Joint Segmentation and Sub-pixel Localization in Structured Light Laryngoscopy

Jann-Ole Henningson[1(✉)], Marion Semmler[2], Michael Döllinger[2], and Marc Stamminger[1]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
jann-ole.henningson@fau.de

[2] Division of Phoniatrics and Pediatric Audiology at the Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany

Abstract. In recent years, phoniatric diagnostics has seen a surge of interest in structured light-based high-speed video endoscopy, as it enables the observation of oscillating human vocal folds in vertical direction. However, structured light laryngoscopy suffers from practical problems: specular reflections interfere with the projected pattern, mucosal tissue dilates the pattern, and lastly the algorithms need to deal with huge amounts of data generated by a high-speed video camera. To address these issues, we propose a neural approach for the joint semantic segmentation and keypoint detection in structured light high-speed video endoscopy that improves the robustness, accuracy, and performance of current human vocal fold reconstruction pipelines. Major contributions are the reformulation of one channel of a semantic segmentation approach as a single-channel heatmap regression problem, and the prediction of sub-pixel accurate 2D point locations through weighted least squares in a fully-differentiable manner with negligible computational cost. Lastly, we expand the publicly available Human Laser Endoscopic dataset to also include segmentations of the human vocal folds itself. The source code and dataset are available at: github.com/Henningson/SSSLsquared

Keywords: Human Vocal Folds · Laryngoscopy · Keypoint Detection · Semantic Segmentation

1 Introduction

The voice is an essential aspect of human communication and plays a critical role in expressing emotions, conveying information, and establishing personal connections. An impaired function of the voice can have significant negative

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_4.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 34–43, 2023.
https://doi.org/10.1007/978-3-031-43987-2_4

impacts on an individual. Malign changes of human vocal folds are conventionally observed by the use of (high-speed) video endoscopy that measures their 2D deformation in image space. However, it was shown that their dynamics contain a significant vertical deformation. This led to the development of varying methods for the 3D reconstruction of human vocal folds during phonation. In these works, especially active reconstruction systems have been researched that project a pattern onto the surface of vocal folds [10,16,19,20]. All these systems need to deal with issues introduced by the structured light system, in particular a) specular reflections from the endoscopic light source which occlude the structured light pattern (cp. Fig. 1), b) a dilation of the structured light pattern through subsurface scattering effects in mucosal tissue, and c) vasts amount of data generated by high-speed cameras recording with up to 4000 frames per second. Furthermore, the introduction of a structured light source, e.g. a laser projection unit (LPU), increases the form-factor of the endoscope, which makes recording uncomfortable for the patient. In current systems, the video processing happens offline, which means that often unnecessarily long footage is recorded to be sure that an appropriate sequence is contained, or—even worse—that a patient has to show up again, because the recorded sequence is not of sufficient quality. Ideally, recording should thus happen in a very short time (seconds) and provide immediate feedback to the operator, which is only possible if the segmentation and pattern detection happen close to real time.

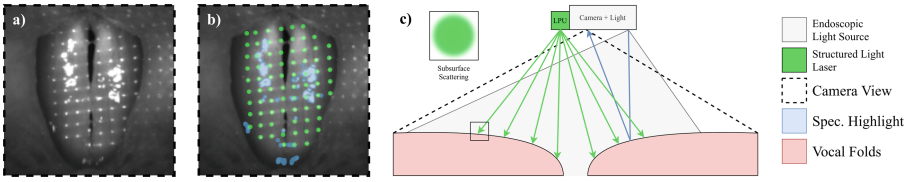

Fig. 1. Recording setup. From left to right: a) single frame of a recorded video containing laser dots and specular reflections, b) same frame with highlighted laser dots and specular reflections, c) illustration of recording setup with endoscope and laser projection unit (LPU) looking at the vocal folds.

To address all these practical issues, we present a novel method for the highly efficient and accurate segmentation, localization, and tracking of human vocal folds and projection patterns in laser-supported high-speed video endoscopy. An overview of our pipeline is shown in Fig. 2. It is based on two stages: First, a convolutional neural network predicts a segmentation of the vocal folds, the glottal area, and projected laser dots. Secondly, we compute sub-pixel accurate 2D point locations based on the pixel-level laser dot class probabilities in a weighted least-squares manner that further increase prediction accuracy. This approach can provide immediate feedback about the success of the recording to the physician, e.g. in form of the number of successfully tracked laser dots. Furthermore, this method can not only be used in vocal fold 3D reconstruction pipelines but also allows for the analysis of clinically relevant 2D features.

2 Related Work

Our work can properly be assumed to be simultaneously a heatmap regression as well as a semantic segmentation approach. Deep learning based medical semantic segmentation has been extensively studied in recent years with novel architectures, loss-functions, regularizations, data augmentations, holistic training and optimization approaches [23]. Here, we will focus on the specific application of semantic segmentation in laryngoscopy.

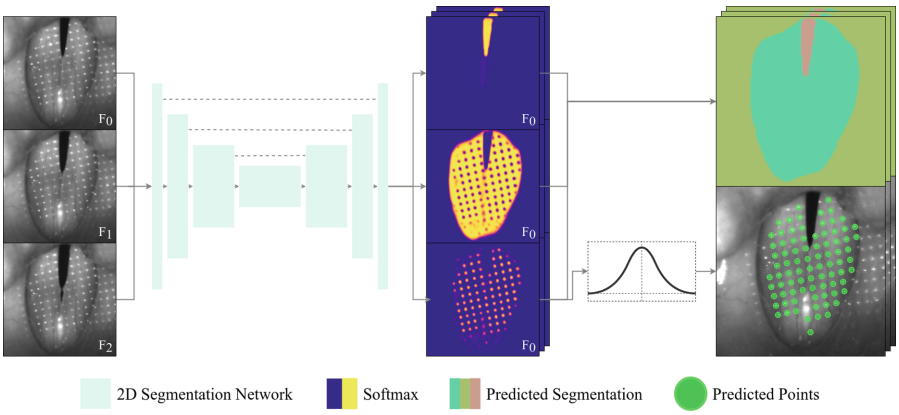

Fig. 2. Our method receives an image sequence and computes a pixel-wise classification of the glottal gap, vocal folds, and laser points. Next, we estimate 2D keypoint locations on the softmaxed laser point output of our model via weighted least squares.

Deep Learning in Laryngoscopy. In the realm of laryngoscopy works involving deep learning are few and far between. Most of these works focus on the segmentation of the glottal gap over time. The glottal dynamics give information about the patients underlying conditions, all the while being an easily detectable feature. Fehling et al. were the first to propose a CNN-based method that also infers a segmentation of the human vocal folds itself [8]. Their method uses a general U-Net architecture extended with Long Short-Term Memory cells to also take temporal information into account. Pedersen et al. [17] use off-the-shelf U-Nets to estimate glottal gap and vocal folds itself. Cho et al. [3] compare different segmentation architectures including CNN6, VGG16, Inception V3 and XCeption. Döllinger et al. [4] have shown that pretraining CNNs for human vocal fold segmentation can boost the respective CNNs performance. To the best of our knowledge, no publication has specifically targeted segmentation and detection in structured light laryngoscopy via deep learning.

Keypoint Detection can generally be separated into regression-based approaches that infer keypoint positions directly from the images [5,22,26] and

heatmap-based approaches that model the likelihood of the existence of a keypoint, i.e. landmark, via channel-wise 2D Gaussians and determining channel-wise global maxima via $argmax()$. This leads to obvious quantization errors, that are addressed in recent works [1,7,25]. Most related to our approach is the work by Sharan et al. that proposes a method for determining the position of sutures by reformulating a single-channel binary segmentation to find local maxima and calculating their positions through general center of gravity estimation [21]. In case of human vocal folds, there have been works regarding laser dot detection in structured light laryngoscopy. However, these suffer from either a manual labeling step [14,16,19,20] or work only on a per-image basis [10], thus being susceptible to artifacts introduced through specular highlights. In a similar vein, none of the mentioned methods apply promising deep learning techniques.

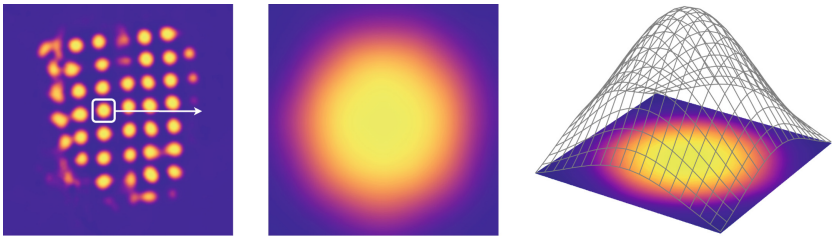

Fig. 3. We extract windows around local maxima through dilation filtering and fit a Gaussian function into the window to estimate sub-pixel accurate keypoints.

3 Method

Given an isotropic light source and a material having subsurface scattering properties, the energy density of the penetrating light follows an exponential falloff [12]. In case of laser-based structured light endoscopy, this means that we can observe a bleeding effect, i.e. diffusion, of the respective collimated laser beams in mucosal tissue. Note that the energy density of laser beams is Gaussian distributed as well, amplifying this effect. Our method is now based on the assumption, that a pixel-level (binary- or multi-class) classifier will follow this exponential falloff in its predictions, and we can estimate sub-pixel accurate point positions through Gaussian fitting (cp. Fig. 3). This allows us to model the keypoint detection as a semantic segmentation task, such that we can jointly estimate the human vocal folds, the glottal gap, as well as the laserpoints' positions using only a single inference step. The method can be properly divided into two parts. At first, an arbitrary hourglass-style CNN (in our case we use the U-Net architecture) estimates semantic segmentations of the glottal gap, vocal folds and the 2D laserdots position for a fixed videosequence in a supervised manner. Next, we extract these local maxima lying above a certain threshold and use weighted least squares to fit a Gaussian function into the windowed regions. In this way, we can estimate semantic segmentations as well as the position of keypoints in a single inference pass, which significantly speeds up computation.

Feature Extraction. Current heatmap regression approaches use the $argmax(.)$ or $topk(.)$ functions to estimate channel-wise global maxima, i.e. a single point per channel. In case of an 31 by 31 LPU this would necessitate such an approach to estimate 961 channels; easily exceeding maintainable memory usage. Thus our method needs to allow multiple points to be on a single channel. However, this makes taking the $argmax(.)$ or $topk(.)$ functions to extract local maxima infeasible, or outright impossible. Thus, to extract an arbitrary amount of local maxima adhering to a certain quality, we use dilation filtering on a thresholded and Gaussian blurred image. More precisely we calculate $\mathbf{I}_T = \mathbf{I} > [T(\mathbf{I}, \theta) * \mathbf{G} \oplus \mathbf{B}]$, where $T(x, y)$ depicts a basic thresholding operation, $\mathbf{G}$ a Gaussian kernel, $\mathbf{B}$ a general box kernel with a 0 at the center, and $\oplus$ the dilation operator. Finally we can easily retrieve local maxima by just extracting every non-zero element of $\mathbf{I}_T$. We then span a window I_{ij} of size $k \in N$ around the non-zero discrete points $\mathbf{p}_i$. For improved comprehensibility, we are dropping the subscripts in further explanations, and explain the algorithm for a single point $\mathbf{p}$. Next, we need to estimate a Gaussian function. Note that, a Gaussian function is of the form $f(x) = Ae^{-(x-\mu)^2/2\sigma^2}$, where $x = \mu$ is the peak, A the peaks height, and σ defines the functions width. Gaussian fitting approaches can generally be separated into two types: the first ones employ non-linear least squares optimization techniques [24], while others use the result of Caruana et al. in that the logarithm of a Gaussian is a polynomial equation [2].

Caruanas Algorithm. As stated previously, Caruanas algorithm is based on the observation that by taking the logarithm of the Gaussian function, we generate a polynomial equation (Eq. 1).

$$\ln(f(x)) = \ln(A) + \frac{-(x-\mu)^2}{2\sigma^2} = \ln(A) - \frac{-\mu^2}{2\sigma^2} + \frac{2\mu x}{2\sigma^2} - \frac{x^2}{2\sigma^2} \tag{1}$$

Note that the last equation is in polynomial form $\ln(y) = ax^2 + bx + c$, with $c = \ln(A) - \frac{-\mu^2}{2\sigma^2}$, $b = \frac{\mu}{\sigma^2}$ and $a = \frac{-1}{2\sigma^2}$. By defining the error function $\delta = \ln(f(x)) - (ax^2 + bx + c)$ and differentiating the sum of residuals gives a linear system of equations (Eq. 2).

$$\begin{bmatrix} N & \sum x & \sum x^2 \\ \sum x & \sum x^2 & \sum x^3 \\ \sum x^2 & \sum x^3 & \sum x^4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum \ln(\hat{y}) \\ \sum x \ln(\hat{y}) \\ \sum x^2 \ln(\hat{y}) \end{bmatrix} \tag{2}$$

After solving the linear system, we can finally retrieve μ, σ and A with $\mu = -b/2c$, $\sigma = \sqrt{-1/2c}$, and $A = e^{a-b^2/4c}$.

Guos Algorithm. Due to the logarithmic nature of Caruanas algorithm, it is very susceptible towards outliers. Guo [9] addresses these problems through a weighted least-squares regimen, by introducing an additive noise term η and reformulating the cost function to (Eq. 3).

$$\epsilon = y[\ln(y + \eta) - (a + bx + cx^2)] \approx y[\ln(y) - (a + bx + cx^2)] + \eta. \tag{3}$$

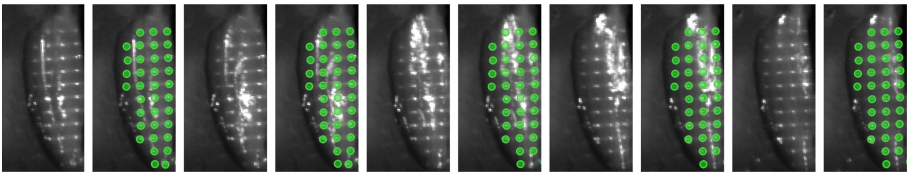

Fig. 4. In-vivo dynamics of a human vocal fold in laser-based structured light endoscopy. Due to the endoscopic light source and moist mucosal tissue, we see an abundance of specular reflections occluding the laserdots. In green: keypoint detection of our approach. It successfully detects laser dots in occluded regions. (Color figure online)

Similarly, by differentiating the sum of ϵ^2, we retrieve a linear system of the form given in Eq. 4.

$$\begin{bmatrix} \sum \hat{y}^2 & \sum x\hat{y}^2 & \sum x^2\hat{y}^2 \\ \sum x\hat{y}^2 & \sum x^2\hat{y}^2 & \sum x^3\hat{y}^2 \\ \sum x^2\hat{y}^2 & \sum x^3\hat{y}^2 & \sum x^4\hat{y}^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum \hat{y}^2 \ln(\hat{y}) \\ \sum x\hat{y}^2 \ln(\hat{y}) \\ \sum x^2\hat{y}^2 \ln(\hat{y}) \end{bmatrix} \tag{4}$$

The parameters of the Gaussian function μ, σ and A can be calculated similar to Caruanas algorithm. Recall that polynomial regression has an analytical solution, where the vector of estimated polynomial regression coefficients is $\beta = (\mathbf{X^T X})^{-1}\mathbf{X^T}\hat{y}$. This formulation is easily parallelizable on a GPU and fully differentiable. Hence, we can efficiently compute the Gaussian coefficients necessary for determining the subpixel position of local maxima. Finally, we can calculate the points position $\hat{\mathbf{p}}$ by simple addition using Eq. 5.

$$\hat{\mathbf{p}} = \boldsymbol{\mu} + \mathbf{p} \tag{5}$$

4 Evaluation

We implemented our code in Python (3.10.6), using the PyTorch (1.12.1) [15] and kornia (0.6.8) [6] libraries. We evaluate our code as well as the comparison methods on an Nvidia RTX 3080 GPU. We follow the respective methods closely for training. For data augmentation, we use vertical and horizontal flipping, affine and perspective transformations, as well as gamma correction and brightness modulation. We opted to use data augmentation strategies that mimic data that is likely to occur in laryngoscopy. We evaluate all approaches using a k-fold cross validation scheme with $k = 5$, on an expanded version of the publicly available Human-Laser Endoscopic (HLE) dataset [10] that also includes segmentations of the human vocal fold itself. We evaluate on subjects that were not contained in the training sets. The data generation scheme for the human vocal fold segmentation is described further below. For our baseline, we use dilation filtering and the moment method on images segmented using the ground-truth labels similar to [10,19,20]. We opt to use the ground-truth labels to show how

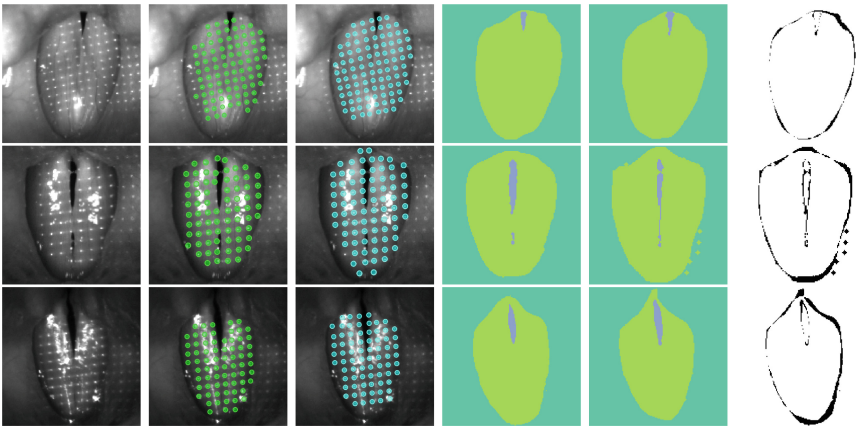

Fig. 5. Qualitative Assessment of the inferred point positions and segmentations. Left to right: Input, Predicted Keypoints, GT Keypoints, Predicted Segmentation, GT Segmentation, Pixelwise Error.

good these methods may become given perfect information. We train our approach via Stochastic Gradient Descent, with a learning rate of 10^{-1} for 100 epochs and update the learning rate via a polynomial learning rate scheduler similar to nnU-Net [11]. For estimating the keypoints we use a window size of 7, a Gaussian blur of size 5 and set the keypoint threshold to 0.7.

HLE++. The HLE dataset is a publicly available dataset consisting of 10 labeled in-vivo recordings of human vocal folds during phonation [10], where each recording contains one healthy subject. The labels include segmentations of the glottal gap, the glottal mid- and outline as well as the 2D image space positions of the laser dots projected onto the superior surface of the vocal folds. We expand the dataset to also include segmentation labels for the vocal folds itself. Due to the high framerate of the recordings, motion stemming from the manually recording physician is minimal and can be assumed to be linear. Thus, to generate the vocal fold segmentation masks, we generate a segmentation mask of the vocal fold region manually and calculate its centroid in the first and last frame of each video. Then, we linearly interpolate between the measured centroids. To account for the glottal gap inside each frame, we set $F_i = F_i \setminus G_i$, where F_i is the vocal fold segmentation at frame i and G_i the glottal segmentation, respectively.

Quantitative and Qualitative Evaluation. Table 1 shows a quantitative evaluation of different neural network architectures that have been used in laryngoscopy [8,18] or medical keypoint detection tasks [21]. In general, we could reformulate this segmentation task as a 3D segmentation task, in which we model the width and height of the images as first and second dimension, and lastly time as our third dimension. However, due to the poor optimization of 3D Convolutions on GPUs [13] and the large amounts of data generated in high-speed video

Table 1. Quantitative evaluation of the precision and F1-score of predicted keypoints, IoU and DICE score of inferred segmentations as well as the inference time for a single image and the frames per second on an Nvidia RTX 3080 GPU.

	Precision↑	F1-Score↑	IoU↑	DICE↑	Inf. Speed(ms)↓	FPS↑
Baseline	0.64	0.6923	✗	✗	✗	✗
U-LSTM [8]	0.70 ± 0.41	0.58 ± 0.32	0.52 ± 0.18	0.77 ± 0.08	65.57 ± 0.31	15
U-Net [18]	**0.92** ± 0.08	**0.88** ± 0.04	**0.68** ± 0.08	**0.88** ± 0.02	4.54 ± 0.03	220
Sharan [21]	0.17 ± 0.19	0.16 ± 0.17	✗	✗	5.97 ± 0.25	168
2.5D U-Net	0.90 ± 0.08	0.81 ± 0.05	0.65 ± 0.06	0.87 ± 0.02	**1.08** ± 0.01	**926**

recordings, this would create a serious bottleneck in real-time 3D pipelines. Thus, we also evaluate a 2.5D U-Net architecture, in which we employ channel-wise 3D convolutions inside the bottleneck as well as the output layers. This allows us to predict segmentations sequence-wise; drastically lowering inference times on a per frame basis while keeping the number of floating point operations minimal. Interestingly, this architecture achieves similar results to a standard U-Net architecture (see Table 1). However, frame jumps can be seen inbetween sequences. Since HLE was generated with a single recording unit, we assume that a properly trained 2D-CNN can infer occluded point positions based on the surrounding laser points as well as the observed topology of the vocal folds itself. However, we believe that it generalizes less well to arbitrary point patterns than a network architecture including temporal information. For the segmentation tasks, we measure the foreground IoU and DICE scores. For the keypoints, we evaluate the precision and the F1-score. We count a keypoint prediction as true positive, when its distance to its closest ground-truth point does not exceed 2 pixels. In Fig. 4 an example of a prediction over 5 frames is given, showing that neural networks can infer proper point positions even in case of occlusions. A further qualitative assessment of point predictions, vocal fold and glottal gap segmentations is given in Fig. 5.

5 Conclusion

We presented a method for the simultaneous segmentation and laser dot localization in structured light high-speed video laryngoscopy. The general idea is that we can dedicate one channel of the output of a general multiclass segmentation model to learn Gaussian heatmaps depicting the locations of multiple unlabeled keypoints. To robustly handle noise, we propose to use Gaussian regression on a per local-maxima basis that estimates sub-pixel accurate keypoints with negligible computational overhead. Our pipeline is very accurate and robust, and can give feedback about the success of a recording within a fraction of a second. Additionally, we extended the publicly available HLE Dataset to include segmentations of the human vocal fold itself. For future work, it would be beneficial to investigate how this method generalizes to arbitrary projection patterns and non-healthy subjects.

Acknowledgements. We thank **Dominik Penk** and **Bernhard Egger** for their valuable feedback.This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant STA662/6-1, Project-ID 448240908 and (partly) funded by the DFG - SFB 1483 - Project-ID 442419336, EmpkinS. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center of the Friedrich-Alexander-Universität Erlangen-Nürnberg.

References

1. Bulat, A., Sanchez, E., Tzimiropoulos, G.: Subpixel heatmap regression for facial landmark localization. In: 32nd British Machine Vision Conference 2021, BMVC 2021, vol. 2021(32), pp. 22–25 (2021). https://arxiv.org/abs/2111.02360
2. Caruana, R.A., Searle, R.B., Shupack, S.I.: Additional capabilities of a fast algorithm for the resolution of spectra. Anal. Chem. **60**(18), 1896–1900 (1988). https://doi.org/10.1021/ac00169a011
3. Cho, W.K., Choi, S.H.: Comparison of convolutional neural network models for determination of vocal fold normality in laryngoscopic images. J. Voice **36**(5), 590–598 (2022). https://doi.org/10.1016/j.jvoice.2020.08.003, https://www.sciencedirect.com/science/article/pii/S0892199720302927
4. Döllinger, M., et al.: Re-training of convolutional neural networks for glottis segmentation in endoscopic high-speed videos. Appl. Sci. **12**(19), 9791 (2022). https://doi.org/10.3390/app12199791, https://www.mdpi.com/2076-3417/12/19/9791
5. Duffner, S., Garcia, C.: A connexionist approach for robust and precise facial feature detection in complex scenes, pp. 316–321 (2005). https://doi.org/10.1109/ISPA.2005.195430
6. Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G.: Kornia: an open source differentiable computer vision library for pytorch. In: Winter Conference on Applications of Computer Vision (2020). https://arxiv.org/pdf/1910.02190.pdf
7. Earp, S.W.F., Samacoïts, A., Jain, S., Noinongyao, P., Boonpunmongkol, S.: Subpixel face landmarks using heatmaps and a bag of tricks. CoRR abs/2103.03059 (2021). https://arxiv.org/abs/2103.03059
8. Fehling, M.K., Grosch, F., Schuster, M.E., Schick, B., Lohscheller, J.: Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional LSTM network. PLOS ONE **15**, 1–29 (2020). https://doi.org/10.1371/journal.pone.0227791
9. Guo, H.: A simple algorithm for fitting a gaussian function [DSP tips and tricks]. IEEE Signal Process. Mag. **28**(5), 134–137 (2011). https://doi.org/10.1109/MSP.2011.941846
10. Henningson, J.O., Stamminger, M., Döllinger, M., Semmler, M.: Real-time 3D reconstruction of human vocal folds via high-speed laser-endoscopy. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. Lecture Notes in Computer Science, vol. 13437, pp. 3–12. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_1
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Meth. **18**(2), 203–211 (2021)

12. Jensen, H.W., Marschner, S.R., Levoy, M., Hanrahan, P.: A practical model for subsurface light transport. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 511–518. SIGGRAPH 2001, Association for Computing Machinery, New York, NY, USA (2001). https://doi.org/10.1145/383259.383319
13. Jiang, J., Huang, D., Du, J., Lu, Y., Liao, X.: Optimizing small channel 3D convolution on GPU with tensor core. Parallel Comput. **113**(C), 102954 (2022). https://doi.org/10.1016/j.parco.2022.102954
14. Luegmair, G., Mehta, D., Kobler, J., Döllinger, M.: Three-dimensional optical reconstruction of vocal fold kinematics using high-speed videomicroscopy with a laser projection system. IEEE Trans. Med. Imaging **34**, 2572–2582 (2015). https://doi.org/10.1109/TMI.2015.2445921
15. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035. Curran Associates, Inc. (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
16. Patel, R., Donohue, K., Lau, D., Unnikrishnan, H.: In vivo measurement of pediatric vocal fold motion using structured light laser projection. J. Voice: Off. J. Voice Found. **27**, 463–472 (2013). https://doi.org/10.1016/j.jvoice.2013.03.004
17. Pedersen, M., Larsen, C., Madsen, B., Eeg, M.: Localization and quantification of glottal gaps on deep learning segmentation of vocal folds. Sci. Rep. **13**, 878 (2023). https://doi.org/10.1038/s41598-023-27980-y
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28, http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a, (arXiv:1505.04597 [cs.CV])
19. Semmler, M., Kniesburges, S., Birk, V., Ziethe, A., Patel, R., Döllinger, M.: 3D reconstruction of human laryngeal dynamics based on endoscopic high-speed recordings. IEEE Trans. Med. Imaging **35**(7), 1615–1624 (2016). https://doi.org/10.1109/TMI.2016.2521419
20. Semmler, M., et al.: Endoscopic laser-based 3D imaging for functional voice diagnostics. Appl. Sci. **7**, 600 (2017). https://doi.org/10.3390/app7060600
21. Sharan, L., et al.: Point detection through multi-instance deep heatmap regression for sutures in endoscopy. Int. J. Comput. Assist. Radiol. Surg. **16**, 2107–2117 (2021). https://doi.org/10.1007/s11548-021-02523-w
22. Sun, P., Min, J.K., Xiong, G.: Globally tuned cascade pose regression via back propagation with application in 2D face pose estimation and heart segmentation in 3D CT images. ArXiv abs/1503.08843 (2015)
23. Ulku, I., Akagündüz, E.: A survey on deep learning-based architectures for semantic segmentation on 2D images. Appl. Artif. Intell. **36**(1), 2032924 (2022). https://doi.org/10.1080/08839514.2022.2032924
24. Ypma, T.J.: Historical development of the Newton-Raphson method. SIAM Rev. **37**(4), 531–551 (1995). http://www.jstor.org/stable/2132904
25. Yu, B., Tao, D.: Heatmap regression via randomized rounding. IEEE Trans. Pattern Anal. Mach. Intell. **44**(11), 8276–8289 (2021)
26. Zhang, J., Liu, M., Shen, D.: Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. IEEE Trans. Image Process. **26**(10), 4753–4764 (2017). https://doi.org/10.1109/TIP.2017.2721106