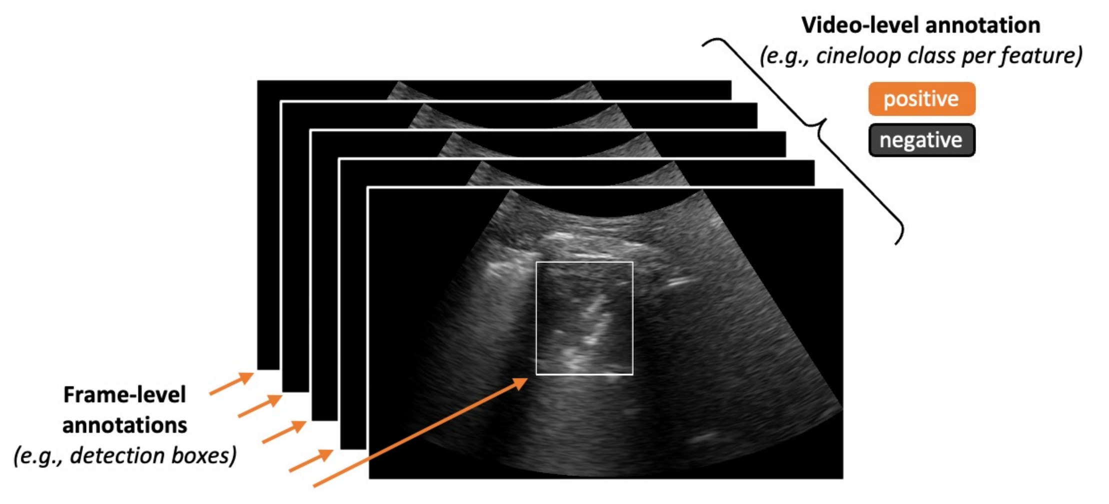
Weakly Semi-supervised Detector-based Video Classification with Temporal Context for Lung Ultrasound

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

[1]Philips Research North America, [2] Global Health Laboratories, [3] Oregon Health & Science University

Introduction



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

Technical challenge:

- Medical imaging tasks often require simultaneous frame detection and video classification.
  - ***Example:*** *Lung ultrasound detection of consolidation and pleural effusion*
- Standard detection models require frame-by-frame annotations for training, which are costly.
- Direct video classifiers do not provide localization, which limits clinical interpretability.

Contributions:

- Address tradeoff between annotation burden and interpretability
- Provide simultaneous detection and classification on medical videos while requiring very limited frame-level supervision
- Introduce a mechanism to aggregate feature representation from spatial to temporal
- Demonstrate real-world effectiveness on a multi-center clinical lung ultrasound dataset

Methods

Proposed framework:

1. ***Frame detection (weakly semi-supervised)[1]:***
   - Stage 1, "Burn-in": supervised initialization
   - Stage 2, Mutual learning: use frame- and video-labeled data for teacher-student training
2. ***Aggregate predictions along tracklets:***
   - Group predicted boxes into tracklets representing temporally connected regions
   - Extract tracklet clips from the original video
3. ***Tracklet classification:***
   - Second-stage network ("trackletNet") for tracklet classification using:
     - enriched dataset of challenging examples from incorrect detector predictions
     - weak semi-supervision using both frame and video annotations
4. ***Video classification:***
   - Based on highest tracklet confidence

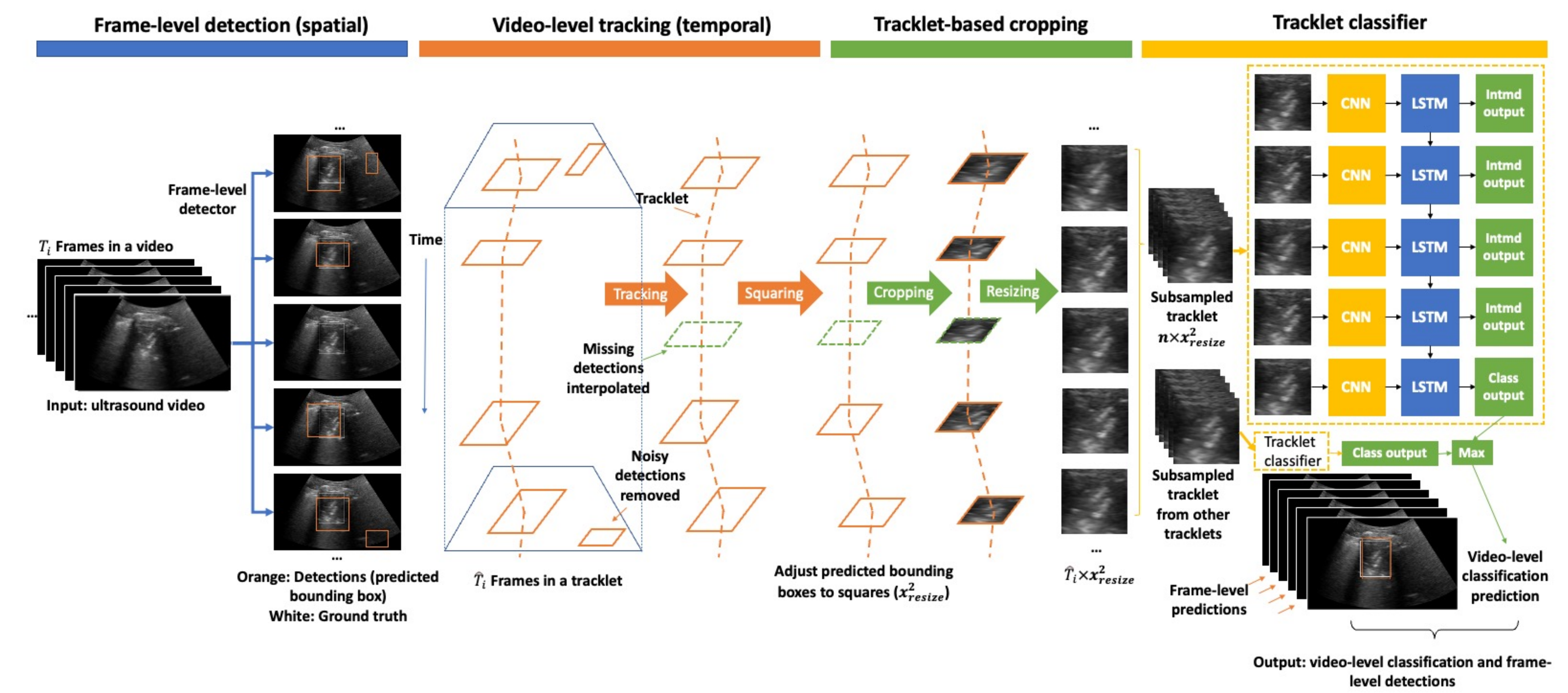


Figure 1. Detection and video classification framework. The method aggregates boxes from a 2D frame detector into tracklets, which are classified using a second-stage (CNN+LSTM) network. Both the frame detector and tracklet classifier are trained via weak semi-supervision using frame- and video-level labels.

Experiments

Data:

- **Multi-center dataset** of 7,712 ultrasound videos from 420 patients at 8 sites (60 to 180 frames per video).
- Training, validation, and testing datasets separated by subject:

	Consolidation	Pleural effusion
Training	99 frame-labeled videos, 6,677 video-labeled videos	80 frame-labeled videos, 9,836 video-labeled videos
Validation	337 videos	273 videos
Testing	599 videos	233 videos

Experiments:

- **Base classification model:** CNN + LSTM
- **Base detector-based classification model:** STN[2]
- **Reduce ROI:** bypass detection step → directly use whole image to train and evaluate trackletNet (Table 1, row 3)
- **Simple rule-based aggregation:** bypass tracklet classifier → classify video based on max detection confidence (Table 2, rows 2 and 3)
- **Remove tracker:** bypass tracking step → directly use frame detection confidences for video classification
- **Remove temporal aggregation by trackletNet:** classify tracklet based on single (central) frame (Table 2, row 4)

Table 1. Frame detection and video classification results for consolidation and pleural effusion.

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. Ablation experiments for consolidation frame detection and video classification.

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**

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

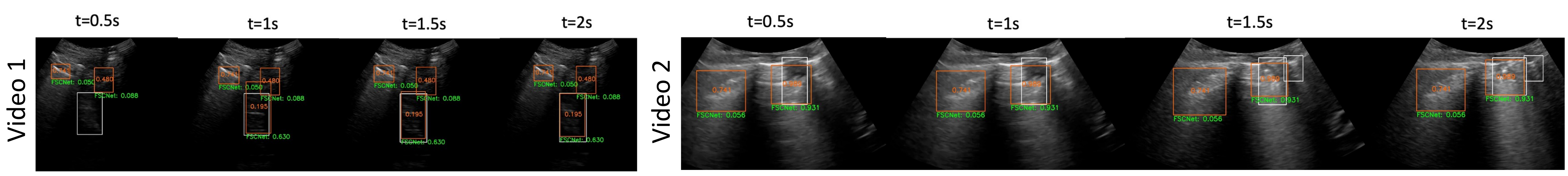


Figure 2. Examples of videos correctly classified by trackletNet but not by frame detector. **White**: ground-truth; **Orange:** detector confidences; **Green:** tracklet confidences.

[1]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.
[2]Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Lui- jten, 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.