Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound

Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, Emanuele Peschiera, Riccardo Trevisan, Giovanni Maschietto, Elena Torri, Riccardo Inchingolo, Andrea Smargiassi, Gino Soldati, Paolo Rota, Andrea Passerini, Ruud J. G. van Sloun, Elisa Ricci, and Libertario Demi

***Abstract*— Deep learning (DL) has proved successful in medical imaging and, in the wake of the recent COVID-19 pandemic, some works have started to investigate DL-based solutions for the assisted diagnosis of lung diseases. While existing works focus on CT scans, this paper studies the application of DL techniques for the analysis of lung ultrasonography (LUS) images. Specifically, we present a novel fully-annotated dataset of LUS images collected from several Italian hospitals, with labels indicating the degree of disease severity at a frame-level, video-level, and pixel-level (segmentation masks). Leveraging these data, we introduce several deep models that address relevant tasks for the automatic analysis of LUS images. In particular, we present a novel deep network, derived from Spatial Transformer Networks, which simultaneously predicts the disease severity score associated to a input frame and provides localization of pathological artefacts in a weakly-supervised way. Furthermore, we introduce a new method based on uninorms for effective frame score aggregation at a video-level. Finally, we benchmark state of the art deep models for estimating pixel-level segmentations of COVID-19 imaging biomarkers. Experiments on the proposed dataset demonstrate satisfactory results on all the considered tasks, paving the way to future research on DL for the assisted diagnosis of COVID-19 from LUS data.**

***Index Terms*— COVID-19, lung ultrasound, deep learning.**

Manuscript received April 28, 2020; revised May 8, 2020; accepted May 10, 2020. Date of publication May 14, 2020; date of current version July 30, 2020. *(Andrea Passerini, Paolo Rota, Ruud J. G. van Sloun, Elisa Ricci, and Libertario Demi are co-first authors.) (Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Paolo Rota, Andrea Passerini, Ruud J. G. van Sloun, Elisa Ricci, and Libertario Demi contributed equally to this work.) (Corresponding author: Libertario Demi.)*

Subhankar Roy and Elisa Ricci are with the Deep and Structured Machine Learning Research Program, Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy, and also with Fondazione Bruno Kessler, 38123 Trento, Italy (e-mail: e.ricci@unitn.it).

Willi Menapace, Enrico Fini, Cristiano Saltori, Paolo Rota, and Andrea Passerini are with the Deep and Structured Machine Learning Research Program, Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: willi.menapace@studenti.unitn.it; enrico.fini@unitn.it; cristiano.saltori@unitn.it; paolo.rota@unitn.it; andrea.passerini@unitn.it).

Sebastiaan Oei, Ben Luijten, Iris Huijben, Nishith Chennakeshava, and Ruud J. G. van Sloun are with the Department of Electrical Engineering, Eindhoven University of Technology, 5612 Eindhoven, The Netherlands (e-mail: r.j.g.v.sloun@tue.nl).

Federico Mento, Alessandro Sentelli, Emanuele Peschiera, Riccardo Trevisan, Giovanni Maschietto, and Libertario Demi are with the Ultrasound Laboratory Trento (ULTRa), Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: federico.mento@unitn.it; libertario.demi@unitn.it).

Elena Torri is with BresciaMEd, 25128 Brescia, Italy (e-mail: elena.torri@gmail.com).

Riccardo Inchingolo and Andrea Smargiassi are with the Department of Cardiovascular and Thoracic Sciences, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, 00168 Rome, Italy (e-mail: riccardo.inchingolo@policlinicogemelli.it; smargiassi.a@gmail.com).

Gino Soldati is with the Diagnostic and Interventional Ultrasound Unit, Valle del Serchio General Hospital, 55100 Lucca, Italy (e-mail: gino.soldati@uslnordovest.toscana.it).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.2994459

I. Introduction

THE rapid global SARS-CoV-2 outbreak resulted in a scarcity of medical equipment. In addition to a worldwide shortage of mouth masks and mechanical ventilators, testing capacity has been severely limited. Priority of testing was therefore given to suspected patients and hospital staff [1]. However, extensive testing and diagnostics are of great importance in order to effectively contain the pandemic. Indeed, countries that have been able to achieve large-scale testing of possibly infected people combined with massive citizen surveillance, reached significant containment of the SARS-CoV-2 virus [2]. The insufficient testing capacity in most countries has therefore spurred the need and search for alternative methods that enable diagnosis of COVID-19. In addition, the accuracy of the current lab test, reverse transcription polymerase chain reaction (RT-PCR) arrays, remains highly dependent on swab technique and location [3].

COVID-19 pneumonia can rapidly progress into a very critical condition. Examination of radiological images of over 1,000 COVID-19 patients showed many acute respiratory distress syndrome (ARDS)-like characteristics, such as bilateral, and multi-lobar glass ground opacifications (mainly posteriorly and/or peripherally distributed) [4], [5]. As such, chest

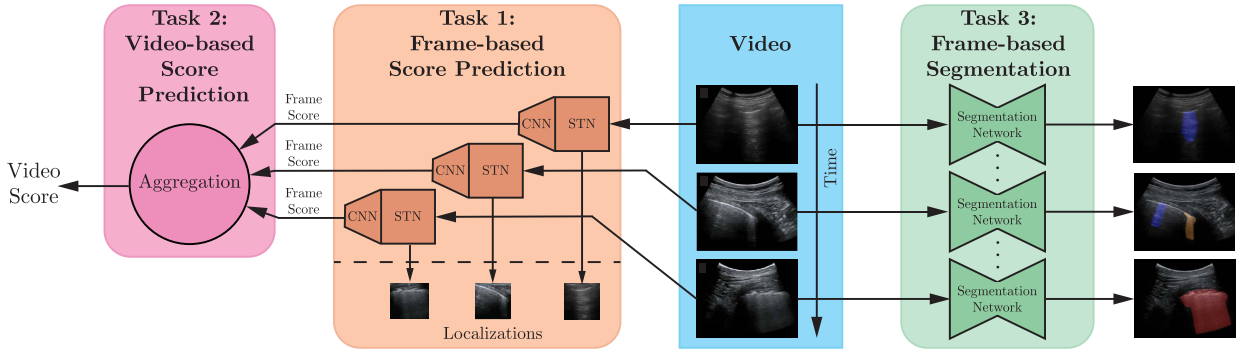

Fig. 1. Overview of the different tasks considered in this work. Given a LUS image sequence, we propose approaches for: (orange) prediction of the disease severity score for each input frame and weakly supervised localization of pathological patterns; (pink) aggregation of frame-level scores for producing predictions on videos; (green) estimation of segmentation masks indicating pathological artifacts.

computed tomography (CT) has been coined as a potential alternative for diagnosing COVID-19 patients [4]. While RT-PCR may take up to 24 hours and requires multiple tests for definitive results, diagnosis using CT can be much quicker. However, use of chest CT comes with significant drawbacks: it is costly, exposes patients to radiation, requires extensive cleaning after scans, and relies on radiologist interpretability.

Lately, ultrasound imaging, a more widely available, cost-effective, safe and real-time imaging technique, is gaining attention. In particular, lung ultrasound (LUS) is increasingly used in point-of-care settings for detection and management of acute respiratory disorders [6], [7]. In some cases, it demonstrated better sensitivity than chest X-ray in detecting pneumonia [8]. Clinicians have recently described use of LUS imaging in the emergency room for diagnosis of COVID-19 [9]. Findings suggest specific LUS characteristics and imaging biomarkers for COVID-19 patients [10]–[12], which may be used to both detect these patients and manage the respiratory efficacy of mechanical ventilation [13]. The broad range of applicability and relatively low costs make ultrasound imaging an extremely useful technique in situations when patient inflow exceeds the regular hospital imaging infrastructure capabilities. Thanks to its low costs, it is also accessible for low- and middle-income countries [14]. However, interpreting ultrasound images can be a challenging task and is prone to errors due to a steep learning curve [15].

Recently, automatic image analysis by machine and deep learning (DL) methods have already shown promise for reconstruction, classification, regression and segmentation of tissues using ultrasound images [16], [17]. In this paper we describe the use of DL to assist clinicians in detecting COVID-19 associated imaging patterns on point-of-care LUS. In particular, we tackle three different tasks on LUS imaging (Fig. 1): frame-based classification, video-level grading and pathological artifact segmentation. The first task consists of classifying each single frame of a LUS image sequence into one of the four levels of disease severity, defined by the scoring system in [12]. Video-level grading aims to predict a score for the entire frame sequence based on the same scoring scale. Segmentation instead comprises pixel-level classification of the pathological artifacts within each frame.

This paper advances the state of the art in the automatic analysis of LUS images for supporting medical personnel in the diagnosis of COVID-19 related pathologies in many directions. **(1)** We propose an extended and fully-annotated version of the ICLUS-DB database [18]. The dataset contains labels on the 4-level scale proposed in [12], both at frame and video-level. Furthermore, it includes a subset of pixel-level annotated LUS images useful for developing and assessing semantic segmentation methods. **(2)** We introduce a novel deep architecture which permits to predict the score associated to a single LUS image, as well as to identify regions containing pathological artifacts in a weakly supervised manner. Our network leverages Spatial Transformers Network (STN) [19] and consistency losses [20] to achieve disease pattern localization and from a soft ordinal regression loss [21] for robust score estimation. **(3)** We introduce a simple and lightweight approach based on uninorms [22] to aggregate frame-level predictions and estimate the score associated to a video sequence. **(4)** We address the problem of automatic localization of pathological artifacts evaluating the performance of state-of-the-art semantic segmentation methods derived from fully convolutional architectures. **(5)** Finally, we conduct an extensive evaluation of our methods on all the tasks, showing that accurate prediction and localization of COVID-19 imaging biomarkers can be achieved with the proposed solutions. Dataset and code are available at *https://iclus-web.bluetensor.ai* and at *https://github.com/mhug-Trento/DL4covidUltrasound*.

II. Related Work

DL has proven to be successful in a multitude of computer vision tasks ranging from object recognition and detection to semantic segmentation. Motivated by these successes, more recently, DL has been increasingly used in medical applications, e.g. for biomedical image segmentation [23] or pneumonia detection from chest X-ray [24]. These seminal works indicate that, with the availability of data, DL can lead to the assistance and automation of preliminary diagnoses which are of tremendous significance in the medical community.

In the wake of the current pandemic, recent works have focused on the detection of COVID-19 from chest CT [25], [26]. In [27], a U-Net type network is used to regress

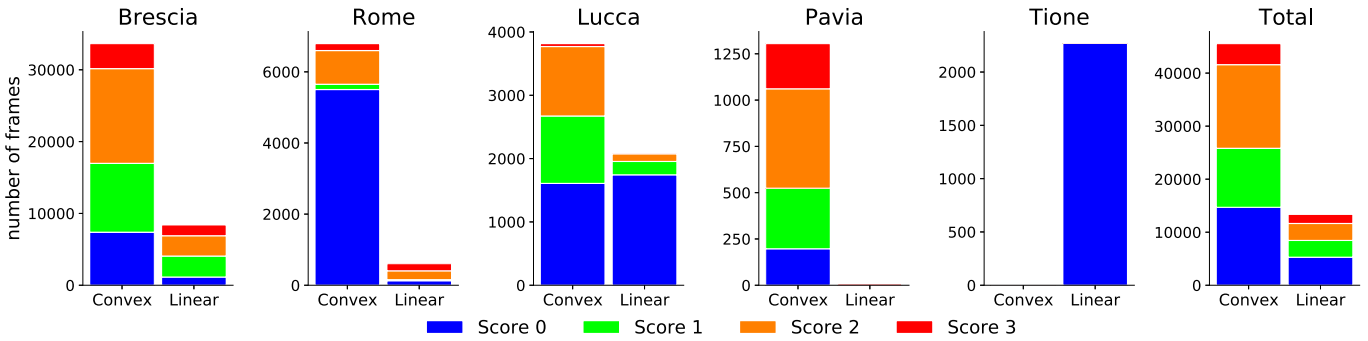

Fig. 2. The distribution of the probes and the scores of frames grouped by hospital and overall statistics.

a bounding box for each suspicious COVID-19 pneumonia region on consecutive CT scans, and a quadrant-based filtering is exploited to reduce possible false positive detections. Differently, in [28] a threshold-based region proposal is first used to retrieve the region of interests (RoIs) in the input scan and the Inception network is exploited to classify each proposed RoI. Similarly, in [29], a VNET-IR-RPN model pre-trained for pulmonary tuberculosis detection is used to propose RoIs in the input CT and a 3D version of Resnet-18 is employed to classify each RoI. However, very few works using DL on LUS images can be found in the literature [30]. A classification and weakly-supervised localization method for lung pathology is described in [17]. Based on the same idea, in [18] a frame-based classification and weakly-supervised segmentation method is applied on LUS images for COVID-19 related pattern detection. Here, Efficientnet is trained to recognize COVID-19 in LUS images, after which class activation maps (CAMs) [31] are exploited to produce a weakly-supervised segmentation map of the input image. Our work has several differences compared to all the previous works. First, while in [18] CAMs are used for localization, in this work we exploit STN to learn a weakly-supervised localization policy from the data (i.e. not exploiting explicit labelled locations but inferring it from simple frame-based classification labels). Second, while in [18] a classification problem is solved, we focus on ordinal regression, predicting not only the presence of COVID-19 related artifacts, but also a score connected to the disease severity. Third, we move a step forward compared to all previous methods by proposing a video-level prediction model built on top of the frame-based method. Finally, we propose a simple yet effective method to predict segmentation masks using an ensemble of multiple state-of-the-art convolutional network architectures for image segmentation. Additionally, the model's predictions are accompanied with uncertainty estimates to facilitate interpretation of the results.

III. ICLUS-DB: Data Collection and Annotation

We here present the Italian COVID-19 Lung Ultrasound DataBase (ICLUS-DB), which currently includes a total of 277 lung ultrasound (LUS) videos from 35 patients, corresponding to 58,924 frames.[1] The data were acquired within different clinical centers (BresciaMed, Brescia, Italy, Valle del Serchio General Hospital, Lucca, Italy, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy, Fondazione Policlinico Universitario San Matteo IRCCS, Pavia, Italy, Tione General Hospital, Tione (TN), Italy) and using a variety of ultrasound scanners (Mindray DC-70 Exp, Esaote MyLabAlpha, Toshiba Aplio XV, WiFi Ultrasound Probes - ATL). Both linear and convex probes were used, depending on necessities. Of the 35 patients, 17 were confirmed positive to COVID-19 by swab technique (49 %), 4 were COVID-19 suspected (11 %), and 14 were healthy and symptomless individuals (40 %).

A recent proposal by Soldati *et al.* describes how specific imaging biomarkers in LUS can be used in the management of COVID-19 patients [12]. Specifically, to evaluate the progression of the pathology, a 4-level scoring system was devised [32], with scores ranging from 0 to 3. Score 0 indicates the presence of a continuous pleural-line accompanied by horizontal artifacts called A-lines [33], which characterize a healthy lung surface. In contrast, score 1 indicates the first signs of abnormality, i.e., the appearance of alterations in the pleural-line in conjunction with vertical artifacts. Scores 2 and 3 are representative of a more advanced pathological state, with the presence of small or large consolidations, respectively. Finally score 3 is associated with the presence of a wider hyperechogenic area below the pleural surface, which can be referred to as "white lung".

A total of 45,560 and 13,364 frames, acquired using respectively convex and linear probes, were labelled according to the scoring system defined above. Of the 58,924 LUS frames forming the dataset, 5,684 were labeled score 3 (10%), 18,972 score 2 (32%), 14,295 score 1 (24%), 19,973 score 0 (34%). A plot showing the distribution of the scores and probes per hospital is shown in Fig. 2. To guarantee objective annotation, the labelling process was stratified into 4 levels: 1) score assigned frame-by-frame by four master students with ultrasound background knowledge, 2) validation of the assigned scores performed by a PhD student with expertise in LUS, 3) second level of validation performed by a biomedical engineer with more than 10 year of experience in LUS and 4) third level of validation and agreement between clinicians with more than 10 years of experience in LUS.

Additionally, a subset of 60 videos sampled across all 35 patients was selected and video-level annotations were provided for them. These annotations use the same scoring

[1] https://iclus-web.bluetensor.ai.

system defined for the frame-level annotations. In order to address subjective biases in the evaluation of the videos, five different clinicians provided their evaluation for each sequence. We assess the complexity of this task by calculating the inter-operator agreement, comparing the evaluation of the predictions of each doctor against the average prediction of the remaining four. The resulting average agreement is about 67% among the available labels.

Finally, for 33 patients, a total of 1,005 and 426 frames respectively acquired using convex and linear probes, were semantically annotated at a pixel-level by contouring the aforementioned imaging biomarkers using the annotation tool LabelMe [34]. For the frames acquired using the linear probe, relative pixel-level occurrences for scores 0, 1, 2, and 3 are 6.4%, 0.080%, 0.67%, and 3.7%, respectively. For the convex probe, these statistics are 1.9%, 0.074%, 1.8%, and 2.1%, respectively. Notably, a large proportion of pixels is not associated to either of these scores. These pixels do not display clear characteristics of a specific class, and are referred to as background (BG). A few images and the corresponding annotations are shown in the supplementary material.

IV. Deep Learning-Based Analysis of LUS Images

This paper tackles several challenges towards the development of automatic approaches for supporting medical personnel in the diagnosis of COVID-19 related pathologies (see Fig. 1). In particular, following the COVID-19 LUS scoring system in [12] we present a novel deep architecture which automatically predicts the pathological scores associated to all frames of a LUS image sequence (Section IV-A) and optimally fuse them to produce a disease severity score at video-level (Section IV-B). We also show that the proposed model automatically identifies regions in an image which are associated to pathological artifacts without requiring pixel-level annotation. Finally, to further improve the accuracy in the automatic detection of disease-related patterns, we also consider a scenario where frames are provided with pixel-level annotations and we propose a segmentation model derived from a state of the art convolutional network architecture (Section IV-C). In the following, we describe the proposed deep learning models.

A. Frame-Based Score Prediction

1) Problem Formulation and Notation: With the purpose of supporting medical personnel in the analysis of LUS images, in this paper we introduce an approach for predicting the presence or the absence of a pathological artifact in each frame of a LUS image sequence and for automatically assessing the severity score of the disease related to such patterns according to the COVID-19 LUS scoring system [12]. We are also interested in the spatial localisation of a pathological artifact in the frame *without assuming any annotation* about such artifact positions in a frame. The weak localization is achieved through the use of Spatial Transformer Networks (STN) [19]. The use of STN stems from the fact that most of the pathological artifacts are concentrated in a relatively small area of the image, and, hence the entire image should not be considered by the network to make predictions. The problem can be formalized as follows.

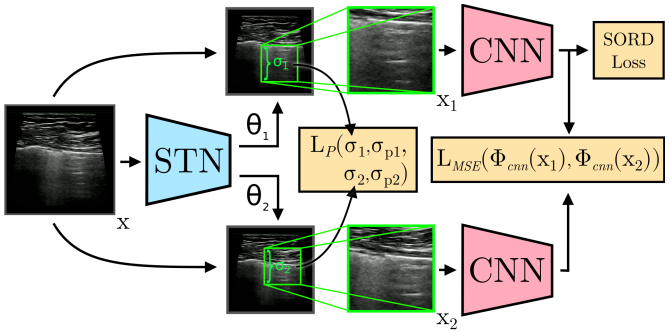

Fig. 3. Illustration of the architecture for frame-based score prediction. An STN modeled by Φ_{stn} predicts two transformations θ_1 and θ_2 which are applied to the input image producing two transformed versions $\mathbf{x}_1$ and $\mathbf{x}_2$ that localize pathological artifacts. The feature extractor Φ_{cnn} is applied to $\mathbf{x}_1$ to generate the final prediction.

Let $\mathcal{X}$ denote the input space (i.e. the image space) and $\mathcal{S}$ the set of possible scores. During training, we are given a training set $\mathcal{T} = \{(\mathbf{x}_n, s_n)\}_{n=1}^{N}$ where $\mathbf{x}_n \in \mathcal{X}$ and $s_n \in \mathcal{S}$.

2) Model Definition: We are interested in learning a mapping $\Phi : \mathcal{X} \rightarrow \mathcal{S}$, which given an input LUS image outputs the associated pathological score label. We model Φ as the composition of two functions $\Phi = \Phi_{stn} \circ \Phi_{cnn}$ where $\Phi_{stn} : \mathcal{X} \rightarrow \mathcal{X}$ estimates an affine transformation and applies it to the input image $\mathbf{x}$ and $\Phi_{cnn} : \mathcal{X} \rightarrow \mathcal{S}$ assigns the score to the transformed image. Intuitively, Φ_{stn} learns to localize regions of interest in the input image and provides Φ_{cnn} with an image crop where information about the score is most salient. Consequently, Φ_{stn} produces as a side effect the localization of pathological artifacts in the frame. The mapping Φ_{cnn} is composed by a convolutional feature extractor and a linear layer with $|\mathcal{S}|$-dimensional output logits. The model Φ_{stn} is implemented as a deep neural network derived STN [19]. Fig. 3 shows an overview of the proposed deep architecture.

In the context of deep learning the generalization capability of a network is of critical importance. To this end, data augmentation has shown to be very effective [35] in improving the performance of a network. Previous works [18] showed that augmenting a dataset composed of LUS images can drastically improve the ability of the network to discriminate healthy and ill patients. Another way to achieve robust predictions is to enforce some consistency between two perturbed versions (colour jitter, dropout, etc.) of the same image [20], [36]. This makes the network produce smoothed predictions by attending to the salient features in an image. Inspired by this idea, we propose to use STN [19] to produce two different crops from a single image and enforce the predictions of the network to be similar. We name our approach Regularised Spatial Transformer Networks (Reg-STN).

STN [19] is a differentiable module that applies a learnable affine transformation to an input image, or more in general to a feature map, conditioned on the input itself. It consists of three parts: (i) a *localization network* that predicts the parameters of the affine transformation, (ii) a *grid generator* which selects the grid co-ordinates in the source image, to be sampled from,

and (iii) a *sampler* that warps the input image based on the transformation, producing the output map.

For what concerns the *localization network*, it is trained to output a transformation matrix θ such that:

$$\begin{pmatrix} \alpha^s \\ \beta^s \end{pmatrix} = \theta \begin{pmatrix} \alpha^t \\ \beta^t \\ 1 \end{pmatrix} \tag{1}$$

where α^s, β^s, α^t, β^t, are the source and target coordinates in the input and output feature map respectively. In principle θ can describe any affine transformation, however, keeping in mind the properties of LUS images we restrict the space of possible transformations to rotation, translation, and isotropic scaling:

$$\theta = \begin{bmatrix} \sigma & r_1 & \tau_\alpha \\ r_2 & \sigma & \tau_\beta \end{bmatrix} \tag{2}$$

In our proposed method, an input image, $\mathbf{x}$ is processed by the Φ_{stn} that predicts two set of transformations θ_1 and θ_2, instead of one θ. Subsequently, the transformations are applied to $\mathbf{x}$, generating cropped images $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. The network Φ_{cnn} is then applied to $\mathbf{x}_1$ and $\mathbf{x}_2$, producing two sets of logits for the same image under different transformations. As a side effect, the intermediate images $\mathbf{x}_1$ and $\mathbf{x}_2$ are produced and can be interpreted as the localization of the pathological artifacts in the input image $\mathbf{x}$. Finally, the $\Phi_{cnn}(\mathbf{x}_1)$ branch then can be trained with any standard supervised classification loss and $(\Phi_{cnn}(\mathbf{x}_1), \Phi_{cnn}(\mathbf{x}_2))$ is trained with a consistency enforcing loss (see below).

3) Loss Definition: As stated before, we are interested in devising a deep network Φ for automatically predicting the 4-level scores identified in [12]. While this problem can trivially be cast within a classification framework, in this paper we argue that ordinal regression [37] is more appropriate as we are interested in predicting labels from an ordinal scale. The rationale behind the choice of ordinal regression is that there exist certain categories that are more correct than others with respect to the true label, as opposed to an independent class scenario, in which the order of the levels does not matter. In fact, errors on low-distance levels should be less penalized with respect to long-distance error. For instance, predicting a severely ill patient (score 3) as healthy (score 0) should be strongly discouraged, while sometimes the difference between score 1 and score 2 can be subtle and the network should not be overly penalized.

While ordinal regression can be implemented resorting on the traditional approach of decomposing the problem assuming a $|\mathcal{S}|$-rank formulation [38], following [21] we introduce a lightweight approach for Soft ORDinal regression (SORD). In practice, we implement an ordinal regression framework by using a carefully devised label smoothing mechanism. Instead of one-hot representations of labels, we encode the ground truth information into soft-valued vectors (SORD vectors) $\hat{s} \in \mathbb{R}^{|S|}$, where $\mathcal{S}$ is the set of possible scores for a frame. Hence, for a frame $\mathbf{x}$ with score $s \in \mathcal{S}$ the i-th element of the SORD vector is computed as follows:

$$\hat{s}_i = \frac{e^{-\delta(s,i)}}{\sum_{j \in \mathcal{S}} e^{-\delta(j,i)}} \tag{3}$$

where δ is a manually defined distance function between scores/levels for which we use square distance multiplied by a constant factor. This formulation produces a smooth probability distribution over $\mathcal{S}$, in which the magnitude of the elements decreases while the distance to the ground truth increases. Encoding ground truth labels as probability distributions seamlessly blends with common classification loss functions that use a softmax output. Therefore, at training time, we simply train the network Φ using cross entropy:

$$\mathcal{L}_{SORD} = -\sum_{i=0}^{|\mathcal{S}|} \hat{s}_i \log \left(\frac{\exp(\Phi(\mathbf{x})_i)}{\sum_j^{|\mathcal{S}|} \exp(\Phi(\mathbf{x})_j)} \right) \tag{4}$$

The result is a loss function that yields a smaller cost for predictions that are in the neighbourhood of the ground truth label, which, in turn generates smaller gradients, hence discouraging drastic updates of the network for small errors. Empirically, we found that our algorithm works best when we increase the distance of score 0 from the others. As mentioned before, this is also validated by the semantics of the scores.

Another desirable property of the network is to extract important semantic features of the input image, in order to enable accurate frame score prediction. This can be strengthened by resorting to a regularization in the form of consistency loss on the two branch predictions $(\Phi_{cnn}(\mathbf{x}_1), \Phi_{cnn}(\mathbf{x}_2))$ with the rationale that two different crops from the same image should have similar predictions. In our case, these two crops are produced by the Φ_{stn}. In details, the consistency loss is defined on the network representations as following:

$$\mathcal{L}_{MSE} = \|\Phi_{cnn}(\mathbf{x}_1) - \Phi_{cnn}(\mathbf{x}_2)\|_2^2 \tag{5}$$

Unfortunately, $\mathcal{L}_{MSE}$ coupled with learnable affine transformations produces degenerate solutions in which the *localization network* of the STN learns to output identical parameters for the affine transformations. In fact, it is enough to impose $\theta_1 = \theta_2$ to minimize $\mathcal{L}_{MSE}$. To prevent this pathological behaviour of the network, we enforce a prior on the parameters of the transformations. In particular, we stimulate the *localization network* to produce reasonably scaled patches by minimizing $|\sigma - \sigma_p|$, where σ_p is a fixed prior. Now, in order to enable the STN into yielding different parameters $\theta_1 \neq \theta_2$, we simply choose $\sigma_{p1} \neq \sigma_{p2}$. Hence, a loss is defined as follows:

$$\mathcal{L}_P = |\sigma_1 - \sigma_{p1}| + |\sigma_2 - \sigma_{p2}| \tag{6}$$

Finally, the proposed Reg-STN model is trained end-to-end minimizing the following joint loss function:

$$\mathcal{L}_{TOT} = \mathcal{L}_{SORD} + \mathcal{L}_{MSE} + \mathcal{L}_P \tag{7}$$

4) Training Strategy: We split the ICLUS-DB dataset into a train and test split. The test split comprises 80 videos from 11 patients, with a total of 10,709 frames. All the frames from the remaining videos are included in the train set. The split is performed at patient level, such that the sets of patients in the training and test set are disjoint. The STN is modeled by a ConvNet similar to [17]. Specifically, we removed the Average Pooling and the output layer and replaced it with two fully connected layers to predict the affine transformation

parameters. The CNN architecture [17] is kept unchanged. The STN and CNN are jointly trained by using the Adam optimizer with an initial learning rate of $1e-4$, a batch size of 64 and trained for 120 epochs. We also used similar data augmentation strategies and learning rate decay as suggested in [17], [18]. We set the values of σ_1 and σ_2 to 0.50 and 0.75 respectively, leveraging the prior knowledge about LUS images that pathological artifacts roughly covers 25% to 50% area of the image.

B. Video-Level Score Aggregation

1) Problem Formulation and Notation: The identification of potentially pathological artifacts in LUS images is a crucial step towards diagnosis support. However, frame-based predictions should be turned into a single video-based score prediction in order to assess the pathological state of a patient. The video-based score aggregation problem can be formalized as follows. Let $\mathbf{v} = \{\mathbf{x}_i\}_{i=1}^{M}$, be a video, $\mathcal{V}$ be the set of videos of any length, and $\mathcal{S}$ the set of scores. The goal of video-level score prediction is learning a mapping $\Psi : \mathcal{V} \rightarrow \mathcal{S}$.

In principle the mapping Ψ could be obtained by taking the maximum score assigned to any frame of the current video because the identification of an artifact of score s in a frame implies that the patient has a severity level of at least s. This hard rule, however, is inapplicable in practice when dealing with machine-predicted scores, as even a single frame-based prediction error could harm the overall prediction. Thus, in this section we propose a more flexible aggregation mechanism devised for predicting the score associated to a video, leveraging the video-level annotations provided in the ICLUS-DB (Section III).

2) Model Definition: In designing the model Ψ, we consider the fact that it needs to operate in a low-data regime, where few videos are provided with annotations as in the current version of the ICLUS-DB. Inspired by the hard rule previously mentioned, we propose a simple strategy that combines frame-level predictions using a parameterized aggregation layer, i.e.:

$$\Psi(\mathbf{v}) = \Psi_U(\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_M)) \tag{8}$$

Here Φ is the frame-level mapping and Ψ_U is an aggregation function based on *uninorms* [39], which are a principled way to soften the hard rule. A uninorm U is a monotonic increasing, commutative and associative mapping from $[0,1]\times[0,1]$ to $[0,1]$ with neutral element $e \in [0,1]$. This means that $U(a,e) = U(e,a) = a$ for all $a \in [0,1]$. If $e = 1$, U is fully non-compensatory (like taking the minimum between a and b), while it is fully compensatory if $e = 0$ (like the taking maximum). Choosing $e \in (0,1)$ allows the uninorm to have a hybrid behaviour. Note that being associative, uninorms can be applied to an arbitrary number of inputs (e.g., $U(a,b,c) = U(U(a,b),c)$). Following [22], we learn the appropriate value for the neutral element e from data. Our aggregation layer takes as input the sequence of frame-based prediction scores $\Phi(\mathbf{x})$, aggregates them along each dimension/score using a uninorm U and returns the softmax of the resulting aggregation as a video-based prediction. The layer has only four parameters, which are the neutral elements for each candidate score $\{0,1,2,3\}$, and it is thus amenable to training with little supervision.

Any uninorm with neutral element e can be written as [39]:

$$U_e(a,b) = \begin{cases} eT(\frac{a}{e}, \frac{b}{e}) & \text{if } a,b \in [0,e] \\ e + (1-e)S(\frac{a-e}{1-e}, \frac{b-e}{1-e}) & \text{if } a,b \in [e,1] \\ \hat{U}(a,b) & \text{otherwise} \end{cases} \tag{9}$$

for a certain choice of T, S and $\hat{U}(a,b)$ such that $\min(a,b) \leq \hat{U}(a,b) \leq \max(a,b)$. The functions T and S are called t-norm and t-conorm respectively, and model the non-compensatory and compensatory behaviour. Different choices for these functions lead to different uninorms. We found the product t-norm $T(a,b) = ab$ (and corresponding t-conorm $S(a,b) = a + b - ab$) to be the most effective choice as it allows the gradient to flow the most. Concerning the function $\hat{U}(a,b)$, common choices are $\min(a,b)$ and $\max(a,b)$, producing the so-called min-uninorms and max-uninorms respectively. We found min-uninorms to be the best choice in our setting (with respect to $\max(a,b)$ but also $\text{mean}(a,b)$), likely because of their fully non-compensatory behaviour in the area of highest discrepancy between frame-based predictions.

3) Loss Definition: The architecture is trained using the SORD loss described in Eq. (5) computed over the video-level prediction.

4) Training Strategy: The frame-based predictor outputs prediction scores with a distribution that differs between the training and the test set. In order not to overfit the video-based predictor on the training scores distribution, we completely separate the training sets of the frame-based and video-based predictor. We train the frame-based predictor on all video sequences $\mathcal{T}$ without any video-based annotation, and evaluate it on the remaining sequences $\mathcal{T}'$. We then train and evaluate the video-based predictor on $\mathcal{T}'$, using a k-fold cross validation procedure ($k = 5$) with splits made at the patient level (i.e. all videos from the same patient are in the same fold). We choose to use as video-level annotations the ones produced by the first annotator, the clinician with the highest expertise. We train our model using an Adam optimizer with learning rate 10^{-2} without weight decay and with no learning rate scheduling. For each epoch, we compute the loss for each train video sequence and accumulate its gradients, performing a single optimization step at the end of each epoch. We train the model for a maximum of 30 epochs and use the loss on the training set to define an early stopping strategy.

Note that the entire architecture including the frame-level component could be trained entirely end-to-end. However, this solution is not effective given the vast disproportion in the amount of supervision at the video and frame levels currently available in ICLUS-DB. We thus trained the aggregation layer after freezing the weights of the frame-based architecture. Full end-to-end training combining frame-based and video-based supervision will be investigated in future work.

C. Semantic Segmentation

1) Problem Formulation and Notation: Let $\mathcal{X} = \mathbb{R}^{i\times j}$ and $\mathcal{Y}$ denote the input (i.e. the image space) and output

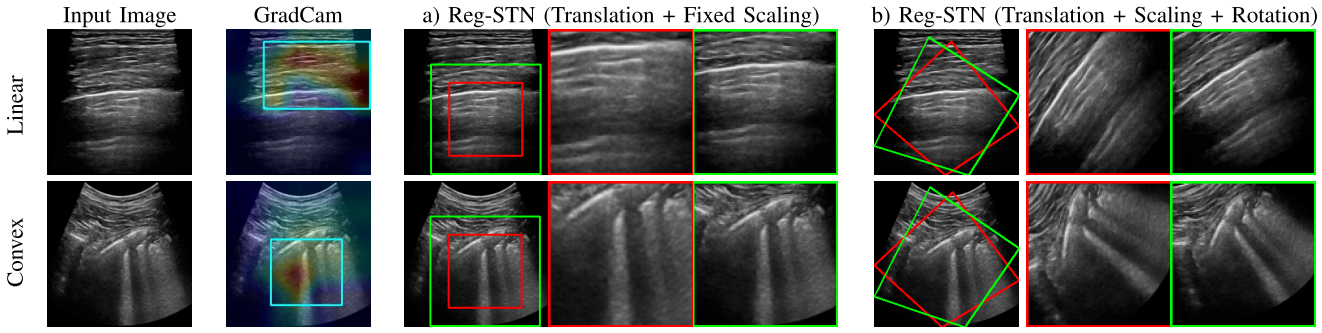

Fig. 4. Examples of the image crops produced by the Reg-STN network. The first column shows input images acquired with linear and convex sensors, respectively. In the second column we report the heatmaps produced by GradCam [44] and the bounding boxes obtained by thresholding. In the remaining columns, original image overlayed with bounding boxes and the two respective crops (in red and green) produced when the Reg-STN models: a) only translation and a fixed scaling; b) all possible transformations viz. translation, scaling and rotation, are shown. In each case the Reg-STN focuses on the most salient parts which contains the pathological artifacts.

(i.e. the segmentation masks) space, respectively. In the earlier presented frameworks for image- and video-based classification, the score set was defined as $\mathcal{S} = \{0, 1, 2, 3\}$. For semantic segmentation we however distinguish five different scores, i.e. the four scores in $\mathcal{S}$, complemented by the background (BG) score, assigned to pixels that were not annotated for showing markers associated with any of the classes in $\mathcal{S}$. As such, $\mathcal{Y} = \{0, 1, 2, 3, \text{BG}\}^{i \times j}$.

2) Model Definition: We are interested in learning a mapping $\mathbf{\Omega} : \mathcal{X} \rightarrow \mathcal{Y}$, which given an input LUS image, outputs the associate pathological segmentation mask. To model $\mathbf{\Omega}$, we compare several network architectures for end-to-end image segmentation, such as the vanilla U-Net [23], and the more recently proposed U-Net++ [40], and Deeplabv3+ [41].

Our baseline U-Net model has three encoding layer blocks, each comprising two convolutional layers with ReLU activations and one maxpool layer (pooling across 2, 2, and 5 pixels in both dimensions, respectively), a latent layer, and a mirrored decoder (where pooling is replaced by nearest neighbour upsampling). We use skip connections between each layer block of the encoder and decoder. To mitigate overfitting we apply dropout ($p = 0.5$) during training at the latent bottleneck of the model. The Unet++ variant leverages the first four encoder blocks of the ResNet50 model [42] to construct a latent space. The latent space is upsampled in the decoder stage by means of transpose 2D convolutional layers. The decoder contains residual blocks, and also exploits skip connections between (same-sized) hidden layer outputs in the ResNet50 encoder and the decoder. The Deeplabv3+ model similarly employs an encoder-decoder structure, where features are extracted using spatial pyramid pooling (i.e. pooling at different grid scales) and atrous convolutions, resulting in decoded segmentation maps with detailed object boundaries.

3) Loss Definition: We adopt a pixel-wise categorical cross-entropy loss between he segmentation masks $g(\mathbf{y}_n)$ and the model predictions $\hat{\mathbf{y}}_n = \mathbf{\Omega}(h(\mathbf{x}_n))$. Functions $g(\cdot)$, and $h(\cdot)$ are pre-processing transformations applied prior to training.

Function $h(\cdot)$ comprises the resizing of all acquired B-mode images to 260×200 pixels, preserving the original aspect ratio of the scans by appropriate zero padding, and subsequent normalization between -1 and 1.

4) Training Strategy: Due to the larger (and more representative) set of pixel-level annotations for the convex probe, compared to the linear probe acquisitions (1,005 and 426 annotations, respectively), we here specifically focus on the convex acquisitions. We split our dataset into a train (70%) and test set (30%) at a patient level, i.e. all movies and frames from one patient fall into a specific set. Among the 1005 frames, a total of 1158 imaging biomarkers were segmented.

During training, we are given a training set of N image-label pairs $\mathcal{T} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$ where $\mathbf{x}_n \in \mathcal{X}$ and $\mathbf{y}_n \in \mathcal{Y}$. The model parameters are learned by back-propagating the earlier defined categorical cross-entropy using the Adam optimizer (default settings), with a learning rate of 10^{-5}. Training was stopped upon convergence of the training loss.

Each training batch consists of 32 B-mode images and their corresponding segmentation masks, which are balanced across patients and scores to avoid biases resulting from the length of the ultrasound scan (number of frames in a single video) and population-level distribution of scores. While these biases generally aid the overall accuracy, they hamper patient-level decision making across demographics.

To promote invariance to common LUS image transformations and thereby improve generalization at inference, each image-label pair is heavily manipulated on-line during training by a set of augmentation functions that were each activated on the image-label pair with a probability of 0.33. The set of augmentation functions, each applied with a randomly sampled strength bounded by a set maximum, consists of: affine transformations (translation (max. $\pm 15\%$), rotation (max. $\pm 15°$), scaling (max. $\pm 45\%$), and shearing (max. $\pm 4.5°$)), multiplication with a constant (max. $\pm 45\%$), Gaussian blurring ($\sigma_{\max} = \frac{3}{4}$), contrast distortion (max. $\pm 45\%$), horizontal flipping ($p = 0.5$), and additive white Gaussian noise ($\sigma_{\max} = 0.015$).

5) Inference: To further boost robustness and performance, we apply model ensembling and calculate the unweighted average over predicted softmax logits of the U-net, U-net++, and Deeplabv3+ models (all trained with data augmentation).

To allow for qualitatively assessment of the uncertainty of the predictions, we produce pixel-level estimates of model uncertainty by using Monte-Carlo (MC) dropout [43]. During

TABLE I

F1 SCORES (%) FOR THE FRAME-BASED CLASSIFICATION UNDER DIFFERENT EVALUATION SETTINGS. SETTING 1 REPRESENTS EVALUATION ON THE FULL TEST SET, SETTING 2 REPRESENTS THE ANALYSIS ON THE TEST SET WITH DROPPED TRANSITION FRAMES AND SETTING 3 REPRESENTS THE ANALYSIS ACCOUNTING FOR INTER-DOCTOR AGREEMENT. THE BASELINE FOR THIS SETTING IS PROVIDED BY THE EVALUATION ON THE SET OF TEST SEQUENCES WITH VIDEO-LEVEL ANNOTATIONS (VIDEO ANN.). BEST AND SECOND BEST F1 SCORES (%) ARE IN BOLD AND UNDERLINES, RESPECTIVELY

Model	Setting 1 Regular Metric	Setting 2 Drop Transition Frames (K)				Setting 3 Inter-doctor Agreement (A)			Avg
		K=1	K=3	K=5	K=7	Video Ann.	A=3	A=4	
CNN+CE	61.6	63.1	64.9	66.3	67.6	74.8	78.0	77.0	69.2
CNN+SORD	<u>63.2</u>	<u>64.8</u>	<u>66.3</u>	<u>67.8</u>	<u>68.9</u>	73.0	76.8	75.8	69.6
Resnet-18+SORD	62.2	63.9	65.5	66.9	67.8	74.5	77.4	76.4	69.3
CNN+STN+SORD	61.0	62.6	63.8	64.8	65.6	**78.4**	**82.2**	**81.4**	<u>70.0</u>
CNN+Random Crop+SORD	61.8	63.0	64.2	65.1	65.9	71.9	74.6	73.5	67.5
CNN+Reg-STN+SORD (Ours)	**65.1**	**66.7**	**68.3**	**69.5**	**70.3**	<u>75.4</u>	<u>78.4</u>	<u>77.5</u>	**71.4**

inference, we stochastically apply dropout in the latent space, yielding multiple point estimates of our class predictions. The amount of variation in the resulting predictions, ultimately provides an indication of uncertainty for every pixel.

V. EXPERIMENTAL RESULTS

A. Frame-Based Score Prediction

To evaluate the performance of our proposed frame-based scoring method and its constituent components we consider the following baselines: i) CNN trained with Cross Entropy loss (CE), ii) CNN trained with SORD, iii) Resnet-18 trained with SORD, iv) STN based CNN trained with SORD; v) CNN + Random Crop + SORD, a CNN trained on SORD with random crops rather than bounding boxes extracted by STN and vi) Our proposed Reg-STN model.

In Table I, we evaluate the performance of our method in terms of F1-score. Since, the annotations in LUS images are quite subjective (see later) we also report results for two additional metrics, which are then defined as Setting 2 and Setting 3, respectively. The metrics are: i) Setting 1 considers the F1 score computed on the entire test set, ii) Setting 2 considers the F1 score computed on a modified version of the test set obtained by dropping, for each video, the K frames before and after each transition between two different ground truth scores, potentially removing ambiguous frames that present characteristics at the boundary between two classes, thereby allowing us to identify the impact of noisy labeling on the performance of the model; and iii) Setting 3, we drop the most challenging videos by using the inter-doctor agreement between the 5 independent video-level annotations. In practice, we only keep in the test set the videos with at least A doctors agreeing on the video-level annotations. For completeness, we report under Setting 3 also the scores obtained on the complete portion of the test set containing video-level annotations (Video Ann.).

As shown in Table I, our proposed Reg-STN trained with SORD beat the baseline models in most of the settings and is the second best in the remaining. On average, Reg-STN performs the best amongst all baselines. This proves the effectiveness of our proposed method for doing frame-based prediction for pathology detection in LUS images. Our experiments were run on a RTX-2080 NVIDIA GPU. As for computational complexity, it takes ∼11 hours to train a CNN + Reg-STN + SORD model on this hardware.

TABLE II

MEAN AND STANDARD DEVIATION OF WEIGHTED F1 SCORE, PRECISION AND RECALL COMPUTED OVER THE FIVE CROSS VALIDATION FOLDS, FOR THE PROPOSED VIDEO-BASED CLASSIFICATION METHOD AND BASELINES

Method	F1 (%)	Precision (%)	Recall (%)
max_argmax	46 ± 21	55 ± 27	49 ± 18
argmax_mean	51 ± 12	56 ± 19	53 ± 09
uninorms	$\mathbf{61} \pm 12$	$\mathbf{70} \pm 19$	$\mathbf{60} \pm 07$

B. Video-Based Score Prediction

We evaluate video-based score prediction in terms of weighted F1 score, Precision and Recall. These are obtained by first computing the metric for each score (zero to three), and then computing the weighted average over scores, where the weight is the fraction of instances having that score. Note that weighted recall corresponds to (multiscore) accuracy, i.e., the fraction of correctly predicted scores over the total number of predictions. Table II reports averages and standard deviations of these metrics over the five folds of the cross validation procedure. We compare our video-level predictor with two standard aggregation methods, *max_argmax* and *argmax_mean*. The former implements the hard rule described in Section IV-B. It labels each frame with the most probable score according to the frame-level predictor, and takes the maximal score along the video. The latter averages frame-level predictions over the video and returns the score with the maximal average. The proposed method outperforms both baselines in terms of F1-score, precision and recall. Table III shows confusion matrices for the three methods, obtained by concatenating the predictions for all folds. As expected, the *max_argmax* hard rule is strongly biased towards predicting the highest score, resulting in bad performance on all other scores. On the other hand, the *argmax_mean* baseline has the best performance in predicting score zero, but performs poorly on the other scores (under-predicting scores one and three and over-predicting score two). The uninorm-based aggregation is more balanced, outperforming each of the baselines on three out of four scores.

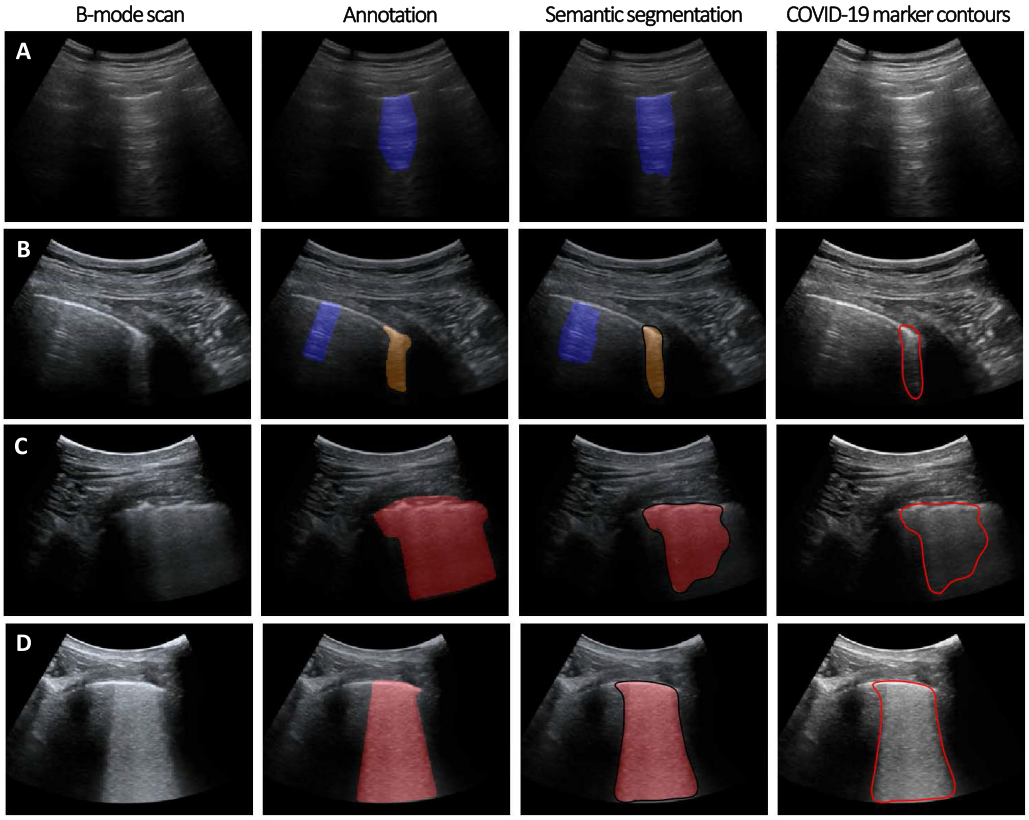

Fig. 5. Four examples of B-mode input image frames (first column), their annotations (second column) including COVID-19 biomarkers (moderate/score 2: orange, severe/score 3: red), and signs of healthy lung (blue). The corresponding semantic segmentations and contours of COVID-19 markers by deep learning are given in the third and fourth colomn, respectively.

TABLE III
CONFUSION MATRICES (%) FOR THE PROPOSED VIDEO-BASED CLASSIFICATION METHOD AND BASELINES

	max_argmax				*argmax_mean*				*uninorms*			
	0	1	2	3	0	1	2	3	0	1	2	3
0	7	9	7	3	16	10	0	0	10	16	0	0
1	0	3	12	5	3	9	9	0	2	17	2	0
2	0	0	17	9	0	3	21	2	0	5	16	5
3	0	0	5	22	0	2	17	9	0	5	5	17

TABLE IV
SEGMENTATION PERFORMANCE IN TERMS OF THE MEAN CATEGORICAL ACCURACY ACROSS ALL PIXELS AND SCORES (ACC.), THE DICE COEFFICIENT FOR THE UNION OF COVID-19-RELATED SCORES (DICE), AND THE MEAN DICE ACROSS SCORES 0, 2, AND 3 (CAT. DICE). SCORE 1 WAS EXCLUDED DUE TO THE LOW NUMBER OF ANNOTATIONS

Network and training strategy	Acc.	Dice	Cat. Dice
(1) U-Net	0.94	0.64	0.47
(2) U-Net with on-line augmentation	0.95	0.69	0.55
(3) U-Net++ with on-line augmentation	0.97	0.72	0.64
(4) Deeplab v3+ with on-line augmentation	0.95	0.71	0.62
Ensemble of (2,3,4)	0.96	0.75	0.65

C. Semantic Segmentation

Fig. 5 shows several illustrative examples of semantic segmentation results of our ensemble network, along with their ground-truth annotations. A quantitative assessment and comparison of segmentation performance for the U-Net, U-Net++, Deeplabv3+, and ensemble models are provided in Table IV. We observe that using on-line augmentation of images and annotations in combination with model ensembling yields a strong performance gain over a baseline U-Net, increasing the Dice coefficient from 0.64 to 0.75 for the union of COVID-19 markers. The ensemble model yields a categorical Dice score of 0.65 (mean across the segmentations for score 0, 2 and 3). This metric was 0.47 for our baseline U-net.

In Fig. 6 we provide a visualization of uncertainty in the predicted segmentations for two example images by plotting the pixel-wise standard deviation yielded by MC dropout across 40 samples. Arrows in (A) indicate a region displaying COVID-19 markers for which ambiguity in the exact shape and extent are well reflected in the pixel-level uncertainty. Arrows in (B) indicate a seemingly false-positive region which was assessed as a high-grade COVID-19 marker by the deep network, and not annotated as such. Interestingly, retrospectively the network output was judged as a true positive by the annotators, showing an area of hyperechogenic lung below the pleural surface [12], which characterizes a high permeability and advanced disease state.

VI. Discussion and Conclusions

A. Frame-Based Score Prediction Evaluation

In Table I we ablate the contribution of the building blocks of our model for frame-based prediction. The replacement of the traditional cross-entropy (CE) with the SORD loss for ordinal regression clearly improves the performance. On the other hand, we found that the addition of STN leads to a drop in the F1-score because of the additional trainable parameters (as many as the CNN) introduced by the STN and the absence of a regularisation. However, STN comes with two positive side effects: (i) it provides weakly supervised localizations without using fine-grained supervision; and (ii) enables the use of consistency-based regularization, which is very beneficial in terms of performance. Our full model, which embeds the STN module, the SORD loss and the proposed consistency loss achieves an F1-score of 65.1, outperforming all the baselines by a large margin. To further investigate if the boost occurs because of the consistency term or the STN, we conducted an experiment using two sufficiently overlapping random crops and enforced consistency loss between the two. Unsurprisingly, the F1-score for CNN + Random Crop + SORD stays much below to our proposed method. We hypothesize that the consistency loss is only useful when the crops cover the area of the artefact.

In contrast to the previous work [18], we found that the use of more complex architectures like ResNet18 does not bring any positive improvement in performance. We reason that this is due to the low intrinsic complexity of the task. Conversely, we suggest that most of the confusion of the model is caused by the noise in both frames and labels. In turn, we believe that this noisiness is due to the subjectivity of the annotation and the presence of ambiguous frames. In fact, frame labels do not take into account that multiple artifacts can be present at a time. This happens mostly when the sensor is moving, causing a transitions from one score to another. In order to highlight the concentration of the errors of our models around transitions, we devise the experimental Setting 2, as shown in Table I, in which we drop frames close to transition points. The results in the Table I show that removing ambiguous frames from the test set dramatically reduces the amount of errors of the model, regardless of the architecture, empirically validating our hypothesis about noisy labeling.

In Table I we also measured how the subjectivity of the annotated scores affects the performance of the model in Setting 3 and discovered that when there is a strong agreement among doctors (more than 2 doctors agree on a score) our network performs notably better, increasing the F1-score by almost 3 points. This suggests that some videos are intrinsically more ambiguous than others. In addition, we found that, on this matter, the network seems to be behave similarly to human annotators, which is a desirable property. Moreover, although it seems counter-intuitive, our experiments point out that the performance of the model does not change much after a certain degree of agreement between doctors ($A = 3$ vs. $A = 4$). This is probably caused by the fact that imposing stronger agreement makes the test set smaller, yielding less statistically significant results.

Finally, we visualize the crops yielded by the STN and illustrate them in Fig. 4. We considered two kind of affine transformations modeled by the Reg-STN in our experiments: i) learnable translation with fixed scaling; and ii) learnable translation, scaling and rotation. We compute an F1-score of 65.9 when the STN models a learnable translation with fixed scaling. In both the cases the STN produces highly localized crops that mostly hinges around the area of pathological artifact. Interestingly, for both convex and linear sensors acquisitions, the Reg-STN learns to ignore the area above the pleura, which is essentially irrelevant for the prediction of a frame. This validates the usefulness of incorporating STN blocks in our frame-based predictor. We also report the heatmaps produced by GradCam [44] for the same images. Qualitatively, GradCam does not always focus on the relevant areas of the image. For example, for the linear probe image displayed in the figure, attention is given to the intercostal tissue layers and not to the areas of the image below the pleural-line, which are the areas of interest for the analsysis of LUS data. Also, we noticed that the quality of the heatmaps deteriorates when the prediction of the network is incorrect. Moreover, we found it hard to produce reasonable boxes from the heatmaps produced by GradCam, since it requires thresholding. For these reasons, we believe that STN produces superior localizations.

B. Video-Based Score Prediction Evaluation

When trained on the annotations by the most expert clinician, video-based classification achieves an F1-score of 61%, a precision of 70% and a recall of 60%. It is noticeable that these values are in line with the low inter-annotator agreement reported in Section III, which together to the small number of samples with video-level annotations can explain the high variance of the scores across folds. We expect that extending our relatively small set of video-level annotations will help counteracting the labeling noise, increase the model performance and reduce its variance.

C. Segmentation Evaluation

Our segmentation model is able to segment and discriminate between areas in B-mode LUS images that contain background, healthy markers and (different stages of) COVID-19 biomarkers at a pixel-level, reaching a pixel-wise accuracy of 96% and a binary Dice score of 0.75. Alongside these segmentations, we provide spatial uncertainty estimates that may be used to interpret model predictions.

Interestingly, and importantly, none of the highest (and most severe) score index annotations in the test set were missed by our model, judged by visual assessment of the resulting segmentations, and by analysing the relative image-level intersections among the corresponding predicted and annotated regions. Moreover, we observed model predictions of COVID-19-positive regions, that had however not been annotated as such. Fig. 6B shows a representative example of such a case. After re-evaluating some of such examples from the test set, together with the annotators, we learned that the annotators were sometimes unsure whether to annotate a

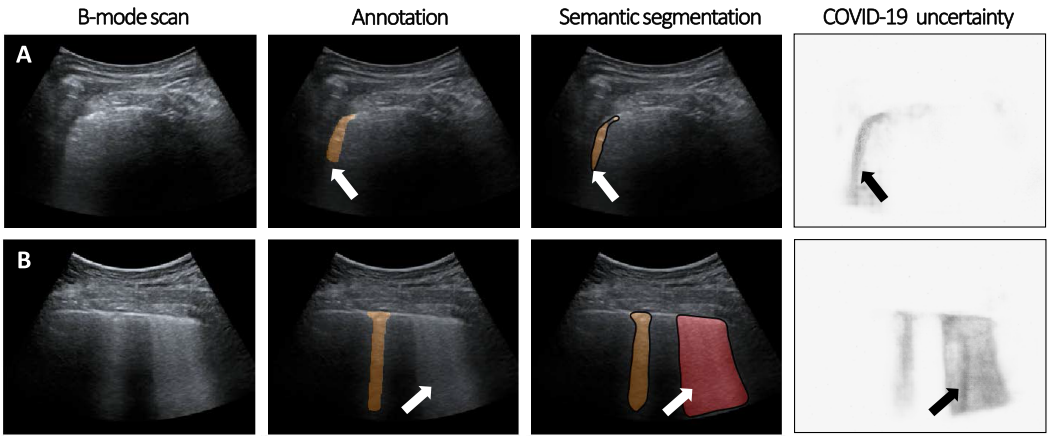

Fig. 6. Two examples (A, B) of class uncertainty in the segmentations, showing B-mode input image frames (first column), annotations (second column), including COVID-19 biomarkers (moderate/score 2: orange, severe/score 3: red), the corresponding semantic segmentations by deep learning (third column), and pixel-level COVID-19 class uncertainty by MC-dropout (fourth column).

region as e.g. score 2 or 3, and therefore decided that the marker was not clear enough to annotate the region at all, leading to the aforementioned discrepancy.

Segmentation performance and extraction of semantics could be further boosted by leveraging temporal structure among frames in a sequential model. Such models could learn from annotations across full videos, or through partial annotations and weak supervision. We leave these extensions of the present method to future work.

D. Limitations of the Dataset

In order to unravel the specific characteristics of this disease, researchers needed to gather as much data from patients as possible. However, due to the enormous impact and rapid spread of infected patients, data gathering in an organized manner proved a challenge. As a result, the precise demographics of the patient group in our database remain unknown.

Ideally, the dataset should be larger, more heterogeneous, and more balanced in term of scores in order to be used for learning accurate deep models. In our case, the data has been collected in a limited set of hospitals, all of them located in Italy. Furthermore, the way data was collected is prone to certain bias, e.g. due to a high patient inflow, the most severe patients were prioritized and assessed, and ultrasound diagnosis was performed on patients with a high clinical suspicion. No subsequent testing was done, resulting in the possible inclusion of false positive cases.

Labels in the ICLUS-DB turned out to be noisy. Furthermore, for frame-based classification and segmentation tasks the inter-operator agreement was not available. The noise can be indirectly observed in Table I, where using only a selection of training samples, performance improves by almost 5%. Extending the database to obtain frame-level labels from multiple annotators would surely lead to more robust models. Finally, the included LUS videos with score 0 are all of healthy patients, and therefore by no means we claim to distinguish between COVID-19 patients and those with different pathologies.

E. Possible Applications

A benefit of using ultrasound is the low risk of cross-infection when using a plastic disposable cover and individually packaged ultrasound gel on a portable handheld machine [45]. This is in contrast with use of CT, for which rooms and systems need to be rigorously cleaned to prevent contamination (and preferably reserved for patients with a high COVID-19 suspicion). LUS can be performed inside the patient's room without need of transportation, making it a superior method for point-of-care assessment of patients.

Moreover, ultrasound renders real-time images and, combined with our DL methods, provides results instantly. It may also directly assist in triage of patients; first-look estimation of the disease's severity and the urgency at which a patient needs to be addressed. In addition, low and middle-income countries, where diagnosis through RT-PCR or CT may not always be available, can particularly benefit from low-cost ultrasound imaging as well [46]. However lack of training on the interpretation of these LUS images [47] could still limit its use in practice. Our proposed DL method may therefore facilitate ultrasound imaging in these countries.

Acknowledgment

The authors would like to thank the Caritro Deep Learning Lab of ProM Facility who made available their GPUs for the current work. They also thank Fondazione VRT for financial support [COVID-19 CALL 2020 Grant #1].

References

[1] WHO. (2020). *Laboratory Testing Strategy Recommendations for COVID-19: Interim Guidance*. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/331509/WHO-COVID-19-lab_testing-2020.1-eng.pdf

[2] R. Niehus, P. M. de Salazar, A. Taylor, and M. Lipsitch, "Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers," medRxiv 2020.02.13.20022707, Feb. 2020.

[3] Y. Yang *et al.*, "Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections," medRxiv 2020.02.11.20021493, Feb. 2020.

[4] S. Salehi, A. Abedi, S. Balakrishnan, and A. Gholamrezanezhad, "Coronavirus disease 2019 (COVID-19): A systematic review of imaging findings in 919 patients," *Amer. J. Roentgenology*, pp. 1–7, Mar. 2020.

[5] A. Bernheim *et al.*, "Chest CT findings in coronavirus disease-19 (COVID-19): Relationship to duration of infection," *Radiology*, Feb. 2020, Art. no. 200463. [Online]. Available: http://pubs.rsna.org/doi/10.1148/radiol.2020200463

[6] F. Mojoli, B. Bouhemad, S. Mongodi, and D. Lichtenstein, "Lung ultrasound for critically ill patients," *Amer. J. Respiratory Crit. Care Med.*, vol. 199, pp. 701–714, Mar. 2019.

[7] R. Raheja, M. Brahmavar, D. Joshi, and D. Raman, "Application of lung ultrasound in critical care setting: A review," *Cureus*, vol. 11, no. 7, p. e5233, Jul. 2019.

[8] Y. Amatya, J. Rupp, F. M. Russell, J. Saunders, B. Bales, and D. R. House, "Diagnostic use of lung ultrasound compared to chest radiograph for suspected pneumonia in a resource-limited setting," *Int. J. Emergency Med.*, vol. 11, no. 1, p. 8, Dec. 2018.

[9] E. Poggiali *et al.*, "Can lung US help critical care clinicians in the early diagnosis of novel coronavirus (COVID-19) pneumonia?" *Radiology*, Mar. 2020, Art. no. 200847.

[10] Q.-Y. Peng, Chinese Critical Care Ultrasound Study Group, X.-T. Wang, and L.-N. Zhang, "Findings of lung ultrasonography of novel corona virus pneumonia during the 2019-2020 epidemic," *Intensive Care Med.*, vol. 46, no. 5, pp. 849–850, May 2020.

[11] G. Soldati *et al.*, "Is there a role for lung ultrasound during the COVID-19 pandemic?" *J. Ultrasound Med.*, Apr. 2020.

[12] G. Soldati *et al.*, "Proposal for international standardization of the use of lung ultrasound for patients with COVID-19: A simple, quantitative, reproducible method," *J. Ultrasound Med.*, Apr. 2020.

[13] K. Stefanidis *et al.*, "Lung sonography and recruitment in patients with early acute respiratory distress syndrome: A pilot study," *Crit. Care*, vol. 15, no. 4, p. R185, 2011.

[14] K. A. Stewart *et al.*, "Trends in ultrasound use in low and middle income countries: A systematic review," *Int. J. MCH AIDS*, vol. 9, no. 1, pp. 103–120, 2020.

[15] L. Tutino, G. Cianchi, F. Barbani, S. Batacchi, R. Cammelli, and A. Peris, "Time needed to achieve completeness and accuracy in bedside lung ultrasound reporting in intensive care unit," *Scandin. J. Trauma, Resuscitation Emergency Med.*, vol. 18, no. 1, p. 44, 2010.

[16] R. J. van Sloun, R. Cohen, and Y. C. Eldar, "Deep learning in ultrasound imaging," *Proc. IEEE*, vol. 108, no. 1, pp. 11–29, Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.02994

[17] R. J. G. van Sloun and L. Demi, "Localizing B-Lines in lung ultrasonography by weakly supervised deep learning, *in-vivo* results," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 4, pp. 957–964, Apr. 2020.

[18] G. Soldati *et al.*, "Towards computer aided lung ultrasound imaging for the management of patients affected by COVID-19," Tech. Rep.

[19] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NIPS*, 2015, pp. 2017–2025.

[20] S. Roy, A. Siarohin, E. Sangineto, S. R. Bulo, N. Sebe, and E. Ricci, "Unsupervised domain adaptation using feature-whitening and consensus loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9471–9480.

[21] R. Diaz and A. Marathe, "Soft labels for ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4738–4747.

[22] V. Melnikov and E. Hüllermeier, "Learning to aggregate using uninorms," in *Proc. ECML*, 2016, pp. 756–771.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[24] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-Rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: http://arxiv.org/abs/1711.05225

[25] D. Dong *et al.*, "The role of imaging in the detection and management of COVID-19: A review," *IEEE Rev. Biomed. Eng.*, early access, Apr. 27, 2020, doi: 10.1109/RBME.2020.2990959.

[26] F. Shi *et al.*, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, early access, Apr. 16, 2020, doi: 10.1109/RBME.2020.2987975.

[27] J. Chen *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: A prospective study," MedRxiv, Tech. Rep., 2020.

[28] S. Wang *et al.*, "A deep learning algorithm using ct images to screen for corona virus disease (COVID-19)," MedRxiv, Tech. Rep., 2020.

[29] X. Xu *et al.*, "Deep learning system to screen coronavirus disease 2019 pneumonia," 2020, *arXiv:2002.09334*. [Online]. Available: http://arxiv.org/abs/2002.09334

[30] S. Liu *et al.*, "Deep learning in medical ultrasound analysis: A review," *Engineering*, vol. 5, no. 2, pp. 261–275, Apr. 2019.

[31] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[32] G. Soldati *et al.*, "Simple, safe, same: Lung ultrasound for COVID-19 (LUSCOVID19)," ClinicalTrials.gov Identifier: NCT04322487, 2020.

[33] G. Soldati, M. Demi, R. Inchingolo, A. Smargiassi, and L. Demi, "On the physical basis of pulmonary sonographic interstitial syndrome," *J. Ultrasound Med.*, vol. 35, no. 10, pp. 2075–2086, Oct. 2016.

[34] K. Wada. (2016). *Labelme: Image Polygonal Annotation With Python*. [Online]. Available: https://github.com/wkentaro/labelme

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[36] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. NIPS*, 2016, pp. 1163–1171.

[37] C. Winship and R. D. Mare, "Regression models with ordinal variables," *Amer. Sociol. Rev.*, vol. 49, no. 4, p. 512, Aug. 1984.

[38] K. Crammer and Y. Singer, "Pranking with ranking," in *Proc. NIPS*, 2002, pp. 641–647.

[39] R. R. Yager and A. Rybalov, "Uninorm aggregation operators," *Fuzzy Sets Syst.*, vol. 80, no. 1, pp. 111–120, May 1996.

[40] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[41] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[43] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016, pp. 1050–1059.

[44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[45] J. C.-H. Cheung and K. N. Lam, "POCUS in COVID-19: Pearls and pitfalls," Tech. Rep., Apr. 2020.

[46] S. Sippel, K. Muruganandan, A. Levine, and S. Shah, "Review article: Use of ultrasound in the developing world," *Int. J. Emergency Med.*, vol. 4, no. 1, p. 72, Dec. 2011.

[47] S. Shah, B. A. Bellows, A. A. Adedipe, J. E. Totten, B. H. Backlund, and D. Sajed, "Perceived barriers in the use of ultrasound in developing countries," *Crit. Ultrasound J.*, vol. 7, no. 1, p. 11, Dec. 2015.