Highlights

Detection-Gated Glottal Segmentation with Zero-Shot Cross-Dataset Transfer and Clinical Feature Extraction

Harikrishnan Unnikrishnan

- Detection gate zeroes output after at most 3 missed frames, removing spurious detections.
- New state of the art on the **GIRAFE** benchmark with 0.809 Dice, beating all published baselines.
- YOLO-Crop variant improves zero-shot transfer to independent BAGLS dataset.
- Coefficient of variation distinguishes voices by pathology after sex control.
- Lightweight pipeline processes full HSV videos in seconds on consumer hardware.

Detection-Gated Glottal Segmentation with Zero-Shot Cross-Dataset Transfer and Clinical Feature Extraction

Harikrishnan Unnikrishnan[a]

[a]*Orchard Robotics, San Francisco, CA, USA*

Abstract

Quantitative analysis of vocal-fold vibration from high-speed videoendoscopy (HSV) requires segmenting the glottal area in each frame, a task that is prohibitively labour-intensive when performed manually. Existing automated methods are typically evaluated on a single dataset and lack a principled strategy for handling frames in which the endoscope is off-target. We propose a *detection-gated* segmentation pipeline that pairs a YOLOv8 glottis detector [1] with a lightweight U-Net segmenter [2]. The detector serves two roles: (*i*) it gates the segmentation output, zeroing the predicted area when no glottis is detected (or after at most 3 consecutive missed frames), thereby removing spurious detections and preventing waveform artefacts in clinical recordings; and (*ii*) in a second variant (YOLO-Crop+UNet), it crops the detected region and rescales it to fill the U-Net's input canvas, providing higher effective resolution at the glottis boundary. On the **GIRAFE** benchmark our U-Net alone achieves a mean Dice of 0.809, and the detection-gated YOLO+UNet pipeline achieves 0.746—both surpassing all three published baselines (InP 0.713, U-Net 0.643, SwinUNetV2 0.621). In a zero-shot cross-dataset experiment on **BAGLS** (3500 frames from a different institution and endoscope, with no **BAGLS** training data), YOLO-Crop+UNet with an optimised detection threshold reaches Dice 0.659 versus 0.588 for a YOLO-free U-Net baseline. Finally, running the pipeline on all 65 **GIRAFE** patients' full video recordings yields a *Glottal Area Waveform* (GAW) per patient. Sex-stratified analysis of seven kinematic features shows that the coefficient of variation of the glottal area significantly distinguishes Healthy from Pathological voices in the female subgroup (p=0.006, Mann–Whitney U), consistent with the established finding that laryngeal pathologies reduce vibratory regularity. Code, trained weights, and evaluation scripts are released at `https://github.com/hari-krishnan/openglottal`.

Keywords: glottal segmentation, high-speed videoendoscopy, YOLOv8, U-Net, glottal area waveform, vocal fold pathology, cross-dataset generalisation

1. Introduction

High-speed videoendoscopy (HSV) enables frame-by-frame observation of vocal fold vibration at several thousand frames per second, making it the gold standard for objective voice assessment in clinical laryngology [3]. The central derived quantity is the *Glottal Area Waveform* (GAW)—the per-frame area of the glottal opening as a function of time—from which kinematic biomarkers such as open quotient, fundamental frequency, and vibration regularity can be computed [4, 3, 5, 6, 7].

Accurate glottal segmentation is the bottleneck. Rule-based methods (active contours, level sets, optical flow) struggle with the wide variability in illumination, endoscope angle, and patient anatomy [5]. Deep learning holds promise for this task, yet results so far are mixed: Andrade-Miranda et al. released the **GIRAFE** dataset—760 expert-annotated HSV frames from 65 patients—and found that the classical InP method (Dice 0.713) still outperformed their U-Net (Dice 0.643) and SwinUNetV2 (Dice 0.621), which the authors attributed to the small training set [8]. Nevertheless, two important gaps remain:

1. **Robustness.** Clinical recordings routinely contain frames in which the glottis is not visible (scope insertion, coughing, endoscope motion) [3]. Existing segmentation models are not equipped to detect this condition and produce spurious non-zero area predictions, corrupting the downstream GAW.
2. **Generalisation.** Published methods are evaluated on a single dataset. Whether the learned representations transfer to images from a different institution, camera system, or patient population is unknown.

We address both gaps with a detection-gated pipeline (Section 3) and evaluate it on two independent public datasets (Section 4). Our contributions are:

- A *detection gate*: YOLO-based glottis detection is used as a binary switch—when YOLO fires, U-Net prediction within the detected bounding box is reported; when it does not, the previous box is held for

at most 3 consecutive frames and then the detection is zeroed. Only by zeroing after this short hold do we remove spurious detections on non-glottis frames (e.g. closed glottis, scope motion) without post-hoc filtering.

- A *crop-zoom variant* (YOLO-Crop+UNet): the detected bounding box is cropped and resized to the full U-Net canvas, providing higher effective pixel resolution at the glottal boundary and improved cross-dataset generalisation.
- *End-to-end GAW analysis*: the pipeline is applied to all 65 **GIRAFE** patients' full recordings and kinematic features are extracted; the coefficient of variation significantly distinguishes Healthy from Pathological groups even after controlling for sex imbalance.

2. Related Work

2.1. Classical glottal segmentation

Early methods employed active contours and level-set evolution seeded by manually placed landmarks [5]. Optical-flow-based trackers and morphological inpainting variants (InP) remained competitive for years due to the limited size of labelled datasets [8].

2.2. Deep learning

The publication of the **GIRAFE** benchmark [8] enabled a rigorous comparison: their U-Net [2] (Dice 0.643) and the transformer-based SwinUNetV2 [9] (Dice 0.621) were both outperformed by the classical InP method (Dice 0.713), which the authors attributed to the small training set size. The **BAGLS** dataset [10] provides 55 750 training and 3500 test frames from multiple endoscope systems without patient-level diagnoses, making it a natural testbed for generalisation. The original **BAGLS** paper validated the benchmark with a U-Net baseline achieving IoU ≈ 0.89 on the test split. Subsequent work demonstrated that a single latent bottleneck channel suffices for accurate glottis segmentation [11], while Fehling et al. incorporated temporal context via a convolutional LSTM encoder–decoder, reaching Dice 0.85 on 13 000 frames from 130 subjects [12]. Kist et al. packaged three quality-tiered segmentation networks into the *Glottis Analysis Tools* (GAT) software for clinical use [13]. These temporal and recurrent approaches improve consistency across

frames but require substantially more GPU memory and training data than frame-level models, limiting their applicability on small datasets such as **GIRAFE**.

2.3. Foundation models

The Segment Anything Model (SAM) [14] and its medical variants [15] have demonstrated strong zero-shot segmentation across diverse imaging domains when provided with a point or bounding-box prompt. However, SAM's ViT-H backbone (636M parameters) is an order of magnitude larger than the pipeline proposed here and requires per-frame prompting, making it impractical for real-time GAW extraction from thousands of HSV frames.

2.4. Detect-then-segment

Two-stage pipelines—region proposal followed by per-region segmentation—are standard in general object segmentation [16] but have not been systematically applied to glottal HSV. Closest to our work, [17] used a bounding-box initialisation for active-contour tracking, but did not gate the output on detection confidence.

2.5. GAW feature analysis

Kinematic features of the GAW are well-established clinical measures [7, 4]. Significant differences in open quotient and fundamental frequency between healthy and disordered voices have been reported, though studies are typically limited to small, single-institution cohorts.

3. Methods

3.1. Datasets

GIRAFE. The **GIRAFE** dataset [8] contains 760 high-speed laryngoscopy frames (256×256 px) from 65 patients (adults and children, healthy and pathological) with pixel-level glottal masks annotated by expert clinicians. Frames are grouped into official training / validation / test splits (600 / 80 / 80 frames; test patients: 57A3, 61, 63, 64). Splits are strictly at the patient level: the 30 training patients, 4 validation patients, and 4 test patients are disjoint sets, ensuring that no patient's anatomy appears in both training and evaluation. Each patient folder also contains the full AVI recording (median length 502 frames at 4000 fps) and a metadata file recording the disorder status (Healthy, Paresis, Polyps, Diplophonia, Nodules, Paralysis, Cysts, Carcinoma, Multinodular Goiter, Other, or Unknown).

BAGLS. The Benchmark for Automatic Glottis Segmentation (**BAGLS**) [10] contains 55 750 training and 3500 test frames from multiple endoscope types and institutions. Image dimensions vary (256×256 to 512×512); each frame is paired with a binary glottal mask. No patient-level labels are provided. Crucially, **BAGLS** was not used in any training step—it serves exclusively as a zero-shot cross-dataset evaluation set.

3.2. Pre-processing

GIRAFE. Images are used at their native 256×256 resolution.

***BAGLS** letterboxing.* Variable-size **BAGLS** frames are letterboxed to 256×256: the longer side is scaled to 256 pixels while maintaining aspect ratio, and the remaining dimension is zero-padded symmetrically. The same transformation is applied identically to the GT mask to maintain spatial correspondence.

3.3. YOLO Glottis Detector

We fine-tune YOLOv8n [1] on bounding boxes derived from the **GIRAFE** training split. GT bounding boxes are computed as the tight enclosing rectangle of each GT mask, then converted to YOLO label format. Training runs for 100 epochs using the default YOLOv8 augmentation pipeline.

At inference time a *TemporalDetector* wrapper provides temporal consistency without the memory overhead of 3D convolutions or recurrent architectures [12]: the detected box centre is drift-clamped to at most 30 pixels per frame, and when the detector misses a frame the previous bounding box is held for at most 3 consecutive misses before the detection is *zeroed* (output set to zero until YOLO fires again). Only by zeroing after this short hold do we remove spurious detections (e.g. stale boxes from closed glottis or scope motion). The box size is updated from each fresh detection. This suppresses spurious jumps caused by reflections or instrument occlusion while allowing the natural opening–closing motion of the glottis. Because all temporal reasoning resides in the lightweight detection wrapper rather than in the segmentation backbone, the U-Net itself remains a standard 2D model that can be trained on the small **GIRAFE** training set (600 frames) without risk of temporal overfitting.

3.4. U-Net Segmenter

We train two U-Net [2] variants with a four-level encoder–decoder (channel widths $32, 64, 128, 256$, 7.76M parameters).

Full-frame U-Net (`unet_only`). Input: 256×256 grayscale frame. Training data: 600 **GIRAFE** training frames with augmentation (random flips, $\pm 30°$ rotation, $\pm 15\,\%$ scale jitter, brightness / contrast / Gaussian blur perturbations). Loss: $0.5 \cdot \mathcal{L}_{\text{BCE}} + 0.5 \cdot \mathcal{L}_{\text{Dice}}$ [18]. Optimiser: AdamW [19], learning rate 10^{-3}, cosine annealing [20] over 50 epochs.

Crop-mode U-Net (`unet_crop`). For each training frame, the YOLO detector is run and the detected bounding box (plus 8 px padding on each side) is cropped and resized to 256×256. The matching GT mask undergoes the same crop-resize. Frames with no YOLO detection are excluded (487 training crops / 77 validation crops retained out of 600 / 80 frames). Training procedure is identical to the full-frame model, saving to a separate checkpoint (`unet_crop.pt`).

3.5. Inference Pipelines

Five pipelines are evaluated (Figure 1):

U-Net only. Run the full-frame U-Net on the 256×256 grayscale input; output the thresholded probability map directly. No detection gate—every frame produces a prediction.

YOLO+UNet. (1) Run the YOLO detector on the full frame. (2) Run the full-frame U-Net on the full frame. (3) Zero the U-Net mask outside the detected bounding box. If YOLO does not fire (or after 3 consecutive misses), output is all-zero, removing spurious detections.

YOLO-Crop+UNet. (1) Run the YOLO detector. (2) Crop the detected region (+8 px padding), resize to 256×256. (3) Run the crop-mode U-Net on the resized crop. (4) Resize the output mask back to the original crop dimensions. (5) Paste into a full-frame zero mask at the detected coordinates. If YOLO does not fire (or after 3 consecutive misses), output is all-zero.

YOLO+Motion. (1) Run the YOLO detector. (2) Feed the grayscale frame and bounding box to a motion-based tracker (YOLO-GuidedVFT [21, 7], extending our frame-differencing approach) that segments the glottis via temporal frame differencing within the detected region. The first frames are used for tracker initialisation and excluded from metrics.

YOLO+OTSU (baseline). Otsu thresholding [22] (inverted, glottis is dark) within the YOLO bounding box. No learned segmentation component.

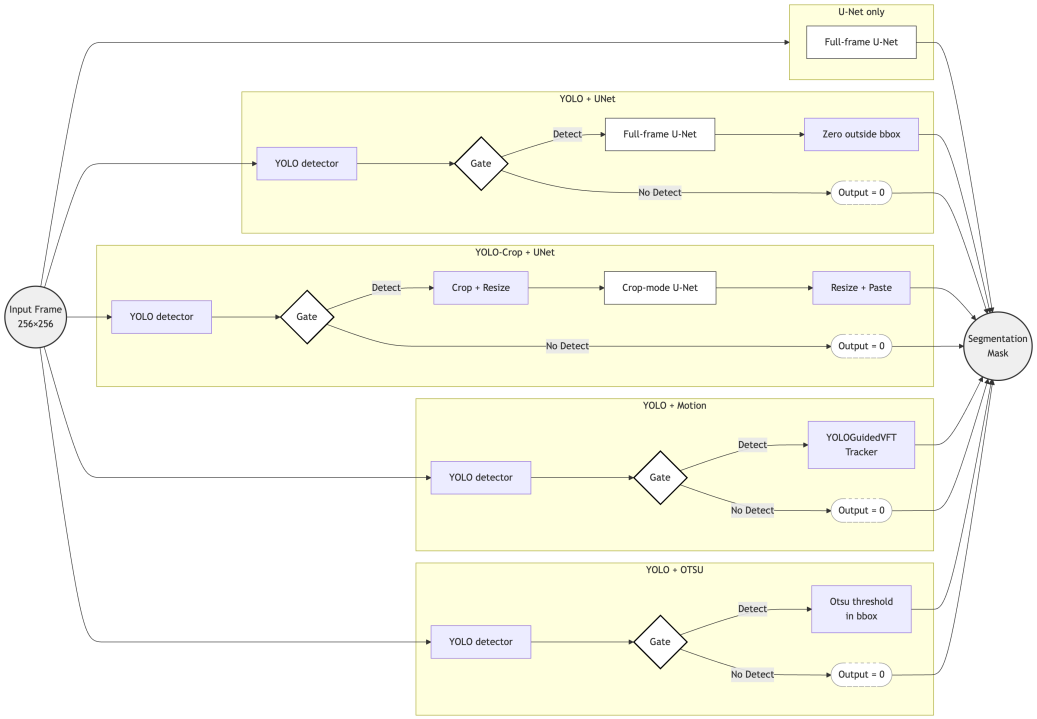

Figure 1: Overview of the five inference pipelines. Input (left) is the 256×256 grayscale frame; each pipeline yields a segmentation mask (right). Solid arrows denote data flow; the gate symbol indicates that the output is set to zero when YOLO does not detect a glottis (or after at most 3 consecutive missed frames), removing spurious detections.

3.6. Glottal Area Waveform Features

For each patient video the YOLO+UNet pipeline is applied to every frame, yielding an area waveform $A(t)$. As in the pipeline definition (Section 3.5), the detector acts as a gate: frames where YOLO does not detect a glottis (or after at most 3 consecutive missed frames the detection is zeroed) contribute zero to the waveform, removing spurious detections and avoiding non-zero area from off-target endoscope views. Seven scalar kinematic features are extracted (Table 1). The fundamental frequency f_0 is estimated from the dominant FFT peak and converted from cycles/frame to Hz using the recording frame rate. Features are compared between Healthy (n=15) and Pathological (n=25) groups using the two-sided Mann–Whitney U test (significance threshold α=0.05); the 25 patients with Unknown or other disorder status are excluded from the group comparison.

Table 1: Kinematic features extracted from the Glottal Area Waveform.

Feature	Description
`area_mean`	Mean glottal area (px^2) over open frames
`area_std`	Standard deviation of area
`area_range`	Max − min area (vibratory excursion)
`open_quotient`	Fraction of cycle with area > 10% of mean
`f0`	Dominant frequency from FFT (Hz)
`periodicity`	Peak autocorrelation at lags 1–50
`cv`	Coefficient of variation (`area_std` / `area_mean`)

4. Experiments

4.1. Evaluation Metrics

- **Det.Recall**: fraction of frames where the YOLO detector fired (relevant for YOLO-gated pipelines; reported as 1.000 for detection-free baselines that always output a prediction).
- **Dice**: $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$, computed per frame then averaged.
- **IoU**: $\text{TP}/(\text{TP} + \text{FP} + \text{FN})$, per frame then averaged.
- **Dice**$\geq$0.5: fraction of frames with Dice ≥ 0.5, a clinical pass/fail threshold [8].

4.2. Implementation Details

All experiments run on Apple M-series hardware (MPS backend). YOLO training: YOLOv8n, 100 epochs, default hyperparameters, image size 256×256. U-Net training: 50 epochs, batch size 16, AdamW (`lr=1e-3`), cosine annealing. Both models trained solely on **GIRAFE** training split.

5. Results

5.1. GIRAFE In-Distribution Evaluation

Table 2 compares our pipelines against the published **GIRAFE** baselines on the 80-frame test split. Our U-Net alone achieves the highest Dice (0.809) and clinical pass rate (Dice$\geq$0.5 = 96.2 %), substantially outperforming all three published methods. The detection-gated YOLO+UNet pipeline reaches

Table 2: Segmentation results on the **GIRAFE** test split (4 patients, 80 frames). Published baselines from [8]. Det.Recall = n/a for methods that do not include a detection stage.

Method	Det.Recall	Dice	IoU	Dice≥0.5
InP [8]	n/a	0.713	n/a	n/a
U-Net [8]	n/a	0.643	n/a	n/a
SwinUNetV2 [8]	n/a	0.621	n/a	n/a
YOLO+OTSU (ours)	0.95	0.230	0.136	2.5%
U-Net only (ours)	n/a	**0.809**	**0.699**	**96.2%**
YOLO+UNet (ours)	0.95	0.746	0.629	83.8%
YOLO-Crop+UNet (ours)	0.95	0.697	0.567	77.5%
YOLO+Motion (ours)	0.95	0.349	0.234	23.5%

Dice 0.746, still surpassing InP (0.713) and SwinUNetV2 (0.621). The gap between U-Net only and YOLO+UNet on **GIRAFE** arises because the YOLO bounding box occasionally clips GT glottis pixels that extend beyond the detected region; this cost is absent without gating. YOLO detects a glottis on 95 % of test frames (Det.Recall = 0.95); the remaining 5 % are zeroed after the 3-frame hold, consistent with occasional closed-glottis or low-confidence frames. However, the detection gate provides essential robustness on real clinical recordings where the endoscope may be off-target (Section 5.3).

The YOLO-Crop+UNet pipeline, trained on YOLO-cropped patches, achieves Dice 0.697—below YOLO+UNet but above both deep-learning baselines from the **GIRAFE** paper. The performance gap relative to YOLO+UNet stems from the **GIRAFE** test frame structure: the 80 test frames are the *first* 20 frames per patient, and the tight YOLO bounding box occasionally clips GT glottis pixels that extend marginally beyond the detected region. Crucially, this limitation is overcome in the cross-dataset setting where the glottis region is larger relative to the frame (Section 5.2).

Figure 2 shows an example of the pipeline output: a montage of 12 annotated frames from patient 1 over one vibratory cycle, with the glottal opening segmented (green) and the detected region boxed (yellow); the numeric label in each frame is the glottal area in pixels^2.

Table 3 breaks down the Dice and IoU scores by test patient. U-Net only is the strongest pipeline on every patient, with the largest margin on patient 57A3 (Dice 0.803 vs. 0.597 for YOLO+UNet), the case where the YOLO bounding box clips glottis pixels most aggressively. Patient 61 is the

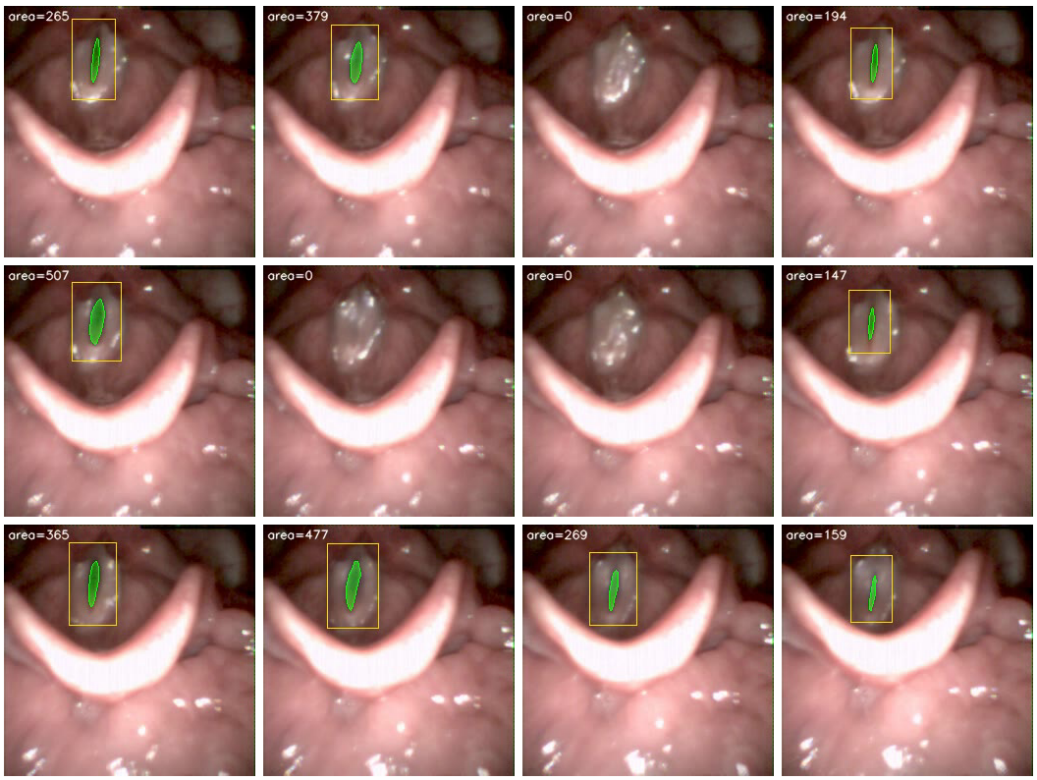

Figure 2: Output of the YOLO+UNet pipeline on 12 evenly spaced frames from one patient (patient 1): glottal mask (green), YOLO bounding box (yellow), and per-frame area. The montage illustrates temporal consistency of the segmentation across the vibratory cycle.

easiest case—all three pipelines exceed Dice 0.83.

5.2. Zero-Shot Cross-Dataset Evaluation on BAGLS

Table 4 reports results on 3500 **BAGLS** test frames. Neither the U-Net nor the YOLO weights were trained on any **BAGLS** data.

The YOLO detector fires on 68.8 % of **BAGLS** frames (Det.Recall = 0.688), confirming a degree of domain shift from the **GIRAFE**-trained detector. On frames where YOLO does not detect, the output is correctly zeroed.

At the default threshold (τ=0.25), YOLO-Crop+UNet achieves the highest Dice (0.609) and Dice≥0.5 (70.3 %), outperforming unguided U-Net inference (0.588 Dice, 67.1 % Dice≥0.5). The plain YOLO+UNet pipeline scores 0.545—below unguided U-Net—because the 31.2 % of frames with no YOLO detection

Table 3: Per-patient Dice / IoU on the **GIRAFE** test split (20 frames each). Best Dice per patient in bold.

	U-Net only		YOLO+UNet		YOLO-Crop+UNet	
Patient	Dice	IoU	Dice	IoU	Dice	IoU
p57A3	**0.803**	0.682	0.597	0.449	0.554	0.404
p61	**0.902**	0.822	0.902	0.822	0.833	0.715
p63	**0.794**	0.697	0.788	0.694	0.748	0.633
p64	**0.736**	0.596	0.690	0.553	0.654	0.516
Mean	**0.809**	0.699	0.746	0.629	0.697	0.567

Table 4: Zero-shot cross-dataset results on **BAGLS** test set (3500 frames). No **BAGLS** data used in training. Det.Recall shown as 1.000 for U-Net only (no YOLO gate; always processes every frame).

Method	Det.Recall	Dice	IoU	Dice≥0.5
U-Net only	1.000	0.588	0.504	67.1%
YOLO+UNet (ours)	0.688	0.545	0.473	61.9%
YOLO-Crop+UNet (ours)	0.688	0.609	0.533	70.3%

receive a zero mask that lowers the mean; when YOLO does detect, restricting the full-frame U-Net output to the bounding box clips some GT pixels. YOLO-Crop+UNet avoids this by rescaling the detected region, giving U-Net higher effective resolution at the glottis boundary and compensating for the missed frames through better per-frame accuracy. Lowering the confidence threshold to τ=0.02 further improves YOLO-Crop+UNet to Dice 0.659 (Table 5).

Confidence threshold sensitivity. The default YOLO confidence threshold (τ=0.25) was inherited from the **GIRAFE** in-distribution setting. Because the **GIRAFE**-trained detector exhibits domain shift on **BAGLS**, many true glottis frames receive detection scores below 0.25 and are incorrectly suppressed. Table 5 reports a single-pass threshold sweep: YOLO inference is run once at τ=0.001 and the confidence scores are thresholded in post-processing. Lowering τ to 0.02 raises YOLO-Crop+UNet detection recall from 68.8 % to 85.9 % and Dice from 0.609 to 0.659 (+0.050), with the clinical pass rate increasing from 70.3 % to 76.3 %. Below τ=0.02 performance plateaus and then degrades as false-positive detections introduce noisy bounding boxes.

Table 5: Effect of YOLO confidence threshold on YOLO-Crop+UNet performance (**BAGLS** test, 3500 frames, zero-shot). YOLO inference is run once; thresholds are applied in post-processing.

τ	Det.Recall	Dice	IoU	Dice≥0.5
0.001	0.943	0.646	0.553	75.0%
0.005	0.917	0.652	0.561	75.7%
0.01	0.895	0.654	0.563	75.8%
0.02	**0.859**	**0.659**	**0.568**	**76.3%**
0.03	0.842	0.656	0.567	76.0%
0.05	0.819	0.652	0.565	75.6%
0.10	0.773	0.641	0.558	74.3%
0.25	0.688	0.609	0.533	70.3%

5.3. Technical Validation: Glottal Area Waveform Features

To validate that the pipeline produces clinically meaningful output, we extract kinematic GAW features from all 65 **GIRAFE** patient recordings and test whether the automatically derived features replicate known group differences between Healthy and Pathological voices. Table 6 reports seven features for the 15 Healthy and 25 Pathological patients (25 patients with Unknown or Other status are excluded). This analysis is exploratory—given the small sample sizes and multiple features tested, we report uncorrected p-values from two-sided Mann–Whitney U tests ($\alpha = 0.05$) without multiple-comparison correction.

The Healthy and Pathological groups are sex-imbalanced: Healthy recordings are 80% female (12F/3M) while Pathological recordings are 56% male (14M/11F; Fisher's exact p=0.025). Because f_0 is strongly sex-dependent (males 100.3 Hz vs. females 223.5 Hz, p<0.001), we report results stratified by sex (Table 6) rather than pooled.

In the female subgroup (12 Healthy vs. 11 Pathological), f_0 does not reach significance (p=0.156), indicating that any apparent difference in the unstratified data is driven by sex composition. In contrast, cv is the only feature that distinguishes groups after stratification:

- **Coefficient of variation (cv, female only)**: 0.95 ± 0.20 (Healthy) vs. 0.57 ± 0.29 (Pathological), $p = 0.006$.

Healthy voices exhibit significantly higher vibration variability—consistent with the established observation that laryngeal pathologies increase vocal fold

Table 6: Glottal area waveform kinematic features: Healthy (H) vs. Pathological (P), stratified by sex. p-values from two-sided Mann–Whitney U; bold = $p<0.05$. The male subgroup has only $n=3$ Healthy recordings and results should be interpreted with caution.

	Female (12 H / 11 P)			Male (3 H / 14 P)		
Feature	H	P	p	H	P	p
area_mean	125.2±43.1	247.8±204.6	0.230	192.1±18.3	172.7±94.0	0.768
area_std	112.9±32.2	118.9±96.0	0.406	142.7±35.0	92.0±66.9	0.197
area_range	336.7±97.6	375.5±272.2	0.559	439.7±86.7	343.1±212.3	0.488
open_quot.	0.760±0.207	0.874±0.131	0.192	0.860±0.145	0.843±0.186	1.000
f_0 (Hz)	241.7±34.8	203.5±73.6	0.156	183.3±75.0	82.5±79.3	0.169
periodicity	0.955±0.008	0.946±0.013	0.255	0.962±0.001	0.900±0.116	0.068
cv	**0.95±0.20**	**0.57±0.29**	**0.006**	0.75±0.19	0.63±0.40	0.509

mass and stiffness, reducing the amplitude of glottal oscillation [7, 4]. In the male subgroup (3 Healthy vs. 14 Pathological), cv shows the same directional trend (0.75 vs. 0.63) but does not reach significance (p=0.509), as expected given the very small Healthy sample. Periodicity approaches significance in males (p=0.068), suggesting it may also distinguish groups with a larger cohort.

6. Discussion

Detection gating as a clinical safety mechanism. The detection gate provides a qualitative benefit that segmentation metrics alone do not capture: after at most 3 consecutive frames without a YOLO detection the output is zeroed, so the GAW is zero-valued when the endoscope has moved away from the glottis (or the glottis is closed), rather than containing artefactual non-zero area from spurious U-Net activations. This matters in practice because a clinician computing open quotient or periodicity over a full recording would otherwise need to manually identify and excise off-target frames—a laborious and subjective step.

Why YOLO-Crop+UNet generalises better. On the in-distribution **GIRAFE** test, YOLO+UNet outperforms YOLO-Crop+UNet. On **BAGLS**, the order reverses. We attribute this to two factors. First, **BAGLS** images span a wider range of glottis sizes and aspect ratios (frames from 256×120 to 512×512 pixels, with the glottis occupying a variable fraction of the image). By

normalising the glottis to fill the U-Net canvas, YOLO-Crop+UNet removes this scale variability and presents a consistent input distribution to the model. Second, the **BAGLS** glottis appears at relatively lower resolution in the full letterboxed frame than in **GIRAFE**; the crop step recovers this resolution.

Experimental direction and data efficiency. A natural alternative would be training on the larger **BAGLS** dataset (55 750 frames) and performing zero-shot transfer to **GIRAFE**. However, we intentionally prioritised the inverse direction for two reasons. First, the clinical objective of this work—technical validation of GAW biomarkers—requires the highest possible segmentation accuracy on the patient-labelled **GIRAFE** recordings. Second, demonstrating that a model trained on only 600 frames can generalise "upwards" to the heterogeneous, multi-institutional **BAGLS** dataset provides a more rigorous test of the pipeline's robustness. This approach proves that the YOLO-Crop+UNet mechanism effectively learns glottal anatomy rather than merely memorising institutional imaging characteristics.

*Why our U-Net outperforms the **GIRAFE** U-Net baseline.* Our U-Net alone achieves Dice 0.809 versus 0.643 for the **GIRAFE** paper's U-Net [8], despite using the same dataset, split, and a comparable augmentation pipeline (rotation, scaling, flipping, Gaussian noise/blur, brightness/contrast). Three training-recipe differences account for the gap: *(i) Grayscale input* (1 channel vs. 3-channel RGB)—the glottal gap is defined by intensity contrast, so colour triples the input dimensionality without adding discriminative signal, making the network harder to train on only 600 frames; *(ii) Combined BCE + Dice loss* versus Dice only—the BCE term supplies stable per-pixel gradients that complement the region-level Dice objective and avoid the gradient instability of pure Dice near 0 or 1; *(iii) Higher learning rate with cosine annealing* (10^{-3} vs. fixed $2{\times}10^{-4}$) and AdamW [19] instead of Adam [23], which together explore the loss landscape more aggressively and converge in 50 epochs to a stronger minimum than 200 epochs at a fixed low rate. These are straightforward engineering choices rather than architectural novelties, yet they yield a +0.166 Dice improvement—underscoring that on small medical-imaging datasets the training recipe matters as much as model design.

Lightweight pipeline vs. foundation models. Foundation models such as SAM [14] and MedSAM [15] offer impressive zero-shot segmentation but require a per-frame bounding-box or point prompt—precisely what our YOLO detector already provides. Using YOLO as the SAM prompter is conceptually possible;

however, SAM's ViT-H encoder (636M parameters, ∼150 ms per frame on GPU) is over 80× larger than our U-Net (7.76M parameters) and would make real-time GAW extraction from clinical recordings (>1000 frames at >1000 fps capture rate) impractical without dedicated hardware. Our U-Net pipeline processes a 502-frame **GIRAFE** patient video in approximately 11 s on consumer hardware (Apple M-series, MPS backend; ∼47 frames/s), well within offline clinical workflow requirements. Exploring SAM-based distillation to further improve U-Net accuracy without sacrificing throughput is an interesting direction for future work.

YOLO detection and confidence tuning. At the default threshold (τ=0.25) the YOLO detector fires on only 68.8 % of **BAGLS** frames. Lowering τ to 0.02 recovers 85.9 % recall and lifts YOLO-Crop+UNet Dice from 0.609 to 0.659 (Table 5). Below τ=0.02, false-positive detections introduce noisy bounding boxes that degrade the crop, so performance peaks at this threshold. Fine-tuning the detector on a small **BAGLS** subset would likely raise recall further and is left as future work.

Technical validation of GAW features. The GAW analysis is not intended as a clinical study of new biomarkers; rather, it serves as a technical validation that the fully automated pipeline reproduces group differences previously established through manual segmentation [7, 4]. Because the **GIRAFE** cohort has a significant sex imbalance (Fisher's exact p=0.025) and f_0 is strongly sex-dependent, Table 6 reports results stratified by sex rather than pooled. The stratified analysis shows that f_0 does not distinguish groups within either sex, confirming the unstratified difference would be driven by sex composition rather than disease status. In contrast, cv—the coefficient of variation of the glottal area waveform—remains the sole feature that survives sex stratification (p=0.006, female only), capturing the reduced vibratory regularity in pathological vocal folds due to increased mass and stiffness [7]. With only 12 Healthy and 11 Pathological female patients and no multiple-comparison correction, this result is exploratory and should be confirmed on a larger, sex-balanced cohort.

Limitations. The **GIRAFE** cohort is small (15 Healthy, 25 Pathological) and sex-imbalanced; the male Healthy subgroup (n=3) is too small for sex-stratified inference. With larger, balanced samples the non-significant features may reach significance. The GAW analysis uses the 4000 fps capture rate of the high-speed videoendoscope for converting f_0 from cycles/frame to

Hz. Finally, the YOLO-Crop+UNet weights are calibrated to a specific bounding-box padding (8 px); using a different padding at inference reduces performance.

7. Conclusion

We presented a lightweight U-Net trained with a carefully tuned recipe (grayscale input, combined BCE + Dice loss, AdamW with cosine annealing) that sets a new state of the art on the **GIRAFE** benchmark (Dice 0.809, Dice$\geq$0.5 = 96.2 %), outperforming all three published baselines and our own detection-gated variants. We further showed that pairing this U-Net with a YOLOv8 glottis detector provides a principled robustness mechanism: the detection gate suppresses spurious predictions on off-target frames, producing clean glottal area waveforms from full clinical recordings. A crop-zoom variant (YOLO-Crop+UNet) achieves the best performance in a zero-shot evaluation on the independent **BAGLS** dataset (Dice 0.659 at optimised threshold τ=0.02), demonstrating that the detection-guided approach generalises across institutions and equipment. As a technical validation, we applied the pipeline to all 65 **GIRAFE** patient recordings and showed that the automatically extracted coefficient of variation of the glottal area waveform significantly distinguishes Healthy from Pathological voices even after controlling for sex imbalance (p=0.006, female subgroup)—confirming that the segmentation quality is sufficient for downstream clinical analysis.

Data and Code Availability

All training and evaluation scripts, trained model weights, and the **GIRAFE** evaluation results JSON are available at `https://github.com/hari-krishnan/openglottal`. The **GIRAFE** dataset [8] is freely available from `https://doi.org/10.5281/zenodo.7962150`. The **BAGLS** dataset [10] is available from `https://zenodo.org/record/3381469`.

Author Contributions

Harikrishnan Unnikrishnan designed the study, implemented the detection-gated pipeline (YOLO glottis detector, U-Net segmenter, and temporal detector with 3-frame hold), conducted the experiments on **GIRAFE** and **BAGLS**, performed the glottal area waveform feature analysis, and wrote the manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

We thank Andrade-Miranda et al. for making the **GIRAFE** dataset publicly available and Gómez et al. for the **BAGLS** benchmark; both datasets were essential to this work.

References

[1] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, version 8.0.0, AGPL-3.0 (2023).
URL https://github.com/ultralytics/ultralytics

[2] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Vol. 9351 of Lecture Notes in Computer Science, Springer, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

[3] D. D. Deliyski, Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution, Folia Phoniatrica et Logopaedica 60 (1) (2008) 33–44. doi:10.1159/000111802.

[4] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, I. M. Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, BioMedical Engineering OnLine 6 (2007) 23. doi:10.1186/1475-925X-6-23.

[5] J. Lohscheller, U. Eysholdt, H. Toy, M. Döllinger, Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics, IEEE Transactions on Medical Imaging 27 (3) (2008) 300–309. doi:10.1109/TMI.2007.903690.

[6] R. R. Patel, D. Dubrovskiy, M. Döllinger, Characterizing vibratory kinematics in children and adults with high-speed digital imaging, Journal of Speech, Language, and Hearing Research 57 (2) (2014) S674–S686. doi:10.1044/2014_JSLHR-S-12-0278.

[7] R. R. Patel, K. D. Donohue, D. Lau, H. Unnikrishnan, In vivo measurement of pediatric vocal fold motion using structured light laser projection, PLOS ONE 11 (4) (2016) e0154586. doi:10.1371/journal.pone.0154586.

[8] G. Andrade-Miranda, M. Hernández-Álvarez, J. I. Godino-Llorente, GIRAFE: Glottal imaging dataset for advanced segmentation, analysis, and facilitative playbacks evaluation, Data in Brief 59 (2025) 111376. doi:10.1016/j.dib.2024.111376.

[9] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: Computer Vision – ECCV 2022 Workshops, Vol. 13803 of Lecture Notes in Computer Science, Springer, 2022, pp. 205–218. doi:10.1007/978-3-031-25066-8_9.

[10] P. Gómez, A. M. Kist, P. Schlegel, D. A. Berry, D. K. Chhetri, R. Montaño, F. Müller, A. Schützenberger, M. Semmler, S. Dürr, D. Eytan, J. Lohscheller, M. Echternach, M. Döllinger, BAGLS, a multihospital benchmark for automatic glottis segmentation, Scientific Data 7 (2020) 186. doi:10.1038/s41597-020-0526-3.

[11] A. M. Kist, J. Zilker, P. Gómez, A. Schützenberger, M. Döllinger, A single latent channel is sufficient for biomedical glottis segmentation, Scientific Reports 12 (2022) 14292. doi:10.1038/s41598-022-17764-1.

[12] M. K. Fehling, F. Grosch, M. E. Schuster, B. Schick, J. Lohscheller, Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional LSTM network, PLOS ONE 15 (2) (2020) e0227791. doi:10.1371/journal.pone.0227791.

[13] A. M. Kist, P. Gómez, A. Schützenberger, M. Döllinger, Glottis Analysis Tools: Open tools for laryngeal high-speed videoendoscopy analysis, Journal of Speech, Language, and Hearing Research (2021). doi:10.1044/2021_JSLHR-21-00064.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, in: IEEE/CVF International Conference on Computer

Vision (ICCV), 2023, pp. 4015–4026. doi:10.1109/ICCV51070.2023.00371.

[15] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nature Communications 15 (2024) 654. doi:10.1038/s41467-024-44824-z.

[16] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.322.

[17] P. Gómez, A. Schützenberger, M. Döllinger, A. M. Kist, Automatic detection and tracking of the glottal gap from high-speed video data, Biomedical Engineering Online (2021).

[18] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: Fourth International Conference on 3D Vision (3DV), 2016, pp. 565–571. doi:10.1109/3DV.2016.79.

[19] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019.
URL https://openreview.net/forum?id=Bkg6RiCqY7

[20] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: International Conference on Learning Representations (ICLR), 2017.
URL https://openreview.net/forum?id=Skq89Scxx

[21] R. R. Patel, K. D. Donohue, H. Unnikrishnan, R. J. Kryscio, Kinematic measurements of the vocal-fold displacement waveform in typical children and adult populations: quantification of high-speed endoscopic videos, Journal of Speech, Language, and Hearing Research 58 (2) (2015) 227–240.

[22] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66. doi:10.1109/TSMC.1979.4310076.

[23] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2015.
URL https://arxiv.org/abs/1412.6980