Weakly Semi-supervised Detector-based Video Classification with Temporal Context for Lung Ultrasound

Gary Y. Li*[1], Li Chen[1], Mohsen Zahiri[1], Naveen Balaraju[1], Shubham Patil[1], Courosh Mehanian[2], Cynthia Gregory[3], Kenton Gregory[3], Balasundar Raju[1], Jochen Kruecker[1], and Alvin Chen[1]

[1]Philips Research North America

[2]Global Health Laboratories

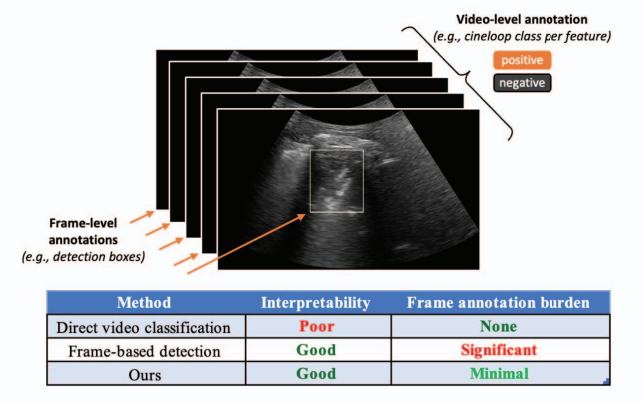
[3]Oregon Health & Science University

Abstract

For many challenging medical imaging tasks involving sequences, video-level labels alone are insufficient to train accurate disease classification models and do not carry information about the locations of relevant features. Alternatively, localization-based models such as detectors offer much stronger interpretability by indicating areas of suspicion, but require comprehensive frame-by-frame annotations by experts. We propose a method to address the trade-off between annotation burden and interpretability by performing simultaneous detection and classification on medical video sequences while requiring very limited frame-level supervision. Specifically, our approach aggregates individual predictions from a detection model into "tracklets" representing temporally consistent regions of pathology along the sequence. The tracklets are classified in a second stage to arrive at an overall video-level prediction. Both the detector and tracklet classifier are trained in a weakly semi-supervised manner using a large amount of video-annotated data alongside a limited set of frame annotations. We apply the approach to several challenging medical imaging tasks, namely localizing and predicting the presence or absence of lung consolidation and pleural effusion in ultrasound videos. We show that, with only a very small amount of additional frame-annotated data, the method provides strong model interpretability through localization and achieves state-of-the-art detection and classification, outperforming both direct video classifiers and comparable frame-based detectors trained without the added temporal context.

1. Introduction

Lung ultrasound (LUS) is a medical imaging technique deployed at the point-of-care to aid in evaluation of pulmonary and infectious diseases. A critical observation of LUS is the presence of consolidations (fluid-filled alveoli), which are a known indicator of lung infection and pneumonia, including COVID-19. Similarly, presence of pleural effusions are associated with pneumonia as well as complications from heart or liver disease. Today, LUS interpretation requires expert manual review of videos (typically comprising 60 to 200 image frames) to identify pathological features that may be very small (<1 cm in the case of consolidations, <0.5 cm in the case of pleural effusions) and difficult to visualize. Expert interpretation requires considering the location, size, and appearance of pathological areas in each frame, and then aggregating the information across the sequence to derive an overall, actionable assessment.



Method	Interpretability	Frame annotation burden
Direct video classification	Poor	None
Frame-based detection	Good	Significant
Ours	Good	Minimal

Figure 1. Illustration of frame-level and video-level annotated ultrasound video data and the trade-offs among model training strategy, data labelling efforts, and interpretability

Deep learning methods have shown promise in providing automated, video-based analysis of the presence or absence of pathology [4, 10, 25, 28, 30, 38, 24]. For example, prior studies have shown that CNN-LSTM networks can extract spatial features and learn temporal dependence in LUS sequences for video-level classification [9, 1, 24]. A limitation of direct video classifiers, however, is that they do not indicate the *locations* of the pathology, which greatly limits its clinical interpretability. Visualization of activation maps [35, 5] may likewise be limited, especially when features are small or ambiguous. Finally, given the large dimension-

*Corresponding author: ye.li@philips.com

ality of 2D+time ultrasound video data, direct classification using video-level supervision alone [10] may be susceptible to overfitting.

Rather than relying on direct video classification, a more interpretable alternative is to aggregate individual frame-level outputs (e.g., detections) to derive the overall video-level prediction. Frame aggregation can be based on simple human-designed rules, such as max or mean pooling across frames [32], or learned from data using models with temporal design [34, 22, 14]. However, detector training requires more comprehensive, frame-by-frame annotation of videos, as shown in Fig. 1; such frame-level annotations are extremely time-consuming and expensive to generate. Multiple works have been attempted to reduce the number of frame-level annotations for detector training by exploiting semi-supervised and weakly supervised methods [19, 31, 16, 11, 36, 21, 7, 33, 27, 37]. In particular, Roy *et al*. [23], trained a classifier to predict COVID-19 severity at the frame level and provide coarse localizations; the frame predictions were aggregated using uninorms [20], a learnable aggregation unit, for final video-level classification. Lin *et al*. [18] proposed a method for lesion detection in ultrasound videos by aggregating video-level lesion classification features and clip-level temporal features; however, the classification is performed per-lesion rather than on the video level.

In general, a major design trade-off is between the interpretability of an ML model and the granularity of the supervision used to train the model. For LUS, better interpretability means more precise localization of pathology, which in turn requires more fine-grained supervision (e.g., frame-level bounding box labels) and a higher annotation burden. Ultimately, it is desirable to use minimal annotation efforts while achieving sufficient model interpretability.

In this work, we tackle the trade-off of annotation burden and model interpretability by proposing a method to perform simultaneous frame detection and video classification while requiring a very limited amount of frame-annotated data. Specifically, our approach aggregates individual predictions from a trained frame detector into "tracklets" (representing temporally consistent regions of pathology along the video sequence), which are then classified in a second step to ultimately arrive at the overall video-level prediction. Critically, both the frame detector and tracklet classifier are trained in a weakly semi-supervised manner using a large amount of "weak" video-annotated data alongside a very limited amount of frame-annotated data.

Our experimental findings demonstrate that: (1) with only a small amount of additional frame-annotated data, we achieve state-of-the-art (SOTA) detection and classification of lung consolidation and pleural effusion; (2) on the detection task, our approach outperforms comparable frame-based detectors trained without the added temporal context; and (3) on the video classification task, our method outperforms direct video classifiers as well as the SOTA detector-based methods[23], while still providing the interpretability of detector-based methods. Finally, we note that the proposed framework is modular and interchangeable; alternative detection, tracking, and classification methods may be introduced to meet specific data and task requirements.

2. Methods

2.1. Overall framework

The overall framework (Fig. 2) comprises four main steps:

1. *Frame detection (weakly semi-supervised):* First, we train any standard object detector (YOLOv5 in this study) to generate predicted bounding boxes for each frame (section 2.2). We apply a weakly semi-supervised teacher-student mutual learning approach [21] to train the detector on combined video-annotated and very limited frame-annotated data.

2. *Aggregating predictions along tracklets*: Next, using the method of [2], we aggregate predicted boxes into tracklets representing temporally connected bounding boxes. The tracklets are used to crop the original video to generate tracklet clips (tracklet-cropped candidate regions) (section 2.3).

3. *Tracklet classification (weakly semi-supervised):* Third, we introduce a dedicated second-stage network (a lightweight CNN+LSTM, denoted as *trackletNet*) for binary classification of tracklet clips (section 2.4). Notably, the trackletNet is trained on an enriched dataset containing challenging examples drawn from incorrect predictions made by the detector. Similarly to the frame detector, the trackletNet is trained on combined frame- and video-annotated data in a weakly semi-supervised manner.

4. *Video classification:* The final video-level classification is simply determined by the maximum predicted confidence among tracklets; this is consistent with how medical ultrasound videos are clinically interpreted.

2.2. Weakly semi-supervised frame detection

Given an input ultrasound video $X \in \mathbb{R}^{H \times W \times T}$ with a frame size of (H, W) and T frames, we train an initial 2D object detector in a weakly semi-supervised (WSS) manner via a teacher-student training procedure designed for temporal image sequences [21]. Specifically, a teacher model was first trained with a small number of frame-annotated videos $D_f = (x_i^f, y_i^f)_{i=1}^{N_f}$ comprising N_f frames x_i^f and corresponding frame labels y_i^f. For consolidation detection, we used frame annotations from 99 LUS videos;

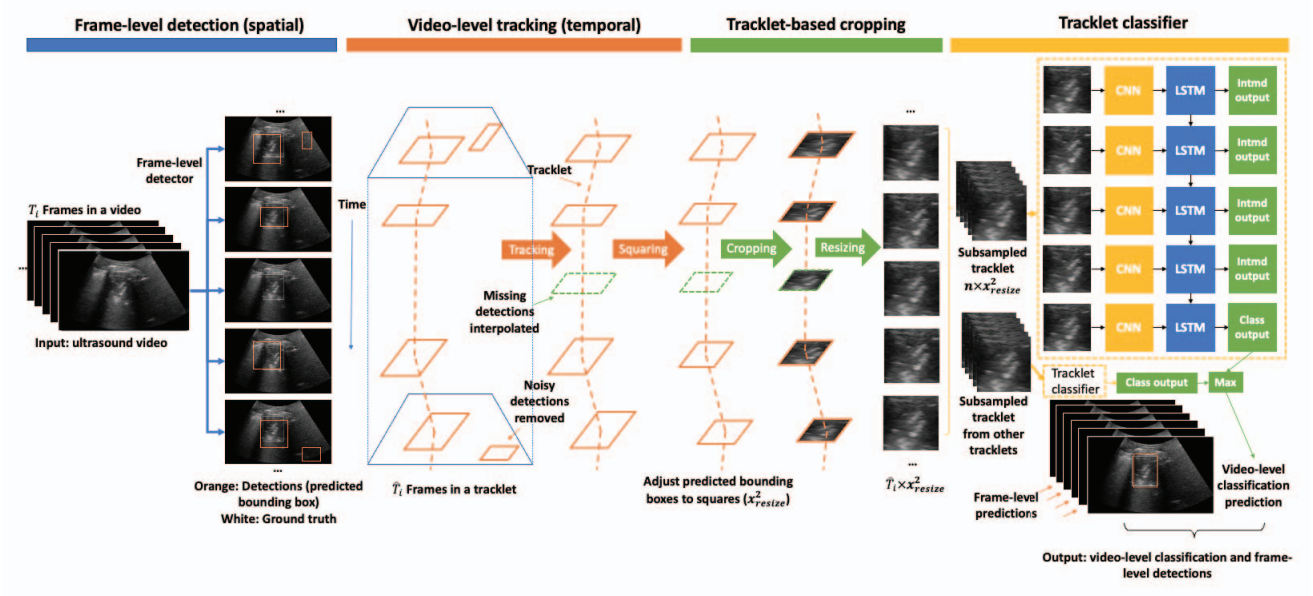


Figure 2. Overall architecture of the proposed framework. To efficiently extract the spatial and temporal features about the underlying pathology, we first use a 2D detector to predict the spatial features on each frame of a video and then connect the predicted spatial features (as given by a bounding box) temporally via a tracking algorithm. After that, we use the temporally refined bounding boxes (tracklets) to crop the original video to generate 3D candidate region(s) within each video. We train a dedicated second-stage network (a lightweight CNN+LSTM model, denoted as *trackletNet*) to classify these 3D candidate regions, minimizing the disruption from the complex and variable background which occupies a large fraction of the input volume. To train the *trackletNet* with fixed number of input frames, we uniformly subsampled n slices from one tracklet-cropped region between random start and end slices. During inference, we use the maximum prediction of all candidate region(s) from a video as the final prediction for video classification.

for pleural effusion detection, we used 80 frame-annotated videos. A student model was then initialized with the same network structure and weights as the teacher.

The teacher and student are jointly optimized by introducing a much larger weakly annotated dataset $D_w = (x^w_{j,1:T_j}, z^w_j)_{j=1}^{N_w}$ comprising N_w videos $x^w_{j,1:T_j}$ and their corresponding video labels z^w_j. T_j is the number of frames for a given video j. During the teacher-student mutual learning process, frame-level "pseudolabels" $\{\hat{y}^w_{j,1:T_j}\}_{j=1}^{N_w}$, with predicted confidences given by $c^w_{j,1:T_j}\}_{j=1}^{N_w}$, are generated by the teacher and used to train the student via backpropagation. Meanwhile, the student's weights are used to update the teacher via exponential moving averaging [13]. Weak supervision is provided at the video level, thus ensuring that the quality of pseudolabels does not degrade during mutual learning. In our experiments, 6677 weakly labeled videos were used for consolidation, and 9836 weakly labeled videos were used for pleural effusion.

The overall detector training loss is a combination of the (fully supervised) frame-labeled loss $\lambda_{f_{\text{label}}}$; (semi-supervised) frame-level pseudolabel loss $\lambda_{f_{\text{pseudo}}}$; and (weakly-supervised) video-level loss $\lambda_{v_{\text{label}}} L_{v_{\text{label}}}$:

$$\begin{aligned} Loss &= \lambda_{f_{\text{label}}} L_{f_{\text{label}}} + \lambda_{f_{\text{pseudo}}} L_{f_{\text{pseudo}}} + \lambda_{v_{\text{label}}} L_{v_{\text{label}}} \\ &= \lambda_{f_{\text{label}}} \sum_{i=1}^{N_f} L_{\text{detection}} \left(\mathbf{x}^f_i, \mathbf{y}^f_i \right) \\ &+ \lambda_{f_{\text{pseudo}}} \sum_{j=1}^{N_w} \sum_{t=1}^{T_j} L_{\text{detection}} (\mathbf{x}^w_{j,t}, \hat{\mathbf{y}}^w_{j,t}) \\ &+ \lambda_{v_{\text{label}}} \sum_{i=1}^{N_w} L_{\text{BCE}} \left(\mathbf{z}^w_j, \frac{\sum_{t=1}^{T_j} \max \left(\mathbf{c}^w_{j,t} \right)}{T_j} \right), \end{aligned} \tag{1}$$

where λ indicates the weights for the three loss terms and $L_{\text{detection}}$ represents any detector loss (such as IoU). $\max \left(\mathbf{c}^w_{j,t} \right)$ is the maximum confidence from all the predicted bounding boxes on video j and frame t.

2.3. Aggregating predictions along tracklets

To arrive at representations of temporally consistent regions of pathology along the video sequence, we generated "tracklets" from each video in $\mathbf{D}_f$ and $\mathbf{D}_w$ by applying object tracking directly on bounding box predictions from the trained teacher model. In our experiment, we used a simple

and efficient SORT tracker [2] as the tracking algorithm, although other techniques may be used [3, 26, 29, 39]. We kept all the tracklets generated from videos in $\mathbf{D}_f$, since reliable tracklet-level labels can be derived from the existing frame labels. On the other hand, true labels for tracklets generated from $\mathbf{D}_w$ are not known (since they could have come from true or false positive detections). Thus, from $\mathbf{D}_w$, only tracklets with sufficient prediction confidences were used subsequently to train the tracklet classifier. The details about prediction confidence filtering is described below.

2.3.1 Filtering tracklets via prediction confidence

Since many standard object detectors (including YOLOv5) produce a large number of low confidence boxes, it is critical to exclude tracklets created from unreliable detections. Given that frame labels were not available for videos in $\mathbf{D}_w$, and thus the true labels for these tracklets are not known, we relied on detection confidences to select tracklets for inclusion in training. Specifically, we introduce two parameters for tracklet selection: τ_{pos} defines the minimum confidence for inclusion of tracklets obtained from positively-labeled videos; τ_{neg} defines the minimum confidence for inclusion of tracklets from negative-labeled videos. The tracklet-level confidence is assigned equal to $\max\left(\mathbf{c}_{j,t}^{w}\right)$, that is, the maximum confidence of all bounding boxes that form the tracklet.

For positive videos, a low τ_{pos} creates many false positive tracklets containing no pathology while a high τ_{pos} cannot provide sufficient candidate regions for classification on positive videos. On the other hand, for negative videos, we can simply use a low value for τ_{neg}, since predicted tracklets from a negative video are certainly false positives. τ_{pos} was empirically chosen from $\mathbf{D}_f$ by comparing the filtered tracklets with frame-level annotations $\left\{\mathbf{y}_i^f\right\}_{i=1}^{N_f}$. For our experiment, we selected $\tau_{\text{pos}} = 0.6$ and $\tau_{\text{neg}} = 0.01$ (see Table 6 for ablation results demonstrating the effect of τ_{pos} on tracklet data distributions) to allow a balanced distribution of true positive, false positive, and false negative tracklets to be sampled for training the trackletNet classifier.

2.3.2 Tracklet cropping and rescaling

Detection bounding boxes within each tracklet are represented as $\left\{\hat{\mathbf{y}}_{i,1:\hat{T}_i}^{f}\right\}_{i=1}^{N_f}$ and $\left\{\hat{\mathbf{y}}_{j,1:T_j}^{w}\right\}_{j=1}^{N_w}$, where $\hat{T}_i$ and $\hat{T}_j$ are the number of frames of a tracklet generated from $\mathbf{D}_f$ and $\mathbf{D}_w$, respectively. $\hat{\mathbf{y}}$ includes the coordinates of bounding boxes represented as box center coordinates $\widehat{\mathbf{y}}_x$, $\widehat{\mathbf{y}}_y$, box width $\widehat{\mathbf{y}}_w$, and height $\widehat{\mathbf{y}}_h$.

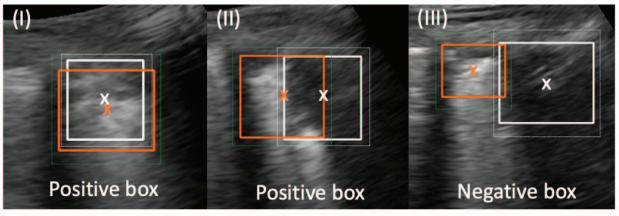


Figure 3. From left to right, the images show an easy positive case (I), a borderline positive case (II), and a negative case (III). The green and white boxes represent the to-be-cropped regions from the predicted and ground truth bounding box, respectively. The orange and white crosshair marks the center of the predicted and ground truth bounding box, respectively.

To allow sufficient spatial context around each tracklet, detection boxes are squared to $\left\{\widehat{\mathbf{y}}'_{i,1:\hat{T}_i}{}^{f}\right\}_{i=1}^{N_f}$ and $\left\{\widehat{\mathbf{y}}'_{j,1:\hat{T}_j}{}^{w}\right\}_{j=1}^{N_w}$, that is,

$$\widehat{\mathbf{y}'_w} = \widehat{\mathbf{y}'_h} = \max\left(\widehat{\mathbf{y}_w},\ \widehat{\mathbf{y}_h}\right) + \varepsilon \tag{2}$$

where ε is a margin added around each detection box within the tracklet. Tracklet images were then resized to $x_{resize} \times x_{resize}$, resulting in rescaled tracklets of dimension $x_{resize} \times x_{resize} \times \hat{T}_i$.

2.3.3 Tracklet label assignment

The class labels for each frame of a rescaled tracklet were determined according to whether the tracklet originated from $\mathbf{D}_f$ or $\mathbf{D}_w$:

Tracklets from $\mathbf{D}_f$*:* Given a frame-level bounding box label $\mathbf{y}_{i,t}^{w} = (\mathbf{y}_{x,i,t}^{w}, \mathbf{y}_{y,i,t}^{w}, \mathbf{y}_{w,i,t}^{w}, \mathbf{y}_{h,i,t}^{w})$ and predicted box $\hat{\mathbf{y}}_{j,t}^{w} = (\hat{\mathbf{y}}_{x,j,t}^{w}, \hat{\mathbf{y}}_{y,j,t}^{w}, \hat{\mathbf{y}}_{w,j,t}^{w}, \hat{\mathbf{y}}_{h,j,t}^{w})$, we assign a positive class label to the predicted box if the center of the predicted and ground-truth boxes (shown in Fig. 3) reside within each other's extents, that is,

$$\begin{cases} \widehat{\boldsymbol{y}}_{x_{i,t}}^{w} - \frac{\widehat{\boldsymbol{y}}_{w_{i,t}}^{w}}{2} \leq \boldsymbol{y}_{x_{i,t}}^{w} \leq \widehat{\boldsymbol{y}}_{x_{i,t}}^{w} + \frac{\widehat{\boldsymbol{y}}_{w_{i,t}}^{w}}{2}, \\ \widehat{\boldsymbol{y}}_{y_{i,t}}^{w} - \frac{\widehat{\boldsymbol{y}}_{h_{i,t}}^{w}}{2} \leq \boldsymbol{y}_{y_{i,t}}^{w} \leq \widehat{\boldsymbol{y}}_{y_{i,t}}^{w} + \frac{\widehat{\boldsymbol{y}}_{h_{i,t}}^{w}}{2}, \\ \boldsymbol{y}_{x_{i,t}}^{w} - \frac{\boldsymbol{y}_{w_{i,t}}^{w}}{2} \leq \widehat{\boldsymbol{y}}_{x_{i,t}}^{w} \leq \boldsymbol{y}_{x_{i,t}}^{w} + \frac{\boldsymbol{y}_{w_{i,t}}^{w}}{2}, \\ \boldsymbol{y}_{y_{i,t}}^{w} - \frac{\boldsymbol{y}_{h_{i,t}}^{w}}{2} \leq \widehat{\boldsymbol{y}}_{y_{i,t}}^{w} \leq \boldsymbol{y}_{y_{i,t}}^{w} + \frac{\boldsymbol{y}_{h_{i,t}}^{w}}{2} \end{cases} \tag{3}$$

Tracklets from $\mathbf{D}_w$*:* As there are no ground truth bounding box labels available, we simply assign all predicted boxes with detection confidences above τ_{pos} (if originating from positive videos) and τ_{neg} (if originating from negative videos) to the video-level class label.

Finally, the overall tracklet-level label is determined from the assigned frame-level box classes. Specifically,

Dataset	Consolidation		Pleural effusion	
	Negative	Positive	Negative	Positive
Train	9427	9012	9373	9062
Val	476	321	159	175
Test	1030	581	138	158

Table 1. Train, validation, and test distributions for trackletNet classifier.

a tracklet is considered positive if it contains at least one positive box (implying high likelihood of pathology along at least a portion of the tracklet). Fig. 3 shows several examples of label assignments for detection boxes within rescaled tracklets.

2.4. Tracklet and video classification

Similarly to the frame detector, the trackletNet is trained using the combined frame-annotated ($\mathbf{D}_f$) and video-annotated ($\mathbf{D}_w$) data in a weakly semi-supervised manner. Using the methods described in section 2.3 above, a total of 20,847 rescaled tracklets were generated from the consolidation dataset, and 19,065 tracklets were generated from the pleural effusion dataset, with approximately equal balance of negative and positive tracklets. Notably, these tracklets represent an enriched dataset containing a high proportion of challenging examples drawn from incorrect (false positive and false negative) predictions made by the detector (e.g., Fig. 3). Furthermore, unlike the individual 2D frames used to train the detector, the rescaled tracklets are spatially localized around suspect pathological regions and contain temporal context.

Table 1 summarizes the distribution of training, validation, and test samples used for our experiments. During training, we applied spatial and temporal augmentation on the rescaled tracklets (described in section 2.5).

As a form of temporal augmentation (and to allow training in batches), we uniformly sampled *n* frames from each rescaled tracklet using randomized start and end slices; similar to [6], we used $n = 5$. As described above, the resulting subsampled tracklets are assigned a positive label if they contain at least one positively labeled prediction box.

During inference, we determined the overall video-level class prediction as the maximum prediction confidence across all the tracklets in a video. As a further improvement, during inference, we removed all tracklets from videos predicted as negative.

2.5. Implementation details

2.5.1 Frame detection

We adopted the Pytorch Ultralytics implementation of YOLO-v5 [15] as the backbone detector due to its accuracy and computational efficiency, although other SOTA detectors can be used in our framework. Teacher model is trained for 10 epochs (burn-in epochs) with frame-level labeled videos alone, then teacher and student models are mutual learning from the weakly labeled videos for 150 epochs. The teacher model checkpoint with the minimum $L_{v_{label}}$ on the validation set is used as the final detector. For weakly semi-supervised training, $\lambda_{f_{label}} = 1$, $\lambda_{f_{pseudo}} = \lambda_{v_{label}} = 0.5$. SORT tracker runs with a minimum length of 5, hit of 3, and IOU of 0.5. $\varepsilon = 20$ is used for the tracklet-cropped regions. We trained the fine-scaled classifier for 500 iterations using binary cross entropy loss and Adam's optimizer with a learning rate of 1e-5. The checkpoint with minimum validation loss is selected for the final test set evaluation.

2.5.2 Tracklet generation and classification

For all experiments, rescaled tracklets were generated using ε equal to 0.1 of the $\max(\widehat{\mathbf{y}_w}, \widehat{\mathbf{y}_h})$ and $x_{resize} = 148$ to accommodate augmentation for an input size of 128×128. For tracklets inferenced on videos from $\mathbf{D}_w$, we used $\tau_{pos} = 0.6$ and $\tau_{neg} = 0.01$. For subsampling, *n* was set to 5.

The trackletNet classifier consisted of a convolutional and temporal feature extraction network (CNN+LSTM). We adopted a lightweight encoder with 5 convolutional blocks each comprising a 3×3 convolutional layer, batch normalization, Leaky ReLU and 2D max pooling. A global average pooling layer was applied as the last layer with 256 dimensions. The LSTM takes the CNN representations (5×256) and aggregates along the temporal dimension.

We trained the trackletNet with spatial augmentations including rotation ($-10°$, $10°$), scaling (0.9,1.1), shearing (-10,10), translation (0.1,0.3), center cropping, and random horizontal flipping, as well as temporal augmentations including randomized reversal of the frame order and randomized start/end positions for subsampling. During inference, tracklets are uniformly sampled to *n*=5 cropped frames.

3. Experiments and results

3.1. Dataset

An extensive retrospective, multi-center clinical dataset of 7712 lung ultrasound videos was used in this work. The data were acquired from 420 patients with suspicion of lung consolidation or other related pathology (e.g., pneumonia, pleural effusion) from 8 U.S. clinical sites between 2017 and 2020. The videos were acquired each at least 3 seconds in length and contain at least 60 frames.

To train the models and assess model performance, 6677/9863 training videos (357/393 subjects), 337/273 validation videos (23/10 subjects), and 599/233 test videos (40/16 subjects) were annotated only for the presence or absence of lung consolidation/pleural effusion at the video

level. Among training videos, 99/80 videos (13/6 subjects) were additionally frame-level labeled for lung consolidation/pleural effusion bounding boxes. All the test set videos were frame-level labeled to evaluate detection performance as the secondary evaluation metric. The data were partitioned at the subject level. Annotation was carried out by a multi-center team of expert physicians with medical training in lung ultrasound. Each video was annotated by two experts and adjudicated by a third expert when a disagreement between the first two annotators occurred. Frame-level annotations were annotated by a single expert.

3.2. Comparison experiments

We first compared our detector-based classification method with a classic temporal frame-level classification model – CNN (with EfficientNet or MobileNet backbone) + LSTM network [1] and a baseline detector-based classification method (where the maximum confidence from detection/tracking is used for video-level classification). We also compared our method with a SOTA video classification method with frame-level supervision [23]. These methods were selected considering the LUS application, similar data/label requirements, and code availability for implementation.

We ran these experiments on two lung pathologies: consolidation and pleural effusion. The summary of video classification test results with and without added frame-level supervision for the consolidation/pleural effusion datasets, are summarized in Table 2. The results showed consistent trends on these experiments on different lung pathologies, with our method achieving both the highest AUC and AP on the test set, indicating its superior performance to SOTA classification method as well as the detector-based classification methods.

The proposed framework not only improves classification accuracy by a large margin but also clearly locates the pathological regions in the video (with an AP of 0.381) not available from direct video classification methods. Some example videos which were correctly classified by our approach but failed by existing direct video classifier or detector-based classifier are shown in Fig. 4.

To test the performance drop of our method, we further reduced the frame-annotated data from 99 videos to only 14 videos with no other changes. We observe that, even with an extremely limited amount of frame-level labeled data (only 3 subjects/14 videos, as shown in the last two rows of Table 2), the proposed method continues to exhibit superior classification performance compared to the baseline and SOTA methods, despite a slight reduction in average precision (AP).

3.3. Ablation experiments

A series of ablation studies were performed (summarized in Table 3 and below).

3.3.1 Contributions of weak supervision, temporal aggregation, and tracklet classification

To evaluate the contribution of reducing region of interest, we removed the detection part, and directly used the whole image as input to train and evaluate the trackletNet (Table 2, row 3). To further understand the impact of frame-level labels, we trained the detector with frame-level labels alone. The results (Table 3, row 1) show that the model trained with frame-level labels alone could not detect or classify well (low AP and AUC), indicating the effectiveness of our WSS detector training method.

To understand the impact of the second-stage trackletNet classification, we used the maximum confidence of all predictions (a simple rule-based aggregation method) for video-level classification (Table 3, rows 2 and 3). To understand the impact of temporal consistency, we remove the SORT tracking step and directly derive video-level classification from the detection outputs without further temporal filtering. Both results show worse detection and classification performances.

We then evaluated the contribution from the temporal aggregation performed by the trackletNet. First, we assessed performance only using a single frame (the center frame) of each tracklet, thus removing the temporal context introduced by the full tracklet (Table 3, row 4). Second, we tested a trackletNet variant without temporal design by using fully connected layer instead of LSTM on the feature embeddings (Table 3, row 5). Both experiments show that the temporal aggregation from LSTM is required to achieve the best classification performance. For completeness, detection results without tracklet-based filtering of detections are also shown (Table 3, rows 7-9).

3.3.2 Model variance using different data splits

In addition to the ablation studies shown in Table 3, we also investigated model variance by running repeated experiments on independent test data. In Table 4, we present results for models trained using 153 frame-annotated videos sampled from a different group of patients compared to the 99 videos used for the experiments in Table 2. We observe consistent trends on the independent test dataset.

3.3.3 Tracklet-level aggregation methods

Additionally, we compared different methods for deriving the overall video-level classification from tracklet predictions (Table 5). Consistent with how medical images are

WSS: weakly semi-supervised; FS: fully-supervised; TR: tracking; FLT: filtering detection results based on tracklet predictions; FLL: frame-level labeled; VLL: video-level labeled;
*: frame-level bounding box label was used to create the frame-level class label; **: experiments performed on consolidation dataset only

Approach	Frame detector	Video/tracklet classifier	# of FLL videos	# of VLL videos	Detection (Test AP_{50})	Classification (Test AUC)
Direct video classifier	N.A.	EfficientNet+LSTM	0	6677 / 9836	N.A.	0.748 / 0.809
	N.A.	MobileNet+LSTM	0	6677 / 9836	N.A.	0.870 / 0.910
	N.A.	CNN+LSTM(video)	0	6677 / 9836	N.A.	0.909 / 0.894
Detector-based video classifier	STN	Uninorms	99*/80*	6677 / 9836	N.A.	0.886 / 0.916
	WSS Yolo+TR	MaxConf	99/80	6677 / 9836	0.345 / 0.334	0.880 / 0.893
	WSS Yolo+TR+FLT	CNN+LSTM(tracklet)	99/80	6677 / 9836	**0.381 / 0.365**	**0.936 / 0.938**
	WSS Yolo+TR	MaxConf	14**	6677	0.318	0.905
	WSS Yolo+TR+FLT	CNN+LSTM(tracklet)	14**	6677	0.369	0.927

Table 2. Experimental comparisons for consolidation and pleural effusion frame detection and video classification.

WSS: weakly semi-supervised; FS: fully-supervised; TR: tracking; FLT: filtering detection results based on tracklet predictions; FLL: frame-level labeled; VLL: video-level labeled

Frame detector	Video/tracklet classifier	# of FLL videos	# of VLL videos	Detection (Test AP_{50})	Classification (Test AUC)
FS Yolo	Max conf	99	0	0.257	0.845
WSS Yolo	Max conf	99	6677	0.329	0.882
WSS Yolo+TR	Max conf	99	6677	0.345	0.880
WSS Yolo+TR	CNN+LSTM (Single frame)	99	6677	0.345	0.905
WSS Yolo+TR	CNN+Dense (Subsampled tracklet)	99	6677	0.345	0.921
WSS Yolo+TR	CNN+LSTM (Subsampled tracklet)	99	6677	0.345	0.936
WSS Yolo+TR+FLT	CNN+LSTM (Single frame)	99	6677	0.371	0.905
WSS Yolo+TR+FLT	CNN+Dense (Subsampled tracklet)	99	6677	0.366	0.921
WSS Yolo+TR+FLT	**CNN+LSTM (Subsampled tracklet)**	99	6677	**0.381**	**0.936**

Table 3. Ablation experiments for consolidation detection and video classification.

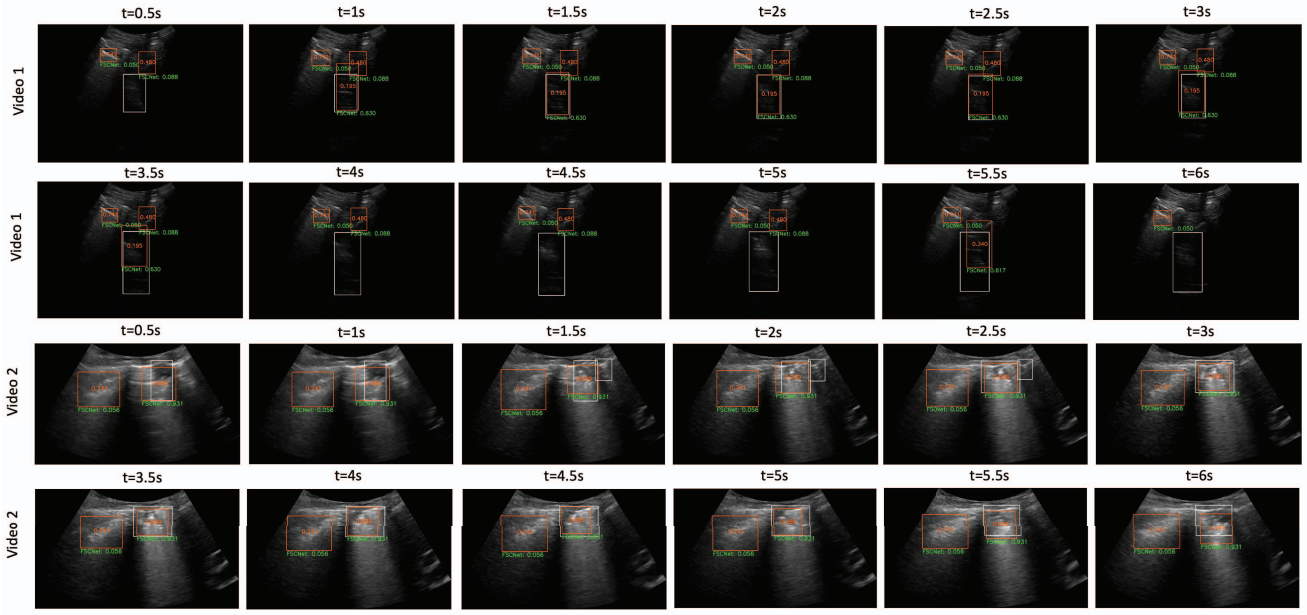


Figure 4. Examples of video frames correctly classified by trackletNet (here labeled as FSCNet with tracklet prediction confidences shown in green) but misclassified by the initial frame-level detector (orange boxes with frame-level detection confidences also shown). Ground-truth boxes are labeled in white.

WSS: weakly semi-supervised; FS: fully supervised; TR: tracking; FLT: filtering detection results based on tracklet predictions; FLL: frame-level labeled; VLL: video-level labeled;

Frame detector	Video/tracklet classifier	# of FLL videos	# of VLL videos	Detection (Test AP_{50})	Classification (Test AUC)
WSS Yolo	Max conf	153	6677	0.360	0.903
WSS Yolo+TR	Max conf	153	6677	0.418	0.897
WSS Yolo+TR	CNN+LSTM (Subsampled tracklet)	153	6677	0.418	0.918
WSS Yolo+TR+FLT	**CNN+LSTM (Subsampled tracklet)**	153	6677	**0.458**	**0.918**

Table 4. Repeated experiments for consolidation detection and video classification on an independent data split. The 153 frame-annotated videos were sampled from a different group of patients compared to the 99 frame-annotated videos used for the experiments in Table 2.

Detection Method	Classification Method	Tracklet Pred. Agg. Method	Test AUC
WSS Yolo +TR+FLT	CNN +LSTM (Subsampled (tracklet)	Mean	0.918
		Sum	0.928
		Max(Proposed)	**0.936**

Table 5. Comparison of methods for deriving overall video-level class from tracklet predictions (experiments performed on the consolidation dataset).

τ_{pos}	Avg. # of FP Tracklets per Pos. Video	Avg. # of TP Tracklets per Pos. Video	%FN Videos	Test AUC
0.2	0.438	1.500	0.0125	0.926
0.4	0.188	1.550	0.0125	0.930
0.6	**0.138**	**1.410**	**0.0750**	**0.936**
0.8	0.025	1.138	0.2000	0.930
0.9	0.013	0.750	0.4750	0.917

Table 6. Summary of statistics of predicted tracklets on frame-level labeled data and video classification test results as a function of different minimum tracklet confidence thresholds. Choice of $\tau_{pos} = 0.6$ leads to the highest accuracy and AUC on the consolidation dataset comprising 99 frame-level labeled videos.

often are clinically interpreted, we find that simply using the maximum predicted confidence among all tracklets results in the strongest video-level classification performance.

3.3.4 Model variance using different τ_{pos}

To evaluate the performance variation on using different τ_{pos}, we generated train, validation, and test data set for the trackletNet using different τ_{pos}s and calculated their corresponding video classification performance on the test set. The summary of video classification test results for different τ_{pos} is available in Table 6. The results reveal slight fluctuations in video-level performance concerning the choice of τ_{pos}. This suggests that our method's performance is not solely reliant on selecting an optimal hyperparameter, demonstrating its robustness.

4. Conclusions

This work introduced a method for simultaneous detection and classification on medical video sequences with very limited frame-level annotation burden. Specifically, our method aggregates individual predictions from a detection model into temporally consistent "tracklet" regions, which are classified in a second stage. Despite the minimal added frame supervision, empirical results on two LUS pathology prediction tasks demonstrate that our method leads to improved video classification compared to existing SOTA models, while also improving the localization performance of the initial detector.

One limitation of the current approach is the need for two separately trained models (frame detector and tracklet classifier), which increases complexity and limits end-to-end training. Future work will investigate intrinsic feature extraction methods that would not require a separately trained second-stage network (e.g., [8, 12]). Such methods may allow temporal information to be directly introduced into the detection step, thereby improving localization at the initial stage. Alternative spatiotemporal aggregation techniques (e.g., based on self-attention [22]) will also be explored.

Another limitation is the need to empirically select hyper-parameters such as τ_{pos} to filter for reliable tracklet candidates. Future work will investigate methods for adaptive tracklet selection, analogous to the pseudo-label filtering problem that occurs at the frame level when training the detector with weak or semi-supervision [17, 16, 19, 31].

Finally, future studies will evaluate the generalizability of the method on other diverse medical imaging datasets.

Acknowledgement

We acknowledge the following clinical contributors: Sourabh Kulhare, Rachel Millin, Meihua Zhu, David O. Kessler, Laurie Malia, Almaz Dessie, Joni Rabiner, Di Coneybeare, Bo Shopsin, Andrew Hersh, Cristian Madar, Jeffrey Shupp, Laura S. Johnson, Jacob Avila, Kristin Dwyer, Peter Weimersheimer, Zohreh Laverriere, Xinliang Zheng, Annie Cao, Katelyn Hostetler, Yuan Zhang, Amber Halse, James Jones, Jack Lazar, Devjani Das, Tom Kennedy, Lorraine Ng, Penelope Lema, and Nick Avitabile.

References

[1] Bruno Barros, Paulo Lacerda, Celio Albuquerque, and Aura Conci. Pulmonary covid-19: learning spatiotemporal features combining cnn and lstm networks for lung ultrasound video classification. *Sensors*, 21(16):5486, 2021.

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[3] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6247–6257, 2020.

[4] Xavier P Burgos-Artizzu, David Coronado-Gutiérrez, Brenda Valenzuela-Alcaraz, Elisenda Bonet-Carne, Elisenda Eixarch, Fatima Crispi, and Eduard Gratacós. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports*, 10(1):10200, 2020.

[5] Li Chen, Jonathan Rubin, Jiahong Ouyang, Naveen Balaraju, Shubham Patil1, Courosh Mehanian, Sourabh Kulhare, Rachel Millin, W. Gregory, Kenton, Cynthia R. Gregory, Meihua Zhu, David O. Kessler, Laurie Malia, Almaz Dessie, Joni Rabiner, Di Coneybeare, Bo Shopsin, Andrew Hersh, Cristian Madar, Jeffrey Shupp, Laura S. Johnson, Jacob Avila, Kristin Dwyer, Peter Weimersheimer, Balasundar Raju, Jochen Kruecker, and Alvin Chen. Contrastive self-supervised learning for spatio-temporal analysis of lung ultrasound videos. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023.

[6] Li Chen, Huilin Zhao, Hongjian Jiang, Niranjan Balu, Duygu Baylam Geleri, Baocheng Chu, Hiroko Watase, Xihai Zhao, Rui Li, Jianrong Xu, et al. Domain adaptive and fully automated carotid artery atherosclerotic lesion detection using an artificial intelligence approach (latte) on 3d mri. *Magnetic Resonance in Medicine*, 86(3):1662–1673, 2021.

[7] Sihong Chen, Weiping Yu, Kai Ma, Xinlong Sun, Xiaona Lin, Desheng Sun, and Yefeng Zheng. Semi-supervised breast lesion detection in ultrasound video based on temporal coherence. *arXiv preprint arXiv:1907.06941*, 2019.

[8] Daniel Cores, Manuel Mucientes, and Víctor M Brea. Roi feature propagation for video object detection. In *ECAI 2020*, pages 2680–2687. IOS Press, 2020.

[9] Ankan Ghosh Dastider, Farhan Sadik, and Shaikh Anowarul Fattah. An integrated autoencoder-based hybrid cnn-lstm model for covid-19 severity prediction from lung ultrasound. *Computers in Biology and Medicine*, 132:104296, 2021.

[10] Laura De Rosa, Serena L'Abbate, Claudia Kusmic, and Francesco Faita. Applications of artificial intelligence in lung ultrasound: Review of deep learning methods for covid-19 fighting. *Artificial Intelligence in Medical Imaging*, 3(2):42–54, 2022.

[11] Yanhua Gao, Bo Liu, Yuan Zhu, Lin Chen, Miao Tan, Xiaozhou Xiao, Gang Yu, and Youmin Guo. Detection and recognition of ultrasound breast nodules based on semi-supervised deep learning: a powerful alternative strategy. *Quantitative Imaging in Medicine and Surgery*, 11(6):2265, 2021.

[12] Tao Gong, Kai Chen, Xinjiang Wang, Qi Chu, Feng Zhu, Dahua Lin, Nenghai Yu, and Huamin Feng. Temporal roi align for video object recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1442–1450, 2021.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[14] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Cannot see the forest for the trees: Aggregating multiple viewpoints to better classify objects in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17052–17061, 2022.

[15] Glenn Jocher. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements. `https://github.com/ultralytics/yolov5`, Oct. 2020.

[16] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. Rethinking pseudo labels for semi-supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1314–1322, 2022.

[17] Wei Li, Yuanjun Xiong, Shuo Yang, Mingze Xu, Yongxin Wang, and Wei Xia. Semi-tcl: Semi-supervised track contrastive representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10399, 2021.

[18] Zhi Lin, Junhao Lin, Lei Zhu, Huazhu Fu, Jing Qin, and Liansheng Wang. A new dataset and a baseline model for breast lesion detection in ultrasound videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 614–623. Springer, 2022.

[19] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021.

[20] Vitalik Melnikov and Eyke Hüllermeier. Learning to aggregate using uninorms. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II 16*, pages 756–771. Springer, 2016.

[21] Jiahong Ouyang, Li Chen, Gary Y Li, Naveen Balaraju, Shubham Patil, Courosh Mehanian, Sourabh Kulhare, Rachel Millin, Kenton W Gregory, Cynthia R Gregory, et al. Weakly semi-supervised detection in lung ultrasound videos. In *International Conference on Information Processing in Medical Imaging*, pages 195–207. Springer, 2023.

[22] Ihor Protsenko, Taras Lehinevych, Dmytro Voitekh, Ihor Kroosh, Nick Hasty, and Anthony Johnson. Self-attention aggregation network for video face representation and recognition. *arXiv preprint arXiv:2010.05340*, 2020.

[23] Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, et al. Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. *IEEE transactions on medical imaging*, 39(8):2676–2687, 2020.

[24] Harshita Sharma, Richard Droste, Pierre Chatelain, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 987–990. IEEE, 2019.

[25] Daniel E Shea, Sourabh Kulhare, Rachel Millin, Zohreh Laverriere, Courosh Mehanian, Charles B Delahunt, Dipayan Banik, Xinliang Zheng, Meihua Zhu, Ye Ji, et al. Deep learning video classification of lung ultrasound features associated with pneumonia. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3102–3111, 2023.

[26] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[27] Jian Wang, Liang Qiao, Shichong Zhou, Jin Zhou, Jun Wang, Juncheng Li, Shihui Ying, Cai Chang, and Jun Shi. Weakly supervised lesion detection and diagnosis for breast cancers with partially annotated ultrasound images. *arXiv preprint arXiv:2306.06982*, 2023.

[28] Yuchen Wang, Zhongyu Li, Xiangxiang Cui, Liangliang Zhang, Xiang Luo, Meng Yang, and Shi Chang. Key-frame guided network for thyroid nodule recognition using ultrasound videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 238–247. Springer, 2022.

[29] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Advances in Neural Information Processing Systems*, 34:726–738, 2021.

[30] Jiang Xie, Ying Li, Xiaochun Xu, Jinzhu Wei, Haozhe Li, Shuo Wu, and Haibing Chen. Cptv: Classification by tracking of carotid plaque in ultrasound videos. *Computerized Medical Imaging and Graphics*, 104:102175, 2023.

[31] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyu Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3040–3049, 2021.

[32] Wufeng Xue, Chunyan Cao, Jie Liu, Yilian Duan, Haiyan Cao, Jian Wang, Xumin Tao, Zejian Chen, Meng Wu, Jinxiang Zhang, et al. Modality alignment contrastive learning for severity assessment of covid-19 from lung ultrasound and clinical information. *Medical image analysis*, 69:101975, 2021.

[33] Tianqi Yang, Nantheera Anantrasirichai, Oktay Karakuş, Marco Allinovi, and Alin Achim. A semi-supervised learning approach for b-line detection in lung ultrasound images. *arXiv preprint arXiv:2211.14050*, 2022.

[34] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[35] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[36] Hongye Zeng, Songhan Ge, Yuchong Gao, Jianhao Zhao, Shenghua Gao, Rui Zheng, et al. Vertmatch: A semi-supervised framework for vertebral structure detection in 3d ultrasound volume. *arXiv preprint arXiv:2212.14747*, 2022.

[37] Yupei Zhang, Joseph Harms, Yang Lei, Tonghe Wang, Tian Liu, Ashesh B Jani, Walter J Curran, Pretesh Patel, and Xiaofeng Yang. Weakly supervised multi-needle detection in 3d ultrasound images with bidirectional convolutional sparse coding. In *Medical Imaging 2020: Ultrasonic Imaging and Tomography*, volume 11319, pages 229–236. SPIE, 2020.

[38] Guojia Zhao, Dezhuag Kong, Xiangli Xu, Shunbo Hu, Ziyao Li, and Jiawei Tian. Deep learning-based classification of breast lesions using dynamic ultrasound video. *European Journal of Radiology*, 165:110885, 2023.

[39] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020.