1. Introduction

Given an single image, humans can readily distinguish plausible subsequent scenes from implausible ones. In the process, we make use of enormous amounts of visual common-sense knowledge about how the world works: people tend to walk forwards and things that are “attached” to people tend to stay attached. In this work, we investigate learning these sorts of dynamic common-sense knowledge and use our knowledge to predict subsequent scenes.

What should be represented in spatiotemporal common-sense? We believe that a model should capture these at least five factors, and incorporate them in our approach. Absolute position. As is common in the spatial context literature [1], we model the positions of objects in absolute terms (e.g., the sun is unlikely to appear below the horizon). Relative position. We also model position in relative terms (e.g., people should be under the sun). Absolute motion. People tend to move in the direction they are facing and soccer balls that are about to be kicked tend to move away from the kicker. We therefore model how each object is likely to move. Relative motion. Much of what separates the plausible from the implausible are joint motions: peoples’ hats and sunglasses tend to move with them. It is therefore not enough to predict each object’s dynamics separately but we must instead model the interdependencies. Attributes. Objects change appearance over time: among other things, for instance, people change pose and facial expression. We illustrate four of these factors in Fig. 1.

Two enormous challenges stand in the way of learning these factors from data: at training time, a lack of data and at inference time, a lack of models capable of recognizing objects of interest (e.g., sunglasses, hands) with sufficient accuracy in unconstrained images. Rather than defer our investigations until these challenges are solved, we turn to abstract scenes [3], which allow us to easily gather large amounts of data with bounding-box-level annotations. These abstract scenes allow us to test multiple models for prediction and determine what aspects of prediction are important to model. Furthermore, we also demonstrate the surprising ability of our model to generalize to natural images.

2. Method

The goal of our model is: given a scene $S_t$, produce a following scene $S_{t+1}$. We achieve this by building a Conditional Random Field (CRF) model that evaluates the plausibility of a scene given the previous scene. Phrasing the problem as a CRF enables us to encode our beliefs about the underlying absolute and relative factors of prediction in a principled way and resolve conflicting beliefs. Each object and its attributes is represented by a node in our model with no special properties given to any particular object and we convert each of the five factors enumerated into potentials: absolute motion, for instance, is a unary potential and relative motion is a pairwise potential. Together, these potentials characterize the likelihood of a scene configuration given the previous scene.

Learning the model: Location potentials do not depend on the previous scene and thus are the empirical distribution observed in the training data. Motion and attributes are dependent on the previous scene, and we learn a mapping from the previous scene to a distribution over motion vectors with a Random Forest.

Inference: At inference time, a multiple-restart algorithm in the style of iterative conditional modes is used to sample new scenes, which the CRF can score.

Features: We use an encoding of the position of the presence and location of objects as well as their attributes.
### 3. Results

**Data:** We gathered a dataset of 5,000 5-scene sequences of abstract scenes containing objects described in [3] using Amazon Mechanical Turk (AMT).

**Baselines:** We compare against three automatic baselines. In No CRF, we use the same features but independently predict the location and attributes of each object, without enforcing relative constraints. In (BoW Copy, Transfer), we use a global model that predicts the whole scene at once. We represent the scene with a bag-of-words representation, find its nearest neighbor, and copy the following scene or transfer the relative motions.

**Qualitative Results:** Given a test scene, we ask AMT users to judge the relative quality of predictions by two methods. Our method is preferred by a substantial margin over all baselines. We show consistently rated results in Fig. 3: frequently, holistic predictions produce continuity errors and independent ones produce implausible joint dynamics.

We show two qualitative results illustrating our model in Fig. 2. Without a priori knowledge of agency, our model learns that hot dogs tend not to move unless there is a person to move them and picnic tables are stationary. Similarly, it learns that people are more likely to move forward than backwards, but not if there are bears.

**Quantitative Results:** We quantitatively evaluate prediction by asking yes/no questions about the prediction and the ground truth. Our method achieves especially good relative performance on joint motion tasks like “Does the boy move with the ball?”, yielding $2.5 \times$ the performance (in F1 score) of the next-best automatic baseline over all objects. This suggests that joint dynamics do not automatically emerge from independent modeling but must instead be explicitly encoded. A more detailed analysis appears in the full-length version of this work [2].

**Natural Images:** Finally, we gather a dataset of 225 Flickr images with bounding-box annotations of our clipart objects. We ask annotators to answer questions like “Will this ball move?” and compare their responses to predictions by the automatic systems. Our system outperforms the other baselines in terms of motion, dramatically so in terms of joint motion. We also show qualitative predictions of people’s attributes by our method in Fig. 4.

### References

