

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/349710246>

# A Framework for Integrating Domain Knowledge into Deep Networks for Lung Ultrasound, and its Applications to COVID-19

Preprint · March 2021

DOI: 10.13140/RG.2.2.19252.17289

CITATIONS

0

READS

742

12 authors, including:



**Oz Frank**

Weizmann Institute of Science

5 PUBLICATIONS 8 CITATIONS

[SEE PROFILE](#)



**Nir Schipper**

5 PUBLICATIONS 8 CITATIONS

[SEE PROFILE](#)



**Gino Soldati**

Valle del Serchio General Hospital

133 PUBLICATIONS 6,089 CITATIONS

[SEE PROFILE](#)



**Andrea Smargiassi**

Catholic University of the Sacred Heart

104 PUBLICATIONS 2,329 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Phd. Thesis [View project](#)



Lung Ultrasound Spectroscopy [View project](#)

# A Framework for Integrating Domain Knowledge into Deep Networks for Lung Ultrasound, and its Applications to COVID-19

Oz Frank, Nir Schipper, Mordehay Vaturi, Gino Soldati, Andrea Smargiassi, Riccardo Inchingolo, Tiziano Perrone, Federico Mento, Libertario Demi, *Member, IEEE*, Meirav Galun, Yonina C. Eldar, *Fellow, IEEE*, and Shai Bagon

Submitted on **March 1<sup>st</sup>**, 2021

**Abstract**—Lung ultrasound (LUS) is a cheap, safe and non-invasive imaging modality that can be performed at patient bed-side. However, to date LUS is not widely adopted due to lack of trained personnel required for interpreting the acquired LUS frames. In this work we propose a framework for training deep artificial neural networks for interpreting LUS, which may promote broader use of LUS. When using LUS to evaluate a patient's condition, both anatomical phenomena (e.g., the pleural line, presence of consolidations), as well as sonographic artifacts (such as A- and B-lines) are of importance. In our framework, we propose to provide a deep neural network not only the raw LUS frames as input, but explicitly inform it of these important anatomical features and artifacts in the form of additional channels containing pleural and vertical artifacts masks. By explicitly supplying this domain knowledge to deep models standard off-the-shelf neural networks can be rapidly and efficiently finetuned to perform well various tasks on LUS data, such as frame classification or semantic segmentation. Our framework allows for a unified treatment of LUS frames captured by either convex or linear probes. We evaluated our proposed framework on the task of COVID-19 severity assessment using the ICLUS dataset. In particular, we finetuned simple image classification models to predict per-frame COVID-19 severity score. We also trained a semantic segmentation model to predict per-pixel COVID-19 severity annotations. Using the combined raw LUS frames and the detected lines for both tasks, our off-the-shelf models performed better than complicated models specifically designed for these tasks, exemplifying the efficacy of our framework.

**Index Terms**—COVID-19, Deep Learning, Image Classification, Lung Ultrasound, Semantic Segmentation

This research was supported by the Weizmann Institute COVID-19 Fund, Miel de Botton and Jean and Terry de Gunzburg Coronavirus Research, Manya Igel Centre for Biomedical Engineering, the Carolito Stiftung, the European Institute of Technology (project Ultra On, EIT Digital 2020) and the Fondazione Valorizzazione Ricerca Trentina (grant 1, COVID-19 2020).

O. Frank, M. Galun, Y. C. Eldar and S. Bagon are with the Weizmann Institute of Science, and with the Weizmann Artificial Intelligence Center (WAIC), Rehovot, Israel (e-mail {first.last}@weizmann.ac.il). N. Schipper is with The Hebrew University of Jerusalem, Israel (e-mail nirschipper4@gmail.com). M. Vaturi is with the Sackler Faculty of Medicine, Tel Aviv University, Israel. G. Soldati is with the Valle del Serchio General Hospital, Italy. A. Smargiassi and R. Inchingolo are with the Fondazione Policlinico Universitario A. Gemelli IRCCS, Italy. T. Perrone is with the Fondazione IRCCS Policlinico San Matteo di Pavia, Italy. F. Mento and L. Demi are with University of Trento, Italy (e-mail: {first.last}@unitn.it).

O. Frank and N. Schipper contributed equally.

## I. INTRODUCTION

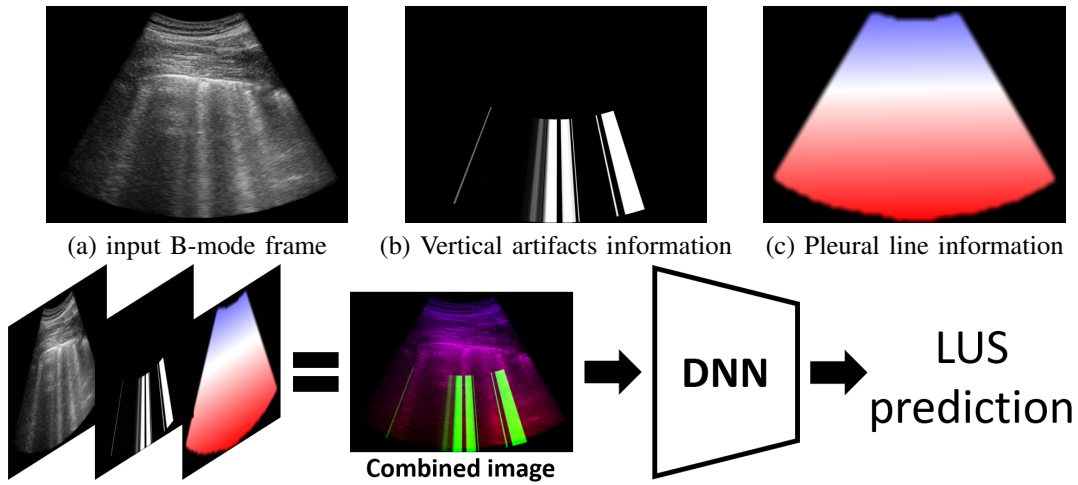
The diagnosis and treatment of respiratory diseases relies on the use of various imaging modalities. Chest CT is considered the imaging gold standard for pulmonary diseases [1], however, it is expensive, non-portable, and exposes the patients to ionising radiation. There is growing evidence showing that lung ultrasound (LUS) can be effectively used as an imaging modality for pulmonary diseases (e.g., [2]–[4]). LUS is a cheap, safe and non-invasive imaging modality that can be performed at patient bed-side.

The recent outbreak of COVID-19 pandemic drove clinicians to use LUS imaging also in emergency rooms [5]. Findings suggest LUS can assist in both detecting COVID-19 patients and monitoring their condition throughout their hospitalization [4], [6]–[10].

When using LUS to evaluate a patient's condition, both anatomical findings (e.g., presence of consolidations, the integrity of the pleural line [11]), as well as sonographic artifacts (such as A-lines and B-lines [12]) are of importance. Examples of these phenomena are shown in Fig. 5. However, spotting these findings and correctly interpreting them requires highly trained personnel. Consequently, to date, LUS is not widely adopted as its potential would reasonably suggest, particularly in the face of dire needs arising in treating patients with the COVID-19 pandemic.

Deep neural networks (DNN) and deep learning (DL) proved to be very powerful tools for accomplishing many challenging tasks, especially in the domain of image understanding. Given enough training examples (*>millions*) and computational resources, deep models can even exceed human performance on specific tasks (e.g. [13]–[15]). There is also growing work on applying DNNs to ultrasound imaging (see [16] and references therein). Nonetheless, when training data is hard to come by, as is often the case with medical imaging, it becomes more challenging to successfully train these complex models.

In this work we propose a framework for training DNNs for interpreting LUS that allows to effectively and efficiently train a DNN on LUS data, even when only several thousands of training examples are available. A similar approach, i.e., augmenting the raw input with additional masks, for analysing chest Xray of COVID-19 patients was proposed by [17]. To achieve this goal, we propose to explicitly enrich the input



**Fig. 1. Our framework for integrating domain knowledge into deep neural networks (DNN) for LUS.** Top: Input frame (a) is augmented with two additional channels containing LUS domain specific knowledge: (b) Automatically detected vertical artifacts (e.g., B-lines, “white lung”). (c) A signed distance mask from the pleural line. Bottom: The concatenation of these three channels (viewed as RGB image) are used as input for the DNN, enhancing the relevant frame regions.

to the model with domain specific knowledge. Specifically, we suggest to inform the model of important anatomical features and sonographic artifacts. We detect the pleural line and vertical artifacts (such as B-lines, “white lung” etc.) as a preprocessing stage. This automatically extracted domain-specific information is then fed, as additional input channels, much like RGB color channels in “natural images”, to a DL model alongside the raw LUS frame. These domain-specific channels allow the model to better tune and attend to relevant features and findings characteristic of this specific domain.

Fig. 1 illustrates our approach. Fig. 1 (top) shows an example of an input LUS frame and the automatically detected vertical artifacts and pleural line channels. The resulting concatenation of these masks and the raw input frame is then used as an input to a standard DNN model (bottom of Fig 1). Explicitly providing the model with this automatically extracted domain knowledge allows using simple off-the-shelf image classification neural network architectures, and rapidly and efficiently finetuning them to perform well on LUS data. Our framework allows to effectively and efficiently train DNN on LUS data, even when only several thousands of training examples are available. Moreover, our framework makes it feasible to train a *single* task-specific DNN model capable of handling LUS frames acquired by either convex or linear probes.

We demonstrate the efficacy of our framework on COVID-19 severity assessment, both on LUS frame classification as well as the task of semantic segmentation. We evaluated our proposed framework using the ICLUS dataset curated by the Ultrasound Laboratory Trento, Italy [18]. We finetuned simple image classification models on the combined raw LUS frames and the detected lines. We also finetune a semantic segmentation model to predict per-pixel COVID-19 annotations. Our finetuned off-the-shelf models performs better than complicated models specifically designed for these tasks.

To summarize, in this work we make the following contributions:

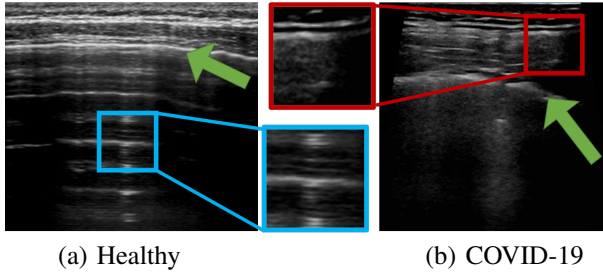
- (a) A widely applicable framework for incorporating LUS domain-specific knowledge into deep neural networks.
- (b) A unified framework capable of handling LUS frames acquired by either linear or convex probes.
- (c) Exceeding state of the art results on the ICLUS COVID-19 severity prediction benchmark, both for LUS frame classification and semantic segmentation.
- (d) A light-weight method for vertical artifacts detection.

This paper is organized as follows: the guiding principles of our framework are outlined in §II, while the specific details of our implementation are provided in §III. We exemplify the efficacy of our framework training DNNs to perform LUS frame classification in §IV and on the task of semantic segmentation in §V. We conclude in §VI.

## II. METHOD

LUS frames have a strong artefactual nature. The pleural line partitions the frame into two parts: the top part showing the exterior tissue and the bottom part showing the aerated lung cavity. Due to the dramatic difference in their acoustical properties, these two regions appear quite differently in LUS. Moreover, this change in acoustical conditions gives rise to sonographic artefacts such as A-lines, B-lines, “white lung” etc. When interpreting LUS one needs to take these unique characteristics into account: For instance, bright horizontal lines can be A-lines if they are *under* the pleural line (Fig. 2 blue), but may account for a completely different findings if they are observed *above* the pleural line (Fig. 2 red), leading to a radically different interpretation. While the presence of A-lines usually suggests healthy condition (Fig. 2a), observing these bright lines above the pleural line may lead to erroneous assessment for the COVID-19 patient in Fig. 2b.

DNN architectures designed for image analysis tasks are oblivious to these idiosyncrasies of LUS. Our framework proposes to make DNNs aware of these idiosyncrasies not by changing their design, but rather by highlighting relevant and



**Fig. 2. Relative location and interpretation.** Visually similar regions in LUS frames may account for very different findings if located above the pleural line (green arrow) or below it. For instance, the blue region below the pleural line shows A-lines, while the red region above the pleural line shows muscle tissue.

salient features via preprocessing of the input LUS frames. Fig. 1 illustrates our approach.

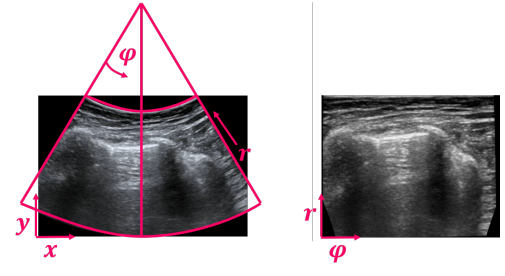
As noted, the pleural line plays a key role in LUS. Therefore, we choose to encode the pleural line information by measuring the *signed* distance of each pixel to the pleural line: negative distance above the pleural line (Fig 1c, blue shade), and positive below (Fig 1c, red shade). Integrating the signed distance into the input channels allows the network to trivially distinguish between the exterior tissue (upper part of the frame, negative signed distance), the lung cavity (lower part of the frame, positive signed distance), and the pleural line region (small distance). Another important phenomenon unique to LUS are vertical artifacts, such as B-lines and “white lung”, indicating loss of aeration of the lung. Their presence usually indicate a pathological condition. We therefore add another channel with a segmentation mask indicating possible locations of vertical artifacts (Fig. 1b). We concatenate the two masks as the second and third channels on top of the original gray-level channel of the input LUS frame. The resulting 3-channel image efficiently encodes LUS-specific information allowing for training DNN models to perform LUS-specific predictions as shown in the bottom of Fig. 1. This multi-channel representation is applicable for LUS frames obtained by either convex or linear probes, allowing to train a *single* task specific DNN capable of handling convex as well as linear frames.

To summarize, our framework preprocesses a LUS frame into a three channel image, similar in structure to an RGB natural image. Therefore, we can take existing DNNs trained on natural 3-channel RGB images (e.g., [19]) and finetune these DNNs to the 3-channel input LUS frames generated by our framework, consisting of the following steps:

*Preprocessing stage:* extract vertical artifacts and pleural line masks from LUS frames, and combine them as additional input channels.

*DNN training stage:* train a DNN on the combined 3-channel input frames.

The proposed framework for processing LUS is quite general: It may be applicable to many LUS frame analysis applications and tasks. It does not dictate the use of any specific DNN architecture. Moreover, it handles LUS frames obtained by either convex or linear probes in a unified manner. Indeed, we demonstrate our framework on two different tasks:



(a) Input convex frame (b) Rectified frame

**Fig. 3. Rectifying convex frames.** (a) Original frame in Cartesian  $x$ - $y$  coordinates and the induced polar  $r$ - $\varphi$  coordinates. (b) The rectified frame according to its polar coordinate system. The transformation from one coordinate system to the other is invertible.

classification and semantic segmentation of COVID-19 severity assessment. Moreover, for each task we train *one* DNN model for both convex and linear frames.

### III. IMPLEMENTATION DETAILS

In the previous section we outlined the concepts on which our framework is based. This section describes a practical implementation of our framework.

#### A. Vertical artifacts estimation

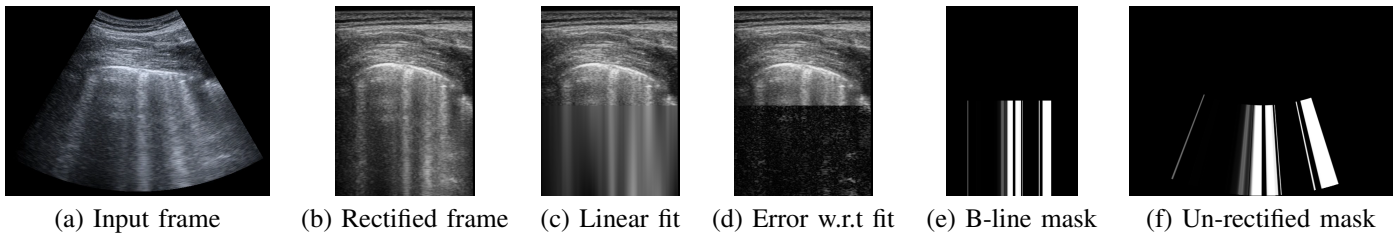
Robustly and accurately detecting vertical artifacts (i.e., B-lines) is a challenging task [12], [20], [21]. Existing approaches were either probe-type specific [20], [21] or cumbersome requiring LUS videos [22]. Here, we do not aim to perfectly solve it, but only to steer the downstream DNN model in the right direction. Therefore, we resort to a simplistic approach, that is fast, unsupervised and provides an informative, albeit noisy, estimation of vertical artifacts.

According to recent developments in LUS [23], [24], vertical artifacts are sonographic signs caused by complex interaction of the multiple scattering phenomena that may form in the presence of an alteration occurring at the lung surface. When forming the LUS frame, the ultrasound signals produced by the multiple scattering events, in case of resonance phenomena, are then interpreted as a bright-vertical-line emitting from the pleural line and aligned along the ultrasound beam axis.

Our approach is based on the observation that the orientation of the vertical artifacts is known and depends only on the probe type used. For a linear probe, these artifacts are exactly vertical. For convex probes, the orientation of these artifacts depends on their polar coordinate: the further away they are from the center of the frame, the more tilted towards the outside they are. Consequently, if we rectify a convex frame according to the polar coordinates induced by the convex probe (see Fig. 3) all vertical artifacts will become strictly vertical, as in frames captured using linear probes. This phenomenon may be observed in Fig. 4a, where the orientation of the vertical artifacts vary according to their polar coordinate. Fig. 4b shows the rectified frame. Note how all these artifacts are now vertical.

The canonical orientation of the vertical artifacts allows us to detect them quite easily. We look for columns at the lower





**Fig. 4. Detecting vertical artifacts as bright columns.** (a) Input. (b) Rectified convex frame according to its polar coordinates (Fig. 3). (c) Fit of the intensities of the lower half of each column with a linear function:  $I_x[y] = a_x \cdot y + b_x$ . (d) Error between the actual intensity and the linear fit. B-lines, “white lung” and similar vertical artifacts have low error. (e) Columns whose linear fit is above threshold  $\tau_{bright}$  and the error is below threshold  $\tau_{err}$  are marked as vertical artifacts. (f) Un-rectify convex frames back to their Cartesian coordinates.

half of the frame that are bright and have relatively low noise (no “speckles”). To find such columns, first, for each column,  $x$ , we fit its intensities,  $I_{xy}$ , with a linear function of the vertical coordinate,  $y$ :  $I_x[y] = a_x \cdot y + b_x$ . Fig. 4c shows the linear fit for each column. We then compute the error between the fit and the actual pixel value,  $\sum_y |a_x \cdot y + b_x - I_{xy}|$  (Fig. 4d). Vertical artifacts have relatively low noise, as opposed to consolidations and speckles. Finally, all columns whose linear fit is above threshold  $\tau_{bright}$  and the error is below threshold  $\tau_{err}$  are marked as vertical artifacts (Fig 4e). The resulting binary mask is then transformed back from polar coordinates to the original Cartesian coordinates to form the vertical artifacts mask for the input frame (Fig. 4). Note that for frames captured using linear probes we do not need to change to polar coordinates, and can do the same processing in the original Cartesian coordinate system.

### B. Pleural line detection

The pleural line is a bright thin line that roughly crosses the frame from side to side (See Fig. 2 and Fig. 5 green). A straight forward approach for pleural line detection was proposed in [20]. They use the Radon transform of the frame and apply an iterative optimization process to locate a distinct and significant bright horizontal line in the Radon space.

Once we obtain the location and orientation of the pleural line from [20] we applied a signed distance transform. We further scale the signed distance such that the distance from the pleural line to the bottom of the frame is equal 1. An example of a pleural line mask is shown in Fig 1c: negative distance at the top part of the frame and positive distance at the bottom.

Both approaches for vertical artifacts and pleural line detection are simple but inexact. Nevertheless, the information they extract is introduced in a “soft” manner to the network via additional input channels. This way the network is trained to reason with this noisy data and distil meaningful cues from it to facilitate better performance on any target task.

### C. DNN Models

Our framework is not restricted to any specific deep neural network architecture. In fact, working with three input channels (i.e., the raw frame, vertical artifacts mask and pleural signed-distance), allows us to use models trained on natural

3-channel RGB images “as-is”. This gives us the flexibility to opt for large models (e.g., ResNet-18 [25]) when accuracy is of importance, or trade it for a light-weighted model (e.g., MobileNetV2 [26]) when computing resources are scarce. For the semantic segmentation task we used DeepLabV3++ [27] model.

All these models were pre-trained on natural RGB images to perform image classification [19] or semantic segmentation [28], and their trained weights are readily available on-line. Once we choose our model, we can use its pre-trained weights except of the last task-specific prediction layer that needs to be trained from scratch.

### D. Finetuning the models

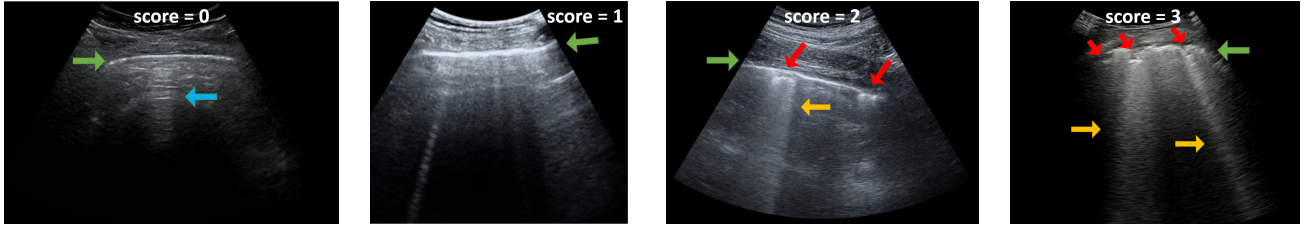
To make our trained model more robust to small changes in the input frames we applied various augmentations to the training data. The set of augmentation functions, each applied with a randomly sampled strength bounded by a set maximum, consists of: affine transformations (translation (max.  $\pm 10\%$ ), rotation (max.  $\pm 23^\circ$ ), scaling (max. 10%)), horizontal flipping (50%). We additionally applied random jittering to the raw gray-level frame channel (contrast (max  $\pm 30\%$ ), brightness (max  $\pm 30\%$ )). To encourage the model to be more robust to the exact location of the pleural line we applied a random global shift to the signed distance channel (max  $\pm 8\%$ ). We used the same augmentations for the image classification and the semantic segmentation tasks.

We implemented our framework using pytorch [29] and used the supplied `torchvision.models` and their pre-trained weights. We finetuned the models using Adam optimizer [30]. We used a fixed learning rate of  $\lambda = 0.0075$  ( $1e-4$  for the semantic segmentation task) and default values  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ .

We used the same loss functions as [18]: the Soft ORdinal (SORD) loss for the classification task and cross-entropy loss for the semantic segmentation task.

## IV. COVID-19 SEVERITY GRADING RESULTS

We evaluated our proposed framework on the task of COVID-19 severity grading. It has been recently shown that LUS can be used for stratification and monitoring of patients with COVID-19 [4], [8]. Soldati *et al.* [4] proposed a 4-level scoring system with scores ranging from 0 to 3. Score 0 indicates a healthy lung characterised by a continuous pleural-line and visible A-lines artifacts. In contrast, score 1



**Fig. 5. COVID-19 Severity scores.** LUS frames exemplifying the severity score of [4] from healthy (score=0, left) to severe (score=3, right). One can observe the pleural line (green), A-lines (blue), subpleural consolidations (red) and vertical artifacts (e.g., B-lines and “white lung”) (yellow). While the pleural line and consolidations are anatomical features, the A-lines and vertical artifacts are sonographic echoes.

indicates first signs of abnormality mostly related to small alterations in the pleural-line, and the appearance of few vertical artifacts. Scores 2 and 3 are representative of a more advanced pathological state, with the presence of small or large consolidations, respectively, and significant presence of vertical artifacts (B-lines and “white lung”). Fig. 5 shows example frames representative of each score. This scoring system is the only imaging protocol and scoring system specific to COVID-19. It has been validated against other imaging protocols [9] and there is evidence of its prognostic value [10].

We use our framework to train a deep model to classify LUS frames to their annotated COVID-19 severity score, and compare the performance of models trained in our framework to previously published results.

#### A. Data

We use the ICLUS dataset [18] curated by the Ultrasound laboratory in Trento, Italy<sup>1</sup>. The ICLUS dataset contains 277 LUS videos of 35 patients, with total of 58,924 frames, out of which 45,560 frames acquired using convex probes and 13,364 frames acquired using linear probes. All frames in the ICLUS dataset were carefully and manually annotated into one of the four severity scores. The annotations were additionally verified by expert clinicians. Nevertheless, Roy *et al.* [18] reported only 67% agreement across annotators per LUS video, emphasising the difficulty of the task. The ICLUS dataset is then split into a train and test set, with the test set comprising of 10,709 frames. The split is performed at patient level: data from any patient is either in the train or test set, but not in both. More details about the ICLUS dataset can be found in [18].

As a pre-processing stage, we computed vertical artifacts masks and pleural line signed distance masks, as described in §III-A and §III-B. We concatenate these two masks to the original input frame to form an input tensor with 3 channels.

#### B. Results

We used our framework to finetune a ResNet-18 model to classify each frame to its annotated severity score. We measured the performance of models trained using our framework in terms of F1-score, which is the harmonic mean of precision and accuracy. Similar to [18] we used two settings: *Setting 1* considers the F1 score computed on the entire test set. *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, thereby allowing us to identify the impact of noisy labeling on the performance of the model. Table I compares our results (3<sup>rd</sup> row) to the results of Roy *et al.* [18] (top rows) using the two settings.

Using the same ResNet-18 architecture our model achieved F1=68.8% score compared to F1=62.2% achieved by the same architecture, but without the explicit use of the vertical artifacts and pleural line information. Moreover, our ResNet-18 model outperforms the CNN-Reg-STN architecture proposed by [18] that was specifically designed to cope with the idiosyncrasies of LUS data. We see that it is more advantageous to incorporate domain knowledge as additional input channels than as deep neural architecture designs. We also presented results of F1=68.7% on this task in [31]. However this required an elaborate use of an ensemble of task specific ResNet-18 models.

We further used the GradCAM method [32] to visually inspect the predictions of ResNet-18 models. Fig. 6 shows a visual comparison of correct classifications by our framework to misclassifications of the baseline of [18]. That is, using the same deep neural architecture (ResNet-18), our model uses all three input channels whereas the baseline [18] uses only the raw frame. The first column shows a frame captured by a linear probe of a healthy patient (score=0). Our model (second row) attends well to the clear region below the pleural line and to the A-lines shown on the left part of the frame. Note that despite the thin vertical artifact falsely detected in the frame, the trained model was able to compensate and ignore it. In contrast, the baseline model falsely predicts severity score=2 and attends to irrelevant regions as the exterior tissue at the top of the frame or the void at the bottom. The second column shows a frame with score=1, misclassified by the baseline as score=0. Our model attends to the pleural line region, which holds important information for the score=1 class. In the third column is a frame labeled as score=2. The baseline model misclassified it as score=3. The GradCAM visualization shows how our model focuses on the vertical artifacts – thanks to the focused input mask, while the baseline model “spreads” all over the bottom part of the frame. Visualizing results for the most severe case, with score=3, on the fourth column, we see that our model successfully focuses on the wide “white lung” region below the pleural line. In contrast, the baseline

<sup>1</sup><https://iclus-web.bluetensor.ai/>

TABLE I  
QUANTITATIVE RESULTS: F1 SCORES FOR PER-FRAME COVID-19 SEVERITY CLASSIFICATION.

Model	Input channels			Settings 1 All Frames	Settings 2 Drop Transition Frames (K)			
					K=1	K=3	K=5	K=7
ResNet-18 (Roy et al)	✓			62.2	63.9	65.5	66.9	67.8
CNN-Reg-STN (Roy et al)	✓			65.1	66.7	68.3	69.5	70.3
<b>Resnet-18 (ours)</b>	✓	✓	✓	<b>68.8</b>	<b>70.2</b>	<b>72.4</b>	<b>73.9</b>	<b>75.2</b>
Ablation	ResNet-18	✓		64.5	65.7	67.6	69.0	70.1
	ResNet-18	✓	✓	64.7	66.1	68.5	70.3	71.8
	ResNet-18		✓	45.1	45.7	46.7	47.4	47.6

model attends to irrelevant regions above the pleural line and thus misclassify this frame as score=2.

These visualizations suggests that our framework, namely, providing a model for LUS analysis with additional pleural line and vertical artifacts masks, is able to effectively steer the model to the relevant regions of the frame: It is able to inspect the pleural line and suspicious regions inside the lung cavity, while paying less attention to the tissue above the pleural line.

### C. Ablation study

*Input masks:* To show the complementary nature of the two additional input channels, we performed an ablation study. We used the same deep neural architecture, namely ResNet-18, and trained it using different combinations of input channels: Once with only the vertical artifacts masks and once with only the pleural line mask. The bottom part of Table I shows the performance of these trained ResNet-18 models. Adding only one additional channel (either vertical artifacts or pleural line channel) helps to increase performance by  $\sim 2\%$  showing that these channels do contain useful information and the model is able to take advantage of it. However, when combining both channels (3<sup>rd</sup> row in Table I) performance increase dramatically to F1=68.8%, significantly exceeding previous methods. Nevertheless, these additional channels cannot replace the raw input frames completely, as suggested by the bottom row of the table. Discarding the raw input frame significantly degrades performance to merely F1=45.1%.

In their work on LUS classification [33] also considered incorporating domain knowledge using relevant semantic segmentation maps as inputs. However, their VGG-Seg model performs worse than their baseline VGG despite the additional semantic information. It seems that overriding the raw input channel in favour of semantic information prevents the model from compensating for inevitable inaccuracies in the input masks. In contrast, we leave the raw LUS frame intact as one of the input channels and only *augment* it with additional domain specific knowledge.

TABLE II  
CHOICE OF DNN ARCHITECTURE: F1 SCORES FOR PER-FRAME COVID-19 SEVERITY CLASSIFICATION USING DIFFERENT DNN ARCHITECTURES.

Model	Settings 1	Settings 2 Drop Transition Frames (K)			
	All Frames	K=1	K=3	K=5	K=7
Resnet-18 (Roy et al)	62.2	63.9	65.5	66.9	67.8
CNN-Reg-STN (Roy et al)	65.1	66.7	68.3	69.5	70.3
<b>Resnet-18 (ours)</b>	<b>68.8</b>	<b>70.2</b>	<b>72.4</b>	<b>73.9</b>	<b>75.2</b>
VGG16 (ours)	65.2	66.7	68.9	70.9	72.4
MobileNetV2 (ours)	67.1	68.3	69.9	71.1	71.9

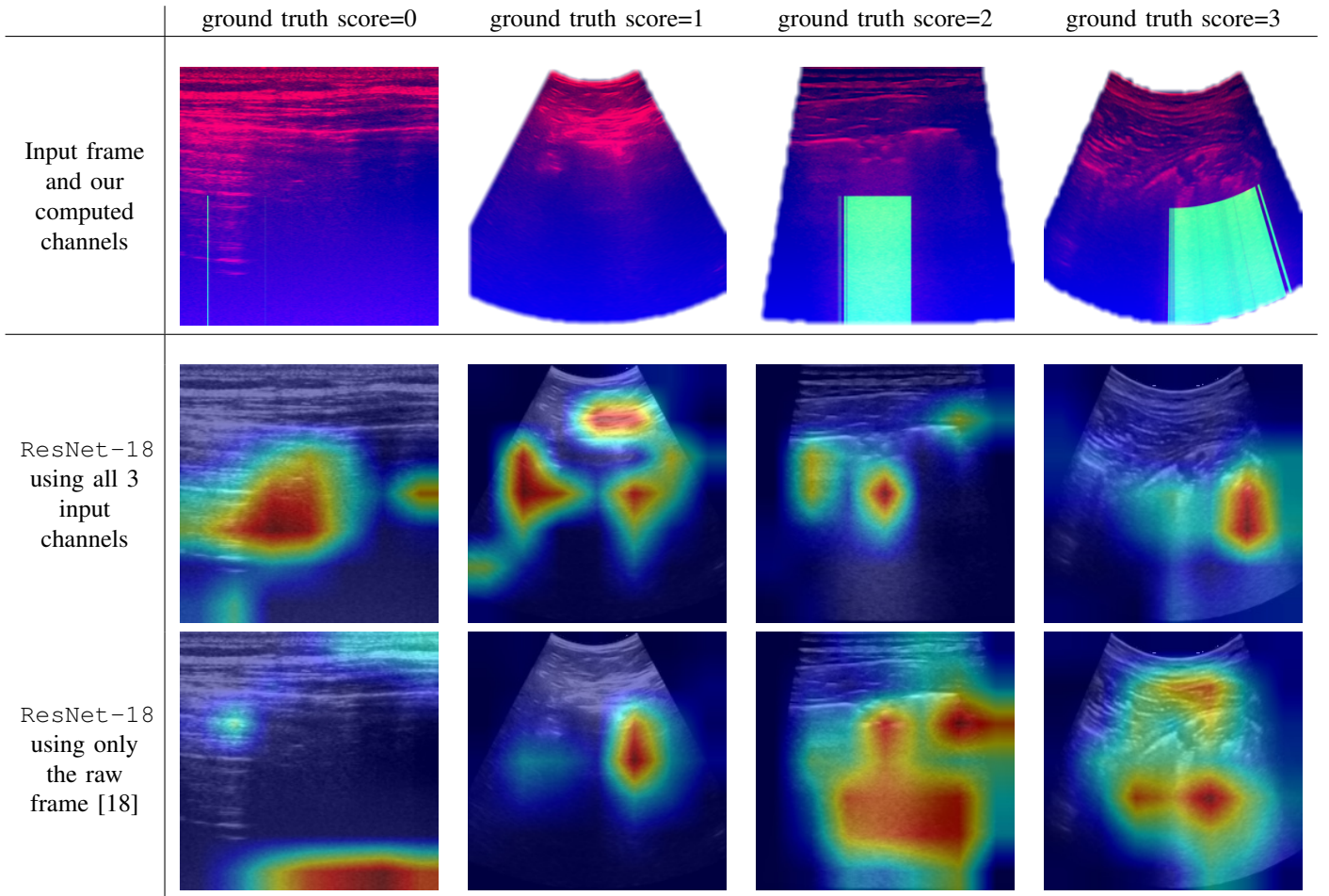
*Choice of backbone:* Our proposed framework for training DNN models for LUS data is not restricted to any specific DNN architecture, but extends to almost any DL architecture. We further experimented with different DNN architectures for the task of COVID-19 severity classification within our framework. We compared a light-weighted MobileNetV2 [26], and also the classic VGG16 [34] architectures. Table II shows the performance of these models. Despite the fact that these are “general-purpose” image classification architectures that were not tailored to the idiosyncrasies of LUS, they perform on-par and even better than the specifically designed architecture of [18].

It is worthwhile noting that even the light-weighted MobileNetV2 model performs very well. This is in contrast to the findings of [33] that reported a significant drop in performance when using a “mobile” architecture for LUS classification. When having no access to domain knowledge thin mobile architecture are indeed likely to fail to extract meaningful features from the raw LUS frames. In contrast, in our framework, domain knowledge is easily accessible via the additional input channels making it easy to exploit even for light-weighted models.

### D. Vertical artifacts masks

Our approach for vertical artifacts detection is extremely simple yet, our experiments suggest it is still effectively





**Fig. 6. Visualizing predictions using GradCAM.** visualizing regions in the frame that influence most the model's prediction. Top row: Overlay of the raw input frame and our estimated pleural line and vertical artifacts masks. Middle row: visualization of correct classifications by our framework – when all input masks are used. Bottom row: visualization of misclassifications by [18] – the same DNN architecture when only the raw input frame is used without the additional input masks.

guiding the downstream model in the right directions.

To get a sense of how informative our masks are we measured the relative width our detected vertical artifacts span, per frame, in the ICLUS dataset. To avoid noisy frame labeling we used the restricted test set of *Settings 2* with  $K = 7$ . Recall that [4] defined the scores such that when vertical artifacts are detected it usually suggests that the frame should not be scored 0, and the wider these artifacts the more severe the condition is. Table III shows the relative width of vertical artifacts detected by our method per-frame. We can see that there is a distinction between frames with score=0,1 and those with score=2,3. However, the variance is quite large, highlighting the limitations of our simplistic approach.

Since B-lines visualization is strongly dependent on the imaging settings and utilized hardware, traditional severity assessment by *counting* B-lines tends to be very subjective, compared to relative span suggested by [4]. Nevertheless, Tab. III also reports counts of vertical artifacts that leads to similar conclusions.

Our simple approach aims at highlighting any vertical artifacts, not just B-lines, consequently, using the Radon-based method of [20], [21] for B-lines detection, attained only  $F1=67.6\%$  [35].

**TABLE III**

DISTRIBUTION OF DETECTED VERTICAL ARTIFACTS ACCORDING TO COVID-19 SEVERITY SCORE OF [4]. THE TABLE SHOWS THE RELATIVE FRAME WIDTH PER FRAME (STD) THE DETECTED VERTICAL ARTIFACTS SPAN, AND THEIR AVERAGE NUMBER. SCORE=0 IS HEALTHY WHILE SCORE=3 INDICATES ACUTE RESPIRATORY CONDITION. VERTICAL ARTIFACTS USUALLY INDICATE A PATHOLOGICAL CONDITION OF THE LUNG, AND THUS APPEAR MORE AS THE SCORE INCREASES.

score	0	1	2	3
<b>Width [%]</b>	<b><math>0.69 \pm 6.0</math></b>	<b><math>0.05 \pm 0.48</math></b>	<b><math>3.27 \pm 6.89</math></b>	<b><math>13.86 \pm 12.14</math></b>
<b># B-lines</b>	$0.11 \pm 1.41$	$0.10 \pm 1.65$	$1.57 \pm 4.40$	$7.36 \pm 6.94$

## V. SEMANTIC SEGMENTATION OF COVID-19 MARKERS

To further demonstrate the wide applicability of our framework, we tested it on a different type of task: training a semantic segmentation DNN to segment LUS frames.




### A. Data

In addition to per-frame COVID-19 severity score annotations, the ICLUS dataset [18] provides detailed pixel-level annotations for the biomarkers indicative of each score. These detailed semantic annotations were provided for 2,154 frames across 33 patient, of which 1,602 frames were captured using



TABLE IV

**SEMANTIC SEGMENTATION RESULTS:** SEGMENTATION PERFORMANCE MEASURED BY ACCURACY ACROSS ALL CATEGORIES (ACC.), THE DICE COEFFICIENT FOR THE UNION OF COVID-19 RELATED SCORED (DICE), AND THE MEAN DICE ACROSS SCORES 0, 2 AND 3 (CAT. DICE) AS IN [18]. ACCURACY AND DICE SCORES ARE HEAVILY BIASED TOWARDS THE DOMINANT BACKGROUND CLASS, WHILE CAT. DICE REFLECTS BETTER THE PERFORMANCE ON THE RELEVANT ANNOTATED PIXELS.

Model		Input			Acc.	Dice	Cat. Dice
							
DeepLabV3++* (Roy et al)		✓			0.95	0.71	<b>0.62</b>
Ensemble* (Roy et al)		✓			0.96	0.75	<b>0.65</b>
DeepLabV3++ (ours)		✓	✓	✓	0.93	0.76	<b>0.70</b>
Ablation	DeepLabV3++	✓			0.92	0.72	<b>0.64</b>
	DeepLabV3++	✓		✓	0.93	0.74	<b>0.70</b>
	DeepLabV3++	✓	✓		0.93	0.74	<b>0.68</b>

\* Note that Roy *et al.* [18] trained and evaluated only on convex frames, while our method used both linear and convex.

convex probes and 552 using linear probes (note that we are using an updated view of the ICLUS dataset with more annotations compared to [18]). We further split the data into train and test sets according to the same patient-level split used in §IV with 1,601 (1,237 convex, 364 linear) training frames. For the frames in the training set, relative pixel-level occurrences for score 0, 1, 2 and 3 are 3.8%, 0.1%, 2.0% and 3.2% respectively. For the test set the relative occurrences are 5.6%, 0.2%, 1.6% and 4.7% respectively. Notably almost 90% of the pixels do not display clear characteristics of any specific class (severity score) and are, therefore, labeled as background. Unlike [18] we treat pixels outside the LUS scan as “ignore” and discard them completely from the training and evaluation (e.g., gray pixels in Fig. 7d).

## B. Results

We used our framework to finetune a DeepLabV3++ semantic segmentation model that was pre-trained on a subset of MS-COCO dataset [28] to predict a pre-pixel severity score according to the semantic annotation of the ICLUS dataset. Table IV shows the segmentation performance of our trained model measured, similar to [18], by accuracy across all categories (Acc.), the Dice coefficient for the union of COVID-19 related scored vs. background (Dice), and the mean Dice across scores 0, 2 and 3 (Cat. Dice). Score=1 is omitted due to its under representation in the pixel annotations. It is worthwhile noting that due to the disproportionate size of the background class (~90%) the accuracy and the Dice measures are quite biased and do not reflect well the performance of the model on the relevant COVID-19 classes. In contrast, the category Dice (Cat. Dice) measure focuses on the relevant labels and reflects better the performance of the model on this specific task.

We train the same DeepLabV3++ architecture, once using our framework with all input channels (e.g., the raw frame, vertical artifacts and pleural-line channels) and once using only the raw input frame as in [18]. We trained and evaluated on frames obtained from both convex and linear probes. Adding

the linear frames slightly improves the Cat. Dice from 0.62 reported by [18] to 0.64 (1<sup>st</sup> and 4<sup>th</sup> rows of Tab. IV). However, adding the additional input channels boost the Cat. Dice score even further to 0.70 (3<sup>rd</sup> row of Tab. IV).

Fig. 7 shows four examples of input frames, the corresponding predicted segmentation masks and the ground truth annotations. The color of the segments indicate their corresponding COVID-19 severity score from blue for score=0 through yellow, and orange all the way to red corresponding to scores 1, 2 and 3 respectively. Note how our trained model is capable of handling both linear (first two rows) as well as convex (bottom rows) frames. Providing a model with the domain specific knowledge in the form of the location of the pleural line allows it to better detect small discontinuities corresponding to the challenging score=1 category (first two rows). Explicitly providing the model with information on vertical artifacts allows it to better classify “white lung” regions as score=3 rather than score=2 (4<sup>th</sup> row).

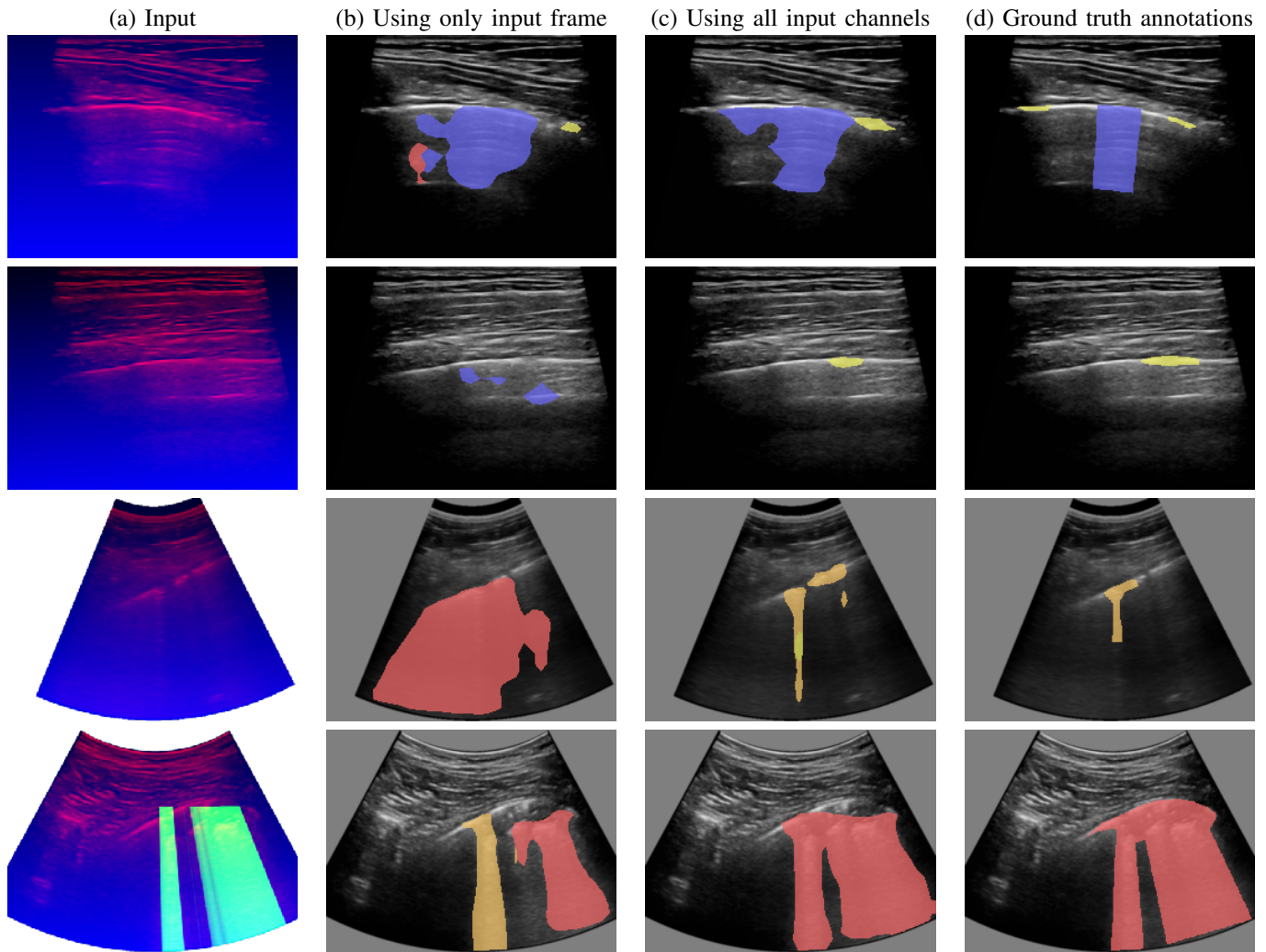
## C. Ablation study

To show the impact of the two additional input channels, we performed an ablation study. We used the same deep neural architecture, namely DeepLabV3++, and trained it using different combinations of input channels: Once with only the vertical artifacts masks and once with only the pleural line mask. The bottom two rows of Table IV show the performance of these two trained DeepLabV3++ models. Adding only the pleural line channel already increase the Cat. Dice score to 0.70 leaving only a difference of 0.02 on the Dice score compared to the model trained using all three channels, showing the power of pleural-line location information for the localization of relevant biomarkers. In contrast, adding only the vertical artifacts channel resulted with a less prominent performance gain: Cat. Dice increased to only 0.68. This is probably due to the inaccuracies in B-line detection of our method already discussed in §IV-D.

## VI. CONCLUSION

In this work we introduced a framework for combining the power of deep neural networks with prior domain knowledge specific to LUS resulting in a fast and efficient way of training DNNs on LUS data. The key insight is to explicitly provide domain-specific knowledge to the models. However, instead of incorporating this knowledge in the form of elaborate and task-specific DNN architecture design, we propose to introduce it as part of the input data. In the context of LUS we demonstrated that informing the model of the location of the pleural line and the presence of vertical artifacts, such as B-lines and “white lung”, can significantly improve performance. Moreover, it allows to treat frames captured by either linear or convex probes in a unified manner – using a *single* DNN for both types of frames. We exemplified the applicability of our framework on COVID-19 severity assessment, both on the task of LUS frame classification as well as the task of LUS semantic segmentation.

Although our masks do not add any external information that is not already “in the pixels”, explicitly extracting it for the



**Fig. 7. Semantic segmentation:** Pixels indicating score=0 are annotated **blue**, score=1 in **yellow**, score=2 in **orange** and score=3 are annotated **red**. Note that the gray pixels outside the LUS scan are ignored. (a) Input frame and the additional channels computed by our framework. (b) Semantic segmentation results of DeepLabV3++ model utilizing only the raw input frames (as in [18]), Cat. Dice=0.64. (c) Segmentation results of the same DeepLabV3++ architecture utilizing all three input channels, Cat. Dice=0.70. (d) Ground truth annotations.

network to use, makes the finetuning process efficient, quick and robust, and can be done on much smaller datasets. We postulate that it would require significantly longer training time and substantially more labeled training examples for deep networks to *automatically* distill this specific domain knowledge directly from the “raw” pixels.

Our proposed framework is, therefore, more widely applicable to LUS than COVID-19 severity prediction. The framework can be used not only to train different DL architectures, but to address other challenges in the analysis of LUS frames. Moreover, our framework does not rely on any *specific* implementation of pleural line or vertical artifacts detection algorithms and allows for incorporating other domain knowledge e.g., A-lines, and more, in a similar manner.

## REFERENCES

- [1] J.-E. Bourcier, J. Paquet, M. Seinger, E. Gallard, J.-P. Redonnet, F. Chedadi, D. Garnier, J.-M. Bourgeois, and T. Geeraerts, “Performance comparison of lung ultrasound and chest x-ray for the diagnosis of pneumonia in the ed,” *The American journal of emergency medicine*, vol. 32, no. 2, pp. 115–118, 2014.
- [2] D. Lichtenstein, I. Goldstein, E. Mourgeon, P. Cluzel, P. Grenier, and J.-J. Rouby, “Comparative diagnostic performances of auscultation, chest radiography, and lung ultrasonography in acute respiratory distress syndrome,” *The Journal of the American Society of Anesthesiologists*, vol. 100, no. 1, pp. 9–15, 2004.
- [3] A. Smargiassi, G. Soldati, E. Torri, F. Mento, D. Milardi, P. D. Giacomo, G. De Matteis, M. L. Burzo, A. R. Larici, M. Pompili *et al.*, “Lung ultrasound for COVID-19 patchy pneumonia: extended or limited evaluations?” *Journal of Ultrasound in Medicine*, 2020.
- [4] G. Soldati, A. Smargiassi, R. Inchingolo, D. Buonsenso, T. Perrone, D. F. Briganti, S. Perlini, E. Torri, A. Mariani, E. E. Mossolani *et al.*, “Proposal for international standardization of the use of lung ultrasound for patients with COVID-19: A simple, quantitative, reproducible method,” *Journal of Ultrasound in Medicine*, 2020.
- [5] E. Poggiali, A. Dacrema, D. Bastoni, V. Tinelli, E. Demichele, P. Matteo Ramos, T. Marcianò, M. Silva, A. Vercelli, and A. Magnacavallo, “Can lung us help critical care clinicians in the early diagnosis of novel coronavirus (covid-19) pneumonia?” *Radiology*, vol. 295, no. 3, pp. E6–E6, 2020.
- [6] G. Soldati, A. Smargiassi, R. Inchingolo, D. Buonsenso, T. Perrone, D. F. Briganti, S. Perlini, E. Torri, A. Mariani, E. E. Mossolani *et al.*, “Is there a role for lung ultrasound during the covid-19 pandemic?” *Journal of Ultrasound in Medicine*, 2020.
- [7] Q.-Y. Peng, X.-T. Wang, L.-N. Zhang, C. C. C. U. S. Group *et al.*,

- “Findings of lung ultrasonography of novel corona virus pneumonia during the 2019–2020 epidemic,” *Intensive care medicine*, p. 1, 2020.
- [8] Y. Lichter, Y. Topilsky, P. Taieb, A. Banai, A. Hochstadt, I. Merdler, A. G. Oz, J. Vine, O. Goren, B. Cohen *et al.*, “Lung ultrasound predicts clinical course and outcomes in COVID-19 patients,” *Intensive care medicine*, pp. 1–11, 2020.
  - [9] F. Mento, T. Perrone, V. N. Macioce, F. Tursi, D. Buonsenso, E. Torri, A. Smargiassi, R. Inchingolo, G. Soldati, and L. Demi, “On the impact of different lung ultrasound imaging protocols in the evaluation of patients affected by coronavirus disease 2019: How many acquisitions are needed?” *Journal of Ultrasound in Medicine*, 2020.
  - [10] T. Perrone, G. Soldati, L. Padovini, A. Fiengo, G. Lettieri, U. Sabatini, G. Gori, F. Lepore, M. Garolfi, I. Palumbo *et al.*, “A new lung ultrasound protocol able to predict worsening in patients affected by severe acute respiratory syndrome coronavirus 2 pneumonia,” *Journal of Ultrasound in Medicine*, 2020.
  - [11] L. Carrer, E. Donini, D. Marinelli, M. Zanetti, F. Mento, E. Torri, A. Smargiassi, R. Inchingolo, G. Soldati, L. Demi *et al.*, “Automatic pleural line extraction and covid-19 scoring from lung ultrasound data,” *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 11, pp. 2207–2217, 2020.
  - [12] R. J. van Sloun and L. Demi, “Localizing b-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results,” *IEEE journal of biomedical and health informatics*, vol. 24, no. 4, pp. 957–964, 2019.
  - [13] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi *et al.*, “International evaluation of an AI system for breast cancer screening,” *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.
  - [14] S. Dodge and L. Karam, “A study and comparison of human and deep learning recognition performance under visual distortions,” in *2017 26th international conference on computer communication and networks (ICCCN)*. IEEE, 2017, pp. 1–7.
  - [15] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, “Deep Convolutional Neural Network for Inverse Problems in Imaging,” *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, 2017.
  - [16] R. J. G. Sloun, R. Cohen, and Y. C. Eldar, “Deep Learning in Ultrasound Imaging,” *Proc. IEEE*, vol. 108, no. 1, pp. 11–29, Jan. 2020.
  - [17] E. Goldstein, D. Keidar, D. Yaron, Y. Shachar, A. Blass, L. Charbinsky, I. Aharon, L. Lifshitz, D. Lumelsky, Z. Neeman *et al.*, “Covid-19 classification of x-ray images using deep neural networks,” *arXiv*, 2020.
  - [18] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli *et al.*, “Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound,” *IEEE Transactions on Medical Imaging*, 2020.
  - [19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009.
  - [20] N. Anantrasirichai, W. Hayes, M. Allinovi, D. Bull, and A. Achim, “Line detection as an inverse problem: application to lung ultrasound imaging,” *IEEE transactions on medical imaging*, vol. 36, no. 10, pp. 2045–2056, 2017.
  - [21] O. Karakuş, N. Anantrasirichai, A. Aguersif, S. Silva, A. Basarab, and A. Achim, “Detection of line artifacts in lung ultrasound images of covid-19 patients via nonconvex regularization,” *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 11, pp. 2218–2229, 2020.
  - [22] H. Kerdegari, P. T. H. Nhat, A. McBride, V. Consortium, R. Razavi, N. V. Hao, L. Thwaites, S. Yacoub, and A. Gomez, “Automatic detection of b-lines in lung ultrasound videos from severe dengue patients,” *arXiv*, 2021.
  - [23] F. Mento and L. Demi, “On the influence of imaging parameters on lung ultrasound b-line artifacts, in vitro study,” *The Journal of the Acoustical Society of America*, vol. 148, no. 2, pp. 975–983, 2020.
  - [24] L. Demi, M. Demi, R. Prediletto, and G. Soldati, “Real-time multi-frequency ultrasound imaging for quantitative lung ultrasound – first clinical results,” *The Journal of the Acoustical Society of America*, vol. 148, no. 2, pp. 998–1006, 2020.
  - [25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
  - [26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
  - [27] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.
  - [28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, 2014, pp. 740–755. [Online]. Available: [cocodataset.org](http://cocodataset.org)
  - [29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems* 32, 2019.
  - [30] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
  - [31] D. Yaron, D. Keidar, E. Goldstein, Y. Shachar, A. Blass, O. Frank, N. Schipper, N. Shabshin, A. Grubstein, D. Suhani *et al.*, “Point of care image analysis for COVID-19,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.
  - [32] 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 *Proceedings of the IEEE international conference on computer vision*, 2017.
  - [33] J. Born, N. Wiedemann, M. Cossio, C. Buhre, G. Brändle, K. Leidermann, A. Aujayeb, M. Moor, B. Rieck, and K. Borgwardt, “Accelerating detection of lung pathologies with explainable ultrasound image analysis,” *Applied Sciences*, vol. 11, no. 2, p. 672, Jan 2021. [Online]. Available: <http://dx.doi.org/10.3390/app11020672>
  - [34] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2015.
  - [35] S. Bagon, M. Galun, O. Frank, N. Schipper, M. Vaturi, G. Zalberg, G. Soldati, A. Smargiassi, R. Inchingolo, E. Torri *et al.*, “Assessment of covid-19 in lung ultrasound by combining anatomy and sonographic artifacts using deep learning,” *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2736–2736, 2020. [Online]. Available: [tinyurl.com/ycxwaerl](https://tinyurl.com/ycxwaerl)