

# Clinically-guided Data Synthesis for Laryngeal Lesion Detection

Chiara Baldini<sup>1,2</sup>, Kaisar Kushibar<sup>3</sup>, Richard Osuala<sup>3</sup>, Simone Balocco<sup>3,4</sup>, Oliver Diaz<sup>3,4</sup>, Karim Lekadir<sup>3,5</sup>, and Leonardo S. Mattos<sup>1</sup>

<sup>1</sup> Biomedical Robotics Laboratory, Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy

`chiara.baldini@iit.it`

<sup>2</sup> Department of Informatics, Bioengineering, Robotics, Systems Engineering, University of Genova, Genoa, Italy

<sup>3</sup> Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Spain

<sup>4</sup> Computer Vision Center, Bellaterra, Spain

<sup>5</sup> Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

**Abstract.** Although computer-aided diagnosis (CADx) and detection (CADE) systems have made significant progress in various medical domains, their application is still limited in specialized fields such as otorhinolaryngology. In the latter, current assessment methods heavily depend on operator expertise, and the high heterogeneity of lesions complicates diagnosis, with biopsy persisting as the gold standard despite its substantial costs and risks. A critical bottleneck for specialized endoscopic CADx/e systems is the lack of well-annotated datasets with sufficient variability for real-world generalization. This study introduces a novel approach that exploits a Latent Diffusion Model (LDM) coupled with a ControlNet adapter to generate laryngeal endoscopic image-annotation pairs, guided by clinical observations. The method addresses data scarcity by conditioning the diffusion process to produce realistic, high-quality, and clinically relevant image features that capture diverse anatomical conditions. The proposed approach can be leveraged to expand training datasets for CADx/e models, empowering the assessment process in laryngology. Indeed, during a downstream task of detection, the addition of only 10% synthetic data improved the detection rate of laryngeal lesions by 9% when the model was internally tested and 22.1% on out-of-domain external data. Additionally, the realism of the generated images was evaluated by asking 5 expert otorhinolaryngologists with varying expertise to rate their confidence in distinguishing synthetic from real images. This work has the potential to accelerate the development of automated tools for laryngeal disease diagnosis, offering a solution to data scarcity and demonstrating the applicability of synthetic data in real-world scenarios<sup>1</sup>.

**Keywords:** Data Synthesis · ControlNet · Endoscopy · Lesion Detection

---

<sup>1</sup> We publicly share our code at <https://github.com/ChiaraBaldini/endoLDMC.git>

## 1 Introduction

Medical imaging is evolving due to the emergence of Computer-aided Diagnosis and Detection (CADx/e) systems, which enable more accurate diagnoses by using deep-learning algorithms to assist clinicians in identifying pathological patterns in medical images. Domains such as radiology, oncology, and general endoscopy benefit from large well-annotated datasets, driving significant improvements [6, 7, 20, 24]. However, in specialized domains like laryngology, the application of such systems remains limited, primarily due to the lack of annotated data. While some public laryngology datasets exist, their annotations are often tailored to specific, narrow tasks, such as evaluating video frame quality [14], analyzing vocal cord mobility [2, 12], performing semantic segmentation [9], or classifying lesions [21]. Moreover, these datasets are typically acquired from a single clinical center and fail to capture the full spectrum of possible laryngeal lesions. This insufficient diversity in data representation and annotation severely limits their applicability in real-world clinical settings, where broader anatomical and pathological variability is required to address complex diagnostic tasks. Although the availability of annotated datasets remains limited, the need for CADx/e tools in laryngology is becoming more evident. This demand is driven by the operator-dependent nature of current diagnostic methods and the inconsistency introduced by subjective assessments. While techniques such as Narrow Band Imaging (NBI) [19] have been introduced to improve the visualization of neoangiogenic changes, their effective use relies on sophisticated endoscopic systems and significant operator expertise. Recently published studies have attempted to develop algorithms for detecting, classifying, and segmenting laryngeal lesions [18]. Nevertheless, only a small portion of them have tested model reliability and robustness on external populations and validated the applicability in the clinical environment due to diverse data scarcity.

Synthetic data generation techniques have become increasingly used by researchers to augment existing datasets and overcome issues related to variability limitations. Recent advances in generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have demonstrated remarkable success in producing realistic synthetic images for medical domains, including general endoscopy [3, 4, 10, 11, 13, 15, 22]. To the best of our knowledge, no studies have explored the synthesis of images for laryngeal lesion diagnosis. The complex anatomical and pathological characteristics of the laryngeal region make it a particularly challenging area for synthetic approaches, requiring advanced generative models to produce clinically relevant images that capture its inherent variability. One approach to ensuring the clinical relevance of synthetic data is to condition the generation process on clinical observations, such as acquisition parameters or spatial information of the lesion. Furthermore, how to maximize the potential of synthetic data by optimally selecting them to improve downstream task performance is rarely explored.

This study addresses the previously identified challenges through the following contributions:

1. Generation of synthetic laryngeal endoscopic image-annotation pairs by leveraging Latent Diffusion Models (LDMs) conditioning by clinical observations.
2. Proposing a synthetic data selection mechanism to improve the performance of a downstream lesion detection task.
3. Validation of the realism of the generated images through a human-observer study.

## 2 Methods

### 2.1 Laryngeal datasets

Laryngeal endoscopic images of different resolutions were collected independently from two hospitals<sup>2</sup> (2014–2023) during routine practice, using flexible or rigid endoscopes<sup>3</sup> in White-Light (WL) and NBI modes. The first medical center involved in the data collection was the Otolaryngology-Head and Neck Unit of the Oncologic Hospital Saint Savvas (Athens, Greece), providing a total amount of 909 laryngeal images, termed as *internal* in the following as they were used for training and internal validation of the LDM and the subsequent downstream task. The second center was the Otolaryngology-Head and Neck Unit of the IRCSS Hospital San Martino (Genova, Italy), where 88 images were extracted from video laryngoscopies. This *external* dataset was considered only for the validation of the downstream task. Each image was labeled with a bounding box that encompassed the surface of each lesion and a class associated with the type of lesion – cyst, granuloma, leukoplakia, polyp, papilloma, Reinke’s edema, squamous cell carcinoma –, according to histopathological confirmation or expert consensus. Such annotations were carried out by expert otorhinolaryngologists at each center and revised by two other experts at the IRCSS Hospital San Martino.

### 2.2 Clinically-guided Latent Diffusion Model

The proposed method generates synthetic images using a Latent Diffusion Model (LDM) for realistic reconstruction and a ControlNet module to preserve clinical relevance.

**Latent Diffusion Model** A Diffusion Model (DM) progressively perturbs the observed data distribution,  $x_0$ , by gradually adding Gaussian noise ( $\epsilon \sim \mathcal{N}(0, 1)$ ) until it converges to a given prior  $x_t$ . This process is known as diffusion. It then learns a reverse diffusion process ( $\epsilon_\theta$ ) to reconstruct and restore the original data, starting from this prior distribution and gradually produces less noisy samples  $x_{t-1}$ ,  $x_{t-2}$ ,  $\dots$ , until reaching the final sample  $x_0$ . The loss function can be written as:

$$L_{DM}(\theta) = \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0,1), t \sim U[1,T]} [\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2]. \quad (1)$$

<sup>2</sup> For each center, the local Institutional Ethics Committee approval was obtained.

<sup>3</sup> Olympus Medical System Corporation, Tokyo, Japan

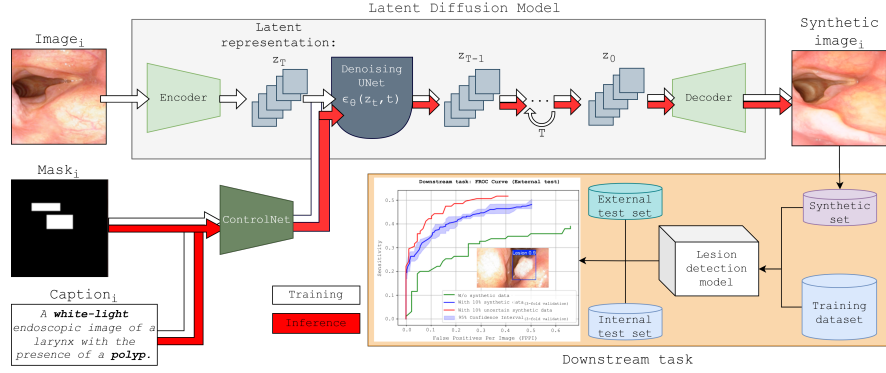


Fig. 1: Overview of the proposed method. During training (white arrow path), a caption with information on the optical modality, the lesion type, and the mask of the annotated bounding boxes were used as inputs to control the learning process of a Latent Diffusion Model (LDM). Realistic synthetic images were generated in the inference step (red arrow path), under the guidance of captions and masks, to enhance the performance of a downstream detection task.

The core concept of LDMs is to apply the diffuse and reverse diffusion processes in the low dimensional latent feature space of the inputs ( $z_t$ ), resulting in reduced complexity [16]. The input image ( $x_0$ ) is fed through an encoder, that outputs a latent vector  $z_0$ . For  $T$  iterations, a diffusion process adds noise to  $z_t$ , and then a denoising process is carried out through a UNet denoiser model. In the last part, a decoder converts the restored latent signal ( $z_0$ ) back to a clean signal in the original image space. The loss can be modified as follows:

$$L_{LDM}(\theta) = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(0,1), t \sim U[1,T]} [\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2]. \quad (2)$$

In this work, pre-trained autoencoder and text encoder models [16] were used following a transfer learning strategy to speed up the LDM training process.

**ControlNet** ControlNet [23] is an extended module that can be added to diffusion models to enable fine-grained control over the image generation process by incorporating additional conditioning inputs. The integration of extra information — such as edge, depth, or segmentation masks - allows the model to generate data that maintains the overall realism provided by the LDM but respects closely the structural or semantic indications of the conditioning signal. This level of control is even more critical in the medical field, especially when replicating pathological conditions, where the structural and spatial characteristics of abnormalities are key factors for clinical interpretation. We integrated detailed observations derived from clinical practice into ControlNet, learning patterns between laryngeal lesions and expert annotations, while also parsing biopsy results. Hence, our approach allows the creation of synthetic images that

contain clinically relevant features. In particular, we decided to pass two different inputs to the ControlNet module. An image caption containing notes about the optical modality used for the acquisition – WL or NBI – and the class of the laryngeal lesion annotated by our clinical partners was provided as the first control condition. An example of a caption is shown in Fig. 1: “A white-light endoscopic image of a larynx with the presence of a polyp.” In addition, a mask of the annotated bounding boxes was included in the conditioning process to provide information about the location of lesions. During inference, the model takes random noise as input and generates a new synthetic endoscopic image of a larynx, having the characteristics of the optical modality specified in the provided caption, with the specific type of lesion located in the desired area of the image.

**Implementation details** To train the LDM model with ControlNet, the *internal* dataset was split into three subsets with a ratio of 80:10:10, resulting in 727 images for training, 91 for validation, and 91 for testing. For each image, a 3-channel mask of the bounding box was obtained starting from the coordinates annotated by experts in YOLO format. A caption was also generated manually as a text prompt for each image, considering the class of the lesion and the modality information from clinical notes, as previously explained. Initially, the LDM model was fine-tuned for 100 epochs, with a batch size of 4, an initial learning rate of 0.00005, and an image resolution of  $640 \times 640$  pixels to generate realistic laryngeal images. Then, while freezing the LDM’s weights, the ControlNet module was integrated and fine-tuned with the same parameters as the previous step, to control lesions’ location and appearance. To test the generalization ability of the LDM, 727 new masks were created via random rotations and scaling operations of those of the *internal* dataset, and were utilized, together with the corresponding captions, to generate a larger set of diversified synthetic images, using ControlNet conditioning and guidance scales empirically set to 1.0. We ran all the experiments using an NVIDIA A100 GPU with 80GB of memory.

### 2.3 Evaluation metrics

Our goal is to generate diverse, clinically relevant synthetic images distinct from the originals to enrich CADx/e training datasets. An evaluation metric that computes the similarity between the distributions of real and generated images, such as the Fréchet Inception Distance (FID) [8], does not fully align with our objectives and may undervalue the contribution of clinically diverse synthetic data. For those reasons, we evaluated the proposed method with the  $FID_{ratio}$  [15], considering the FID between real and synthetic samples ( $FID_{rs}$ ) relative to the natural variability in real data ( $FID_{rr}$ ), which produced a high FID value due to diversity. It was calculated as  $FID_{ratio} = 1 - (FID_{rs} - FID_{rr})/FID_{rs}$ . The Inception Score (IS) [17] was also computed to evaluate data diversity compared to real images. Moreover, two other approaches were adopted to analyze the

advantages of the synthetic images and validate their quality, as detailed in the following.

**Downstream task** We explored the practical utility of the generated synthetic data by augmenting the available *internal* laryngeal dataset. Specifically, we combined the synthetic images with the real ones to train a lesion detection model that can learn to handle a wider variety of cases, improving its robustness. As lesion detector, we utilized the pre-trained YOLOv8 nano model from the Ultralytics GitHub repository<sup>4</sup>. This model is identical to the one currently in use in a hospital that collaborates with our team, and was chosen due to its capability of working efficiently in real-time. We trained the detector for 100 epochs, first only with the real images of the *internal* dataset and then gradually added 5%, 10%, 20%, 40%, and 80% synthetic samples. Taking into account the percentage that achieved the best results, we carried out a 3-fold cross-validation by repeating the detection learning process with 3 different sets of synthetic data and the same real set. However, as demonstrated by the cross-validation experiment, the random selection of synthetic samples can slightly impact performance variability. To address this issue, we defined an Uncertainty Estimation (UE) strategy for extending the training dataset with “challenging samples” identified through detection prediction uncertainty. Considering the 3 models from cross-validation, detection predictions were obtained for all synthetic data. The variance between the confidence scores returned by the 3 models was calculated for each image. After sorting the samples in descending order of variance, only the top 10% were selected, as they represent the most “challenging samples” to detect. For all the experiments, we applied the AdamW optimizer, an early-stopping strategy of 50 epochs, a batch size of 32, an initial learning rate of 0.001, and the default data augmentation. The detection performance was tested - in terms of precision, recall, and Average Precision@IoU=0.5 ( $AP_{IoU=0.5}$ ) - both on the *internal* test set and on the unseen *external* data.

**Human-observer study** We conducted a clinical investigation about the quality of the images by selecting a subset of 20 images, 10 real and 10 synthetic, and asking 5 of our clinical collaborators, with 7, 6, 11, 6, and 8 years of experience, to blindly rate the realism on a “Likert” scale from “strongly disagree” to “strongly agree” concerning the sentence “*This is a REAL image*”. We assigned probability values to each possible vote, with the aim of computing the Area Under the ROC Curve (AUC) of each user and evaluating their accuracy in correctly identifying real samples. In particular, the “Likert” scale - [Strongly disagree, Disagree, Slightly disagree, Slightly agree, Agree, Strongly agree] - was translated into the following probability values: [0.05, 0.23, 0.41, 0.50, 0.77, 0.95] [1, 5].

---

<sup>4</sup> <https://github.com/ultralytics>

### 3 Results and Discussion

#### 3.1 Generating images and annotations pairs

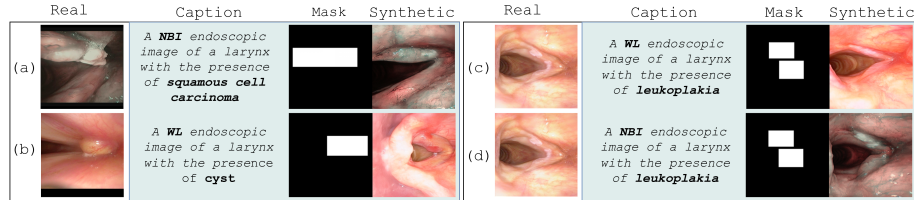


Fig. 2: Starting from real images (1st column), a caption containing relevant clinical information (2nd column) and the manual-annotated mask of the lesion bounding boxes (3rd column) were used to generate new samples (4th column).

Fig. 2 presents qualitative results of the image generation process, highlighting the ability of our method to produce realistic laryngeal images. The generated data reflect variations in vocal cord anatomy, including different lesion types – squamous cell carcinoma, cyst, leukoplakia – and acquisition settings – WL, NBI –, emphasizing the model’s capacity to reconstruct complex and diverse patterns. As shown in cases (c) and (d) of Fig. 2, we observed that the same mask can be associated with diverse synthetic outputs by using different captions. The  $FID_{ratio}$  metric was 0.836 as the difference between the  $FID_{rs}$  and  $FID_{rr}$  was small, while the IS was  $3.840 \pm 0.270$  for real samples and  $4.240 \pm 0.270$  for synthetic data, with both metrics confirming a high variety of the synthetic data.

#### 3.2 Improving detection by adding synthetic images

In Fig. 3a, we present the results of the downstream detection task. When testing on both the *internal* and *external* test sets, the introduction of synthetic data resulted in average performance improvements. On the *internal* test set,  $AP_{@IoU=0.5}$  peaked at 0.895 with 10% synthetic data, after which the performance reached a plateau. The performance trend of the *external* test matched the expectation, indicating that the highest performance could be reached with the addition of 10% synthetic data. However, the performance values fluctuate more due to inherent differences between training and test data. The advantage of adding the 10% of randomly selected synthetic data was further confirmed by the results of the 3-fold cross-validation (Table 1), with an average  $AP_{@IoU=0.5}$  increment regarding the use of only real data of +8.2% and +16.0% for *internal* and *external* tests, respectively. Nevertheless, the highest performance was observed with the uncertainty-based selection strategy, i.e. +9.0% and +22.1% for

*internal* and *external*  $AP_{@IoU=0.5}$ , respectively, meaning that extending the existing dataset with “challenging samples” can substantially improve the model’s ability to detect hard and diverse lesions and to generalize on unseen images.

### 3.3 Real versus synthetic: human-observer study

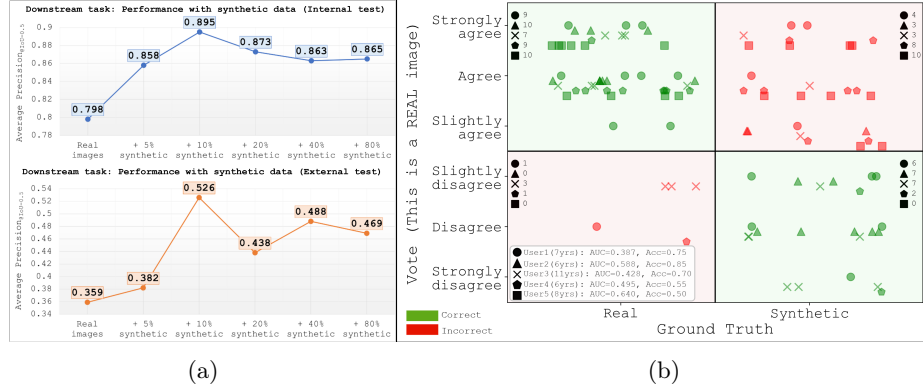


Fig. 3: (a) Evolution of the Average Precision $@IoU=0.5$  by gradually adding randomly selected (fold 1) synthetic images to the training dataset and testing the performance both internally and externally, showing peaks at + 10% synthetic samples. (b) Votes from otorhinolaryngologists assessing the realism of images, with confidence levels represented on the vertical axis and marker colors indicating whether the classification of images as real or synthetic was correct.

In Figure 3b, the responses of the enrolled otorhinolaryngologists were displayed in green for each correct real or synthetic image identified and red for

Table 1: Analysis of the impact of adding 10% synthetic data to the real training dataset, either randomly selected or based on the Uncertainty Estimation (UE) strategy.

Train dataset	Internal test			External test		
	Precision	Recall	$AP_{@IoU=50}$	Precision	Recall	$AP_{@IoU=50}$
Real	0.761	0.763	0.798	0.498	0.376	0.359
Real+10% synthetic						
Random selection (fold 1)	0.848	0.851	0.895	0.614	0.526	0.526
Random selection (fold 2)	0.828	0.889	0.900	0.660	0.463	0.518
Random selection (fold 3)	0.809	0.789	0.846	0.631	0.468	0.515
Mean $\pm$ std	0.828 $\pm$ 0.016	0.843 $\pm$ 0.041	0.880 $\pm$ 0.024	0.635 $\pm$ 0.019	0.485 $\pm$ 0.028	0.519 $\pm$ 0.004
Top uncertain selected	0.905	0.833	0.888	0.800	0.464	0.580



incorrect answers. On the y-axis, the predictions were discriminated by the confidence level of the user’s answer. Among all users, the AUC and accuracy values were  $0.507 \pm 0.095$  and  $0.670 \pm 0.129$ , respectively. Indeed, we can empirically affirm the realism of the synthetic data as clinicians struggled to recognize synthetic samples, often classifying them as real cases.

## 4 Conclusion

This study introduced an LDM and ControlNet-based framework for generating synthetic laryngeal image-annotation pairs, pioneering data generation in laryngology. We showed that the addition of challenging synthetic samples determined via uncertainty estimation improves lesion detection, and successfully validated the clinical realism of generated samples through expert reviews. Future work will further improve generation diversity, optimize sample selection and lesion-specific synthesis in privacy-preserving settings, and evaluate the utility of synthetic data for other tasks like lesion classification.

**Acknowledgments.** This work was partially funded by the European Union – NextGenerationEU and by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.5, project “RAISE - Robotics and AI for Socio-economic Empowerment” (ECS00000035), and Horizon Europe project AIRCARE, GA no. 101137426. It was also supported by AIMED (PID2023-146786OB-I00) from the Ministry of Science and Innovation of Spain, and the Artificial Intelligence and Bio-Medical Applications (AIBA) consolidated research group (2021SGR01094) of the Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR).

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

## References

1. Alyafi, B., Diaz, O., Marti, R.: Dcgans for realistic breast mass augmentation in x-ray mammography. In: Medical Imaging 2020: Computer-Aided Diagnosis. vol. 11314, pp. 473–480. SPIE (2020)
2. Andrade-Miranda, G.X., Godino Llorente, J.I., Julian D., A.L.: GIRAFE: Glottal imaging dataset for advanced segmentation, analysis, and facilitative playbacks evaluation (1.0) (2024)
3. Bajhaiya, D., Unni, S.N., Koushik, A.: Deep learning-powered generation of artificial endoscopic images of gi tract ulcers. *iGIE* **2**(4), 452–463 (2023)
4. Diamantis, D.E., Gatoula, P., Iakovidis, D.K.: Endovae: Generating endoscopic images with a variational autoencoder. In: 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). pp. 1–5. IEEE (2022)
5. Garrucho, L., Kushibar, K., Osuala, R., Diaz, O., Catanese, A., Del Riego, J., Bobowicz, M., Strand, F., Igual, L., Lekadir, K.: High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. *Frontiers in oncology* **12**, 1044496 (2023)

6. Guo, Z., Xie, J., Wan, Y., Zhang, M., Qiao, L., Yu, J., Chen, S., Li, B., Yao, Y.: A review of the current state of the computer-aided diagnosis (cad) systems for breast cancer diagnosis. *Open Life Sciences* **17**(1), 1600–1611 (2022)
7. Harris, M., Qi, A., Jeagal, L., Torabi, N., Menzies, D., Korobitsyn, A., Pai, M., Nathavitharana, R.R., Ahmad Khan, F.: A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. *PloS one* **14**(9), e0221339 (2019)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017)
9. Laves, M.H., Bicker, J., Kahrs, L.A., Ortmaier, T.: A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *International journal of computer assisted radiology and surgery* **14**, 483–492 (2019)
10. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 230–240. Springer (2024)
11. Liu, S., Chen, Z., Yang, Q., Yu, W., Dong, D., Hu, J., Yuan, Y.: Polyp-gen: Realistic and diverse polyp image generation for endoscopic dataset expansion. *arXiv preprint arXiv:2501.16679* (2025)
12. Low, D.M.: Vocal fold paralysis dataset and code (2021)
13. Martyniak, S., Kaleta, J., Dall’Alba, D., Naskręt, M., Płotka, S., Korzeniowski, P.: Simuscope: Realistic endoscopic synthetic dataset generation through surgical simulation and diffusion models. *arXiv preprint arXiv:2412.02332* (2024)
14. Moccia, S., Vanone, G.O., De Momi, E., Mattos, L.S.: Nbi-Infframes (2018)
15. Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., et al.: medigan: a python library of pre-trained generative models for medical image synthesis. *Journal of Medical Imaging* **10**(6), 061403–061403 (2023)
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 10674–10685 (2021)
17. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. *Advances in neural information processing systems* **29** (2016)
18. Sampieri, C., Baldini, C., Azam, M.A., Moccia, S., Mattos, L.S., Vilaseca, I., Peretti, G., Ioppi, A.: Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: A guide for physicians and state-of-the-art review. *Otolaryngology–Head and Neck Surgery* **169**(4), 811–829 (2023)
19. Watanabe, A., Taniguchi, M., Tsujie, H., Hosokawa, M., Fujita, M., Sasaki, S.: The value of narrow band imaging for early detection of laryngeal cancer. *European archives of oto-rhino-laryngology* **266**, 1017–1023 (2009)
20. Wen, D., Khan, S.M., Xu, A.J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A.K., Liu, X., et al.: Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health* **4**(1), e64–e74 (2022)

21. Yin, L., Liu, Y., Pei, M., Li, J., Wu, M., Jia, Y.: Laryngoscope8: Laryngeal image dataset and classification of laryngeal disease based on attention mechanism. *Pattern Recognition Letters* **150**, 207–213 (2021)
22. Yoon, D., Kong, H.J., Kim, B.S., Cho, W.S., Lee, J.C., Cho, M., Lim, M.H., Yang, S.Y., Lim, S.H., Lee, J., et al.: Colonoscopic image synthesis with generative adversarial network for enhanced detection of sessile serrated lesions using convolutional neural network. *Scientific reports* **12**(1), 261 (2022)
23. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
24. Zhu, S., Gao, J., Liu, L., Yin, M., Lin, J., Xu, C., Xu, C., Zhu, J.: Public imaging datasets of gastrointestinal endoscopy for artificial intelligence: a review. *Journal of Digital Imaging* **36**(6), 2578–2601 (2023)