Average maximal F1 score on the OOD dataset % of the "blue" instance in the training data 0.67 0.83 0 0.5 0.61 0.9 0.67 25 0.8 50 0.67 0.67 0.88 0.7 0.67 0.89 75 0.6 0.67 0.89 100 -0.5 25 50 75 100 % of the field background in the training data