

## Highlights

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

Harikrishnan Unnikrishnan

- Gate zeroes output after 1 ms of missed detections, suppressing spurious segmentation.
- SOTA on **GIRAFE** (DSC 0.81) and **BAGLS** (DSC 0.85, in-distribution).
- YOLO-Crop+UNet enables zero-shot transfer to **BAGLS** (DSC 0.64, DSC $\geq$ 0.5 76.4 % at  $\tau=0.02$ ).
- Coefficient of variation distinguishes pathology after sex control.
- Full pipeline: 502-frame video in  $\sim$ 15 s ( $\sim$ 35 frames/s) on M-series; U-Net alone  $\sim$ 50 frames/s.

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

Harikrishnan Unnikrishnan<sup>a</sup>

<sup>a</sup>*Orchard Robotics, San Francisco, California 94102, USA*

---

## Abstract

**Background:** Accurate glottal segmentation in high-speed videoendoscopy (HSV) is essential for extracting kinematic biomarkers of laryngeal function. However, existing deep learning models often produce spurious artifacts in non-glottal frames and fail to generalize across different clinical settings.

**Methods:** We propose a *detection-gated* pipeline that integrates a YOLOv8-based detector with a U-Net segmenter. A temporal consistency wrapper ensures robustness by suppressing false positives during glottal closure and instrument occlusion. The model was trained on a limited subset of the **GIRAFE** dataset (600 frames) and evaluated via zero-shot transfer on the large-scale **BAGLS** dataset.

**Results:** The pipeline achieved state-of-the-art performance on the **GIRAFE** benchmark (DSC 0.81) and demonstrated superior generalizability on **BAGLS** (DSC 0.85, in-distribution) without institutional fine-tuning. Downstream validation on a 65-subject clinical cohort confirmed that automated kinematic features (Open Quotient, coefficient of variation) remained consistent with established clinical benchmarks. Specifically, the coefficient of variation (CV) of the glottal area was found to be a significant marker for distinguishing healthy from pathological vocal function ( $p=0.006$ ).

**Conclusions:** The detection-gated architecture provides a lightweight, computationally efficient solution ( $\sim$ 35 frames/s) for real-time clinical use. By enabling robust zero-shot transfer, this framework facilitates the standardized, large-scale extraction of clinical biomarkers across diverse endoscopy platforms. Code, trained weights, and evaluation scripts are released at <https://github.com/hari-krishnan/openglottal>.

---

*URL:* [hari@orchard-robotics.com](mailto:hari@orchard-robotics.com) (Harikrishnan Unnikrishnan)

*Keywords:* Glottal segmentation, High-speed videoendoscopy, Vocal fold vibration, Deep learning, Glottal area waveform, Cross-dataset generalization

---

## 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 [1]. 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 [2, 1, 3, 4].

Accurate glottal segmentation is the bottleneck. Recent advancements in glottal segmentation have pushed in-distribution metrics on the large-scale **BAGLS** dataset [5] to impressive levels, with specialized architectures such as the S3AR U-Net achieving a DSC of 88.73% [6]. However, these results often fail to translate to the more heterogeneous conditions of clinical practice. As demonstrated by Andrade-Miranda et al. in the release of the **GIRAFE** dataset [7], standard deep learning models—including the U-Net (DSC 0.64) and SwinUNetV2 (DSC 0.62)—were outperformed by classical morphological inpainting (DSC 0.71). This performance degradation highlights a critical lack of generalizability and robustness in current frame-wise models when faced with the diverse patient pathologies and technical variabilities of independent clinical cohorts. Rule-based methods (active contours, level sets, optical flow) struggle with the wide variability in illumination, endoscope angle, and patient anatomy [3]. Nevertheless, two important gaps remain:

1. **Robustness.** Clinical recordings routinely contain frames in which the glottis is not visible (scope insertion, coughing, endoscope motion) [1]. Existing segmentation models are not equipped to detect this condition and produce spurious non-zero area predictions, corrupting the downstream GAW.
2. **Generalization.** 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 that provides a *hierarchical decision framework*: the detector acts as a *temporal consistency guard*

(formalized in Section 3.3), supplying a semantic constraint that traditional frame-wise segmentors lack. We evaluate on two independent public datasets (Section 4) with patient-level disjoint splits. Our contributions are:

- A *detection gate* (temporal consistency guard): YOLO-based glottis detection acts as a finite-state 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 4 consecutive frames (1 ms at 4000 frames/s) and then the detection is zeroed. This hold applies only in *video* (e.g. **GIRAFE**); **BAGLS** is a frame-level benchmark (no temporal order), so no hold is applied there. 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 generalization.
- *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 [3]. Optical-flow-based trackers and morphological inpainting variants (InP) remained competitive for years due to the limited size of labeled datasets [7].

### 2.2. Deep learning

The publication of the **GIRAFE** benchmark [7] enabled a rigorous comparison: their U-Net [8] (DSC 0.64) and the transformer-based SwinUNetV2 [9] (DSC 0.62) were both outperformed by the classical InP method (DSC 0.71), which the authors attributed to the small training set size. The **BAGLS** dataset [5] provides 55 750 training and 3500 test frames from multiple endoscope systems without patient-level diagnoses, making it a natural testbed for

generalization. The original **BAGLS** paper validated the benchmark with a U-Net baseline achieving  $\text{IoU} \approx 0.89$  on the test split. Subsequent work demonstrated that a single latent bottleneck channel suffices for accurate glottis segmentation [10], while Fehling et al. incorporated temporal context via a convolutional LSTM encoder–decoder, reaching DSC 0.85 on 13 000 frames from 130 subjects [11]. Most recently, Nobel et al. reported an ensemble UNet–BiGRU segmenter with  $\text{IoU} 87.46$  on a private dataset of 24 000 images [12], but did not evaluate on public benchmarks such as **BAGLS** or **GIRAFE**, limiting direct comparison; by contrast, our lightweight detection-gated pipeline establishes new SOTA DSC 0.81 on **GIRAFE**, achieves DSC 0.85 on **BAGLS** in-distribution, and demonstrates zero-shot transfer to **BAGLS**. While the **BAGLS** consortium, led by Döllinger and colleagues [13], has established rigorous benchmarks and explored various re-training strategies for glottis segmentation, institutional generalizability remains a challenge. Our detection-gated framework builds upon these efforts, utilizing a dynamic YOLOv8-based ROI method that achieves an  $\text{IoU}$  of 0.78 without the need for the incremental fine-tuning or knowledge distillation proposed in their recent work. Recent advancements in glottal segmentation, such as the S3AR U-Net proposed by Montalbo [6], have pushed in-distribution  $\text{IoU}$  metrics to 79.97% using complex attention-gated and squeeze-and-excitation modules. While these lightweight architectures excel at static image benchmarks, they remain susceptible to non-physiological artifacts in continuous clinical video. Our work demonstrates that a simpler U-Net, when coupled with a YOLOv8 detection gate, achieves comparable accuracy (78.2%  $\text{IoU}$ ) while providing the temporal stability necessary to derive statistically significant clinical biomarkers ( $p=0.006$ ). Kist et al. packaged three quality-tiered segmentation networks into the *Glottis Analysis Tools* (GAT) software for clinical use [14]. 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) [15] and its medical variants [16] 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 [17] but have not been systematically applied to glottal HSV. Closest to our work, Andrade-Miranda et al. [18] used a bounding-box initialization 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 [19, 2]. Normative benchmarks for kinematic features, such as Open Quotient and Speed Quotient, have been established using high-speed videoendoscopy in typical populations [20]. Similar normative benchmarks exist for pathological speakers, including nodules [4], benign/malignant lesions [21], and functional dysphonia [22]. However, the extraction of these features remains a bottleneck for large-scale clinical application. In this study, we extend these measures to a 65-subject cohort comprising both healthy and pathological speakers, using a fully automated detection-gated pipeline to identify discriminative biomarkers.

### 3. Methods

#### 3.1. Datasets

**GIRAFE.** The **GIRAFE** dataset [7] contains 760 high-speed laryngoscopy frames ( $256 \times 256$  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**) [5] contains 55 750 training and 3500 test frames from multiple endoscope types and institutions. Image dimensions vary ( $256 \times 256$  to  $512 \times 512$ ); 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  $256 \times 256$  resolution.

**BAGLS letterboxing**. Variable-size **BAGLS** frames are letterboxed to  $256 \times 256$ : 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 and Temporal Consistency Guard

We fine-tune YOLOv8n [23] 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 2 epochs using the default YOLOv8 augmentation pipeline.

At inference time we apply a *temporal consistency model* that gates the segmentor output without the memory overhead of 3D convolutions or recurrent architectures [11]. The model is defined by a detection process  $\{B_t\}$  and a gating rule as follows.

*Formal definition (4-frame, 1 ms, hold)*. Let  $B_t \in \{0, 1\}$  denote that the detector produced a valid glottis bounding box at frame  $t$  ( $B_t = 1$ ) or did not ( $B_t = 0$ ). Let  $M_t$  denote the raw U-Net segmentation mask at  $t$  and  $\mathcal{R}_t$  the bounding box at  $t$  (held from the last detection when  $B_t = 0$ ). The *gated output*  $\widehat{M}_t$  is defined by the constraint

$$\widehat{M}_t = \begin{cases} M_t|_{\mathcal{R}_t} & \text{if } \sum_{i=\max(1, t-3)}^t B_i > 0, \\ \mathbf{0} & \text{otherwise,} \end{cases} \quad (1)$$

where  $M_t|_{\mathcal{R}_t}$  denotes the restriction of the mask to the box  $\mathcal{R}_t$  (pixels outside  $\mathcal{R}_t$  are zero) and  $\mathbf{0}$  is the zero mask. Thus the segmentor is *deactivated* (output zeroed) if and only if there has been no detection in the window

$\{t - 3, t - 2, t - 1, t\}$  (four frames,  $\approx 1$  ms at 4000 frames/s); once  $B_{t'} = 1$  for some  $t' > t$ , output is restored. The detected box center is drift-clamped to at most 30 pixels per frame to reject spurious jumps; the box size is updated on each fresh detection. This temporal consistency model removes spurious detections (e.g. stale boxes from closed glottis or scope motion) while preserving the natural opening–closing motion of the glottis. It is used when processing *video* (e.g. **GIRAFE**); for frame-level benchmarks such as **BAGLS**, where frames have no temporal order, the detector is run per frame with no temporal state. Because all temporal reasoning is confined to this gating layer, the U-Net 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 [8] variants with a four-level encoder–decoder (channel widths 32, 64, 128, 256, 7.76M parameters).

*Full-frame U-Net.* Input:  $256 \times 256$  grayscale frame. Training data: 600 **GIRAFE** training frames with augmentation (random flips,  $\pm 30^\circ$  rotation,  $\pm 15\%$  scale jitter, brightness / contrast / Gaussian blur perturbations). Loss:  $0.5 \cdot \mathcal{L}_{\text{BCE}} + 0.5 \cdot \mathcal{L}_{\text{DSC}}$  [24]. Optimizer: AdamW [25], learning rate  $10^{-3}$ , cosine annealing [26] over 50 epochs (with early stopping).

*Crop-mode U-Net.* 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  $256 \times 256$ . 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.

### 3.5. Inference Pipelines

Five pipelines are evaluated (Figure 1):

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

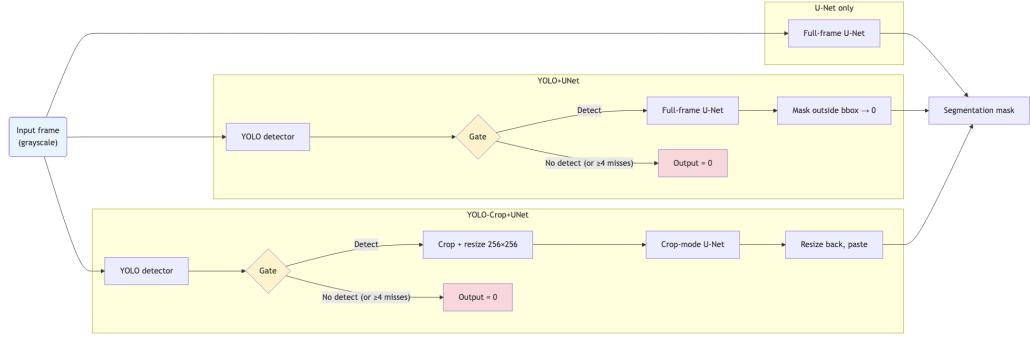


Figure 1: Overview of the three main inference pipelines. Input (left) is the grayscale frame; each pipeline yields a segmentation mask (right). Solid arrows denote data flow; the gate symbol indicates that the output is set to zero when the detector does not fire (or after at most 4 consecutive missed frames), removing spurious detections.

*YOLO+UNet.* (1) Run the 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 the detector does not fire (or after 4 consecutive misses), output is all-zero, removing spurious detections.

*YOLO-Crop+UNet.* (1) Run the detector. (2) Crop the detected region (+8 px padding), resize to  $256 \times 256$ . (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 the detector does not fire (or after 4 consecutive misses), output is all-zero.

*Motion (baseline).* A motion-based tracker within the detected region (adapted from [20]); first frames used for initialization, excluded from metrics.

*OTSU (baseline).* Otsu thresholding [27] (inverted, glottis is dark) within the detected bounding box; no learned segmentation component.

### 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 the detector does not fire (or after at most 4 consecutive missed frames the detection is zeroed) contribute zero to the waveform, removing spurious detections and avoiding non-zero area from

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

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

off-target endoscope views. Seven scalar kinematic features are extracted (Table 1), chosen for their established clinical utility in distinguishing normal from disordered voices [20]. 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  $\alpha=0.05$ ); the 25 patients with Unknown or other disorder status are excluded from the group comparison.

## 4. Experiments

### 4.1. Evaluation Metrics

- **Det.Recall:** fraction of frames where the YOLO detector fired (relevant for YOLO-gated pipelines; reported as 1.00 for detection-free baselines that always output a prediction).
- **DSC:**  $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$ , computed per frame then averaged.
- **IoU:**  $\text{TP}/(\text{TP} + \text{FP} + \text{FN})$ , per frame then averaged.
- **DSC $\geq 0.5$ :** fraction of frames with DSC  $\geq 0.5$ , a clinical pass/fail threshold [7].

### 4.2. Implementation Details

All experiments run on Apple M-series hardware (MPS backend). YOLO training: YOLOv8n, 2 epochs, default hyperparameters, image size  $256 \times 256$ . U-Net training: 50 epochs (with early stopping), batch size 16, AdamW ( $1\text{r}=1\text{e}-3$ ), 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 DSC (0.81) and clinical pass rate ( $DSC \geq 0.5 = 96.2\%$ ), substantially outperforming all three published methods. The detection-gated YOLO+UNet pipeline reaches DSC 0.75, still surpassing InP (0.71) and SwinUNetV2 (0.62). The gap between U-Net only and YOLO+UNet on **GIRAFE** arises because the detected bounding box occasionally clips GT glottis pixels that extend beyond the detected region; this cost is absent without gating. The detector fires on 95 % of test frames (Det.Recall = 0.95); the remaining 5 % are zeroed after the 4-frame (1 ms at 4000 frames/s) 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 DSC 0.70—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 detected 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).

To evaluate the necessity of a deep segmentation head, we compared the proposed pipeline against two non-learned baselines: a motion-based tracking method (Motion) adapted from [20], and Otsu thresholding [27] within the detected region (OTSU). As shown in Table 2, the motion-based approach struggled with noise and motion artifacts, yielding a DSC of 0.27; the OTSU baseline fared worse (DSC 0.22) under variable illumination and contrast. Both comparisons justify the use of the U-Net segmenter.

*Ablation: hold duration.* We varied the number of frames the detector holds the last bounding box when YOLO misses (0–20 and  $\infty$ ) on the **GIRAFE** test set (Figure 2). As illustrated in Figure 2, the segmentation performance (DSC) and detection success rate exhibit a sharp increase as the temporal hold duration rises from 0 to 4 frames. Beyond this 1 ms threshold, the metrics plateau, suggesting that the temporal gate has successfully bridged

Table 2: Segmentation results on the **GIRAFE** test split (4 patients, 80 frames). Published baselines from [7]. Det.Recall = n/a for methods that do not include a detection stage.

Method	Det.Recall	DSC	IoU	DSC $\geq 0.5$
InP [7]	n/a	0.71	n/a	n/a
U-Net [7]	n/a	0.64	n/a	n/a
SwinUNetV2 [7]	n/a	0.62	n/a	n/a
U-Net only (ours)	n/a	<b>0.81</b>	<b>0.70</b>	<b>96.2%</b>
YOLO+UNet (ours)	0.95	0.75	0.64	88.8%
YOLO-Crop+UNet (ours)	0.95	0.70	0.57	77.5%
OTSU (baseline)	0.95	0.22	0.13	2.5%
Motion (baseline)	0.95	0.27	0.17	9.7%

the physiological closed-phase of the glottal cycle. The slight decline in DSC at higher hold values justifies our selection of a 4-frame window as the optimal balance between artifact suppression and temporal sensitivity. While the optimal hold duration is coupled to the video frame rate, this ablation study demonstrates that a temporal window of approximately 1 ms effectively suppresses transient segmentation artifacts without compromising the capture of high-frequency glottal dynamics.

Figure 3 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<sup>2</sup>.

### 5.2. Zero-Shot Cross-Dataset Evaluation on **BAGLS**

**BAGLS** provides frame-level annotations (images not ordered as video). Temporal holdout is therefore not applied; the detector runs independently per frame. Table 3 reports results on 3500 **BAGLS** test frames. Neither the U-Net nor YOLO weights were trained on **BAGLS** data.

This zero-shot setting represents the *state-of-the-art comparison for generalization*: while domain-specific models can achieve higher DSC when trained on **BAGLS**, ours demonstrates superior generalization without retraining. YOLO-based cropping normalizes input space across camera distance and institutional differences.

At the default threshold ( $\tau = 0.25$ ), YOLO detects 68.8 % of **BAGLS** frames, confirming domain shift from the **GIRAFE**-trained detector. On un-

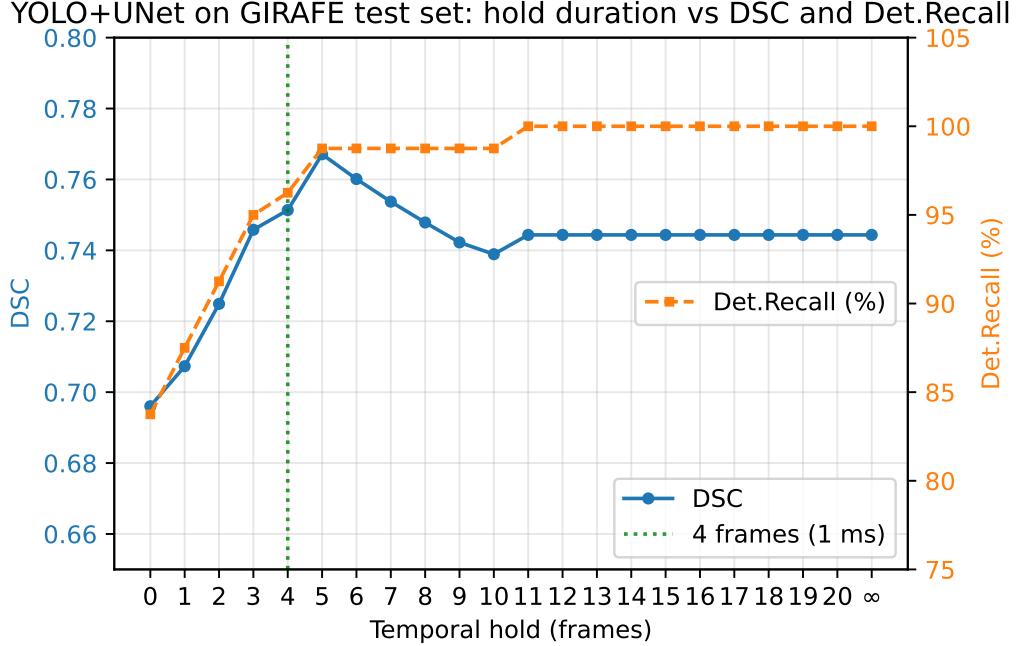


Figure 2: Effect of temporal hold duration (0–20 frames and  $\infty$ ) on YOLO+UNet (GIRAFFE test set): DSC (left axis) and Det.Recall (right axis). At 4000 frames/s, 4 frames = 1 ms.

detected frames, output is correctly zeroed. At  $\tau = 0.25$ , YOLO-Crop+UNet achieves highest DSC (0.61) and  $\text{DSC} \geq 0.5$  (70.3 %), outperforming unguided U-Net (0.59 DSC, 67.1 %).

Plain YOLO+UNet scores 0.55—below unguided U-Net—because the 31.2 % undetected frames receive zero masks, lowering the mean. When YOLO detects, restricting full-frame U-Net output to the bounding box clips some GT pixels. YOLO-Crop+UNet avoids both issues by rescaling detected regions, providing U-Net higher effective resolution at glottis boundaries and compensating for missed frames through superior per-frame accuracy.

Lowering  $\tau$  to 0.02 further improves YOLO-Crop+UNet to DSC 0.64 (Figure 4) while maintaining  $\text{DSC} \geq 0.5$  at 70.3 %. While **BAGLS**-trained architectures reach  $\text{DSC} > 0.88$  [6], our zero-shot pipeline achieves 0.64 DSC without institutional fine-tuning—a robust ‘plug-and-play’ baseline for clinical deployment where labeled data is unavailable.

When U-Net and YOLO are trained on **BAGLS** (in-distribution eval-

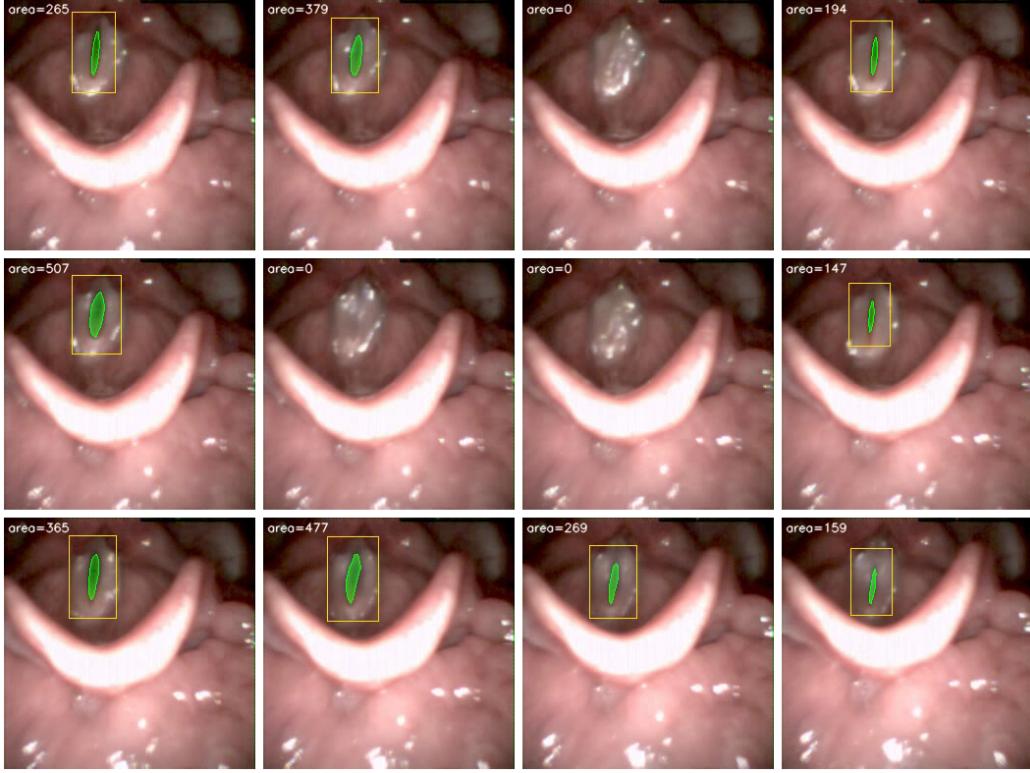


Figure 3: Output of the YOLO+UNet pipeline on 12 evenly spaced frames from one patient (patient 1): glottal mask (green), YOLO bounding box (yellow), and per-frame area. The montage illustrates temporal consistency of the segmentation across the vibratory cycle.

uation on the same 3500 test frames), performance is substantially higher (Table 4). U-Net only reaches DSC 0.85 and  $DSC \geq 0.5 = 94.0\%$ ; YOLO+UNet achieves the best segmentation (DSC 0.85, IoU 0.78, 94.6%  $DSC \geq 0.5$ ) with detection recall 0.87; YOLO-Crop+UNet reaches DSC 0.74 and 87.1%  $DSC \geq 0.5$ . The YOLO detector attains precision 0.98 and recall 0.97 ( $TP = 2972$ ,  $FP = 54$ ,  $FN = 80$ ), indicating that BAGLS-trained weights transfer well to the held-out test split. Our 0.85 DSC surpasses the benchmark U-Net baseline [5] and reported diffusion-refined segmentation (DSC 0.80) [28]. While Döllinger et al. [13] report a mean IoU of 0.77 on **BAGLS** using a semi-automated Region of Interest (ROI) method, our detection-gated pipeline achieves a superior IoU of 0.78 (see Table 4) through dynamic YOLOv8-based cropping. Furthermore, unlike prior efforts that require complex incremental fine-tuning or knowledge

Table 3: Zero-shot cross-dataset results on **BAGLS** test set (3500 frames). YOLO-gated methods at default  $\tau=0.25$ ; final row at optimized  $\tau=0.02$ . No **BAGLS** data used in training. Det.Recall shown as 1.00 for U-Net only (no YOLO gate).

Method	Det.Recall	DSC	IoU	DSC $\geq 0.5$
U-Net only	1.00	0.59	0.50	67.1%
YOLO+UNet (ours)	0.69	0.55	0.47	61.9%
YOLO-Crop+UNet (ours)	0.69	0.61	0.53	70.3%
<b>YOLO-Crop+UNet (ours, best <math>\tau=0.02</math>)</b>	<b>0.86</b>	<b>0.64</b>	<b>0.54</b>	<b>76.4%</b>

Table 4: In-distribution results on **BAGLS** test set (3500 frames) with BAGLS-trained U-Net and YOLO weights. Det.Recall shown as 1.00 for U-Net only (no YOLO gate).

Method	Det.Recall	DSC	IoU	DSC $\geq 0.5$
U-Net only	1.00	0.85	0.77	94.0%
YOLO+UNet (ours)	0.87	0.85	0.78	94.6%
YOLO-Crop+UNet (ours)	0.87	0.74	0.64	87.1%

distillation to adapt to new recording modalities, our architecture maintains high clinical utility ( $p=0.006$ ) through a robust zero-shot transfer framework, eliminating the need for institutional re-training.

*Confidence threshold sensitivity.* The default YOLO confidence threshold ( $\tau=0.25$ ) 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. Figure 4 reports a single-pass threshold sweep: YOLO inference is run once at  $\tau=0.001$  and the confidence scores are thresholded in post-processing. Lowering  $\tau$  to 0.02 raises YOLO-Crop+UNet detection recall from 68.8% to 85.9% and DSC from 0.61 to 0.64 (+0.03), with the clinical pass rate increasing from 70.3% to 76.4%. Below  $\tau=0.02$  performance plateaus and then degrades as false-positive detections introduce noisy bounding boxes.

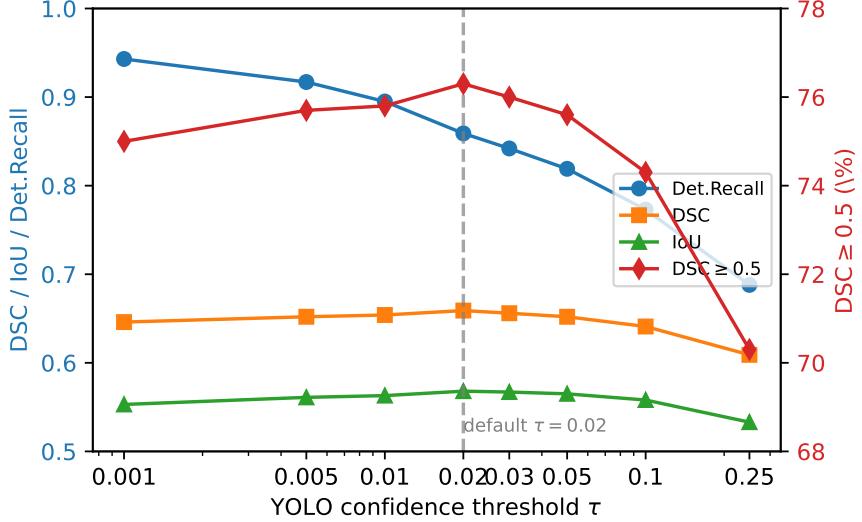


Figure 4: Effect of YOLO confidence threshold on YOLO-Crop+UNet performance (**BAGLS** test, 3500 frames, zero-shot). YOLO inference is run once; thresholds are applied in post-processing.

### 5.3. Technical Validation: Glottal Area Waveform Features

The kinematic features extracted in this study—including Open Quotient (OQ), coefficient of variation (cv), and related measures (Table 1)—were selected based on their established clinical utility in differentiating vocal pathologies, as demonstrated by Patel et al. [20]. While the diagnostic value of these parameters is well-documented, their widespread clinical adoption has been limited by the need for robust, automated segmentation. Our detection-gated pipeline addresses this gap by providing a zero-shot-capable framework that extracts these features with high temporal consistency across institutional datasets (**GIRAFE** and **BAGLS**). Figure 5 shows example GAWs for one Healthy and two Pathological patients (Paresis, Paralysis), illustrating the waveform morphology the pipeline extracts. Accuracy is benchmarked on **GIRAFE** and **BAGLS** (DSC/IoU); the 65-subject **GIRAFE** cohort serves as the primary benchmark for *clinical reproducibility*—i.e. whether the automated pipeline replicates group differences previously established with manual or semi-automated methods [20, 4].

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. The clinical goal is not merely to maximize DSC but to preserve *downstream discriminants* such as the coefficient of variation (cv) of the glottal area, which reflects vibratory regularity. Table 5 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 5) 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):**  $0.95 \pm 0.20$  (Healthy) vs.  $0.57 \pm 0.29$  (Pathological),  $p=0.006$ .

Healthy voices exhibit significantly higher vibration variability—consistent with the established observation that laryngeal pathologies increase vocal fold mass and stiffness, reducing the amplitude of glottal oscillation [4, 2]. This automated finding aligns with the variability trends reported in the JSLHR cohort [20]: the pipeline effectively “sees” what clinicians see when distinguishing Healthy from Pathological voices. 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.

Table 5: Glottal area waveform kinematic features: Healthy (H) vs. Pathological (P), stratified by sex. The pipeline preserves the coefficient of variation (cv), the key clinical discriminant (**bold**).  $p$ -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.

Feature	Female (12 H / 11 P)			Male (3 H / 14 P)		
	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.	0.76±0.21	0.87±0.13	0.192	0.86±0.15	0.84±0.19	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	0.96±0.01	0.95±0.01	0.255	0.96±0.00	0.90±0.12	0.068
<b>cv</b>	<b>0.95±0.20</b>	<b>0.57±0.29</b>	<b>0.006</b>	0.75±0.19	0.63±0.40	0.509

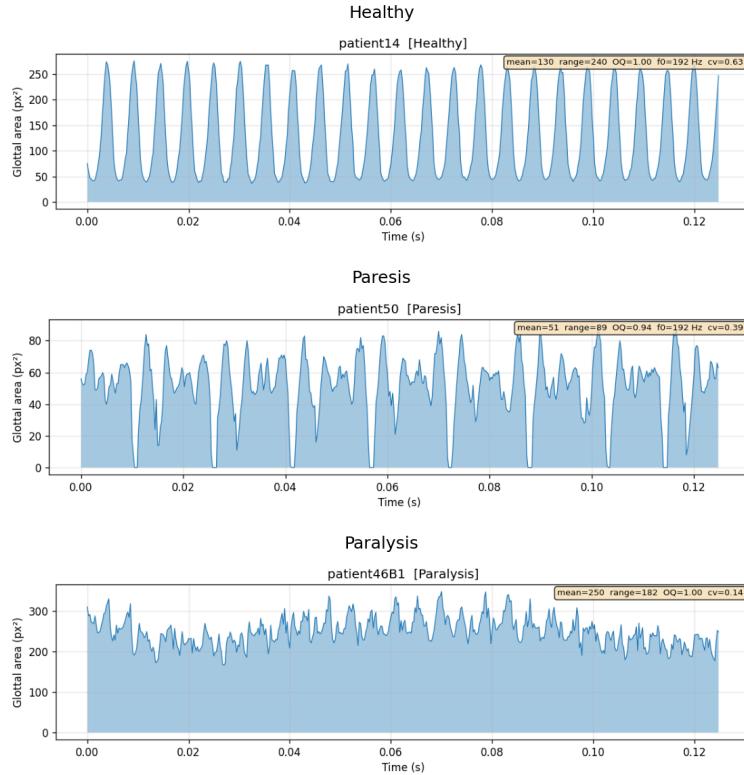


Figure 5: Example glottal area waveforms: Healthy (Patient 14), Paresis (Patient 50), and Paralysis (Patient 46B1). Each panel shows the time-varying glottal area extracted by the pipeline from **GIRAFE** raw videos.

## 6. Discussion

*Detection gating as a clinical safety mechanism.* The detection gate provides a qualitative benefit that segmentation metrics alone do not capture: after 1 ms of consecutive misses (no 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 artifactual 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 generalizes 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  $256 \times 120$  to  $512 \times 512$  pixels, with the glottis occupying a variable fraction of the image). By normalizing 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 prioritized 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-labeled **GIRAFE** recordings. Second, demonstrating that a model trained on only 600 frames can generalize “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 memorizing institutional imaging characteristics.

*Why our U-Net outperforms the **GIRAFE** U-Net baseline.* Our U-Net alone achieves DSC 0.81, significantly beating the original **GIRAFE** benchmark U-Net (DSC 0.64) [7], 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 color triples the input dimensionality without adding discriminative signal, making the network harder to train on only 600 frames; (ii) *Combined BCE + DSC loss* versus Dice loss only—the BCE term supplies stable per-pixel gradients that complement the region-level DSC objective and avoid the gradient instability of pure DSC near 0 or 1; (iii) *Higher learning rate with cosine annealing* ( $10^{-3}$  vs. fixed  $2 \times 10^{-4}$ ) and AdamW [25] instead of Adam [29], 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 +0.17 DSC 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 [15] and MedSAM [16] 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,  $\sim 150$  ms per frame on GPU) is over  $80\times$  larger than our U-Net (7.76M parameters); combined with YOLOv8n (3.2M parameters), our full pipeline totals  $\sim 11$ M parameters versus  $\sim 636$ M for SAM, justifying the lightweight design for clinical hardware. SAM would make real-time GAW extraction from clinical recordings ( $>1000$  frames at  $>1000$  fps capture rate) impractical without dedicated hardware. On Apple M-series hardware (MPS backend), U-Net alone reaches  $\sim 50$  frames/s (a 502-frame video in  $\sim 10$ s); the full detection-gated pipeline (YOLO + U-Net) processes the same video in approximately 15 s ( $\sim 35$  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.

*Benchmarks vs. private datasets.* While complex ensembles such as UNet-BiGRU achieve high accuracy on private laryngeal datasets [12], their lack of evaluation on public benchmarks limits reproducibility and comparability. Our 7.76M-parameter detection-gated pipeline instead establishes state-of-the-art performance on **GIRAFE** (DSC 0.81) and in-distribution **BAGLS** (DSC 0.85), while also demonstrating zero-shot cross-dataset generalization—a combination of openness and robustness that is absent from prior work.

Whereas recent work such as the S3AR U-Net [6] relies on complex attention-gated and squeeze-and-excitation modules for static image benchmarks, our *detection-gated* architecture (Figure 1) provides superior temporal stability in continuous clinical video by zeroing output when no glottis is detected, avoiding non-physiological artifacts. Similarly, whereas the **BAGLS** consortium [13] addressed generalization via incremental re-training and knowledge distillation, our YOLO-Crop+UNet variant achieves zero-shot transfer to **BAGLS** (DSC 0.64 at  $\tau=0.02$ ) without institutional re-training, reducing the deployment burden for new recording modalities.

*YOLO detection and confidence tuning.* At the default threshold ( $\tau=0.25$ ) the YOLO detector fires on only 68.8 % of **BAGLS** frames. Lowering  $\tau$  to 0.02 recovers 85.9 % recall and lifts YOLO-Crop+UNet DSC from 0.61 to 0.64 (Figure 4). Below  $\tau=0.02$ , 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 replicates the group differences (Healthy vs. Pathological) established through manual or semi-automated analysis in the literature [20, 4, 2]. The key result is that the coefficient of variation (cv) significantly distinguishes Healthy from Pathological voices ( $p=0.006$ , female subgroup)—demonstrating that the pipeline is not merely accurate at the pixel level (DSC/IoU) but yields *clinically useful* biomarkers. **GIRAFE** and **BAGLS** are the primary benchmarks for segmentation accuracy (DSC/IoU); the 65-subject **GIRAFE** cohort is the primary benchmark for *clinical reproducibility* of those kinematic findings. Because the **GIRAFE** cohort has a significant sex imbalance (Fisher’s exact  $p=0.025$ ) and  $f_0$  is strongly sex-dependent, Table 5 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 [4]. The automated cv result thus aligns with the variability trends reported by

Patel et al. [20], demonstrating that the pipeline “sees” what clinicians see when distinguishing normal from disordered voices. 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.

## 7. Conclusion

We presented a lightweight U-Net trained with a carefully tuned recipe (grayscale input, combined BCE + DSC loss, AdamW with cosine annealing) that sets a new state of the art on the **GIRAFE** benchmark (DSC 0.81,  $DSC \geq 0.5 = 96.2\%$ ), 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 (DSC 0.64 at optimized threshold  $\tau=0.02$ ), demonstrating that the detection-guided approach generalizes across institutions and equipment. When U-Net and YOLO are trained on **BAGLS**, the pipeline attains DSC 0.85 on the full **BAGLS** test set (YOLO+UNet), surpassing the benchmark baseline [5] and diffusion-refined methods [28], establishing strong in-distribution results on that benchmark as well. 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). Validation thus goes beyond pixel-level metrics (DSC): the pipeline replicates established clinical group differences (Healthy vs. Pathological) and preserves the coefficient of variation as the key discriminant for vocal pathology [20, 4]. By providing a fully automated, detection-gated, zero-shot-capable system, the pipeline makes these clinically validated kinematic findings *clinically scalable*.

## 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 repository README describes dataset splits (training/validation/test) for both **GIRAFE** and **BAGLS** and explains how to run the detection-gated pipeline (YOLO detector, U-Net segmenter, and evaluation scripts). The **GIRAFE** dataset [7] is freely available from <https://zenodo.org/records/13773163>. The **BAGLS** dataset [5] is available from <https://zenodo.org/records/3762320>.

## Declaration of Competing Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

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.

## Ethical Statement

The author confirms that this study was conducted using only secondary, de-identified data from publicly available research benchmarks (**BAGLS** and **GIRAFE**). As the research involved the analysis of pre-existing, non-identifiable datasets and did not involve direct interaction with human subjects or the collection of private health information, it was deemed exempt from institutional review board (IRB) approval in accordance with standard ethical guidelines for secondary data analysis. The original data collection for the **BAGLS** and **GIRAFE** datasets was conducted under the ethical oversight of their respective contributing institutions, and this study adheres to their terms of use and the principles of the Declaration of Helsinki.

## References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] R. R. Patel, K. D. Donohue, H. Unnikrishnan, R. J. Kryscio, Effects of vocal fold nodules on glottal cycle measurements derived from high-speed videoendoscopy in children, *PLOS ONE* 11 (4) (2016) e0154586, nodules vs. typically developing children; kinematic features. doi:10.1371/journal.pone.0154586.
- [5] 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.
- [6] F. J. P. Montalbo, S3AR U-Net: A separable squeezed similarity attention-gated residual U-Net for glottis segmentation, *Biomedical Signal Processing and Control* 92 (2024) 106047. doi:10.1016/j.bspc.2024.106047.  
URL <https://www.sciencedirect.com/science/article/pii/S1746809424001058>
- [7] 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.2025.111376.  
URL <https://doi.org/10.1016/j.dib.2025.111376>

- [8] 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.
- [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] 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.
- [11] 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.
- [12] S. M. N. Nobel, S. M. M. R. Swapno, M. R. Islam, M. Safran, S. Alfarhood, M. F. Mridha, A machine learning approach for vocal fold segmentation and disorder classification based on ensemble method, *Scientific Reports* 14, pMCID: PMC11758383; PMID: 38910146. Ensemble UNet-BiGRU segmentation (IoU 87.46%); no evaluation on public BAGLS or GIRAFE benchmarks. (2024). doi:10.1038/s41598-024-64987-5.  
URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11758383/>
- [13] M. Döllinger, T. Schraut, L. A. Henrich, D. Chhetri, M. Echternach, A. M. Johnson, M. Kunduk, Y. Maryn, R. R. Patel, R. Samlan, et al., Re-training of convolutional neural networks for glottis segmentation in endoscopic high-speed videos, *Applied Sciences* 12 (19) (2022) 9791. doi:10.3390/app12199791.  
URL <https://doi.org/10.3390/app12199791>
- [14] A. M. Kist, P. Gómez, D. Dubrovskiy, P. Schlegel, M. Kunduk, M. Echternach, R. Patel, M. Semmler, C. Bohr, S. Dürr, A. Schützenberger,

- M. Döllinger, A deep learning enhanced novel software tool for laryngeal dynamics analysis, *Journal of Speech, Language, and Hearing Research* 64 (6) (2021) 1889–1903, *pub* 2021 May 17. PMID: 34000199. doi:10.1044/2021\_JSLHR-20-00498.
- [15] 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.
  - [16] 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.
  - [17] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.322.
  - [18] G. Andrade-Miranda, J. I. Godino-Llorente, L. Moro-Velázquez, J. A. Gómez-García, An automatic method to detect and track the glottal gap from high speed videoendoscopic images, *Biomed. Eng. Online* 14 (2015) 100. doi:10.1186/s12938-015-0096-3.  
URL <https://doi.org/10.1186/s12938-015-0096-3>
  - [19] R. R. Patel, K. D. Donohue, D. Lau, H. Unnikrishnan, In vivo measurement of pediatric vocal fold motion using structured light laser projection, *Journal of Voice* 27 (4) (2013) 463–472. doi:10.1016/j.jvoice.2013.03.004.
  - [20] 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. doi:10.1044/2015\_JSLHR-S-14-0242.
  - [21] J. Hoffman, M. Barańska, E. Niebudek-Bogusz, W. Pietruszewska, Comparative evaluation of high-speed videoendoscopy and laryngovideostroboscopy for functional laryngeal assessment in clinical practice, *Journal of Clinical Medicine* 14 (5) (2025) 1723, pMCID: PMC11899851.

doi:10.3390/jcm14051723.

URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11899851/>

- [22] T. Schraut, A. Schützenberger, T. Arias-Vergara, M. Kunduk, M. Echternach, S. Dürr, J. Werz, M. Döllinger, Machine learning based assessment of hoarseness severity: a multi-sensor approach centered on high-speed videoendoscopy, *Frontiers in Artificial Intelligence* 8 (2025). doi:10.3389/frai.2025.1601716.  
URL <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1601716>
- [23] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, version 8.0.0, AGPL-3.0 (2023).  
URL <https://github.com/ultralytics/ultralytics>
- [24] 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.
- [25] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019, pp. 1–22.  
URL <https://openreview.net/forum?id=Bkg6RiCqY7>
- [26] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: International Conference on Learning Representations (ICLR), 2017, pp. 1–16.  
URL <https://openreview.net/forum?id=Skq89Scxx>
- [27] 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.
- [28] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, Y. Xu, MedSegDiff: Medical image segmentation with diffusion probabilistic model, in: Medical Imaging with Deep Learning, Vol. 227 of Proceedings of Machine Learning Research, PMLR, 2024, pp. 1623–1639.  
URL <https://proceedings.mlr.press/v227/wu24a.html>

- [29] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2015, pp. 1–15.  
URL <https://arxiv.org/abs/1412.6980>