Much of the impressive progress in object detection is built on the methodologies of statistical machine learning, which makes use of large training datasets. Consider the benchmark results of the well-known PASCAL VOC object challenge over the past 5 years (above). We see a clear trend in increased performance over the years as methods have gotten better and training datasets have become larger. In this work, we ask a meta-level question about the field: will continued progress be driven faster by increasing amounts of training data or the development of better object detection models?

To answer this question, we collected a massive training set that is an order of magnitude larger than existing collections such as PASCAL [4]. We follow the dominant paradigm of scanning-window templates trained with linear SVMs on HOG features [1, 2, 5, 6], and evaluate detection performance as a function of the amount of positive training data \( N \) and the model complexity \( K \), where \( K \) is measured by the number of mixture components capturing variations in object sub-categories, 3D viewpoint, etc.

We found there is a surprising amount of subtlety in scaling up training data sets in current systems. For a given model, one would expect performance to generally increase with the amount of data, but eventually saturate. Empirically, we found the bizarre result that off-the-shelf implementations often decrease in performance with additional data! One would also expect that to take advantage of additional training data \( N \), it is necessary grow the model complexity \( K \). However, we often found scenarios in which performance was relatively static even as model complexity and training data grew (Fig 2).

In this paper, we offer explanations and solutions for these phenomena. First, we found it crucial to set model regularization as a function of the amount of training data \( N \) using cross-validation, a standard technique not typically deployed in current object detection systems. Second, existing strategies for discovering subcategory structure, such as clustering aspect ratios [5] and appearance features [3] may not suffice. We found this was related to the inability of classifiers to deal with “polluted” data when mixture labels were improperly assigned (Fig. 3). Increasing model complexity \( K \) is thus only useful when mixture components capture the “right” sub-category structure (Fig. 4). Finally, we found that is was easier to capture the “right” structure with compositional representations; we show that one can implicitly encode an exponentially-large \( K \) by composing parts together, yielding substantial performance gains over explicit mixture models (Fig.5). We conclude that there is currently little benefit to simply increasing training dataset sizes. But there may be significant room to improve current representations and learning algorithms, even when restricted to existing feature descriptors.