INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Weakly supervised glottis segmentation in high-speed videoendoscopy using bounding box labels

*Varun Belagali, Achuth Rao M V, Prasanta Kumar Ghosh*

Electrical Engineering, Indian Institute of Science, Bengaluru 560012, India
varunbelagali98@gmail.com, achuthraomv@gmail.com, prasantg@iisc.ac.in

## Abstract

Analysis of vocal fold vibration in high-speed videoendoscopy can aid in the assessment of voice disorders. Glottis segmentation is a preliminary step of this analysis. Previous deep learning approaches have focused on fully supervised learning methods for glottis segmentation which require pixel-level annotation. Collection of pixel-level annotated data is time consuming and tedious. To overcome this challenge, in this work, we explore the use of bounding box labels for weakly supervised glottis segmentation. As such, bounding box labels are relatively easier to annotate. The proposed method uses multiple instance learning to leverage bounding box labels in the form of bag labels. The method outperforms the baseline method (trained with bounding box as mask) by 0.20 in terms of dice score, and matches the performance of fully supervised version after fine-tuning.

**Index Terms**: Glottis segmentation, weakly supervised learning, high-speed videoendoscopy.

## 1. Introduction

A comprehensive diagnostic voice evaluation is a multidimensional process that consists of several perceptual and instrumental evaluations [1]. Laryngeal imaging has been the norm for diagnostic voice evaluation and includes procedures like direct laryngoscopy, videostroboscopy, VideoKymography (VKG), and high-speed videoendoscopy (HSV) [2, 3, 4]. In day-to-day clinical practice, HSV is being widely used to monitor the changes in vocal fold status as a response to treatment procedures being used. This shift in practice paradigm has resulted in an increase in the use of vocal imaging procedures, the amount of imaging data generated, and a strong need for quantitative measures of the changes that are induced as a result of ongoing treatment. For accurate clinical diagnosis and, even more so, for clinical research, quantified yet interpretable parameters are required. To achieve a completely automated quantitative method based on HSV recordings, the essential first step is, the segmentation of the glottal area from these images. Since HSV recordings consist of thousands of images, manual segmentation of the image data is not feasible. Therefore, automated segmentation algorithms are necessary to allow accurate, robust, and efficient segmentation of the glottal area. Figure 1 represents the glottal area (red contour) drawn over a sample video frame. Figure 2 indicates sample video frames and pixel-level labels for segmentation.

Many deep learning-based approaches have been proposed for glottis segmentation in the literature. However, training these deep learning models require pixel-level annotated data. Rao et al. [5] proposed a method that uses fully connected deep neural networks (DNN) for glottis segmentation. This

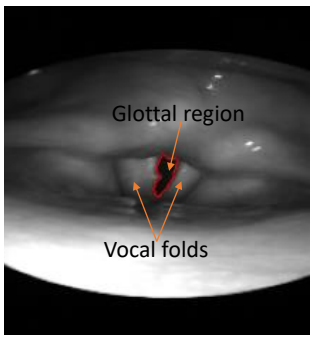


Figure 1: *Sample video frame indicating glottal region (red contour) and vocal folds.*

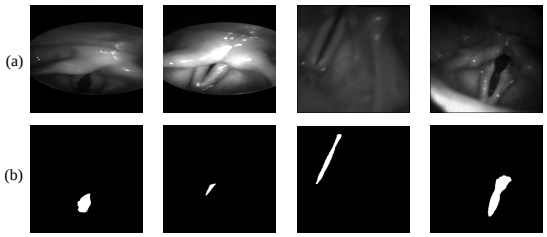


Figure 2: *Sample video frames (a) and corresponding ground truth pixel-level labels indicating glottis segment (b).*

method treats segmentation as a pixel-wise classification problem. To classify each pixel, a feature vector is formed using a $3 \times 3$ image patch around that pixel. This feature vector is passed through a DNN with sigmoid activation. This method can fail to account for the global context as the receptive field ($3 \times 3$) for the segmentation is low. Convolutional neural networks are well suited for segmentation when compared to fully connected DNN. The most popular CNN for segmentation is SegNet [6]. SegNet is a fully convolutional encoder decoder network. Previous work [7] used Segnet for glottis segmentation. The work proposed a two-step localization and segmentation network built using SegNet. The segmentation accuracy depends on the localization network, and bad localization can lead to poor segmentation results. Another predominant CNN used for image segmentation is U-Net [8], a network specifically designed for medical imaging. U-Net is a fully convolutional network with skip connections. The glottis segmentation using U-Net has been proposed in [9]. Further, a deep convolutional LSTM network has been proposed for glottis seg-


10.21437/Interspeech.2023-983

mentation in endoscopic laryngeal high-speed videos [10]. The LSTM is used to take advantage of the temporal relationship between sequential frames. The usage of re-training of neural networks for new recording modalities has been explored in the work [11].

The neural network approaches proposed in the literature use supervised learning and need pixel-level annotated data. The collection of pixel-level annotated data for segmentation is a time-consuming and arduous task as it requires boundary tracing. Medical imaging faces additional challenges as annotation needs to be done only by the domain experts [12]. In the case of glottis segmentation, experts are Speech Language Pathologists. Weakly supervised techniques can help us overcome this arduous process of pixel-level labeled data collection required for glottis segmentation. Bounding box level labeling is one of the labeling method that can be used for weakly supervised learning. Hsu Cheng-Chun et al. [13] used multiple instance learning [14] for instance segmentation. Bounding box labels have not been explored for glottis segmentation. In our work, we propose a method that uses multiple instance learning (MIL) to train U-Net end to end for glottis segmentation with bounding box as weak supervision. We use the BAGLS dataset [9] which contains 59250 HSV images. The proposed weakly supervised model achieves an dice score of 0.75, and outperforms the baseline (trained with bounding box as mask) by 0.20. This indicates the importance of MIL formulation to achieve weakly supervised glottis segmentation. To further improve results, the U-Net is trained (fine-tuned) with a small subset of pixel-level lableled data.

## 2. Proposed method for weakly supervised glottis segmentation

We propose an approach that uses the bounding box level annotation to learn glottis segmentation. The segmentation task is treated as a MIL problem to train the U-Net [8]. In this section we provide details regarding the U-Net architecture, formulation of segmentation in terms of MIL, usage of ground truth labels, and loss functions. Figure 4 represents the overview of the proposed method.

### 2.1. U-Net

U-Net is a convolutional neural network which has gained great success in medical imaging [8]. In this work, we use U-Net for glottis segmentation which has also been used in the previous works [9, 15]. U-Net contains standard encoder decoder structure with skip connections between encoder and decoder layers. In this paper, we maintain the same architecture as [9] which contains four encoder and four decoder layers with final layer as sigmoid activation.

### 2.2. MIL formulation of glottis segmentation

MIL is a training method that deals with data represented in the form of sets. These sets are called bags [16]. The supervised labels are only provided for the bag and not for each instance in the bag. A bag is labelled as positive for the classification problem if it consists of at least one instance belonging to the positive class. A bag is labelled negative if it consists of only negative class instances. This type of representation can be used for weakly supervised learning where only the bag label is known. In our case, the weak supervision is in the form of the bounding box.

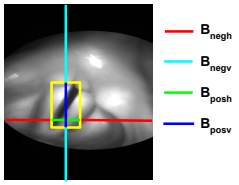


Figure 3: *Bag formation using a sample image. $B_{negh}$ and $B_{negv}$ represent sample horizontal and vertical negative bags. $B_{posh}$ and $B_{posv}$ represent sample horizontal and vertical positive bags.*

Given a bounding box level annotation, every pixel outside the bounding box belongs to the background. A pixel inside the bounding box can belong to either glottis or the background. In terms of MIL, pixels outside the bounding box form negative bags as all such pixels belong to the background (negative class). The pixels inside the bounding box form positive bags as some pixels belong to glottis (positive class) and other pixels belong to background (negative class). We treat every horizontal line and vertical line of pixels that are outside the bounding box as a negative bag. Each horizontal and vertical line of pixels inside the bounding box is considered as a positive bag. Here, the line of pixels is a bag and each pixel on the line is an instance. Figure 3 indicates sample positive and negative bags. Equations [1,2,3,4] indicate the formation of bags. I is a M $\times$ N image. *Box* represents M $\times$ N bounding box mask. $B_{negH}$ and $B_{negV}$ indicate horizontal and vertical negative bags respectively. $B_{posH}$ and $B_{posV}$ indicate horizontal and vertical positive bags respectively.

$$B_{negH} = \{B_{negh}\}_{h=0}^{M-1}$$
$$B_{negh} = \{I_{h,j} | Box_{h,j} = 0\}_{j=0}^{N-1} \quad (1)$$

$$B_{negV} = \{B_{negv}\}_{v=0}^{N-1}$$
$$B_{negv} = \{I_{i,v} | Box_{i,v} = 0\}_{i=0}^{M-1} \quad (2)$$

$$B_{posH} = \{B_{posh}\}_{h=0}^{M-1}$$
$$B_{posh} = \{I_{h,j} | Box_{h,j} = 1\}_{j=0}^{N-1} \quad (3)$$

$$B_{posV} = \{B_{posv}\}_{v=0}^{N-1}$$
$$B_{posv} = \{I_{i,v} | Box_{i,v} = 1\}_{i=0}^{M-1} \quad (4)$$

### 2.3. Training U-Net using MIL

Glottis segmentation is a pixel-wise binary classification task. Sigmoid is used as the activation of the final layer of U-Net. The output of each sigmoid unit is in the range (0,1). We form the negative and positive bags as described in section 2.2. The U-Net is trained on bag level classification accuracy as we only have knowledge regarding bag level ground truth labels generated from bounding box. Ground truth label of negative bags is zero and positive bags is one. Since U-Net generates pixel-level probabilities, there is a need to convert these pixel-level probabilities into bag probabilities. We derive the probability of a bag belonging to the positive class as the maximum of sigmoid outputs in the final layer corresponding to pixels (instances) belonging to the bag. This can be generated by horizontal and vertical max-pooling of sigmoid outputs as indicated in Figure 4. Equations [5,6,7,8] represent the derivation of predicted

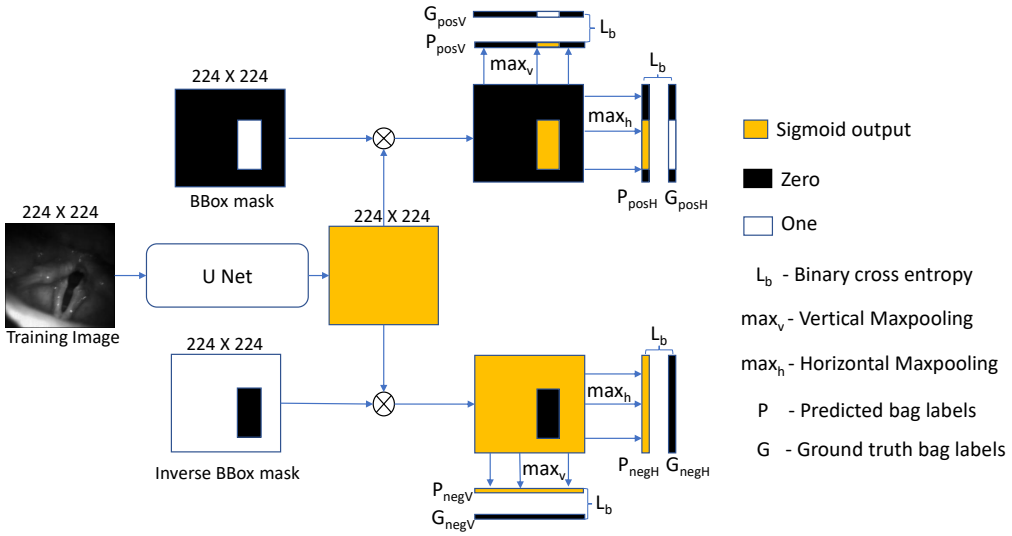


Figure 4: *Training U-Net with MIL.*

bag labels from M $\times$ N sigmoid output P. $P_{negH}$ and $P_{negV}$ indicate predicted labels of horizontal and vertical negative bags respectively. $P_{posH}$ and $P_{posV}$ indicate predicted labels of horizontal and vertical positive bags respectively. The binary cross-entropy [17] is used as a bag level loss function. MIL trained with a bag level loss might have poor instance-level accuracy (segmentation) [16]. To tackle this, in addition to bag level loss, pairwise loss function is used to enforce neighbourhood consistency. This loss function is calculated between each pixel and it's 3 $\times$ 3 neighbourhood pixels. Equation 9 indicates the loss function calculated between pixel i and pixel j, where P is the sigmoid output of U-Net and I is the input image.

$$P_{negH} = \{P_{negh}\}_{h=0}^{M-1}$$
$$P_{negh} = max\{P_{h,j}|Box_{h,j} = 0\}_{j=0}^{N-1} \quad (5)$$

$$P_{negV} = \{P_{negv}\}_{v=0}^{N-1}$$
$$P_{negv} = max\{P_{i,v}|Box_{i,v} = 0\}_{i=0}^{M-1} \quad (6)$$

$$P_{posH} = \{P_{posh}\}_{h=0}^{M-1}$$
$$P_{posh} = max\{P_{h,j}|Box_{h,j} = 1\}_{j=0}^{N-1} \quad (7)$$

$$P_{posV} = \{P_{posv}\}_{v=0}^{N-1}$$
$$P_{posv} = max\{P_{i,v}|Box_{i,v} = 1\}_{i=0}^{M-1} \quad (8)$$

$$PairwiseLoss = (1 - (I_i - I_j)^2) * (P_i - P_j)^2 \quad (9)$$

# 3. Experiments

## 3.1. Experimental Setup

We train the U-Net with the bounding box level labels using multiple instance learning with Bag loss and Pairwise loss functions. We term this model as MIL. For baseline, we use the U-Net model trained with bounding box mask as segmentation mask using dice loss [17]. This model is termed as BoxM. Further, to compare the MIL with fully supervised version, we train the U-Net proposed in [9] with pixel-level annotated labels using dice loss. This model is termed as FullSup. BoxM, MIL, and FullSup use the same U-Net model, but each model is trained using different labels and loss functions. We show that MIL performs better than BoxM, and matches performance of FullSup when fine tuned with small fraction of pixel-level labelled data. The importance of using pairwise loss function to boost the segmentation accuracy is also shown. The models are implemented in Keras [18] with TensorFlow [19] as the backend.

## 3.2. Dataset

We use the dataset presented in the work [9]. This data set consists of 59,250 high-speed videoendoscopy images collected from 640 subjects. We divide these 640 subjects into 5 folds for cross-validation. Each fold contains 11850 images. The training data consists of 3 folds, validation data consists of 1 fold, and the remaining 1 fold is used as test data. The dataset does not contain bounding box labels, hence, we generate them using the ground truth segmentation mask annotation. A tight bounding box is derived from the pixel-level annotation such that the box intersects the mask's outermost pixel in all four directions.

## 3.3. Evaluation Metric

Following the previous works [5, 7, 10, 15], we use dice score [20] as the evaluation metric. Equation 10 indicates the formulation of dice score between the prediction (P) and the ground truth (G), where both P and G are binary images of same size. Class one indicates pixel belongs to glottis and class zero represents background pixel. N indicates the number of pixels with value one. During the evaluation, pixel-level annotation is used as ground truth for all the models.

$$Dice(P, G) = \frac{2 \times N(P \cap G)}{N(P) + N(G)} \quad (10)$$

### 3.4. Results

#### 3.4.1. MIL performance

We compare the results achieved by BoxM, MIL, and FullSup. Figure 5 represents the box plot of the dice score achieved by all methods across 5 folds. The average dice scores achieved by BoxM, MIL, and FullSup are 0.55 ($\pm$0.22), 0.75 ($\pm$0.19), and 0.84 ($\pm$0.20) respectively. MIL clearly outperforms BoxM indicating that the MIL formulation of segmentation using the bounding box is effective compared to directly using the bounding box as mask for segmentation. Figure 6 indicates the sample outputs of BoxM, MIL, and FullSup. As seen in Figure 6, BoxM tends to predict box shaped segments rather than the glottis shape. FullSup forms the upper bound for MIL performance as the FullSup is trained with pixel-level annotations and it uses the same U-Net architecture as MIL. On the other hand, MIL uses only the bounding box level labels for training. The exact glottis edge information is not used in the training which can lead to a small number of incorrect predictions at the edges as seen in Figure 6. But, despite the lack of information about the exact edges during the training phase, MIL is able to achieve a dice score of 0.75 on the test data.

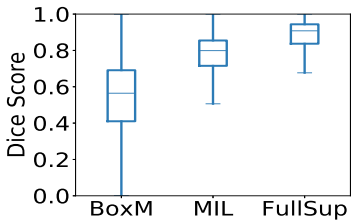


Figure 5: *Boxplot of dice score achieved by BoxM, MIL, and FullSup.*

MIL results can be improved by fine-tuning it with a small percentage of pixel-level labeled images as it has to refine a small number of incorrect predictions near the glottal edges. Table 1 indicates dice score achieved by BoxM and MIL models after fine-tuning with p% of pixel-level labeled data. FullSup trained with p% of training data is also presented. As p increases, the dice score achieved by MIL and BoxM improves. For p $\geq$ 0.5, MIL matches FullSup trained with whole data.

Table 1: *Dice score of BoxM and MIL after fine tuning with p% (= N number of images) of pixel-level labelled data. The corresponding FullSup model trained with same data is also presented.*

| | | Dice Score | | |
|---|---|---|---|---|
| p% | N | BoxM | MIL | FullSup |
| 0 | 0 | 0.55 | **0.75** | – |
| 0.01 | 4 | 0.60 | **0.78** | 0.01 |
| 0.05 | 23 | 0.78 | **0.80** | 0.40 |
| 0.1 | 47 | 0.80 | **0.81** | 0.01 |
| 0.5 | 237 | 0.78 | **0.83** | 0.01 |
| 1.0 | 474 | 0.75 | **0.83** | 0.21 |

#### 3.4.2. Importance of pairwise loss function

We compare the dice score achieved by MIL trained without pairwise loss (NoPair) and with pairwise loss (Pair). Figure 7b represents the box plot of dice score achieved by NoPair and Pair. The average dice scores of NoPair and Pair are 0.65 ($\pm$ 0.18) and 0.75 ($\pm$ 0.19) respectively. Pair outperforms NoPair indicating the importance of pairwise loss function. Figure 7a represents the sample output of NoPair and Pair. Without pairwise loss, MIL tends to learn only the discriminative parts of the glottis because of the max pooling function used to calculate bag level probabilities (Equations [5-8]). Pairwise loss function helps in learning less discriminative parts by adding spatial consistency constraint.

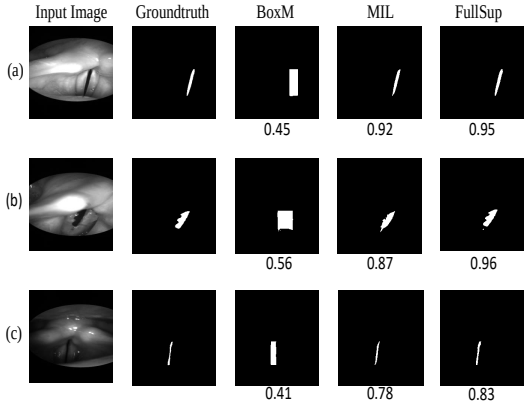


Figure 6: *Sample predictions (a, b, c) of BoxM, MIL, and FullSup along with corresponding input image and ground truth. Dice score is indicated at the bottom of the prediction.*

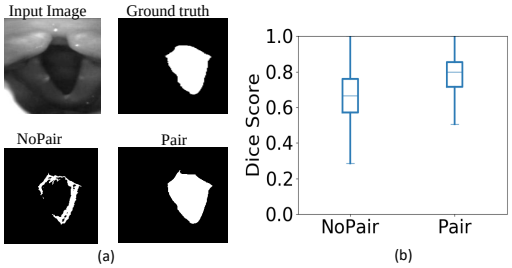


Figure 7: *Sample prediction of NoPair and Pair along with corresponding input image and ground truth (a). Boxplot of dice score achieved by NoPair) and Pair (b).*

## 4. Conclusion

The major challenge of deep learning approach for glottis segmentation is the need of pixel-level labelled data which is expensive to collect as it requires labelling from experts. This challenge posits the need to explore deep learning methods that use less or weakly labelled data. One such approach is weakly supervised learning. We propose a weakly supervised glottis segmentation method using bounding box labels which are relatively easy to collect. The proposed method uses multiple instance learning technique where the bounding box label are converted to bag labels. The proposed MIL outperforms the baseline method, and matches the performance of fully-supervised method after fine-tuning. In future work, we would like to explore weakly supervised learning using image-level labels (whether glottis opening is present or not). Another direction will be to work on self-supervised learning which requires exploration on pretext task formulation.

# 5. References


[1] A. Behrman, "Common practices of voice therapists in the evaluation of patients," *Journal of Voice*, vol. 19, no. 3, pp. 454–469, 2005.

[2] J. L. Cutler and T. Cleveland, "The clinical usefulness of laryngeal videostroboscopy and the role of high-speed cinematography in laryngeal evaluation," *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 10, no. 6, pp. 462–466, 2002.

[3] Q. Qiu, H. Schutte, L. Gu, and Q. Yu, "An automatic method to quantify the vibration properties of human vocal folds via videokymography," *Folia Phoniatrica et Logopaedica*, vol. 55, no. 3, pp. 128–136, 2003.

[4] D. D. Deliyski, P. P. Petrushev, H. S. Bonilha, T. T. Gerlach, B. Martin-Harris, and R. E. Hillman, "Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 1, pp. 33–44, 2008.

[5] A. Rao MV, R. Krishnamurthy, P. Gopikishore, V. Priyadharshini, and P. K. Ghosh, "Automatic glottis localization and segmentation in stroboscopic videos using deep neural network," in *Proc. Interspeech 2018*, 2018, pp. 3007–3011. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2572

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[7] V. Belagali, A. Rao, P. Gopikishore, R. Krishnamurthy, and P. K. Ghosh, "Two step convolutional neural network for automatic glottis localization and segmentation in stroboscopic videos," *Biomedical Optics Express*, vol. 11, no. 8, pp. 4695–4713, 2020.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[9] P. Gómez, A. M. Kist, P. Schlegel, D. A. Berry, D. K. Chhetri, S. Dürr, M. Echternach, A. M. Johnson, S. Kniesburges, M. Kunduk *et al.*, "Bagls, a multihospital benchmark for automatic glottis segmentation," *Scientific data*, vol. 7, no. 1, pp. 1–12, 2020.

[10] M. K. Fehling, F. Grosch, M. E. Schuster, B. Schick, and J. Lohscheller, "Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional lstm network," *Plos one*, vol. 15, no. 2, p. e0227791, 2020.

[11] 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*, vol. 12, no. 19, p. 9791, 2022.

[12] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," *Classification in BioApps*, pp. 323–350, 2018.

[13] C.-C. Hsu, K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised instance segmentation using the bounding box tightness prior," *Advances in Neural Information Processing Systems*, vol. 32, pp. 6586–6597, 2019.

[14] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[15] D. Degala, A. Rao, R. Krishnamurthy, P. Gopikishore, V. Priyadharshini, T. Prakash, and P. K. Ghosh, "Automatic glottis detection and segmentation in stroboscopic videos using convolutional networks," 2020.

[16] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[17] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2020, pp. 1–7.

[18] F. Chollet *et al.*, "Keras," 2015.

[19] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[20] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.