T2ID-CAS: Diffusion Model and Class Aware Sampling to Mitigate Class Imbalance in Neck Ultrasound Anatomical Landmark Detection

UCF College of Medicine
UNIVERSITY OF CENTRAL FLORIDA

Manikanta Varaganti[1], Amulya Vankayalapati[2], Nour Awad[2], Farhan Fuad Abir[3], Gregory R. Dion[2], and Laura J. Brattain[1,3]

University of Cincinnati College of Medicine

[1]Department of Computer Science, University of Central Florida, Orlando, FL, USA
[2]Department of Otolaryngology Head Neck Surgery, University of Cincinnati College of Medicine, OH, USA
[3]Department of Internal Medicine, University of Central Florida College of Medicine, Orlando, FL, USA

Introduction

- **Neck ultrasound (NUS)** is widely used for real-time, non-invasive assessment of airway structures, supporting procedures.
- **Deep learning (DL)-based object detection models** can automate anatomical landmark identification in US images, improving speed and consistency.
- **Class imbalance** hinders detection of critical but underrepresented structures like tracheal rings and vocal folds.

Hypothesis

Combining text-to-image diffusion and class-aware sampling can significantly improve detection of underrepresented anatomical classes in ultrasound.

Dataset Description

- The research was approved by the Institutional Review Board (IRB).
- NUS was collected from **10 adults** (3 Male/ 7 Female, Average Age 52.6 ±14.5) using a Terason uSmart 3200t ultrasound device.
- Multiple **10-second** transverse cineloops per subject were recorded covering key airway regions.
- There are **6 classes:** Thyroid cartilage, cricoid cartilage, strap muscles, thyroid lobes, vocal folds, tracheal rings.
- Fig. 1 shows the class imbalance of the dataset resembling long-tailed distribution.
- **7,464 frames** were resized to **320×320 pixels** for model training.

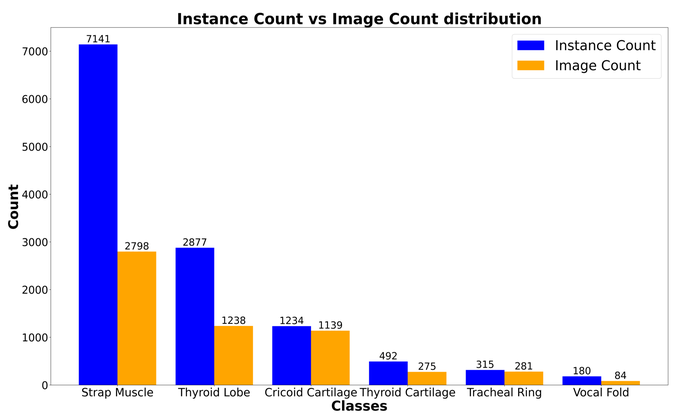

Fig. 1: Long-tailed distribution of instance and image counts per class in the neck US dataset.

Method

T2ID-CAS integrates three strategies (Fig. 2) to mitigate class imbalance in NUS.

1. **Text-to-Image Latent Diffusion + LoRA**
   - Fine-tuned Stable Diffusion XL (SDXL) using Low Range Adaptation (LoRA) on **840** annotated ultrasound images of vocal folds and tracheal rings
   - Generated **600** synthetic images (512×512) using class-specific text prompts
   - Evaluated using **FID** (Fréchet Inception Distance), **IS** (Inception Score), and **CLIP Score** (Contrastive Language-Image Pretraining.
2. **Data Sampling and Augmentation**
   - **Mosaic + Mixup:** Combines multiple images and blends image-label pairs.
   - **Repeat Factor Sampling (RFS):** Increases sampling frequency of rare classes by duplicating images based on inverse class frequency.
   - **Class-Aware Sampling (CAS):** Ensures balanced training by sampling each class equally during mini-batch construction.
3. **Detection Model – YOLOv9s**
   - Used lightweight YOLOv9-small model for fast and efficient anatomical landmark detection.
   - Trained under multiple setups: baseline, data augmentations, CAS, and CAS + synthetic data (T2ID-CAS).

All experiments were conducted on an NVIDIA H100 GPU with 81GB memory.

Results

A. Image Generation (SDXL)

Table 1: Evaluation of the synthetic images

Model	Class	FID ↓	IS ↑	CLIP Score ↑
SD v1-4	Vocal Fold	9.55	17.13	26.31
SDXL-LoRA		**5.54**	**17.62**	**29.81**
SD v1-4	Tracheal Ring	18.12	11.04	28.66
SDXL-LoRA		**16.11**	**18.18**	**30.40**

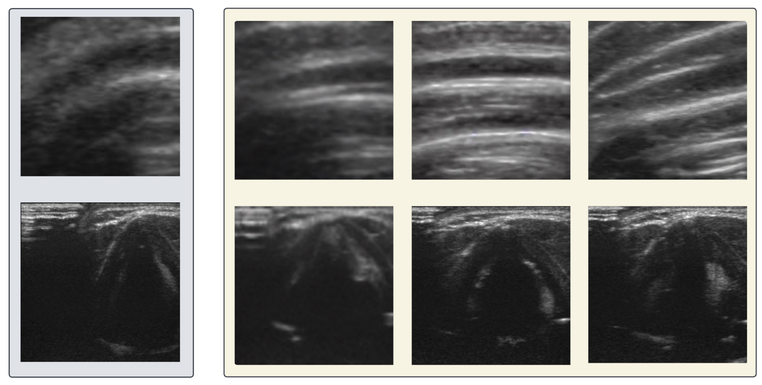

Fig. 3: Comparison between original images and synthetic images by SDXL. The synthetic images show close resemblance to the ground truth ones.

B. Object Detection

Table 2: Performance metrics of different strategies evaluated on YOLOv9s

Strategy	Overall	Tracheal Ring	Vocal Fold
Baseline	66.0%	38.5%	75.6%
Mosaic + Mixup	66.5%	36.9%	74.2%
Repeat Factor Sampling (RFS)	65.7%	37.1%	74.9%
Class-Aware Sampling (CAS)	84.3%	63.4%	95.0%
Baseline + SDXL	75.2%	82.1%	94.6%
RFS + SDXL	74.9%	81.0%	94.7%
T2ID-CAS (Ours)	**88.2%**	**90.5%**	**98.2%**

Conclusion

- **T2ID-CAS** mitigates class imbalance using SDXL-generated images and CAS.
- It achieved over **22%** accuracy gain, specially for tracheal rings and vocal folds.
- It has the potential to enhance the precision and safety in ultrasound-guided airway management.
- In the future, we plan to test on larger datasets and explore text prompt optimization for improved synthetic image generation.

References

1. D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.
2. A. Osman and K. M. Sum, "Role of upper airway ultrasound in airway management," *Journal of Intensive Care*, vol. 4, no. 1, p. 52, 2016.

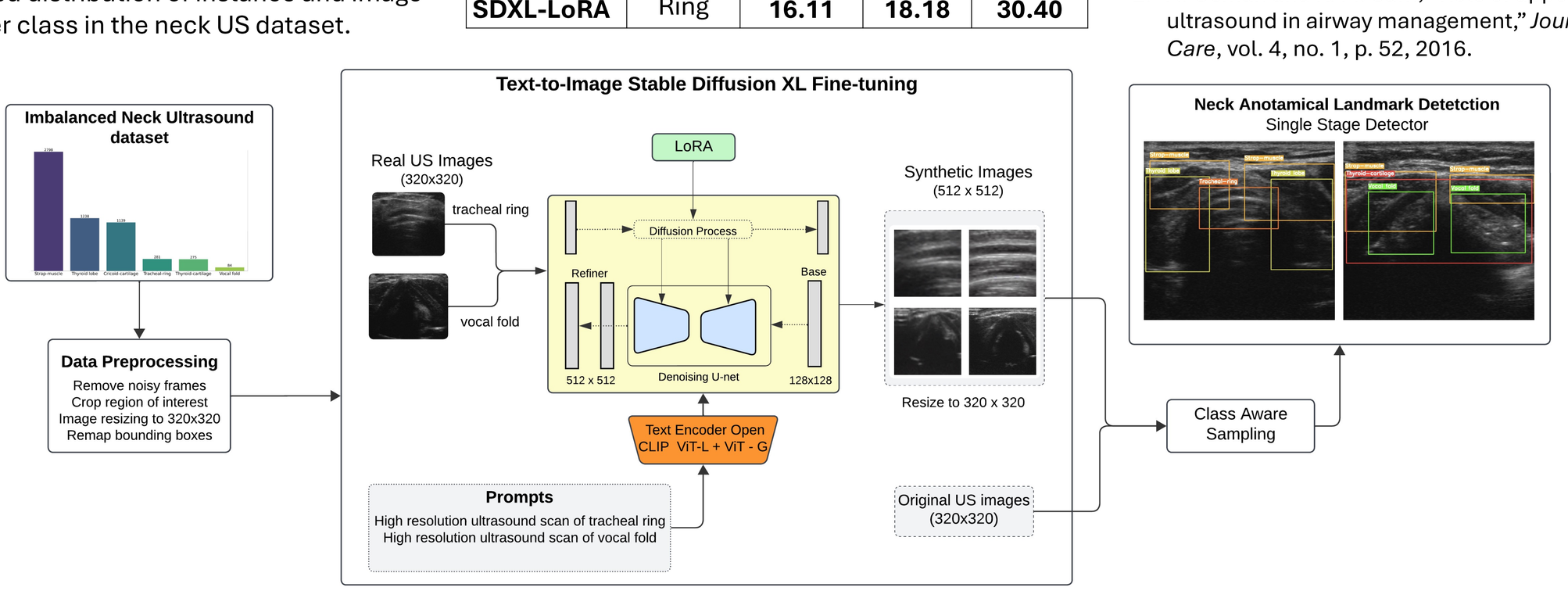

Fig. 2: Overview of the proposed T2ID-CAS framework.