IJCARS manuscript No.
(will be inserted by the editor)

A Dataset of Laryngeal Endoscopic Images with Comparative Study on Convolution Neural Network Based Semantic Segmentation

Max-Heinrich Laves · Jens Bicker · Lüder A. Kahrs · Tobias Ortmaier

Received: date / Accepted: date

Abstract *Purpose* Automated segmentation of anatomical structures in medical image analysis is a prerequisite for autonomous diagnosis as well as various computer and robot aided interventions. Recent methods based on deep convolutional neural networks (CNN) have outperformed former heuristic methods. However, those methods were primarily evaluated on rigid, real-world environments. In this study, existing segmentation methods were evaluated for their use on a new dataset of transoral endoscopic exploration.
Methods Four machine learning based methods SegNet, UNet, ENet and ErfNet were trained with supervision on a novel 7-class dataset of the human larynx. The dataset contains 536 manually segmented images from two patients during laser incisions. The Intersection-over-Union (IoU) evaluation metric was used to measure the accuracy of each method. Data augmentation and network ensembling were employed to increase segmentation accuracy. Stochastic inference was used to show uncertainties of the individual models. Patient-to-patient transfer was investigated using patient-specific fine-tuning.
Results In this study, a weighted average ensemble network of UNet and ErfNet was best suited for the segmentation of laryngeal soft tissue with a mean IoU of 84.7 %. The highest efficiency was achieved by ENet with a mean inference time of 9.22 ms per image. It is shown that 10 additional images from a new patient are sufficient for patient-specific fine-tuning.
Conclusion CNN-based methods for semantic segmentation are applicable to endoscopic images of laryngeal soft tissue. The segmentation can be used for active constraints or to monitor morphological changes and autonomously detect pathologies. Further improvements could be achieved by using a larger dataset or training the models in a self-supervised manner on additional unlabeled data.

Max-Heinrich Laves · Jens Bicker · Lüder A. Kahrs · Tobias Ortmaier
Appelstraße 11A
30167 Hannover, Germany
Tel.: +49 511 762 19617
Fax: +49 511 762 19976
E-mail: laves@imes.uni-hannover.de

Keywords Computer Vision · Larynx · Vocal folds · Soft tissue · Open access dataset · Machine learning · Patient-to-patient fine-tuning

1 Introduction

It is anticipated that the examination of laryngeal endoscopic images will allow early detection of pathologies [5]. Vocal folds, as the main functional organ within the larynx, are sensitive structures for surgery. Computer vision has the potential to assist the physician in restoring or preserving the voice. This can be achieved by combining augmented reality [31], robotics [10] and laser surgery [30] with image processing methods. Segmentation of laryngeal images is one of the most important components for successful application of such a system.

1.1 Outline

This paper discusses the automated segmentation of laryngeal soft tissue. Therefore, we investigate state-of-the-art segmentation methods, all based on deep convolutional neural networks, and compare them on a novel manually annotated in vivo dataset of human vocal folds. Subsequent to the description of the implementation differences of the selected models, the test setup is defined and results are reported. The article concludes with a discussion of limitations of the chosen approaches and gives an outlook on possible improvements.

1.2 Related Work

Image segmentation is the task of partitioning an image into several non-intersecting *coherent* parts and is an important step of early vision [22]. The ultimate goal in medical image segmentation is recognizing real-world objects in image data fully automatically, with high accuracy and efficiency. However, early methods were not capable of doing this and demanded a human supervisor to initialize a method, check the result, or even to correct the segmentation afterwards [20]. These techniques are based on manually selected low-level image characteristics such as grayscale thresholds, pixel color, edge detection or region growing [9, 21, 26] and are therefore strongly affected by image noise or illumination changes [22]. Additionally, these procedures do not perform a semantic assumption of the segmented areas.

More sophisticated methods are based on mathematical models and rely on finding the parameters with which the output of the model minimizes an error or energy function. Some of these approaches require ground truth examples prior to or during the segmentation task. In atlas-based segmentation, for example, a manually segmented "atlas", which acts as a-priori anatomical information, is matched onto the input image [6]. This turns the segmentation problem to a registration problem. The error-minimizing transformation between the images is also applied to the a-priori segmentation, which results in the segmentation of the input image. A good result requires a large atlas database, which must be

considered during segmentation time, thus making atlas-based segmentation not applicable in environments with real-time demands.

Artificial neural networks (ANN) have been used in medical image segmentation for a long time and are characterized by their robustness to image noise and real-time capable output due to their massively parallel structure [6, 19, 21, 22, 27]. The performance of early ANNs with low numbers of neural layers were not superior compared to other methods. However, due to the recent progress in the field of convolutional neural networks, new segmentation methods with outstanding performance have been created. On image classification tasks, CNN-based approaches already exceed human-level performance [14]. Long et al. first proposed a fully convolutional network (FCN) which was trained pixel-to-pixel and exceeded the former state-of-the-art by up to 20 % on the PASCAL VOC 2012 dataset [18]. This led to the emergence of many new CNN-based segmentation methods. In the following, we will focus on potential real-time capable network architectures, as the segmentation result will be used later for intra-operative robot and laser control. Therefore, we will subsequently discuss four selected methods, which already showed promising results on other datasets.

One focus of this research includes the automatic detection of pathological and non-pathological areas. In laryngeal scenes, segmentation on high-speed or stroboscopic videos is done to automatically extract the contours of the vocal folds to analyze the fold vibrations [1] and the glottal space to characterize glottal closure and other vocal fold pathologies [21]. Researchers have developed algorithms for the classification of laryngeal tumors based on narrow band imaging (NBI) and surrounding blood vessel structures [4]. Others have used classifications based on a combination of the vocal fold shape and vascular pattern [34]. It has also been attempted to correlate voice pathologies with endoscopic videos of the vocal fold [23]. Furthermore, high-speed videos are analyzed with wavelet-based phonovibrograms and it is shown that a distinction between malignant and precancerous vocal fold lesions is possible [35]. Recently, automatic detection of vocal fold tumor tissue has been performed with confocal laser endomicroscopy and convolutional neural networks [2]. Nevertheless, we see a lack of publicly available datasets of laryngeal (vocal fold) micro- and endoscopic images associated with comparative segmentation results.

2 Materials and Methods

First, our dataset of segmented vocal fold images will be presented. Second, the machine learning methods used in this study are briefly described and compared. At the end of this section, the evaluation environments are defined.

2.1 The Vocal Folds Dataset

All subsequently described models are trained and evaluated on a dataset, containing 536 manually segmented in vivo color images of the larynx during two different resection surgeries with a resolution of 512×512 pixels. The images have been captured with a stereo endoscope (VSii, Visionsense, Petach-Tikva, Israel) in an in vivo laryngeal surgery and have been used in prior studies [32]. They are

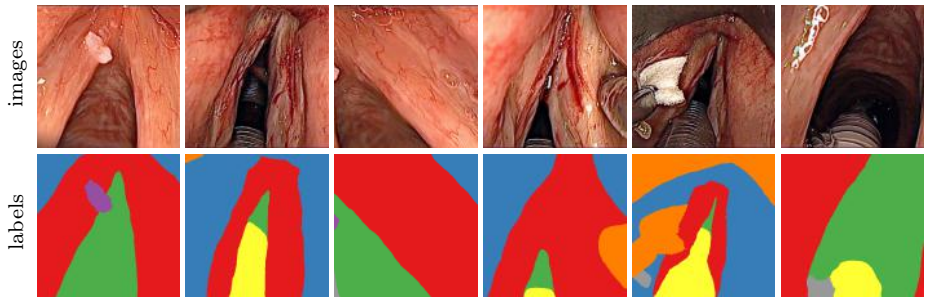

Fig. 1: First row: Examples from vocal folds dataset. Second row: Manually segmented ground truth label maps with classes *vocal folds* (red), *other tissue* (blue), *glottal space* (green), *pathology* (purple), *surgical tool* (orange), *intubation* (yellow) and *void* (gray). The grayscale label maps have been colorized for better visualization.

categorized in the 7 different classes *void*, *vocal folds*, *other tissue*, *glottal space*, *pathology*, *surgical tool* and *intubation* with indices {0, 1, 2, 3, 4, 5, 6}, respectively, which is represented by the gray values of the label maps (see Fig 1). The dataset consists of 5 different sequences from two patients (named SEQ1–4 from patient 1 and SEQ5–8 from patient 2). The sequences have following characteristics:

- SEQ1: pre-operative with clearly visible tumor on vocal fold, changes in translation, rotation, scale, no instruments visible, without intubation
- SEQ2: pre-operative with clearly visible tumor, visible instruments, changes in translation and scale, with intubation
- SEQ3–4: post-operative with removed tumor, damaged tissue, changes in translation and scale, with intubation
- SEQ5–7: pre-operative with instruments manipulating and grasping the vocal folds, changes in translation and scale, with intubation
- SEQ8: post-operative with blood on vocal folds, instruments and surgical dressing, with intubation

Subsequent images have a temporal contiguity as they are sampled uniformly from videos. To reduce inter-frame correlation, images were extracted from the original videos only once per second. In the comparative study SEQ4–SEQ6 were not used due to high similarity to SEQ3 and SEQ7 respectively, as they do not offer any additional variance to the dataset. Segmentations have been manually created on a pen display (DTK-2241, K. K. Wacom). Figure 2 shows the distribution of the annotated pixels per class. The dataset is publicly available[1] and will be extended in the future.

2.2 Network Architectures

UNet is a fully convolutional U-shaped network architecture for biomedical image segmentation [29]. The first part is inspired by FCN [18] and acts as an encoder.

[1] https://github.com/imesluh/vocalfolds

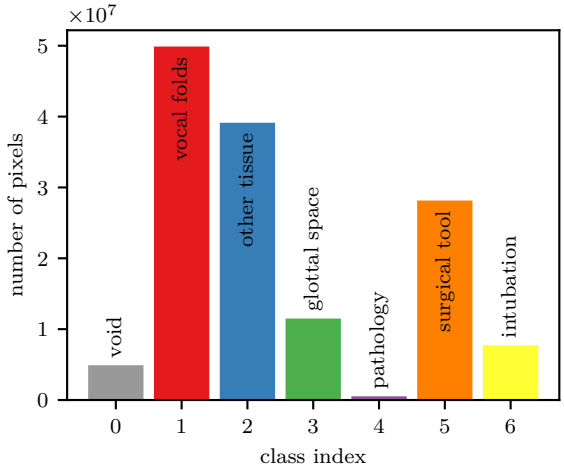

Fig. 2: Number of annotated pixels per class in the dataset.

It consists of repeated 3×3 convolutions. The feature maps are saved for later use before downsampling with 2×2 max-pooling. Dropout is applied at the end of the encoder. To generate a dense segmentation output, unlike simple upsampling in FCN, the encoder is inverted, and pooling is replaced by transposed convolution (up-conv). This acts as a decoder and creates a nearly symmetrical encoder-decoder network. In addition, the stored encoder feature maps are concatenated after upsampling and fed into the convolution of the decoder. UNet has a large number of feature maps and therefore the largest number of learnable parameters of the networks investigated in this study (see Tab. 2). Unlike the original authors, padding of 1 and a final upsampling layer is added to the encoder to obtain matching input and output dimensions.

SegNet is a deep, fully convolutional neural network with a symmetrical encoder and decoder architecture [3]. Each encoder layer in the network performs 3×3 convolutions to create a set of feature maps. After batch normalization, the feature maps are max-pooled to downsample. The indices of the max-pooling layer are saved for later upsampling. On the other side, the corresponding decoder first upsamples its input by using the saved max-pooling indices from encoding. In contrast to UNet, transferring only max-pooling indices instead of full feature maps reduces memory consumption. For later stochastic inference, dropout layers with $p_i = 0.5$ after every pooling/upsampling layer in encoder and decoder stages are added. The resulting architecture is also referred to as Bayesian SegNet [15].

ENet is an architecture with a reduced number of parameters, especially created for low latency tasks in mobile or embedded devices [24]. In contrast to SegNet, it has an asymmetric structure with a large encoder and a small decoder. The network consists of a repeating basic module, which first splits its input into two branches. The main branch performs three convolutions with batch normalization and a final spatial dropout with $p_i = 0.01$ in the first stage and $p_i = 0.1$ afterwards. The outer 1×1 convolutions reduce and increase the number of feature maps respectively, forming a "bottleneck" around the inner convolution. The side branch just copies or copies and downsamples (max-pools) the input, which acts

as a shortcut to the main branch. As in SegNet, the max-pooling indices are saved for later upsampling by max-unpooling. In the end, both branches are reunited by element-wise addition.ENet has the lowest number of trainable parameters in this study (see Tab. 2).

ErfNet stands for "Efficient Residual Factorized Network" and tries to provide a compromise between accuracy and efficiency [28]. It is composed of layers, which are similar to the aforementioned bottleneck modules, but instead the 2D convolutions are factorised into 1D convolutions. The resulting layers are called non-bottleneck-1D and drastically reduce computational cost and the number of parameters. The network architecture forms an encoder-decoder structure similar to ENet. The downsampling modules are the same as the initial module of ENet. Dropout is included in all non-bottleneck-1D layers with $p_i = 0.3$ in the last stage and $p_i = 0.03$ before of that. Instead of max-unpooling, the upsampling block contains transposed convolution.

2.3 Evaluation Setup

In order to evaluate the previously described segmentation methods, performance were assessed on our in vivo vocal fold dataset. At first, all models were implemented in Python using the PyTorch [25] library with CUDA backend. In order to train the models, the dataset was split as follows. For reducing inter-frame correlation in the subsets, the chronologically first 50 % of SEQ1–3 and SEQ7–8 were used as training set (200 images), the subsequent 25 % as validation set (100 images) and the last 25 % as test set (100 images). This seperates training and test data over time as much as possible. The models were trained on the training set and inter-training accuracy was evaluated on the validation set. The test set was left out for final performance assessment. An additional training setup using only data from patient 1 is described in Section 2.6. As accuracy metric, the popularly accepted Intersection-over-Union (IoU) metric is used:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

with the number of true positive (TP), false positive (FP) and false negative (FN) pixels. The IoU metric was calculated for each of the 7 classes independently.

All models were trained in an end-to-end manner using the Adam [16] stochastic gradient descent optimizer with an initial learning rate of $\ell = 1 \cdot 10^{-3}$, and a batch size of 6. The learning rate was reduced by a factor of 10^{-1} when observing saturation of the validation error (reduce-on-plateau). According to [16], the decay rates of the first and second order moments were set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The training objective is to minimize a loss function, which acts as a dissimilarity measure between a prediction $\hat{\mathbf{y}}$ of the model for input image $\mathbf{x}$ and the corresponding ground truth label map $\mathbf{y}$. The prediction $\hat{\mathbf{y}}$ is a tensor and expected to have the shape $\mathbb{R}^{C \times H \times W}$ with the height H, width W, and number of classes C of the input image. For every pixel of the input image, a probability distribution containing the estimated class probabilities is received. In order to match the output dimensions of the models, one-hot encoding scheme was used, in which a ground truth label map $\mathbf{y}$ is reshaped to $\mathbb{R}^{C \times H \times W}$. After that, the ground truth distribution for each pixel contained only one entry with 1 and otherwise

0, assigning a unique class to each pixel. For later evaluation on the test set, the weight configuration with the lowest loss value on the validation set was chosen (early stopping) after training on a single GPU (GeForce GTX 1080 Ti, Nvidia Corp., Santa Clara, CA, USA).

As loss function, a weighted negative log-likelihood (or cross entropy) function was chosen, which is defined for a distribution $\mathbf{p} = \hat{\mathbf{y}}_{h,w}$ of a single pixel of true class c as

$$l(\mathbf{p}, c) = -\mathbf{w}_c \log\left(\frac{\exp(\mathbf{p}_c)}{\sum_{j=0}^{C-1} \exp\left(\mathbf{p}_j\right)}\right) \tag{2}$$

with weight vector

$$\mathbf{w}_c = \frac{\sum_{j=0}^{C-1} N(j)}{C \cdot N(c)} \in \mathbb{R}^C \ , \tag{3}$$

where $N(c)$ is the total number of occurrences of pixels with class c in the whole dataset. Vector $\mathbf{w}$ contains a weight for each class. This is done to give less occurring classes, such as "pathology", more weight when calculating the loss. Otherwise, omitting classes with small areas would result in a relatively small overall error. For one prediction $\hat{\mathbf{y}}$ the total loss becomes

$$L(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} l\left(\hat{\mathbf{y}}_{h,w}, \mathbf{y}_{h,w}\right) \ . \tag{4}$$

2.4 Transfer Learning, Data Augmentation and Network Ensembling

Additionally, all networks were pre-trained on the publicly available Cityscapes segmentation dataset [7]. The Cityscapes-fine subdataset employed contains 5,000 urban steet scenes with fine annotations for 30 classes. After pre-training, the final layer of each network was replaced by a new layer for taking the different number of classes into account and a full fine-tuning on our dataset was performed.

Data augmentation was used to increase the size of the training set to 2,000 images by horizontal flipping and rotating between -10° and +10°. Vertical flipping and further rotations were avoided as this is unlikely in a real enviroment. All networks were trained with and without data augmentation.

To further improve segmentation accuracy, the trained networks were combined as an anverage ensemble with trainable weights for each class and network. Ensemble averaging was reported to significantly improve prediction accuracy and and better generalization [13]. Our combiner

$$\hat{\mathbf{y}}\left(\mathbf{x}\right) = \sum_{j=1}^{p} \boldsymbol{\alpha}_j \circ \hat{\mathbf{y}}_j\left(\mathbf{x}\right) \tag{5}$$

producing the final output of the ensemble network is defined by the sum of the weighted network outputs, where $\circ$ denotes the entrywise Hadamard product. The weights $\boldsymbol{\alpha}_j$ were optimized by fixing the weights of the individual models and training the ensemble combiner on the whole training set again. Pairwise ensembling of the networks ($p = 2$) and an ensemble of all networks ($p = 4$) were investigated, which resulted in 7 different ensemble combinations. One may think

that pairwise ensembling cannot be superior to an ensemble of all networks since the combiner is trained to find the optimal combination. However, Zhou et al. [37] have shown that ensembling *many* instead of *all* available models can provide better results.

2.5 Model Uncertainty

In safety-critical tasks, such as medical imaging, it is important to know the confidence of the prediction of a model, especially when using the information for (semi-)autonomous robot or laser control. According to [11], the class probabilities produced by a softmax function approximate relative probabilities between class labels and give no overall measure of the model's uncertainty. Therefore, prediction uncertainty was evaluated by Monte Carlo sampling with dropout at test time, following the framework of [15] (also called *stochastic inference*). This gives an approximation of the distribution of the softmax class probabilities for every pixel of the model prediction. The *variance* $\sigma^2_{c,h,w}$ of each softmax probability was used as a per-class uncertainty and the *mean of the variances* $\bar{\sigma}^2_{h,w}$ with respect to c as an overall uncertainty. During Monte Carlo sampling the dropout probabilities were set to be the same as during training.

2.6 Patient-to-Patient Generalization

In order to assess how well patient-independent generalization is possible, a training on the sequences of patient 1 solely and subsequent testing on patient 2 were performed. Since *pathology* does not occur in patient 2, it was neglected for this.

To additionally follow the idea of patient-specific fine-tuning [36], a minimal number of images from the beginning of the sequences of patient 2 were added to the training set. The models were fine-tuned on this extended dataset and tested on the latter half of sequences of patient 2. The key idea behind this is, that with a low number of pre-operatively acquired and manually segmented images of new patients, the intra-operative segmentation accuracy can be improved patient-specific. It has been shown that even sparsely scribbled annotations can significantly improve segmentation accuracy in this case [36]. To identify the manifold of additional images needed for patient-to-patient translation, the number of images added to the training set is increased by steps of 5.

3 Results

In the following, the results of the different segmentation methods from the aforementioned evaluation setup are presented. Table 1 summarize the results by showing the per-class and mean segmentation accuracy on the test set measured with the Intersection-over-Union metric.

When trained from scratch (without data augmentation and pre-training), the highest mean IoU value was achieved by UNet, which also has the highest per-class IoU for *pathology* and *intubation* (see Tab. 1 (a)). After full fine-tuning, UNet and ErfNet achieved better results, with ErfNet now performing best (see

Tab. 1 (b)). However, SegNet and ENet dit not benefit from pre-training, which resulted in lower per-class mean results, except for *glottal space*. All networks showed improved results by dataset augmentation (see Tab. 1 (c)). In this case, ErfNet again obtained best mean IoU. SegNet provided the worst mean results on all test scenarios.

Although it has been shown that in medical image analysis, fine-tuning of a pre-trained CNN can outperform training from scratch [33], varying findings were observed. ErfNet and UNet benefited from pre-training by increaded mean IoU by approx. 5.1 % and 2.5 %, respectively. However, this was not the case for ENet and SegNet, where a decrease in mean IoU was observed by approx. 5.3 % and 13.2 %, respectively.

In general, the class *pathology* appears to be worst recognized. This can be explained by the fact that this class has a significantly lower occurrence in the whole dataset (see Fig. 2), even lower than *void*, which has not been considered for accuracy assessment. However, good results were achieved for classes *vocal folds*, *other tissue*, *surgical tool* and *intubation*.

Table 1 (d) shows, that network-and-class-wise average ensembling clearly improved segmentation results. The networks were ensembled after individual training on the augmented dataset. Except for ENet+SegNet configuration, all ensembles performed better in comparison to their individual models.

Figure 4 visualizes class label predictions for selected example images from the test set after training on the augmented dataset. As expected from the IoU values, the prediction of SegNet has major errors in classes *pathology*, *vocal folds* and *other tissue*. The results of UNet are better in general, but it provided visible errors on the edges of two adjacent areas. ENet and ErfNet both achieved good results, with ErfNet also resolving edges and smaller areas well. When zooming into the label maps, it is visible that ENet has a strong aliasing effect on the edges, whereas ErfNet results in smoother edges. The ensemble results are reported for ErfNet+UNet configuration, which provided overall best results. Please see the supplemental video material associated with this publication.

The prediction uncertainties in Fig. 4 give an estimate on how confident the model is for a specific pixel of the selected images. UNet and SegNet use normal dropout for stochastic inference, whereas ENet and ErfNet use spatial dropout (turning off full feature maps). In contrast to normal dropout, there is currently no proof of spatial dropout acting as Bayesian approximation [11]. This must be taken into account when considering the uncertainty maps. It is worth mentioning that ENet and ErfNet have a higher uncertainty in general and especially on *other tissue*. A possible explanation for this could be the low numbers of parameters having less redundancies compared to UNet and SegNet (see Tab. 2). Furthermore, it can be observed that with the exception of ErfNet, all models show a high uncertainty for *pathology*. On the other hand, all models show low uncertainties for *vocal folds*, which can be interpreted as good generalization for this class.

Figure 5 illustrates the recorded mean loss during training on both the training and validation sets. All models converged in our training setup. An increase in validation loss can be observed as soon as the training loss gets close to $\bar{L} = 0$, which indicates overfitting. Early stopping was used to prevent this. The corresponding epochs are marked with an arrow in Fig 5.

Figure 3 shows the patient-to-patient generalization of the different networks and the effect of patient-specific fine-tuning. All networks benefit greatly from

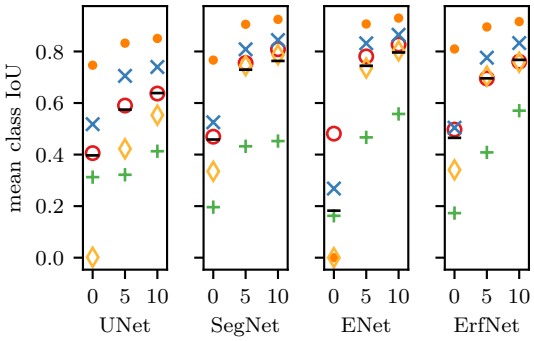

Fig. 3: Results of patient-to-patient transfer and patient-specific fine-tuning. The x-axis indicates number of images used from patient 2. The class IoU is denoted by symbols for *vocal folds* (○), other tissue (×), *glottal space* (+), *surgical tool* (●), *intubation* (◇) and the mean IoU is denoted by black bar (−).

additional data from patient 2, resulting in an average IoU increase of 31.0 %-points with 5 additional images and another 5.6 %-points with 10 additional images. The latter results are already comparable to those in Tab. 1 and therefore no further images were used for patient-specific fine-tuning.

Table 2 shows the mean inference times including memory transfer to the GPU. All models are able to process the images efficiently with at least 20 frames per second (fps). ENet and ErfNet, however, perform significantly better, with ENet achieving the best performance (108.5 fps).

4 Conclusion and outlook

In this paper, a novel dataset of laryngeal surgery images with ground truth segmentation maps has been introduced. The dataset was used for a comparative study of recent CNN-based segmentation methods. It has been shown that an ensemble of ErfNet and UNet yielded the most successful results on segmenting endoscopic image data. Schoob et al. [30] reported that image-based online control of incision lasers for soft tissue undergoing motion is feasible with an imaging pipeline running at 73.5 Hz. Therefore, both ENet and ErfNet are well-suited in terms of efficiency for later use in autonomous or robot aided interventions. The reported laser incision error of their approach was 0.12–0.21 mm. With monoscopic images, we cannot directly compare our segmentation accuracy to this and will address this in future work. However, with sufficient safety margins, segmentation can be used to define active constrains.

The raw data, from which our dataset images are chosen, consist of stereo images. Currently, the dataset only includes the left images. By segmenting both the left and the right stereo images, a metric accuracy value for the segmentation task can be stated, if calibration data is available.

When comparing UNet and SegNet to ENet and ErfNet, two main characteristics differ. The first difference is that the first two have a symmetrical auto-encoder structure, while the latter two have a significantly smaller decoder. It is assumed

Network	Label Index						mean
	1	2	3	4	5	6	
UNet	76.3	74.0	60.0	**64.5**	87.3	**79.3**	**73.6**
SegNet	68.8	69.7	46.2	63.2	86.1	69.5	67.3
ENet	**79.6**	**79.8**	67.4	51.3	86.8	69.5	72.4
ErfNet	75.6	76.8	**70.1**	49.9	**90.9**	77.4	73.5
mean	75.1	75.1	60.9	57.2	87.8	73.9	

(a) training from scratch

Network	Label Index						mean
	1	2	3	4	5	6	
UNet	72.6	71.2	72.4	71.9	88.8	**79.4**	76.1
SegNet	65.3	66.7	42.7	20.8	76.5	52.5	54.1
ENet	74.5	75.0	66.0	32.0	88.9	66.2	67.1
ErfNet	**78.5**	**78.9**	**73.4**	**74.5**	**90.0**	76.2	**78.6**
mean	72.7	73.0	63.6	49.8	86.1	68.6	

(b) pre-training and fine-tuning

Network	Label Index						mean
	1	2	3	4	5	6	
UNet	76.9	75.5	**71.9**	64.7	90.1	**81.6**	76.8
SegNet	73.6	72.6	68.2	58.3	87.1	74.2	72.3
ENet	81.2	80.5	65.7	75.7	88.6	78.7	78.4
ErfNet	**85.4**	**86.0**	70.5	**75.9**	**90.5**	81.2	**81.6**
mean	79.3	78.7	69.1	68.7	89.1	78.9	

(c) training on augmented dataset

Ensemble	Label Index						mean
	1	2	3	4	5	6	
ENet+ErfNet	**85.7**	**85.8**	72.7	83.3	90.9	82.3	83.5
ENet+SegNet	79.2	78.1	70.3	69.2	89.2	80.5	77.8
ENet+UNet	82.3	81.7	73.8	84.3	91.7	84.2	83.0
ErfNet+SegNet	83.1	82.9	73.5	77.9	90.6	82.6	81.8
ErfNet+UNet	84.9	84.9	**77.1**	84.4	**92.5**	**84.4**	**84.7**
UNet+SegNet	78.5	77.0	75.9	69.2	91.0	81.2	78.8
All	84.9	84.1	76.0	**86.6**	92.1	84.3	**84.7**
mean	82.7	82.1	74.2	79.3	91.1	82.8	

(d) ensemble configurations

Table 1: Per-class and mean IoU (%) on the test set for different training scenarios. Bold numbers denote best results.

Network	trainable parameter	512×512	
		ms	fps
UNet	31,032,200	47.2	21.18
SegNet	29,447,624	45.7	21.9
ENet	392,420	**9.22**	**108.5**
ErfNet	2,064,508	11.1	90.1

Table 2: Total number of trainable parameters and mean inference times on a single GeForce GTX 1080 Ti (including data to GPU transfer).

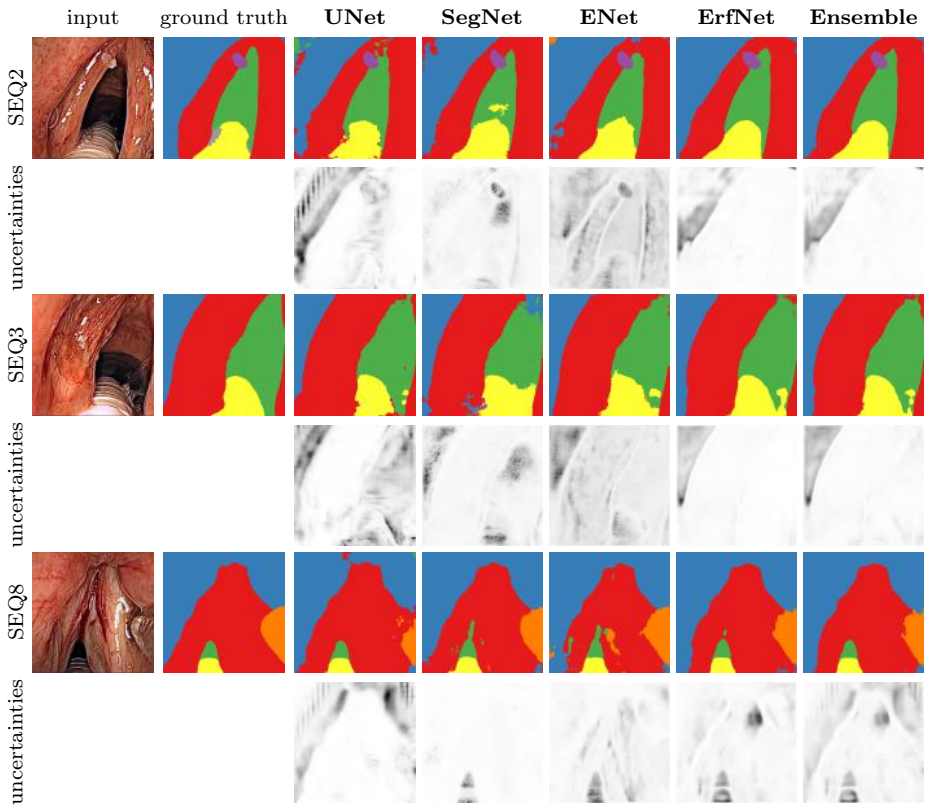

Fig. 4: Qualitative results of the different networks on exemplary images of the test set sequences and corresponding prediction uncertainties of all classes (gray level denotes variance with white where $\bar{\sigma}^2_{h,w} = 0$). The results were generated by selecting the training state that provided best results on the validation set.

that a higher performance can be attributed to the encoder, while the decoder only upsamples the output of the encoder [24]. The second difference is long-term connections between encoder and decoder layers, bypassing the deeper structures, by copying whole feature maps in UNet or by memorizing the pooling indices in SegNet. However, ENet and ErfNet do not have such connections. The residual units bypass only one layer.

With pre-training, the performance of ErfNet and UNet were improved on our dataset. However, not all networks benefited from pre-training. A possible explanation for this can be the different image domains, as the Cityscapes dataset contains segmented street scenes. This underlines the need for publicly available medical segmentation datasets. In addition, an extension of the dataset to different pathologies can enable automated diagnosis. Further improvement regarding our dataset and the generation of ground truth information for medical imaging is required.

The disadvantage of our dataset up to now is the low number of patients and the unbalanced class occurrences. The most important class *pathology* is the least

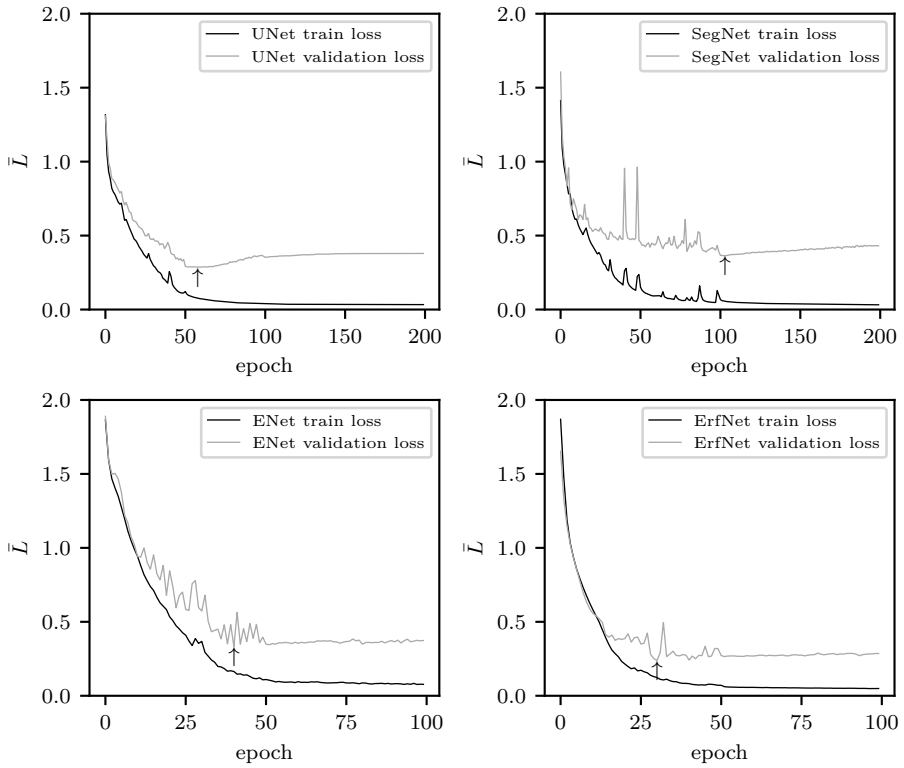

Fig. 5: Mean losses during training of individual networks on augmented dataset. The arrows denote selected epochs for early stopping.

common. Besides using a class-weighted cross entropy, specific loss functions for extreme class imbalances can be used, such as the focal loss [17]. Ideally, performance of networks in the medical imaging domain should be evaluated on data from patients not included in the training set. Therefore, results from Tab. 1 can be considered for comparing segmentation performances but do not show the ability of inter-patient generalization for real clinical use. Results from Fig. 3 indicate, that training with data from one patient is not sufficient for a well-generalized segmentation model in a laryngeal environment. However, a patient-specific fine-tuning with only 10 additional images from a new patient already seem to be sufficient for good segmentation results. These images could be taken during a pre-operative examination and used for fine-tuning after sparse manual annotation.

Ongoing work will therefore focus on enlarging the dataset by investigating self-supervised segmentation methods, where a human expert accepts, discards or corrects the segmentation prediction. Beyond that, improvements can be achieved by modelling the uncertainty, especially for small datasets like ours [15], by combining CNN-based segmentation with tracking [12] or by using additional unlabeled data [8].

Acknowledgements We thank Giorgio Peretti from the Ospedale Policlinico San Martino, University of Genova, Italy, for providing us with the in vivo laryngeal data used in this study. We would also like to thank James Napier from the Institute of Lasers and Optics, University of Applied Sciences Emden-Leer, Germany, for his thorough proofreading of this manuscript.

Disclosure of potential conflicts of Interest

Conflict of Interest The authors declare that they have no conflict of interest.

Funding This research has received funding from the European Union as being part of the ERFE OPhonLas project.

Formal Consent The endoscopic video images were acquired by Prof. Giorgio Peretti (Director of Otorhinolaryngology at Ospedale Policlinico San Martino, University of Genova). Patients gave their written consent for the procedure and the use of the data. No further approval is necessary for such endoscopic recordings. The videos were anonymized made available inside the µRALP consortium for further usage. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

1. Allin, S., Galeotti, J., Stetten, G., Dailey, S.H.: Enhanced snake based segmentation of vocal folds. In: IEEE International Symposium on Biomedical Imaging: Nano to Macro, vol. 1, pp. 812–815 (2004). DOI 10.1109/ISBI.2004.1398662
2. Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., Bohr, C., Neumann, H., Stelzle, F., Maier, A.: Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning. Scientific Reports **7**(1) (2017). DOI 10.1038/s41598-017-12320-8
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017). DOI 10.1109/TPAMI.2016.2644615
4. Barbalata, C., Mattos, L.S.: Laryngeal Tumor Detection and Classification in Endoscopic Video. IEEE Journal of Biomedical and Health Informatics **20**(1), 322–332 (2016). DOI 10.1109/JBHI.2014.2374975
5. Barkmeier-Kraemer, J.M., Patel, R.R.: The next 10 years in voice evaluation and treatment. Semin Speech Lang **37**(03), 158–165 (2016). DOI 10.1055/s-0036-1583547
6. Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., Cuadra, M.B.: A review of atlas-based segmentation for magnetic resonance brain images. Computer Methods and Programs in Biomedicine **104**(3), e158–e177 (2011). DOI 10.1016/j.cmpb.2011.07.015
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016). DOI 10.1109/CVPR.2016.350
8. Creswell, A., Pouplin, A., Bharath, A.A.: Denoising adversarial autoencoders: classifying skin lesions using limited labelled training data. IET Computer Vision **12**(8), 1105–1111 (2018). DOI 10.1049/iet-cvi.2018.5243
9. Doignon, C., Graebling, P., de Mathelin, M.: Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature. Real-Time Imaging **11**(5), 429–442 (2005). DOI 10.1016/j.rti.2005.06.008

10. Friedrich, D.T., Scheithauer, M.O., Greve, J., Duvvuri, U., Sommer, F., Hoffmann, T.K., Schuler, P.J.: Potential Advantages of a Single-Port, Operator-Controlled Flexible Endoscope System for Transoral Surgery of the Larynx. Annals of Otology, Rhinology & Laryngology **124**(8), 655–662 (2015). DOI 10.1177/0003489415575548
11. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Proc. 33rd Int. Conf. Mach. Learn., vol. 48, pp. 1050–1059 (2016)
12. García-Peraza-Herrera, L.C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Poorten, E.V., Stoyanov, D., Vercauteren, T., Ourselin, S.: Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. Lect. Notes Comput. Sci. **10170 LNCS**, 84–95 (2017). DOI 10.1007/978-3-319-54057-3_8
13. Hashem, S.: Optimal Linear Combinations of Neural Networks. Neural Networks **10**(4), 599–614 (1997). DOI 10.1016/S0893-6080(96)00098-6
14. He, K., Zhang, X., Ren, S., Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: IEEE Int. Conf. Comput. Vis., pp. 1026–1034 (2015). DOI 10.1109/ICCV.2015.123
15. Kendall, A., Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: Advances in Neural Information Processing Systems 30, pp. 5574–5584 (2017)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv e-prints (2014). URL `https://arxiv.org/abs/1412.6980`
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. arXiv e-prints arXiv:1708.02002 (2017)
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3431–3440 (2015). DOI 10.1109/CVPR.2015.7298965
19. Noble, J.A., Boukerroui, D.: Ultrasound image segmentation: a survey. IEEE Transactions on Medical Imaging **25**(8), 987–1010 (2006). DOI 10.1109/TMI.2006.877092
20. Olabarriaga, S.D., Smeulders, A.W.M.: Interaction in the segmentation of medical images: A survey. Medical Image Analysis **5**, 127–142 (2001). DOI 10.1016/S1361-8415(00)00041-4
21. Osma-Ruiz, V., Godino-Llorente, J.I., Sáenz-Lechón, N., Fraile, R.: Segmentation of the glottal space from laryngeal images using the watershed transform. Computerized Medical Imaging and Graphics **32**(3), 193–201 (2008). DOI j.compmedimag.2007.12.003
22. Pal, N.R., Pal, S.K.: A review on image segmentation techniques. Pattern Recognition **26**(9), 1277–1294 (1993). DOI 10.1016/0031-3203(93)90135-J
23. Panek, D., Skalski, A., Zielinski, T., Deliyski, D.D.: Voice pathology classification based on High-Speed Videoendoscopy. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 735–738 (2015). DOI 10.1109/EMBC.2015.7318467
24. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. ArXiv e-prints (2016). URL `http://arxiv.org/abs/1606.02147`
25. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: 31st Conference on Neural Information Processing Systems (NIPS) (2017). URL `https://openreview.net/forum?id=BJJsrmfCZ`
26. Phung, S.L., Bouzerdoum, A., Chai, D.: Skin segmentation using color pixel classification: analysis and comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(1), 148–154 (2005). DOI 10.1109/TPAMI.2005.17
27. Rajab, M., Woolfson, M., Morgan, S.: Application of region-based segmentation and neural network edge detection to skin lesions. Computerized Medical Imaging and Graphics **28**(1), 61–68 (2004). DOI 10.1016/S0895-6111(03)00054-5
28. Romera, E., Álvarez, J.M., Bergasa, L.M., Arroyo, R.: ERFNet : Efficient Residual Factorized ConvNet for Real-time Semantic Segmentation. IEEE Trans. Intellent Transp. Syst. **19**(1), 263–272 (2018). DOI 10.1109/TITS.2017.2750080
29. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241 (2015)
30. Schoob, A., Kundrat, D., Kahrs, L.A., Ortmaier, T.: Stereo vision-based tracking of soft tissue motion with application to online ablation control in laser microsurgery. Medical Image Analysis **40**, 80–95 (2017). DOI 10.1016/j.media.2017.06.004

31. Schoob, A., Kundrat, D., Lekon, S., Kahrs, L.A., Ortmaier, T.: Color-encoded distance for interactive focus positioning in laser microsurgery. Optics and Lasers in Engineering **83**, 71 –79 (2016). DOI 10.1016/j.optlaseng.2016.03.002
32. Schoob, A., Laves, M.H., Kahrs, L.A., Ortmaier, T.: Soft tissue motion tracking with application to tablet-based incision planning in laser surgery. Int. J. Comput. Assist. Radiol. Surg. **11**(12), 2325–2337 (2016). DOI 10.1007/s11548-016-1420-5
33. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging **35**(5), 1299–1312 (2016). DOI 10.1109/TMI.2016.2535302
34. Turkmen, H.I., Karsligil, M.E., Kocak, I.: Classification of laryngeal disorders based on shape and vascular defects of vocal folds. Computers in Biology and Medicine **62**, 76–85 (2015). DOI https://doi.org/10.1016/j.compbiomed.2015.02.001
35. Unger, J., Lohscheller, J., Reiter, M., Eder, K., Betz, C.S., Schuster, M.: A Noninvasive Procedure for Early-Stage Discrimination of Malignant and Precancerous Vocal Fold Lesions Based on Laryngeal Dynamics Analysis. Cancer Research **75**(1), 31–39 (2015). DOI 10.1158/0008-5472.CAN-14-1458
36. Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T.: Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Transactions on Medical Imaging **37**(7), 1562–1573 (2018). DOI 10.1109/TMI.2018.2791721
37. Zhou, Z.H., Wu, J., Tang, W.: Ensembling neural networks: Many could be better than all. Artificial Intelligence **137**(1), 239–263 (2002). DOI 10.1016/S0004-3702(02)00190-X