z, Fan K, Cruciani L, Fontana M, Muraglia L, *et al.* Toward human-out-of-the-loop endoscope navigation based on context awareness for enhanced autonomy in robotic surgery. *IEEE Trans. Med. Robot. Bionics* 2024, 6(3):1116–1124.

[11] Zhang J, Liu L, Xiang P, Fang Q, Nie X, *et al.* AI co-pilot bronchoscope robot. *Nat. Commun.* 2024, 15(1):241.

[12] Zhao W, Queralta JP, Westerlund T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Canberra, Australia, December 1–4, 2020, pp. 737–744.

[13] Li Y, Lyu Q, Yang J, Salam Y, Wang W. A hybrid framework using diffusion policy and residual RL for force-sensitive robotic manipulation. *IEEE Robot. Autom. Lett.* 2025, 10(10):10266–10273.

[14] Güitta-López L, Güitta-López L, Boal J, López-López ÁJ. Sim-to-real transfer via a style-identified cycle consistent generative adversarial network: zero-shot deployment on robotic manipulators through visual domain adaptation. *Eng. Appl. Artif. Intell.* 2025, 159:111510.

[15] Chen W, Xu Y, Chen Z, Zeng P, Dang R, *et al.* Bidirectional sim-to-real transfer for GelSight tactile sensors with CycleGAN. *IEEE Robot. Autom. Lett.* 2022, 7(3):6187–6194.

[16] Guo L, Nahm W. A cGAN-based network for depth estimation from bronchoscopic images. *Int. J. Comput. Assist. Radiol. Surg.* 2024, 19(1):33–36.

[17] Gondokaryono R, Haiderbhai M, Suryadevara SA, Kahrs LA. Learning nonprehensile dynamic manipulation: Sim2real vision-based policy with a surgical robot. *IEEE Robot. Autom. Lett.* 2023, 8(10):6763–6770.

[18] Mai Z, Zeng C, Wang N, Yang C. A vision-based humanoid compliant skill transfer framework: application to robotic cutting tasks. *Biomimetic Intell. Robot.* 2026, p. 100280.

[19] Torbunov D, Huang Y, Yu H, Huang J, Yoo S, *et al.* UVCGAN: U-net vision transformer cycle-consistent GAN for unpaired image-to-image translation. In *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, USA, January 2–7, 2023, pp. 702–712.

[20] Josifovski J, Gu S, Malmir M, Huang H, Auddy S, *et al.* Safe continual domain adaptation after sim-to-real transfer of reinforcement learning policies in robotics. *arXiv* 2025, arXiv:2503.10949.

[21] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, *et al.* Analyzing and improving the image quality of StyleGAN. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, June 13–19, 2020, pp. 8110–8119.

[22] Sun Y, Lu Y, Lu H, Zhao Q, Wang S. Multimodal unsupervised image-to-image translation without independent style encoder. In *Proceedings of the 28th International Conference on Multimedia Modeling (MMM)*, Phu Quoc, Vietnam, June 6–10, 2022, pp. 624–636.

[23] Yuan C, Shi Y, Feng Q, Chang C, Liu M, *et al.* Sim-to-real transfer of robotic assembly with visual inputs using CycleGAN and force control. In *Proceedings of the 2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Jinghong, China, December 5–9, 2022, pp. 1426–1432.

[24] Wu Y, Pan L, Wu W, Wang G, Miao Y, *et al.* RL-GSBridge: 3D Gaussian splatting based real2sim2real method for robotic manipulation learning. In *Proceedings of the 2025 International Conference on Robotics and Automation (ICRA)*, Atlanta, USA, May 19–23, 2025, pp. 192–198.

[25] Jiang Z, Zhu H, Liu R, Hu X, Zhang W. Social GAN guided multimodal trajectory prediction and MPC for autonomous driving. *Robot Learn.* 2025, 2(2):22–44.

[26] Samak C, Samak T, Li B, Krovi V. Sim2Real diffusion: leveraging foundation vision language models for adaptive automated driving. *IEEE Robot. Autom. Lett.* 2025, 11(1):177–184.

[27] Lin T, Sachdev K, Fan L, Malik J, Zhu Y. Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids. *arXiv* 2025, arXiv:2502.20396.

[28] Li B, Wei R, Xu J, Lu B, Yee CH, *et al.* 3D perception based imitation learning under limited demonstration for laparoscope control in robotic surgery. In *Proceedings of the 2022 International Conference on Robotics and Automation (ICRA)*, Philadelphia, USA, May 23–27, 2022, pp. 7664–7670.

[29] Oriyama M, Hartono P, Sawada H. Human-in-the-loop transfer learning in collision avoidance of autonomous robots. *Biomimetic Intell. Robot.* 2025, 5(1):100215.

[30] Peloso A, Damiano R, Zhang X, Bicchi A, Votta E, *et al.* Imitation learning for path planning in cardiac percutaneous interventions. *IEEE Trans. Biomed. Eng.* 2025, 72(8):2379–2387.

[31] Wang H, Long Y, Chen Y, Yip HC, Scheppach M, *et al.* Learning dissection trajectories from expert surgical videos via imitation learning with equivariant diffusion. *Med. Image Anal.* 2025, 103:103599.

[32] Leng J, Zhu T, Huang Z, Li R, Zhou X, *et al.* Deep reinforcement learning-based dynamic path planning of flexible needle in robotic puncturing. *Biomim. Intell. Robot.* 2026:100290.

[33] Caddeo GM, Maracani A, Alfano PD, Piga NA, Rosasco L, *et al.* Sim2real bilevel adaptation for object surface classification using vision-based tactile sensors. In *Proceedings of the 2024 International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, May 13–17, 2024, pp. 15128–15134.

[34] Ng WY, Li Y, Pan T, Sun Y, Dou Q, *et al.* A simultaneous polyp and lumen detection framework toward autonomous robotic colonoscopy. *IEEE Trans. Med. Robot. Bionics* 2024, 6(1):163–174.

[35] He X, Zhang S, Chu J, Jia T, Yu L, *et al.* Intuition-guided reinforcement learning for soft tissue manipulation with unknown constraints. *Cyborg Bionic Syst.* 2025, 6:0114.

[36] Yao T, Wang H, Lu B, Ge J, Pei Z, *et al.* Sim2real learning with domain randomization for autonomous guidewire navigation in robotic-assisted endovascular procedures. *IEEE Trans. Autom. Sci. Eng.* 2025, 22:13842–13854.

[37] Kim Y, Iturrate I, Sloth C, Kim H. Safety-ensured robotic control framework for cutting task automation in endoscopic submucosal dissection. *IEEE Access* 2025, 13:102803–102814.

[38] Xie D, Chen W, Zhao J, Song X, Wang K, *et al.* Surgical instruments hyalinization: occlusion removal in minimally invasive endoscopic surgery. *Biomim. Intell. Robot.* 2023, 3(3):100105.

[39] Park T, Efros AA, Zhang R, Zhu JY. Contrastive learning for unpaired image-to-image translation. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, Glasgow, UK, August 23–28, 2020, pp. 319–345.

[40] Yang L, Kang B, Huang Z, Zhao Z, Xu X, *et al.* Depth anything V2. *Adv. Neural Inf. Process. Syst.* 2024, 37:21875–21911.

[41] Ozyoruk KB, Gokceler GI, Bobrow TL, Coskun G, Incetan K, *et al.* EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Med. Image Anal.* 2021, 71:102058.

[42] Borrego-Carazo J, Sanchez C, Castells-Rufas D, Carrabina J, Gil D. BronchoPose: an analysis of data and model configuration for vision-based bronchoscopy pose estimation. *arXiv* 2022, arXiv:2204.11982.

[43] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, Online, December 6–12, 2020, pp. 6840–6851.

[44] Wu CH, la Torre FD. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, October 1–6, 2023, pp. 7378–7387.

[45] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 22–29, 2017, pp. 2223–2232.