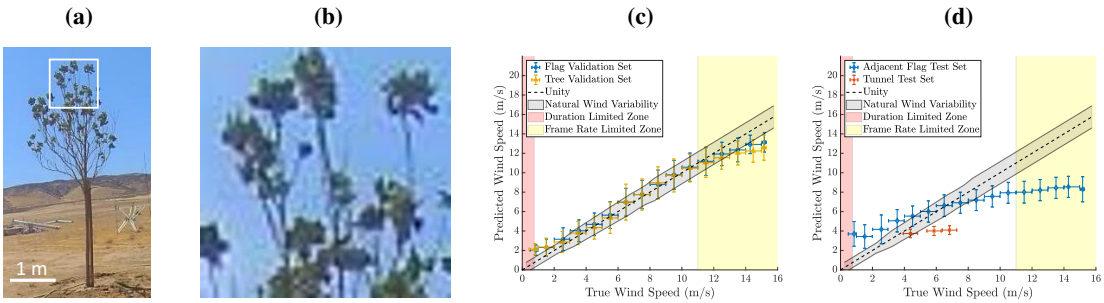

Figure 1: a) Tree used for expanded dataset (sample crop region shown in white box); b) Example of tree canopy model input; c) Validation set results (horizontal error bars represent range of wind speeds within each bin); d) Test set results

To all reviewers: The primary question raised in the reviews was the extent to which the results would generalize beyond the checkered flags in the initial data set. To directly address this concern, we have retrained the model using videos of both checkered flags and a tree that we planted at the field site (Fig. 1a, 1b). The model performs well on the validation sets of the flag and the tree (Fig. 1c), demonstrating the potentially broad application of the presented method. We also corrected an error in the original adjacent flag test set, which results in a flatter profile, especially in the high wind speed range, where we expect the model to be Nyquist-limited by the frame rate (Fig. 1d). This correction does not change the conclusions of the paper.

Reviewer 1: While we do not claim a novel ML architecture, we believe that our work is appropriate for the Applications subject area at NeurIPS because we make a novel application of known techniques to an important problem. Installing an anemometer to monitor a single location costs 2000-3000 USD, and even then only offers measurements at one location. This application of ML potentially enables spatially resolved measurements within a full 2-D scene, as opposed to existing single-point anemometer measurements. To clarify, while we have installed flags and trees at a field site to collect initial training and test data, the application of this method would occur using pre-existing structures in the environment of interest, such as flags and trees. Therefore, the only cost of this method is in the recording device. A standard camera phone provides sufficient resolution, hence the cost of this method is dramatically lower per measurement point. We have also collected an unprecedented dataset, comprising over 50 hours of raw video data of flapping flags and trees in the field, which will be available for others to use. Although other model types could also be well suited for this task, we chose a CNN+LSTM model to leverage transfer learning and parameter sharing to minimize training time, especially because of the demanding data collection and pre-processing required for this work. The LSTM model also allows for different clip lengths to be used in future iterations, which may be important if different frame rates or durations are necessary to capture physics (Sec. 5.1.1). 1-minute averages were used for labels because of the highly turbulent and variable flow. The anemometer and flag are spatially separated, so the instantaneous measurements made by the anemometer do not correspond to exact instantaneous speeds experienced by the flag.

Response to reviewer 2: Vertical error bars represent one standard deviation of the mean prediction in each wind speed bin. This is used to estimate the model capabilities over the range of wind speeds. Horizontal error bars were added to show ranges of true speeds within each bin. For the turbulence fluctuation band, 2-second averages were chosen to match the clip length. Although the anemometer is situated higher than the flag, turbulence intensity typically decreases with height in the atmospheric boundary layer, so this band represents a conservative estimate for the fluctuations at the flag height. There are plausible explanations for why the test set predictions lie in a narrower range than the validation set predictions. For the tunnel test set, the flag length is shorter (0.37 m), which means the model may be limited by physics at even lower speeds (Section 5.1.1). In all sets, we still see monotonically increasing predictions in the frame rate limited zone, suggesting that frame rate only partially limits model capabilities. Predictions for the corrected adjacent flag test set are also flatter, which suggests that the model may be partially over-fit to the specific flag and tree it has been trained on (i.e. relying partly on specific features of those objects). The effect of overfitting will become less significant as additional field data is collected for the flags and tree training sets.

Response to reviewer 3: In response to the reviewer's comments, we have significantly expanded our training and validation sets to include tree data, which is qualitatively different from the original flag videos. Results suggest that the presented technique can indeed be extended to other objects given the additional data. Although wind speed labels are 1-minute averages, the problem is framed as a regression, with a regression output layer that allows for the model to predict any wind speed as opposed to a specific class. We quantify predictions using RMSE for given wind speed ranges in Table 2, and the error bars on the markers indicate bin-specific performance estimates. If accepted, the paper will be revised to include a more extensive literature review on wind speed estimation, and the reviewer's specific comments will be incorporated.