Segmentation of endoscopy images of anterior nasal cavity using deep learning

Nonpawith Phoommanee[1], Peter J. Andrews[2,3], Terence S. Leung[1]
[1]Department of Medical Physics and Biomedical Engineering, University College London, London, WC1E 6BT. [2]Department of Rhinology and Facial Plastic Surgery, Royal National Throat, Nose and Ear Hospital, London, WC1E 6DG. [3]UCL Ear Institute, University College London, WC1X 8EE.

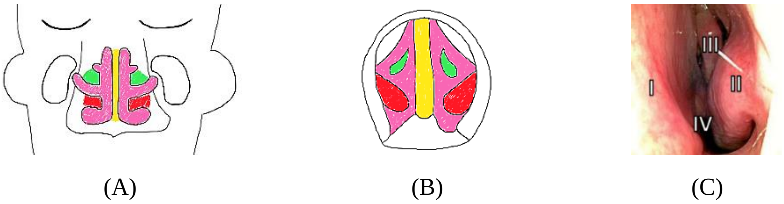
ABSTRACT

Nasal obstruction (NO), which affects one-third of the adult population, is characterized by a blockage in the nasal cavity. Rhinologists commonly employ nasal endoscopy (NE) in the differential diagnosis of NO, along with a focused history and other examinations such as skin prick tests and CT scans. This study aims to establish NE as a reliable standalone diagnostic tool, eliminating the necessity for CT scans and skin prick tests in the diagnosis of NO. However, currently, there is a lack of objective methods to quantify the severity of NO. To address this problem, we used deep learning to identify the anatomical structures of the anterior nasal cavity, which will then be graded by an objective grading system. In this paper, we evaluated the performance of various deep learning methods (DeepLabv3+, MaskFormer, and Mask2Former) with different pre-trained backbones (ResNet-101 - CNN-based, and Swin-Tiny - transformer-based), for semantic segmentation of the anterior nasal cavity. Sixty-two participants were examined with NE before and after using a nasal decongestant. For model training and validation, 608 images from 46 participants were utilized, and 171 images from 16 participants were reserved for testing. The fine-tuned Mask2Former with low-light image enhancement achieved a mean intersection-over-union of 81.7% and 61.2% on the validation and testing sets, respectively. These findings represent the first successful semantic segmentation of key anatomical structures within the anterior nasal cavity. These segmented structures will serve as the basis for classifying the severity of NO and diagnosing NO conditions, enabling AI-based consultations in primary care settings such as general practices and remote locations, where access to ENT expertise may be limited.

Keywords: nasal obstruction, segmentation, deep learning, transfer learning, low-light image enhancement

1. INTRODUCTION

Nasal obstruction (NO) results in difficulty in breathing through the nose due to a blockage in the nasal cavity. This is a common concern in otolaryngology practices and affects approximately one-third of adults worldwide[1]. Chronic rhinosinusitis (CRS) and allergic rhinitis (AR) are common mucosal nasal diseases. Structural changes in the nasal cavities, such as deviated nasal septum (DNS) and nasal valve collapse, are also significant contributors to NO. The NO symptoms include exacerbation of obstructive sleep apnea and decreased work efficiency.



(A) (B) (C)

Figure 1. The anterior nasal cavity in a frontal plane (A) and a nasal endoscopic view (B), and an endoscopic image of the left side of the nose (C), highlighting the key structures: Septum (I = yellow), IT (II = red), MT (III = green), and INV region (IV = pink).

The anterior nasal cavity is located inside the nose, as shown in Figure 1. The septum separates it into the left and right parts along the sagittal plane. Lateral nasal wall is on the side of the nose and includes key structures: the septum, inferior turbinate (IT), middle turbinate (MT), and internal nasal valve (INV). Adequate airflow and normal breathing require nasal resistance, which is critical for respiratory physiology. The IT, MT, and INV regulate airway resistance. It is important to

Medical Imaging 2024: Computer-Aided Diagnosis, edited by Weijie Chen,
Susan M. Astley, Proc. of SPIE Vol. 12927, 1292706 · © 2024
SPIE · 1605-7422 · doi: 10.1117/12.2691427

note that NO can occur if the INV area becomes narrower and contributes to about two-thirds of the total nasal airway resistance[2].

Nasal endoscopy (NE) is a common diagnostic tool used in the differential diagnosis of NO, along with clinical history and other examinations, including skin prick tests and CT scans. NE involves the use of a flexible, thin tube enclosing a small camera and bright lighting to examine the nasal passages, including the anterior part. This imaging technique provides high magnification and illumination, enabling clinicians to explain the management strategy to patients and pinpoint the exact location of bleeding and swelling in the nasal cavity.

NE is also considered the gold standard for diagnosing DNS, as the anatomical shape of the septum observed through the device is utilized in many DNS classification systems. Additionally, the presence of diseased mucosa, such as nasal polyps, mucopurulent secretions, or inflamed mucosa, indicates a possible diagnosis of CRS in patients, both with and without nasal polyps[3]. Bluish discoloration and mucosal edema further suggest a diagnosis of AR[4].

Currently, skin prick tests and CT scans are considered the gold standard for diagnosing AR and CRS, respectively. However, this study aims to establish NE as a reliable standalone diagnostic tool, eliminating the need for CT scans and skin prick tests in the diagnosis of NO. Differentiating between AR and CRS patients is achieved through the administration of nasal decongestants (ND). The CRS patient group exhibited a lower percentage change in unilateral nasal inspiratory peak flow (uNIPF) after applying nasal decongestant compared to the non-CRS group. Conversely, the AR patient group showed an opposite trend[5].

Considering that the INV area is associated with NIPF, we propose implementing an objective grading system to assess the severity of NO. This grading system would build upon existing subjective grading systems that consider the visibility of the MT during NE. By evaluating the severity of NO on each side of the nose, both pre- and post-ND, it becomes possible to diagnose whether the patient has at least one of the following conditions: AR, CRS, or DNS.

However, there is currently a lack of objective methods to quantify the severity of NO. Therefore, the introduction of an objective method based on NE could prove highly beneficial in transitioning from utilizing NE in secondary care to incorporating portable endoscopy in primary care settings. This transition becomes particularly significant in situations where an Ear, Nose and Throat (ENT) specialist is unavailable or when early detection of NO is crucial. By implementing this objective method, a standardized approach to assessing NO severity, diagnosing conditions, and determining appropriate treatment interventions can be established.

The accurate identification and quantification of anatomical structures in the anterior nasal cavity, to be utilized in the proposed objective grading system, require effective image segmentation. Convolutional neural network (CNN)-based methods have proven to be robust and effective in handling imaging noise, making them well-suited for image analysis. Furthermore, transformer models have garnered attention in the field of image segmentation due to their self-attention mechanism, enabling them to capture long-range dependencies between pixels and effectively handle multi-scale information. These advantageous properties of transformers have generated significant interest in the medical community for their adaptation to medical image segmentation tasks.

This paper demonstrates the segmentation of key anatomical structures of the anterior nasal cavity, which will be valuable for the diagnosis of NO. We aim to explore the state-of-the-art semantic segmentation techniques applied to NE data from different body regions. We will evaluate the performance of deep learning networks in segmenting the anterior nasal anatomy. This work contributes to the study of NO as follows: 1) We collected and annotated an NE-UCLH dataset from 62 participants with or without NO, providing valuable resources for future research. 2) We developed and compared deep learning models that accurately performed semantic segmentation of the anatomy of the anterior nasal cavity.

2. METHODS

2.1 Data collection

Sixty-two participants were recruited at Royal National ENT and Eastman Dental Hospitals, including 7 controls and 55 patients. The patient group comprised 31 males and 24 females, with a mean age of 45.7 years (95% CI 41.9-49.5), while the control group included 2 males and 5 females, with a mean age of 39.0 years (95% CI 30.8-47.2). Controls aged 18 years or above had no history of rhinological conditions or symptoms. Based on their medical history and visual inspection, patients were diagnosed by an otolaryngologist with one or more of the following conditions: AR, CRS, or DNS. We obtained full ethical approval from the London - City & East Research Ethics Committee, with reference number 15/LO/0187, and written consent was obtained from all participants.

We administered questionnaires, including VAS, NOSE, and SNOT-23, as well as demographic and clinical history forms. Each participant underwent skin prick tests, NIPF, acoustic rhinometry, and NE. We recorded NIPF for bilateral and unilateral airflows in three attempts using a modified Youlten flow meter (Clement Clark International). The nasal cavity's minimum cross-sectional area (MCA) was measured using the A1 Acoustic Rhinometer (GM Instruments, Kilwinning). We recorded the nasal cavities using a Video Naso-Pharyngo-Laryngoscope (VNL9-CP) (Pentax Medical), and the videos were processed using VIVIDEO Video Processor (CP-1000) (Pentax Medical). We also applied a nasal decongestant spray (xylometazoline hydrochloride 0.1% w/v) and repeated the measures and VAS questionnaire responses 10 minutes later to investigate the decongestant effect.

2.2 Datasets

Up to two frames were extracted for each nasal side from each video: one frame showed the complete IT, and the other showed the most noticeable MT. If there were no or minimally visible nasal hair, 10 and 20 frames before and after these frames were selected. All images of the right nasal side were flipped horizontally since nasal asymmetry is minor in most cases. This dataset is presented in two configurations: the pre-processed version, utilizing low-light image enhancement (LLE) with alpha blending, and the original non-LLE version.

The segmentation and frame classification were performed by the first author, who received training from an ENT consultant with over 10 years of experience in ENT training (the second author). The segmentation map consisted of six classes: septum, IT, MT, polyp, others, and airway, visually represented by different gray values in the segmentation maps (as shown in Figure 2). The visibility of the MT was graded on a scale (grades 0-2) proposed by Patel et al.[6], where Grade 0 indicates an unblocked view of the MT (no NO), Grade 1 indicates that the MT cannot be fully seen (moderate NO), and Grade 2 indicates complete invisibility of the MT (severe NO).

For model training and evaluation, we selected every 4th participant as the testing set, while the remaining participants were assigned to the training set. The NE-UCLH dataset comprised 608 images from 46 participants for the training and validation sets, while the testing set included 171 images from 16 participants.

2.3 Implementation details

The network was implemented in Python 3.10 using PyTorch 2.0.1 with CUDA 11.7 backend. We utilized the following deep learning networks: DeepLabv3+[7], MaskFormer[8], and Mask2Former[9], on different pre-trained backbones: ResNet-101 (CNN-based)[10], and Swin-Tiny (transformer-based)[11] on either PASCAL VOC 2012[12] or ADE20k[13] dataset, then fine-tuned on our NE-UCLH dataset. The Adam optimization algorithm was used to update the network weights, with an initial learning rate of 0.0001, decaying by 0.3 every 10 epochs. The gradient decay factor was set to $\beta_1 = 0.9$, and the squared gradient decay factor was set to $\beta_2 = 0.999$. The training was conducted on a single GPU (NVIDIA GeForce RTX 3050 Ti, Nvidia Corp., Santa Clara, CA, USA) for 30 epochs with a mini-batch size of 1. L2 normalization with a weight decay rate of 10^{-4} was applied for regularization. Early stopping was used to prevent overfitting if the validation accuracy did not improve for 4 consecutive epochs. The weight configuration of the epoch with the lowest validation loss was chosen.

For validation and testing, a 5-fold cross-validation was performed on the training set. Data augmentation was used through rotations between -5° and +5°, while avoiding vertical flipping and additional rotations as they were inappropriate for clinical images. We used all trained networks from cross-validation with the testing set for testing. The mean Intersection over Union (mIoU) evaluation metric was used to quantify pixel-wise overlap between semantic segmentation results and ground truth. Frames-per-second (fps) measurement was conducted on an RTX 3050 GPU with a mini-batch size of 1.

3. RESULTS AND DISCUSSION

Table 1: Deep learning methods compared for semantic segmentation on NE-UCLH dataset with LLE.

Method	Backbone	Pre-trained dataset	Crop size	#Params.	mIoU_{val} (% ± SD)	$\text{mIoU}_{\text{test}}$ (% ± SD)	fps
DeepLabv3+[7]	R101	PASCAL VOC 2012	513 × 513	59M	76.0 (3.4)	54.0 (3.9)	9.21
MaskFormer[8]	R101	ADE20k	512 × 512	60M	54.6 (8.4)	44.7 (9.8)	14.35
	Swin-T	ADE20k	512 × 512	**42M**	74.2 (13.8)	57.6 (5.4)	**14.55**
Mask2Former[9]	Swin-T	ADE20k	512 × 512	47M	**81.7** (4.7)	**61.2** (4.0)	6.28

Table 1 presents a comparison of the performance of various semantic segmentation models on the NE-UCLH dataset with LLE. Among these models, Mask2Former achieved the best mIoU of 81.7% on the validation set and 61.2% on the testing set. Additionally, it displayed lower variability in mIoU scores compared to MaskFormer. A paired samples t-test revealed a significant difference between the mIoU of Mask2Former and other methods for the testing set ($p < 0.05$ for all pairs). However, when the Swin-T backbone was considered, MaskFormer exhibited a slightly lower mIoU of 74.2% on the validation set and 57.6% on the testing set. Despite this, it had fewer parameters compared to DeepLabv3+ and MaskFormer with the ResNet-101 backbone, leading to faster inference speed.

Due to its marginal performance advantage, we chose to utilize the Mask2Former network to compare its performance with the NE-UCLH dataset without LLE, as outlined in Table 2. In this scenario, it achieved a slightly reduced mIoU of 80.2% on the validation set ($p = 0.130$) and 57.6% on the testing set ($p = 0.027$). After applying LLE, there was an overall improvement in per-class mIoU observed across most classes, particularly evident in the polyp class on the testing set. The model's limited generalization to the testing set, especially with polyp classes, implied overfitting to the training data, likely due to the minimal presence of polyp pixels (0.7% of the dataset).

Table 2: Per-class mIoU (%) ± SD of Mask2Former for different configurations on the NE-UCLH dataset.

Dataset	Configuration	Per-class mIoU						mIoU
		Septum	IT	MT	Polyp	Others	Airway	
Val	Non-LLE	96.8 (0.7)	90.2 (2.0)	83.6 (2.0)	**69.9** (9.9)	79.4 (2.2)	61.4 (2.0)	80.2 (4.4)
	LLE	**97.3** (0.6)	**91.2** (1.5)	**84.0** (1.2)	69.2 (11.0)	**81.9** (1.0)	**66.5** (2.0)	**81.7** (4.7)
Test	Non-LLE	91.5 (0.6)	**79.4** (1.3)	56.4 (2.0)	7.6 (6.3)	**67.0** (2.1)	45.8 (2.4)	57.6 (4.0)
	LLE	**91.8** (0.4)	75.7 (2.2)	**63.5** (3.5)	**19.3** (8.1)	64.3 (3.7)	**52.6** (0.5)	**61.2** (4.0)

Both Mask2Former and MaskFormer with a transformer backbone, as shown in Figure 2, enhanced edge localization by integrating long-range pixel dependencies. The outcomes across all networks closely approximated the ground truth in Grade 0 image due to the high visibility of MT. However, all networks exhibited error instances involving mucus occurrence. In Grade 1 image, differences among the IT boundaries were noticeable. Furthermore, only Mask2Former effectively separated classes, preventing one label from encroaching onto another. In Grade 2 image, especially with the occurrence of polyp class, suboptimal performance was observed in all networks. Remarkably, Mask2Former produced the most comprehensive representation of polyps, as expected based on the results in Table 2.

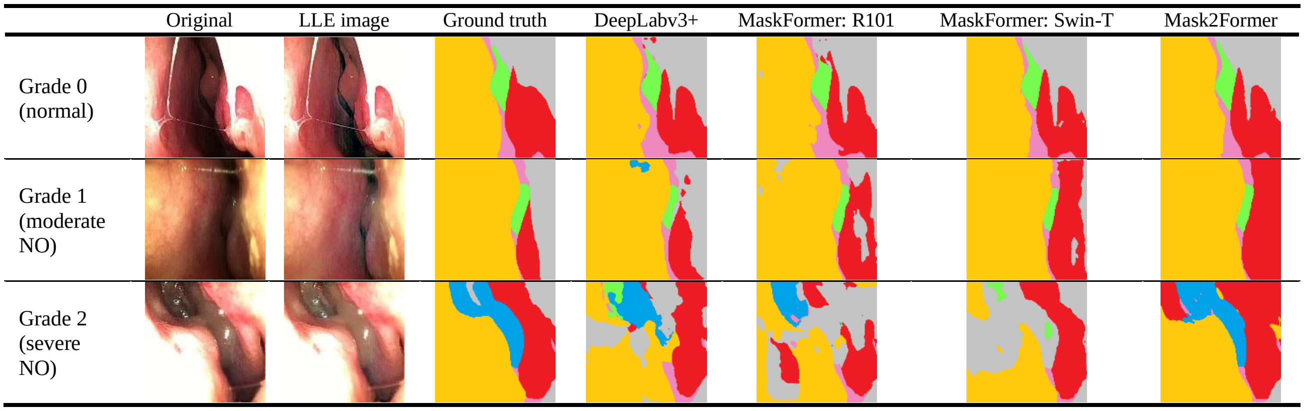


Figure 2. Example images extracted from the testing set, showcasing distinct grades of NO. The figure includes the corresponding ground truth segmentation maps featuring classes including septum (yellow), IT (red), MT (green), polyp (blue), others (gray), and airway (pink), along with semantic segmentations generated by various deep learning methods.

Based on the aforementioned performance, refining the model's hyperparameters and achieving improved results with a larger, polyp-focused dataset is necessary to enhance the performance, especially model generalization, of anatomical segmentation in the anterior nasal cavity. However, the current strategies for frame-based semantic segmentation are deemed suitable, as this method demands less computational resources compared to video semantic segmentation. Additionally, the reported inference time is already suitable for real-time ENT examinations in primary care settings where live labeling of key anatomical structures is required. Further studies will involve the evaluation of extracting clinically validated explainable features based on the generated segmentation for the automatic grade classification of NO.

4. CONCLUSIONS

This paper presents the first successful attempt to achieve good-quality semantic segmentation of key anatomical structures in the anterior nasal cavity. The outcomes of this segmentation serve a crucial purpose in grade classification of NO and diagnosis of common conditions of NO. Future steps involve dataset expansion, validation of the grading system, and the establishment of a hybrid framework involving machine learning networks and a rule-based expert system for the differential diagnosis of NO. We will explore transfer learning for adapting to portable endoscopy. This study holds significant potential for AI-based consultations in primary care settings: general practices and remote locations where access to ENT expertise may be limited.

REFERENCES

[1] Valero, A., Navarro, A. M., Del Cuvillo, A., et al., “Position paper on nasal obstruction: Evaluation and treatment,” J Investig Allergol Clin Immunol 28(2), 67–90 (2018).

[2] Hsu, D. W. and Suh, J. D., “Anatomy and Physiology of Nasal Obstruction,” Otolaryngol Clin North Am 51(5), 853–865 (2018).

[3] Fokkens, W. J., Lund, V. J., Hopkins, C., et al., “European Position Paper on Rhinosinusitis and Nasal Polyps 2020.” (2020).

[4] Ziade, G. K., Karami, R. A., Fakhri, G. B., et al., “Reliability Assessment of the Endoscopic Examination in Patients with Allergic Rhinitis,” Allergy and Rhinology 7(3) (2016).

[5] Li, C. H., Kaura, A., Tan, C., Whitcroft, K. L., Leung, T. S. and Andrews, P., “Diagnosing nasal obstruction and its common causes using the nasal acoustic device: A pilot study,” Laryngoscope Investig Otolaryngol 5(5), 796–806 (2020).

[6] Patel, B., Virk, J. S., Randhawa, P. S. and Andrews, P. J., “The internal nasal valve: a validated grading system and operative guide,” European Archives of Oto-Rhino-Laryngology 275(11), 2739–2744 (2018).

[7] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H., “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation” (2018).

[8] Cheng, B., Schwing, A. G. and Kirillov, A., “Per-Pixel Classification is Not All You Need for Semantic Segmentation” (2021).

[9] Cheng, B., Misra, I., Schwing, A. G., et al., “Masked-attention Mask Transformer for Universal Image Segmentation,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2022-June, 1280–1289, IEEE Computer Society (2022).

[10] He, K., Zhang, X., Ren, S. and Sun, J., “Deep Residual Learning for Image Recognition” (2015).

[11] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” Proceedings of the IEEE International Conference on Computer Vision, 9992–10002, Institute of Electrical and Electronics Engineers Inc. (2021).

[12] Everingham, M. and Winn, J., “The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit” (2012).

[13] Zhou, B., Zhao, H., Puig, X., et al., “Scene parsing through ADE20K dataset,” Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-January, 5122–5130, Institute of Electrical and Electronics Engineers Inc. (2017).