# Adversarial Diffusion Model for Domain-Adaptive Depth Estimation in Bronchoscopic Navigation

Yiguang Yang[1], Guochen Ning[2(✉)], Changhao Zhong[3], and Hongen Liao[1]

[1] School of Biomedical Engineering, Tsinghua University, Beijing, China
[2] School of Clinical Medicine, Tsinghua University, Beijing, China
ningguochen@tsinghua.edu.cn
[3] State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China

**Abstract.** In bronchoscopic navigation, depth estimation has emerged as a promising method with higher robustness for localizing camera and obtaining scene geometry. While many supervised approaches have shown success for natural images, the scarcity of depth annotations limits their deployment in bronchoscopic scenarios. To address the issue of lacking depth labels, a common approach for unsupervised domain adaptation (UDA) includes one-shot mapping through generative adversarial networks. However, conventional adversarial models that directly recover the image distribution can suffer from reduced sample fidelity and learning biases. In this study, we propose a novel adversarial diffusion model for domain-adaptive depth estimation on bronchoscopic images. Our two-stage approach sequentially trains a supervised network on labeled virtual images, and an unsupervised adversarial network that aligns domain-invariant representations for cross-domain adaptation. This model reformulates depth estimation at each stage as an iterative diffusion-denoising process within the latent space for mitigating mapping biases and enhancing model performance. The experiments on clinical sequences show the superiority of our method on depth estimation as well as geometry reconstruction for bronchoscopic navigation.

**Keywords:** Bronchoscopic Navigation · Domain-adaptive Depth Estimation · Adversarial Diffusion Model.

## 1 Introduction

Lung cancer is the leading cause of global cancer incidence and mortality, with an estimated 2.20 million new diagnoses and 1.79 million deaths per year [29]. Early diagnosis of lung cancer is critical for improving patient outcomes, especially in the case of peripheral pulmonary nodules [26]. Compared with transbronchial needle aspiration and thoracic surgery, bronchoscopy offers a safer alternative for airway-related diagnosis and treatment with reduced burden on patients [12]. However, in clinical scenarios, bronchoscopic intervention remains a challenging task for pulmonologists due to the complex structure of airways. The desire to access difficult-to-reach peripheral lesions with minimal complications drives the development of navigation systems for bronchoscopic procedures.

Bronchoscopic navigation systems based on electromagnetic (EM) tracking [17, 25], designed to provide real-time positions using EM sensors, have been investigated for camera localization over the decades. Despite having shown a certain level of success in diagnostic biopsy, EM-based methods suffer from CT-to-body divergence [19] caused by tissue deformation and sensory distortion, limiting their clinical efficacy of diagnosis and treatment.

Vision-based approaches, in contrast to EM-based ones, have been proposed to tackle the above CT-to-body divergence [21]. Various studies have focused on feature-based image-CT registration [13, 32] and reconstruction-based techniques [2, 31], yielding some promising results. However, visual factors such as inconsistent illumination and texture-scarce surface are still principle challenges that hinder their navigational accuracy.

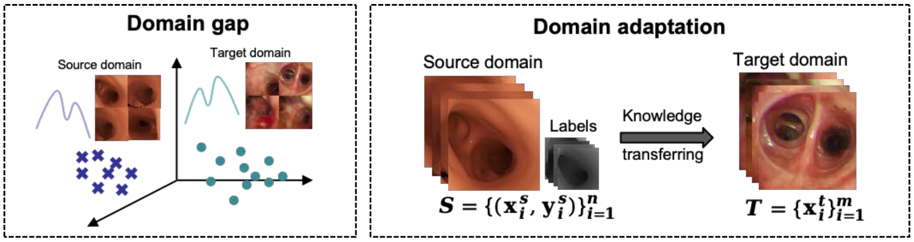

**Fig. 1.** Cross-domain gap between simulated and realistic images can cause performance decline on depth estimation. Unsupervised Domain Adaptation (UDA) bridges the gap by leveraging a label-rich source domain to solve tasks on a related unlabeled target domain, improving model performance without annotations in the real domain.

Due to the enhanced robustness against such visual obstacles, depth estimation has gained attention from researchers as it directly recovers geometrical structure from images. The recent progress in deep-learning techniques has inspired many works on endoscopic depth estimation to adopt training strategies. Since ground-truth (GT) depth maps are hardly available in bronchoscopic scenarios, [3] trained the network on virtual image-depth pairs for supervised learning. However, the domain gap between virtual and realistic images can lead to a performance decline during the inference phase.

Unsupervised domain adaptation (UDA) is proposed to address this problem by reducing distribution gaps between labeled source domain and unlabeled target domain [5, 10], as shown in Fig. 1. Motivated by the UDA works centered on cross-domain image translation, [20] utilized an image-level transfer network before employing the depth estimator for domain adaptation. [27] proposed a network based on CycleGAN [34] for bronchoscopic depth recovery. In additional to an extra computational cost incurred by such image-to-image translation, maintaining task-specific features can also be challenging. Feature-level domain adaptation overcomes these limitations by aligning domain-invariant features without explicitly reconstructing targeted images [30]. Among mainstream feature-level alignment methods, adversarial learning has become a well-proven approach for domain adaptation. [11] utilized an adversarial pipeline with source and target images learned in a sequential manner. Despite their power, conventional GAN-based models employ a rapid one-shot sampling process without any intermediate step, which inherently makes the network susceptible to mapping biases [23]. In addition, plain GANs are prone to suffer from training instability and mode collapse [16]. Diffusion models, as a promising alternative, have recently been adopted to enhance sampling reliability in generative tasks [4, 9]. Yet, their potential for tackling UDA problem in the context of medical images remains largely unexplored.

In this study, we propose a novel two-stage adversarial diffusion model for domain-adaptive depth estimation on bronchoscopic images. To address the lacking of GT depths from the real domain, we initially train a network using virtual images along with their simulated GT depths. In the second stage, an adversarial framework is leveraged to learn domain-invariant representations for an accurate feature-level adaptation. Moreover, our model redefines depth estimation at each stage as a gradual denoising-diffusion process with the guidance of bronchoscopic visual conditions, which mitigates the learning bias associated with GAN-based models. We conduct both qualitative and quantitative tests on real clinical videos to validate our findings. The experimental results demonstrate the superiority of our adversarial-diffusion strategy over non-transfer baselines and the state-of-the-art methods on depth estimation for bronchoscopic navigation.

## 2 Method

Consider a source dataset $S$ with virtual images $X_s$ and corresponding depth labels $Y_s$, and a target dataset $T$ that contains only target images $X_t$. We aim to train our network using $X_s$, $Y_s$, and $X_t$ for depth estimation in the target dataset. To accomplish this, We propose an adversarial-diffusion framework that involves two training stages: supervised learning and unsupervised domain-adaptive learning. This process is depicted in Fig. 2, where the switch signifies the transition between the different training stages.

### 2.1 Two-stage adversarial domain-adaptive framework

During the first training stage, virtual image-depth pairs from the source domain are used for supervised learning. Motivated by encoder-decoder structures favored for visual tasks, we utilize a ResNet-50 [8] encoder to extract features and construct visual conditions, and a decoder with deconvolution for depth recovery. The decoder consists of a sequence comprising a 3x3 deconvolution, a 3x3 convolution, and ends with a Sigmoid [22] activation function for normalization.

To mitigate the learning bias associated with conventional adversarial models, a diffusion-denoising process is introduced to gradually recover depth from

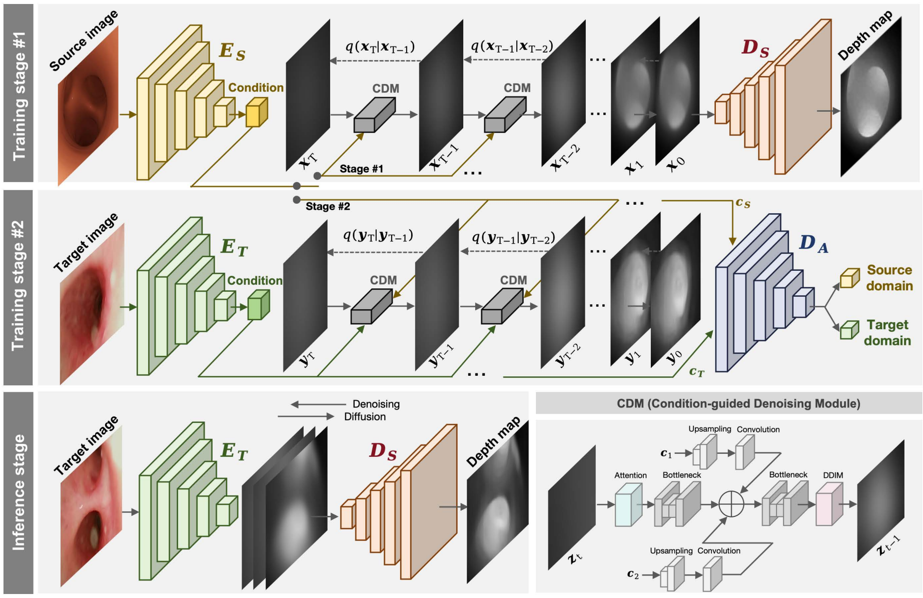

**Fig. 2.** Overview of the proposed framework. In the first stage, we use virtual dataset to train an encoder $E_S$ and a decoder $D_S$. In the second stage, we integrate the pre-trained encoder and a new encoder $E_T$ from the real domain in an adversarial manner. For inference, $E_T$ and $D_S$ are connected to perform depth estimation on real images.

random noise. Given the issue of generating high-resolution depth maps with various constraint-based methods, we adopt a strategy inspired by latent diffusion [24]. Our model perform both the diffusion process $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ and denoising process $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{c})$ within an encoded latent space (See Section 2.2). The refined depth latent $\boldsymbol{x}_0 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times k}$, with latent dimension $k$, is mapped to the depth estimation $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 1}$ through the depth decoder.

In addition to single-step optimization within the latent space, the encoder and decoder are also trained by directly minimizing the pixel-wise depth loss:

$$\mathcal{L}_{pixel}(d_s, \hat{d}_s) = \sqrt{\frac{1}{T}\sum_i (d_s^i - \hat{d}_s^i)^2}, \tag{1}$$

where $d_s$, $\hat{d}_s$ denote the depth prediction and GT , $T$ stands for the total number of valid pixels of virtual images. The supervised loss is formulated by a weighted sum of the single-step iterative loss and the pixel loss:

$$\mathcal{L}_{super} = \lambda_1 \mathcal{L}_{iter}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{c}_s) + \lambda_2 \mathcal{L}_{pixel}(d_s, \hat{d}_s), \tag{2}$$

where $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ denote latent distributions, $\boldsymbol{c}_s$ represents the visual condition extract from $E_S$, $\lambda_1$ and $\lambda_2$ are experimentally set to 0.4 and 0.6.

In the second stage, we connect the pre-trained encoder $E_S$ to train a new encoder $E_T$ that extracts features from real images. During inference, the encoder

$E_T$ is incorporated with the previously obtained decoder $D_S$ for depth estimation. To minimize the distance between cross-domain distributions, we utilize a discriminator $D_A$ from the standard GAN [7] to distinguish feature-level conditions from each domain:

$$\mathcal{L}_{adv_D}(E, \boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{x}_s \sim \boldsymbol{X}_s}[\log D(E_S(\boldsymbol{x}_s))] - \mathbb{E}_{\boldsymbol{x}_t \sim \boldsymbol{X}_t}[\log(1 - D(E_T(\boldsymbol{x}_t)))], \quad (3)$$

where $E_S$ and $E_T$ are optimized in an adversarial manner. The total loss is the sum of adversarial $\mathcal{L}_{adv_D}$ and iterative diffusion $\mathcal{L}_{iter}$ :

$$\mathcal{L}_{UDA} = \mathcal{L}_{adv_D}(E_S, E_T, \boldsymbol{c}_s, \boldsymbol{c}_t) + \mathcal{L}_{iter}(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \boldsymbol{c}_s, \boldsymbol{c}_t), \quad (4)$$

where $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ denote latent distributions, $\boldsymbol{c}_s$ and $\boldsymbol{c}_t$ represent visual conditions from the source and target domain, respectively.

### 2.2 Diffusion-denoising process for depth recovery

As described in Section 2.1, we reframe the depth estimation as a diffusion-denoising process to recover images blurred with random Gaussian noise. The diffusion process $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ is defined as:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1 - \bar{\alpha}_t)\boldsymbol{I}), \quad (5)$$

where noise is progressively added to the initial distribution $\boldsymbol{x}_0$ to obtain latent noisy samples $\boldsymbol{x}_t$, with step $t \in \{0, 1, \ldots, T\}$. $\bar{\alpha}_t$ is the cumulative product defined by $\bar{\alpha}_t := \prod_{s=0}^{t}(1 - \beta_s)$, where $\beta_s$ denotes the noise variance.

In the denoising step, neural network $\boldsymbol{\varepsilon}(\boldsymbol{x}_t, t, \boldsymbol{c})$ is optimized to predict $\boldsymbol{x}_{t-1}$ from $\boldsymbol{x}_t$ with visual conditions $\boldsymbol{c}$ guided, reversing the noise addition process:

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{c}) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\varepsilon}(\boldsymbol{x}_t, t, \boldsymbol{c}), \sigma_t^2\boldsymbol{I}), \quad (6)$$

where $\sigma_t^2$ denotes the transition variance. For faster denoising, we introduce DDIM [28] to trade off computation for sample quality. The denoising model $\boldsymbol{\varepsilon}(\boldsymbol{x}_t, t, \boldsymbol{c})$ is trained by minimizing the loss between diffusion results and denoising predictions:

$$\mathcal{L}_{iter} = \|\boldsymbol{x}_{t-1} - \boldsymbol{\varepsilon}(\boldsymbol{x}_t, t, \boldsymbol{c})\|^2 \quad (7)$$

To integrate the visual condition $\boldsymbol{c} \in \mathbb{R}^{\frac{H}{64} \times \frac{W}{64} \times k}$ from encoders $E_s$ and $E_T$ into the latent space, we develop a Condition-guided Denoising Module (CDM) illustrated in Fig. 2. The condition $\boldsymbol{c}$ is processed by an upsampling and convolution layer, and projected to the same shape of latent $\boldsymbol{z}_t \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times k}$. The projected condition is element-wise summed with latent $\boldsymbol{z}_t$ that refines through a self-attention block and a standard bottleneck layer [8] that maintain local features. The denoising output $\boldsymbol{z}_{t-1}$ is generated by employing a bottleneck, and the DDIM inference process directed by preset diffusion parameters $\alpha$ and $\beta$.

## 3 Experiments

**Datasets.** We validate our approach on two bronchoscopic datasets for unsupervised domain adaptation:

*Source domain* The virtual dataset includes 12,824 image-depth pairs for supervised learning. Detailed in the Appendix, airway trees are segmented from patient CT scans and utilized to compute the 3D airway mesh. Given the ray-casting renderer provided by Unity, we generate GT depth maps by simulating a virtual camera with the intrinsic parameters on our bronchoscope. These generated image-depth pairs are then used for the supervised learning of our proposed network, with a division of four-fifths for training and one-fifth for validation.

*Target domain* Our human recordings include three clinical sequences with a total of 8,921 frames, each with an original resolution of $400 \times 400$. To quantitatively evaluate our method, we navigate the virtual camera through segmented airways to simulate all possible views. We validate and handpick 142 virtual image-depth pairs that align with their target domain counterparts to serve as ground-truth for quantitative evaluation.

**Implementation details.** Our model is trained under PyTorch 1.9 framework. At each training stage, input images are resized to a resolution of $256 \times 256$ with their horizontal or vertical flips randomly sampled for augmentation. We initially train the supervised model with batch size of 64 for 100 epochs using the virtual image-depth pairs. During the second stage, we train the targeted encoder with batch size of 16 for 200 epochs, with a linear warm-up strategy applied during the first one-fifth iterations. The Adam optimizer [14] is utilized with values $\beta_1$ and $\beta_2$ of 0.9 and 0.999, and a learning rate of $1 \times 10^{-5}$. In diffusion process, the timesteps for training and inference are set to 1000 and 20. For the comparison against image-level adaptation, a CycleGAN [34] is trained to perform vanilla transfer. To ablate our diffusion strategy, we remove the entire latent space for all stages, rendering the model as a conventional adversarial framework.

**Qualitative results.** Fig. 3 visualizes a comparison of the proposed approach against other methods on the clinical sequences. Performing a direct image-to-image translation, the vanilla method shows capability simply within deepest locations, leaving widespread errors in other areas. This observation can be attributed to visual differences that primarily occur on the sides of lumens where textures and lightning conditions vary significantly across domains. Although the feature-level adversarial model improves the performance of depth estimation, the model still suffers from inaccurate shapes and over-sharp edges due to its one-shot sampling nature. Qualitative results illustrate that our proposed framework offers more fine-grained control on generating complex and smooth structures through successive diffusion steps, recovering the depth map reasonably well on both bifurcations and sidewalls.

**Quantitative results.** Table 1 presents an ablation study on adaptive strategies and denoising modules. The vanilla method shows a severe performance drop compared to feature-level adaptive methods. This decline can be explained that

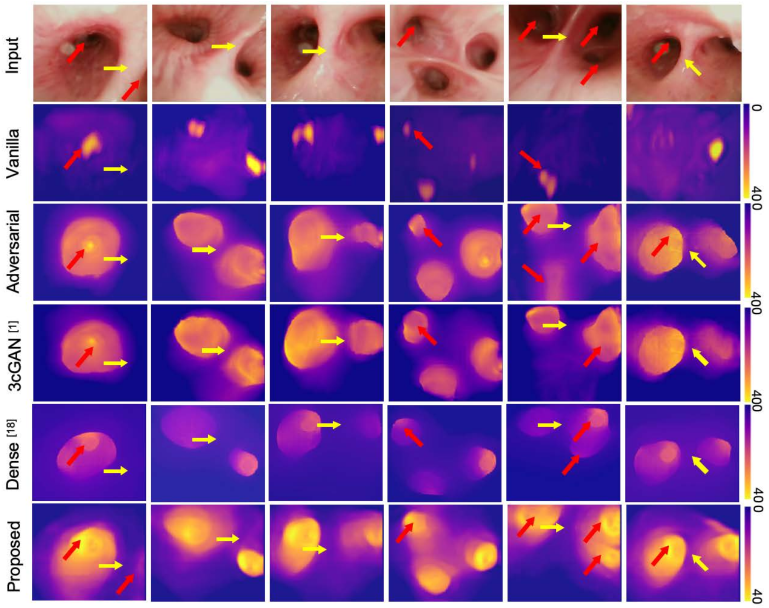

**Fig. 3.** Qualitative comparison among different domain-adaptive methods on real bronchoscopic images. Red and yellow arrows show areas for detailed comparison on lumens and carinas. Depth maps are scaled in $mm$.

style transfer models assume similar spatial distribution across domains, which is impractical in this context due to the lack of standardized rendering configurations. The proposed method outperforms the plain adversarial framework across all metrics, demonstrating that iterative refinements from the diffusion strategy potentially reduce one-shot mapping biases and generate accurate depth maps. For ablation study on CDMs, we replace each module with a plain combination of upsampling layers that align feature maps' shapes and DDIM that recovers depth distribution. Quantitative results show that CDMs with attention mechanism and bottlenecks better preserves the scene's structure.

**Table 1.** Ablation study on adaptive strategies and denoising modules. RMSE, MAE, and REL are metrics quantifying errors, $\delta$ denotes the threshold accuracy [6, 15]. A-D represents our proposed adversarial-diffusion strategy.

| **Method** | **Error metrics ↓** | | | **Accuracy metrics ↑** | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | REL | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| Vanilla | 36.0982 | 25.5377 | 0.5209 | 0.2123 | 0.4880 | 0.6215 |
| Adversarial | 11.3365 | 10.0919 | 0.2649 | 0.6287 | 0.8451 | 0.9520 |
| A-D w/o CDM | 8.0241 | 7.2980 | 0.2105 | 0.8062 | 0.9531 | 0.9774 |
| A-D w/ CDM | **6.0822** | **4.1155** | **0.1729** | **0.8243** | **0.9635** | **0.9848** |

**Table 2.** Quantitative results of the proposed method against state-of-the-art methods using in natural and endoscopic scenes.

| Method | Error metrics ↓ | | | Accuracy metrics ↑ | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | REL | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| mono [6] | 16.8295 | 13.4934 | 0.3291 | 0.6410 | 0.8577 | 0.9505 |
| AdaDepth [15] | 14.9204 | 11.9009 | 0.2929 | 0.6725 | 0.8668 | 0.9577 |
| TransDepth [33] | 12.9050 | 10.8566 | 0.2702 | 0.7024 | 0.9020 | 0.9662 |
| 3cGAN [1] | 13.2942 | 11.7459 | 0.2819 | 0.6890 | 0.8706 | 0.9623 |
| Dense [18] | 10.0921 | 9.2453 | 0.2650 | 0.7602 | 0.9212 | 0.9704 |
| Proposed | **6.0822** | **4.1155** | **0.1729** | **0.8243** | **0.9635** | **0.9848** |

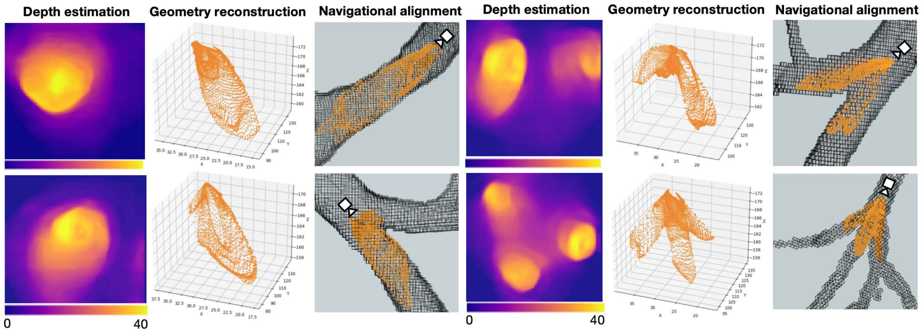

**Fig. 4.** Examples of geometry reconstruction from predicted depth maps. The middle column displays point clouds projected with camera intrinsics, while the right column exhibits spatial alignments between pre- and intra-procedural structures.

Furthermore, our method achieves the best performance against various non-transfer baselines and state-of-the-art approaches on endoscopic depth estimation, as shown in Table 2. Showing some merits on endoscopic depth perception, [1] poses extra training burden and instability by involving six generators and discriminators, and [18] leveraging 3D structure from sinus images tends to underperform in feature-scarce bronchoscopic environments.

**Reconstruction visualizations.** Geometry reconstruction serves as a prerequisite of bronchoscopic navigation pipelines. We re-project the depth images with bronchoscope's intrinsics to create geometrical point clouds. The generated point clouds are then manually registered to the airway structure from pre-procedural CT scans, as shown in Fig.4. Visualizations confirm that our method accurately captures the scene's structure across various morphologies of carinas.

## 4 Conclusion

In this work, we propose an adversarial diffusion model for domain-adaptive depth estimation on bronchoscopic images. Based on the observation of the domain gap between virtual and real images in bronchoscopy, the model introduces a two-stage strategy and learns domain-invariant representations for an accurate feature-level adaptation. Moreover, our model redefines depth estimation as a gradual denoising-diffusion process with the guidance of bronchoscopic visual conditions, which reduces the learning bias and generates detailed depth maps. Experiments on clinical sequences show the effectiveness of the proposed method on depth estimation as well as geometry reconstruction, demonstrating its potential for bronchoscopic navigation pipelines.

**Acknowledgments.** This study was funded by National Key Research and Development Program of China (2022YFC2405200), National Natural Science Foundation of China (82027807, 62201315, U22A2051) and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Banach, A., King, F., Masaki, F., Tsukada, H., Hata, N.: Visually navigated bronchoscopy using three cycle-consistent generative adversarial network for depth estimation. Medical image analysis **73**, 102164 (2021)
2. Chen, L., Tang, W., John, N.W., Wan, T.R., Zhang, J.J.: Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. Computer methods and programs in biomedicine **158**, 135–146 (2018)
3. Chen, R.J., Bobrow, T.L., Athey, T., Mahmood, F., Durr, N.J.: Slam endoscopy enhanced by adversarial depth prediction. arXiv preprint arXiv:1907.00283 (2019)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
5. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
6. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4893–4902 (2019)

11. Karaoglu, M.A., Brasch, N., Stollenga, M., Wein, W., Navab, N., Tombari, F., Ladikos, A.: Adversarial domain feature adaptation for bronchoscopic depth estimation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 300–310. Springer (2021)
12. Kebbe, J., Abdo, T.: Interstitial lung disease: the diagnostic role of bronchoscopy. Journal of Thoracic Disease **9**(Suppl 10), S996 (2017)
13. Khare, R., Higgins, W.E.: Image-based global registration system for bronchoscopy guidance. In: Medical Imaging 2011: Visualization, Image-Guided Procedures, and Modeling. vol. 7964, pp. 145–158. SPIE (2011)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Kundu, J.N., Uppala, P.K., Pahuja, A., Babu, R.V.: Adadepth: Unsupervised content congruent adaptation for depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2656–2665 (2018)
16. Kushwaha, V., Nandi, G., et al.: Study of prevention of mode collapse in generative adversarial network (gan). In: 2020 IEEE 4th Conference on Information & Communication Technology (CICT). pp. 1–6. IEEE (2020)
17. Lavasani, S.N., Farnia, P., Najafzadeh, E., Saghatchi, S., Samavati, M., Abtahi, H., Deevband, M., Ahmadian, A.: Bronchoscope motion tracking using centerline-guided gaussian mixture model in navigated bronchoscopy. Physics in Medicine & Biology **66**(2), 025001 (2021)
18. Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE transactions on medical imaging **39**(5), 1438–1447 (2019)
19. Luo, X., Mori, K.: A discriminative structural similarity measure and its application to video-volume registration for endoscope three-dimensional motion tracking. IEEE transactions on medical imaging **33**(6), 1248–1261 (2014)
20. Mahmood, F., Chen, R., Durr, N.J.: Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE transactions on medical imaging **37**(12), 2572–2581 (2018)
21. Mirota, D.J., Ishii, M., Hager, G.D.: Vision-based navigation in image-guided interventions. Annual review of biomedical engineering **13**, 297–319 (2011)
22. Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S.: Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378 (2018)
23. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging (2023)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
25. Schwarz, Y., Mehta, A.C., Ernst, A., Herth, F., Engel, A., Besser, D., Becker, H.D.: Electromagnetic navigation during flexible bronchoscopy. Respiration **70**(5), 516–522 (2003)
26. Shaller, B.D., Gildea, T.R.: What is the value of electromagnetic navigation in lung cancer and to what extent does it require improvement? Expert review of respiratory medicine **14**(7), 655–669 (2020)
27. Shen, M., Gu, Y., Liu, N., Yang, G.Z.: Context-aware depth and pose estimation for bronchoscopic navigation. IEEE Robotics and Automation Letters **4**(2), 732–739 (2019)

28. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Advances in neural information processing systems **33**, 12438–12448 (2020)
29. Thai, A.A., Solomon, B.J., Sequist, L.V., Gainor, J.F., Heist, R.S.: Lung cancer. The Lancet **398**(10299), 535–554 (2021)
30. van Tulder, G., de Bruijne, M.: Unpaired, unsupervised domain adaptation assumes your domains are already similar. Medical Image Analysis **87**, 102825 (2023)
31. Visentini-Scarzanella, M., Sugiura, T., Kaneko, T., Koto, S.: Deep monocular 3d reconstruction for assisted navigation in bronchoscopy. International journal of computer assisted radiology and surgery **12**, 1089–1099 (2017)
32. Wang, C., Oda, M., Hayashi, Y., Kitasaka, T., Itoh, H., Honma, H., Takebatake, H., Mori, M., Natori, H., Mori, K.: Anatomy aware-based 2.5 d bronchoscope tracking for image-guided bronchoscopic navigation. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(4), 1122–1129 (2023)
33. Yang, G., Tang, H., Ding, M., Sebe, N., Ricci, E.: Transformer-based attention networks for continuous pixel-wise prediction. In: Proceedings of the IEEE/CVF International Conference on Computer vision. pp. 16269–16279 (2021)
34. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)