SAFER: Fine-grained Activity Detection by Compositional Hypothesis Testing

Tae Soo Kim Yi Zhang Zihao Xiao Michael Peven Weichao Qiu Jin Bai
Alan Yuille Gregory D. Hager

Abstract

The performance of a vision system on a set of public benchmarks does not necessarily indicate the expected performance in many practical applications. Indeed, practical use case such as video surveillance requires a vision system to parse visual information at a much finer granularity than in most widely used benchmarks, be flexible enough to deal with the 'open-set' nature of the domain and be transparent by design such that system performance is explainable and easily modified as needed.

In this work, we present SAFER, Slots and Fillers for Explainable Reasoning, for untrimmed activity detection. SAFER is designed to tackle such demands. SAFER models a large space of fine-grained activities using a small set of detectable entities (slots) and their interactions (fillers). Such a design scales effectively with concurrent developments of slot detectors involving object detectors, object parsers and more. Moreover, as SAFER defines a decompositional structure of activities into detectable slots, simulation can be used for training slot detectors when the desired slot is otherwise unavailable.

We demonstrate that SAFER extends easily due to its compositional nature, is more interpretable and most importantly, generalizes more effectively to unseen test samples. Our evaluations on the challenging DIVA dataset for activity detection in surveillance show that SAFER generalizes more effectively to unseen videos at test time, improving the available deep CNN baseline on the DIVA test set evaluation by 11.4 % with no end-to-end training of activity classes. Code is available here: (omitted during blind-review process).

1. Introduction

As with many areas of computer vision, the pattern recognition capabilities of convolutional neural networks (CNNs) have led to significant progress on many benchmarks for activity analysis [2, 17, 30, 33, 35, 38, 39]. However, these benchmarks typically have one or more of the following properties: 1) they focus on classification of pre-segmented video rather than spatial and temporal detection of a class; 2) they are often focused on coarse-level analysis (e.g. "a person walking") rather than fine-grained relational analysis (e.g. "a person dragging a heavy cart"); and 3) they typically are architected around a pre-defined set of labels supported by a large and relatively balanced set of examples for those labels. Taken together, this has led to an emphasis on ever more advanced regression-style approaches, whereby a neural network model is trained from data and tested (scored) on held-out examples from the same label set.

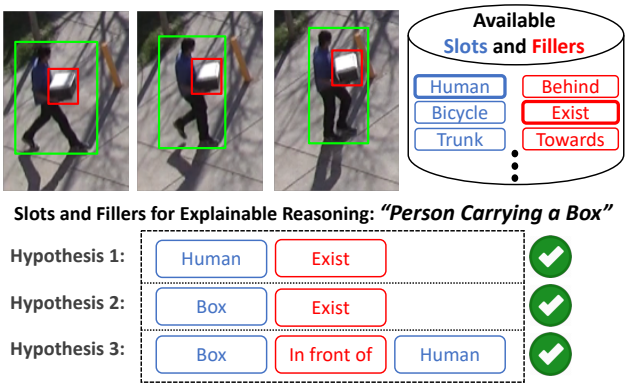


Figure 1: SAFER is a structured modeling approach where activities are defined as time-series of actors, objects, object parts and interactions. By composing a set of available *slots* (detectable entities) and *fillers* (interaction of slots) , activities can be quickly defined and detected through interpretable hypothesis testing.

There are many cases where these assumptions break down, either because the application demands fine-grained detection of a potentially combinatorial number of activities, and/or because the problem at hand is an "open-set" problem where new labels may be defined at test time, and/or because sufficient labeled data is difficult to come by. For example, in surveillance video, it is quite common to perform retrospective analysis where the goal is to find examples of specific activities in a zero-shot manner, e.g. "locate instances where a light brown package is being

placed under a car by a man wearing a gray parka."

In this paper, we present SAFER, a compositional Slots And Fillers for Explainable Reasoning approach to activity detection that is designed to address these limitations. SAFER models activities as explicit, structured time-series models of objects, object parts, and relationships among them. Intuitively, slots denote object-level attributes (e.g. person, bicycle, car), and fillers are the time-series evolution of the presence and location of these attributes. Activities are detected through compositional hypothesis testing on Slots and Fillers. For example, *Riding a Bicycle* is defined by a specific spatial and temporal co-occurrence of a human and a bicycle for an extended period of time.

By driving activities to more fundamental units, a small number of image-level attributes can be (re)used to define different activities – e.g. a person entering or exiting a vehicle share the same set of Slots (car, car door, human) but express opposite temporal behavior (fillers). More generally, many activities of interest can be modeled as dynamic interactions of objects – cooking, surgery, construction, and repair, to name a few. Not only does this play to the power of a compositional framework that allows re-use of modular components, but it also supports zero-shot, i.e. at test time, detection of novel activities, and models are more interpretable as compared to purely data-driven deep models. Moreover, we demonstrate that such structured modeling approach shows better generalization behavior compared to end-to-end models. Finally, an additional advantage of SAFER is that it provides structured means of breaking complex activities down into simple image components which can be trained using relatively few labels, and which can sometimes be systematically boosted using simulated data [18, 31].

In the remainder of this paper, we describe the structure of SAFER and illustrate its use using the DIVA dataset[1] which is a challenging activity detection benchmark with surveillance videos across multiple scenes. We demonstrate the effective generalization performance of SAFER on three evaluation settings. We provide evaluation 1) on the provided validation set, 2) on Leave-One-Scene-Out evaluation which measures SAFER's potential to generalize to novel scenes and 3) on the reserved DIVA test set where unseen samples from novel scenes compose the majority of the data. In summary, our contributions are:

- The formal definition of a compositional framework for activity detection.
- The development of a hybrid deep-learned and hypothesis testing framework for detection.
- The evaluation of this framework with emphasis on generalization to completely new scenes where we show over 10% decrease in probability of a missed detection compared to a published end-to-end baseline on the benchmark.

[1] https://actev.nist.gov/

2. Related Work

A wide range of prior work exists on activity detection in the computer vision community but we identify the following categories of approaches as the most relevant to the presented work: i) pattern recognition based deep CNN approaches for activity analysis, ii) higher level reasoning of activities through structured analyses and iii) vision models exploiting simulation as the means obtaining labeled data for training.

Deep CNN based approaches for activity analysis: End-to-end CNN-based pattern recognition models dominate the well known public video benchmarks [2, 17, 30, 33, 35, 38, 39]. However, when the problem domain of model deployment is wildly different from the representative set found in the available training data, such models fail to generalize to the desired domain. In fact, an activity detection system [36] that has proven to be effective across multiple popular public benchmarks such as [10, 14] does not generalize well to activity detection in surveillance environments as demonstrated in this work and in [7]. Spatio-temporal action localization approaches such as [6, 16] alleviate the assumption that the activities are already contained in the observed video (hence performing activity detection). Nonetheless, most instances are either YouTube videos or structured recordings where the action foreground covers a relatively large portion of the observed image sequence. Where as in our application in the DIVA dataset, only 3 % of the image consists of relevant foreground [7]. The most relevant prior work is [7] as the authors exploit object detections to generate activity proposals on which a variant of I3D [2] is trained in an end-to-end fashion to classify the spatio-temporally localized images. Our model is similar in that objects (slots in our formulation) form the basis of activity proposals but SAFER framework further models activities as direct composition of slots and fillers instead of learning to classify them using end-to-end CNNs. In our formulation, end-to-end pixel wise mapping happens at the slot level, not at the activity class level which we believe transfers more effectively given completely novel samples.

Higher level reasoning of activities through structured analyses: Abstracting out lower level visual information with higher order attributes has demonstrated effective knowledge transfer [19, 21, 37]. Attributed based models for video understanding [29, 34] share the common motivation. In SAFER formulation, higher level abstraction happens at the slot level. For example, in our formulation, we treat opened-door as a detectable slot but it can alternatively be defined as an attribute of a vehicle. Module network [1] and its variants [12, 13, 15, 22] present a compositional and modular reasoning of visual information albeit for a different application. Temporal Modular Net-

work [20] extends module networks to the video domain and demonstrates efficient re-use of visual concepts by exploiting compositionality of the approach. SAFER formulation is by design compositional where we detect activities by composing available slots and fillers.

Synthetic data for training: Given the ability to produce fine-grained scene annotations at scale, simulation has become a popular source of obtaining labeled data for training models in wide variety of applications involving object detection [8], segmentation [26, 28] and object parsing [18, 32]. Development of tools for simulation [3, 25] and domain adaptation approach such as [11] offer strong motivation to exploit simulation further. SAFER defines activities through slots where simulation can be used to obtain labeled training data when the slot detector is otherwise unavailable. We believe slot synthesis is a more effective use of simulation than simulating fine-grained interactions of multiple objects [27].

3. Methods

As with recent methods for object detection [4, 5, 9], we frame activity detection as a sequence of first generating activity *proposals* followed by the filtering and refinement of those proposals into the final activity detections. In the following sections, we first describe proposal generation and then illustrate filtering and refinement starting with a single slot with a trivial filler then gradually extend to activities involving interactions of actors (objects) including spatial and the temporal relations between them.

3.1. Proposals

Given an image sequence $I = [I_1, I_2, \ldots, I_T]$, we represent an activity proposal A as a temporal sequence of bounding box locations, $A = [B_t ... B_{t^*}], 1 \leq t < t^* \leq T$ within I. Each bounding box B in turn consists of a collection of one or more slot-specific bounding boxes, B^j, $j \in \mathcal{S}$ where $\mathcal{S}$ is the set slots which may include objects, object parts, or object configurations for which a detector is available. For each slot $j \in C$ and corresponding bounding box B^j, a slot detector provides a detection score $S_j(I_t, B_t^j) = P(j|I_t(B^j))$ for that class. In practice, $S_j(I_t, B_t^j)$ will be the output of a CNN with input $I_t(B^j)$.

The SAFER formulation requires spatio-temporally localized regions of video. Temporal grouping (tracking) of each slot j or existing action foreground segmentation approaches such as [6, 36] provide an initial set of activity proposals $\mathcal{A}^H = \{A_1, \ldots A_N\}$ where N is the total number of activity proposals.

3.2. Evaluating Proposals with Slots and Fillers

3.2.1 Basic Hypothesis Test for Presence of Object

To establish the basic framework, suppose we have an activity hypothesis $A = [B_1^j ... B_T^j]$ consisting of a sequence of t temporally grouped bounding boxes. We wish to evaluate the likelihood of A being generated by slot j and a filler defining the persistence of slot j throughout the proposal. Consider the following two hypotheses:

1. A is due to j (i.e. human exists).
2. A is due to $\neg j$ (i.e. human does not exists).

The likelihood of the first hypothesis can be written as:

$$P(A|j) = \prod_{t=1}^{T} P(I(B_t^j)|j) \tag{1}$$

and the alternative hypothesis as:

$$P(A|\neg j) = \prod_{t=1}^{T} P(I(B_t^j)|\neg j) \tag{2}$$

We can not explicitly compute the probability distributions $P(I(B_t^j)|j)$ or $P(I(B_t^j)|\neg j)$. However, we only wish to do model selection and thus compare the ratio of the probabilities. This can be performed by the log-likelihood ratio-test:

$$\prod_{t=1}^{T} P(I(B_t^j)|j) > \prod_{t=1}^{T} P(I(B_t^j)|\neg j) \tag{3}$$

$$\sum_{t=1}^{T} \log \frac{P(I(B_t^j)|j)}{P(I(B_t^j)|\neg j)} > \tau \tag{4}$$

This leads to a formulation that the sum of log likelihood ratios or log evidence should be greater than a design parameter τ. Applying Bayes rule to the ratio in Eq. 4

$$\begin{aligned} \frac{P(I(B_t^j)|j)}{P(I(B_t^j)|\neg j)} &= \frac{P(j|I(B_t^j))P(\neg j)}{P(\neg j|I(B_t^j))P(j)} \\ &= \frac{P(j|I(B_t^j))}{1 - P(j|I(B_t^j))} \cdot \frac{P(\neg j)}{P(j)} \end{aligned} \tag{5}$$

This means that in the SAFER framework, we need to be able to compute $P(j|I(B_t^j)) = S_j(I_t, B_t^j)$ to perform hypothesis testing for presence of slot j given a proposal A. As discussed earlier, $S_j(I_t, B_t^j)$ is an output of a detector for slot j. It is worth noting the final term is the prior odds ratio which represents the quality of proposals.

3.2.2 Co-occurrence of Slots

Co-occurrence of slots contains potentially very discriminative information for activity detection and can be composed

in a straight forward manner. For example, a proposal sequence that contains both a human and a bicycle is highly likely to be a *Riding* sequence from the DIVA dataset where as a sequence with human and a heavy object is likely a *Transport Heavy Carry* instance.

Consider trying to detect the co-occurrence of slots j and k. Given activity proposal A, we test the following hypotheses:

1. A is due to j and k.
2. X is due to $\neg j$ and $\neg k$.
3. X is due to j and $\neg k$.

We omit the fourth case ($\neg j$ and k) because we assume that j is the main object of interest (for example, the human in many interactions), which we first test for using the results of the previous subsection. Similar to the derivation in Eq. 4, we compare the first and the second hypotheses:

$$\sum_{t=1}^{T} \log \frac{P(I(B_t^j)|j)P(I(B_t^k)|k)}{P(I(B_t^j)|\neg j)P(I(B_t^k)|\neg k)} > \tau \tag{6}$$

Using Bayes rule again we derive:

$$\begin{aligned} &\sum_{t=1}^{T} \log \frac{S_j(I_t, B_t^j)}{1 - S_j(I_t, B_t^j)} + \\ &\sum_{t=1}^{T} \log \frac{S_k(I_t, B_t^k)}{1 - S_k(I_t, B_t^k)} + C > \tau \end{aligned} \tag{7}$$

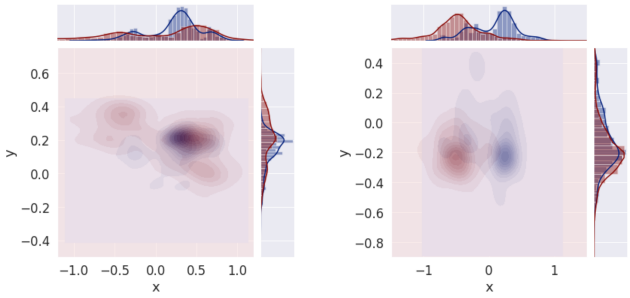
where C is a constant factor (sum of each object's prior log ratios). The additional hypothesis test between 1 and 3 yields:

$$\begin{aligned} &\sum_{t=1}^{T} \log \frac{P(I(B_t^j)|j)P(I(B_t^k)|k)}{P(I(B_t^j)|j)P(I(B_t^k)|\neg k)} > \tau \\ &= \sum_{t=1}^{T} \log \frac{S_k(I_t, B_t^k)}{1 - S_k(I_t, B_t^k)} > \tau \end{aligned} \tag{8}$$

The intuition for the additional test in Eq. 8 is that the evidence of the first hypothesis may be larger than the second hypothesis but the third hypothesis is true. Returning to the *Riding* example, such scenario would be the 'Human without a bicycle' case and without the second test, false positive detections containing only humans in the proposal may be a common failure mode. Within the SAFER framework, co-occurrences of more than two slots is a straightforward extension of the previous derivation and can be composed easily.

3.2.3 Spatial Relations Between Slots

We extend the aforementioned case to include a filler condition modeling the spatial relation of slots. Suppose a particular spatial configuration $C(B_t^j, B_t^k)$ exists for slots j and k. For example, $C(B_t^j, B_t^k)$ may represent a relationship that j is 'in front of' of k.



(a) (b)

Figure 2: Effects of different designs of filler modeling spatial relation of slots. (a) Distributions of relative positions of humans and objects, (b) conditioned on the direction of human motion. Blue: Distribution computed for *Transport Heavy Carry* from DIVA, Red: from *Pulling*. Modeling spatial relations between humans and objects provides strong cues for activity detection. Best viewed in color.

Then, we can learn the probability distributions $P(B_t^k|B_t^j, C(B_t^j, B_t^k) = 1)$ and $P(B_t^k|B_t^j, C(B_t^j, B_t^k) = 0)$. With this, spatial relations between slots can be composed as an additional term in the hypothesis test by adding the log-likelihood ratio:

$$\sum_{t=1}^{T} \log \frac{P(B_t^k|B_t^j, C(B_t^j, B_t^k) = 1)}{P(B_t^k|B_t^j, C(B_t^j, B_t^k) = 0)} \tag{9}$$

The intuition here is that the spatial relations between slots may discriminate activities with identical slots. For example, *Transport Heavy Carry* and *Pulling* activities by definition consist of identical slots (a human and a heavy object) but the spatial relation between the slots can discriminate the two classes. In implementation, motion cue information is used to extract approximate relative position between objects and humans. Figure 2 illustrates the effect of modeling spatial relations to distinguish between the two classes.

3.2.4 Temporally Dynamic Slots: Appearing and Disappearing

Consider an activity proposal A which contains a temporal state change in slot j. More concretely, it is a sequence where we observe an attribute j until some time T^* and do not detect j for the remainder of the proposal (j disappearing) or vice-versa (j appearing). The derivation is similar to Section 3.2.1 except the sum of evidence is temporally split

into two portions:

$$\sum_{t=1}^{T^*} \log\left(\frac{S_j(I_t, B_t^j)}{1 - S_j(I_t, B_t^j)} \cdot \frac{P(\neg j)}{P(j)}\right) + \sum_{t=T^*+1}^{T} \log\left(\frac{1 - S_j(I_t, B_t^j)}{S_j(I_t, B_t^j)} \cdot \frac{P(j)}{P(\neg j)}\right) > \tau \tag{10}$$

The term $P(j)$ represents the prior of the attribute j and assuming an unbiased distribution of attributes such that $\frac{P(j)}{P(\neg j)} = 1$, we can rewrite Eq. 10 as follows:

$$\sum_{t=1}^{T} V_c^j(t) \cdot \log\left(\frac{S_j(I_t, B_t^j)}{1 - S_j(I_t, B_t^j)}\right) > \tau$$
$$V_c^j(t) = \begin{cases} 1 & \text{, if } \ t < T^* \\ -1 & \text{, else} \end{cases} \tag{11}$$

where $V^j \in \mathbb{R}^T$ is a vector representing a filler that models disppearance of j. V^j is the expected temporal behavior of j or the *template* sequence of j. Note that the sum in 3.2.1 is a specific case of Eq. 11 where V^j is equal to 1 for all timesteps.

3.3. Learning

In SAFER, activities are detected by composition of hypothesis tests that are defined by slots and fillers. This allows activities to be defined 'on-the-fly' without end-to-end training. Instead, we map activity pixels to class labels given appropriate detectors for slots and the means to compose temporal interactions with fillers. Therefore, learning process only involves gathering or training a set of detectors $P(j|I(B^j))$, i.e. slot detectors, which are typically simpler and thus require fewer labeled examples. High level slots such as detecting a person or a car can be trained from publicly available data sets. However, more detailed slots describing a particular state of object parts are sometimes needed to enable SAFER to detect activities at a more granular level. When the desired slot detector is unavailable due to a lack of sufficient labeled data, we exploit simulation to train a detector. A full list of the slots, fillers and activity definitions employed in this paper is provided in the supplementary material.

4. Activity Detection Evaluation

The ability of a model to generalize to unseen data is arguably the most important metric when evaluating a model built for activity detection. Our model was designed under the assumption that invariance to the observed scene is the most desirable parameter. Here, we present results of our model in two scenarios: 1) under the provided benchmark, evaluating on videos from the **same** scenes (camera locations) in the training set, albeit at different times, and 2) in a custom Leave-One-Scene-Out (LOSO) evaluation, using videos from **disparate** scenes for training and testing. We show how SAFER outperforms deep CNN baselines by a larger margin on the LOSO evaluations than in the standard setup, leading us to conclude that it has a higher capacity for generalization to novel data. Furthermore, we demonstrate that the interpretable latent state in SAFER allows to improve results in a fashion that can't be done on black-box deep models.

Figure 3: In the DIVA dataset, variation across scenes are severe but instances within the same scene share visual similarities. Experiment design to explicitly test for detection performance on unseen scenes. The figure depicts the LOSO-0000 experimental setup. Yellow: training, Green: model selection, Blue: Inference. Best viewed in color.

4.1. DIVA Dataset Evaluation Settings

The DIVA dataset is an untrimmed activity detection dataset that requires both spatial and temporal localization of predefined set of activities. The videos originate from the VIRAT dataset [24] and annotations more suitable for activity detection were collected by the IARPA DIVA program. The DIVA dataset consists of 64 training and 54 validation videos with full set of annotations. 96 videos are reserved for the held out test set such that associated annotations are made unavailable.

DIVA test evaluation favors models that generalize effectively to unseen data given that 74 of the 96 videos are from novel scenes that are not part of the provided available 5 scenes during training. As illustrated in Figure 3, there exists large variation across different scenes including noticeable discrepancies in viewpoint, scale, level of occlusion and actor identities. However, intra-scene activity instances are visually comparable. Under such conditions, end-to-end models are more likely to overfit the available training data. To explicitly test for model generalization, we perform experiments on the LOSO setup. An illustration of the LOSO experimental setup is presented in Figure 3. Under LOSO, a sample at test time strictly comes from a held-out scene.

	P_{miss}@.15rfa		P_{miss}@1rfa	
	TRN	SAFER	TRN	SAFER
Closing	**0.909**	0.955	**0.720**	0.765
ClosingTrunk	**0.810**	0.952	**0.667**	0.762
Entering	0.676	**0.667**	0.577	**0.183**
Exiting	0.831	**0.789**	0.738	**0.231**
OpenTrunk	**0.591**	0.818	**0.409**	0.818
Opening	0.921	0.921	**0.756**	0.771
HeavyCarry	0.677	**0.484**	0.452	**0.323**
Pull	0.739	**0.304**	0.304	**0.174**
Riding	0.864	**0.545**	0.456	**0.364**
Talking	0.805	**0.683**	0.683	**0.414**
Mean P_{miss}	0.782	**0.709**	0.576	**0.480**

Table 1: Performance on provided validation set of the DIVA dataset which contains identical set of scenes as the train set.

4.2. Activity Detection Metric with Operational Use In Surveillance

We adopt the same metric as [7, 23] that is the probability of missed detection (P_{miss}) at multiple false alarm rates (rate of false-alarm, *rfa*). First, one-to-one correspondence between pairs of ground truth and system output activity instance is found. Then, any unmatched detected prediction becomes a false positive detection while any ground truth instance without a prediction match is considered a miss detection. The optimal matching is found by the Hungarian Algorithm. We refer to the TRECVID 2017 [23] for details about the evaluation metric. The evaluation software used in this work is provided by the independent evaluation team through Github[2].

4.3. System Performance

DIVA Validation (Mixed Scene) Evaluations: We first provide our activity detection results on the DIVA validation set in Table 1 that compares the baseline Temporal Relation Network (TRN) [38] approach to SAFER. There does not exist published or publicly available **per-activity** baseline results for this dataset at the time of the submission thus requiring the additional TRN baseline. We selected a representative subset of the activities involving object interactions in the "Winter 18 ActEV Prize Challenge" evaluation defined by the ActEV Challenge. Both TRN and SAFER are evaluated on a set of identical proposals. We provide probability of missed detection, P_{miss}, at two operating points, 0.15 *rfa* and 1 *rfa*, for each activity.

We observe consistent improvements in activity detection performance on both operational points but the overall gap in performance is not significant. We believe under the mixed scene set up, TRN baseline performs relatively well given the visual similarly between samples during training and inference.

	Val. @ .15 rfa	Val. @ 1 rfa	Test @ .15 rfa	LOSO @ 1 rfa
R-C3D [36]	0.863	0.720	0.907	x
TR-I3D [7]	**0.618**	**0.441**	x	x
TRN [38]	0.717	0.574	0.872	0.853
SAFER	0.709	0.535	**0.793**	**0.563**

Table 2: Summary of comparisons across available, implemented baselines with SAFER under different evaluations. Evaluations on LOSO and the DIVA test set favors SAFER which generalizes more effectively than other approaches but TR-I3D [7] still provides a very strong baseline for the validation partition evaluation. For activities not modeled with SAFER, we merge predictions of TRN to produce a complete set of predictions.

Leave-One-Scene-Out (LOSO) Evaluations: We believe one of the major benefits of SAFER for activity detection is its ability to generalize to entirely unseen data and mixed scene evaluation fails to explicitly measure this dimension of the model. Thus, quantitative experimental results for LOSO evaluation of both models are reported in Table 3.

Under LOSO evaluation, the gap in performance between the TRN baseline and SAFER is considerably larger compared to the mixed scene results in Table 1. In fact, the TRN baseline fails to generalize at all for many activity classes and such generalization behavior is observed consistently across multiple scenes. At the same operating point of 1 *rfa*, SAFER outperforms the TRN baseline by 0.289 points on average whereas the gap is 0.095 in the mixed scene set up. The average of mean P_{miss} across five scenes for SAFER is 0.563 at 1 *rfa*

DIVA Test Evaluation: The DIVA test set favors models that generalizes more effectively as the evaluation is mostly performed on samples from novel scenes. To further demonstrate that SAFER performs well on de-novo instances, we report our DIVA test set experimental results using the public ActEV leaderboard[3] in Table 2. We demonstrate that SAFER improves the baseline approach of [36] by 0.114 which we believe is due to SAFER's higher capacity for generalization. Compared to presented quantitative analyses in this work, per-activity results are not made available through the leaderboard system. The reported metric is a weighted average of P_{miss} at 0.15 *rfa*. To the best of our knowledge, there does not yet exist published results consis-

[2]https://github.com/usnistgov/ActEV_Scorer/tree/v0.3.0

[3]https://actev.nist.gov

Held out scene:	0000		0002		0400		0401		0500	
Model:	TRN @ 1 rfa	SAFER @ 1 rfa	TRN @ 1 rfa	SAFER @ 1 rfa	TRN @ 1 rfa	SAFER @ 1 rfa	TRN @ 1 rfa	SAFER @ 1 rfa	TRN @ 1 rfa	SAFER @ 1 rfa
Closing	0.977	**0.636**	0.905	**0.889**	0.762	**0.643**	0.892	**0.814**	0.857	0.857
Closing_Trunk	1	**0.500**	0.833	0.833	0.900	**0.500**	0.967	**0.867**	n/a	n/a
Entering	1	**0.143**	0.733	**0.578**	0.684	**0.105**	0.792	**0.226**	0.333	**0**
Exiting	1	**0.095**	0.950	**0.750**	0.882	**0.411**	0.827	**0.269**	1	**0.429**
Open_Trunk	0.857	0.857	1	**0.833**	1	**0.769**	1	**0.967**	n/a	n/a
Opening	0.953	**0.535**	0.900	0.900	0.829	**0.634**	0.926	**0.863**	1	**0.667**
HeavyCarry	0.852	**0.556**	1	1	1	**0.625**	0.742	0.742	1	1
Pull	0.875	**0.375**	1	1	0.684	**0.474**	1	**0.703**	n/a	n/a
Riding	0	0	0.826	**0.304**	0.333	0.333	1	**0.455**	0.600	0.600
Talking	1	**0.821**	1	**0.550**	1	**0**	1	**0.625**	n/a	n/a
Mean P_{miss}	0.792	**0.451**	0.910	**0.673**	0.825	**0.449**	0.913	**0.653**	0.827	**0.592**

Table 3: Comparison of TRN and SAFER activity detection performance under the LOSO experimental setup. We observe larger performance gains due to better generalization of SAFER. n/a refers to the case when there does not exist a single instance of that activity in the scene.

(a) (b) (c)

Figure 4: Without modeling the spatial relations between slots, we cannot distinguish Pull from HeavyCarry. (a) Instance of Pull classified as HeavyCarry. (b) Instance of HeavyCarry classified as Pull. Both cases are detected correctly with spatial reasoning. (c) Failure case where a HeavyCarry instance is classified as Pull even with the additional spatial constraint. Note that the location of the heavy object is relatively lower than human which is common for Pulling instances.

(a) (b) (c)

Figure 5: Examples of false positive *Entering* and *Exiting* detections from SAFER due to missed slot detection.

	no spatial constraint	spatial constraint
Pull	0.565	**0.304**
Heavy Carry	0.645	**0.484**

Table 4: Detection performance of Pulling and Heavy Carry at P_{miss}@0.15 *rfa* on the DIVA validation data.

tent with the public leaderboard evaluations other than the [36] baseline result provided by the organizers of the leaderboard. To generate a complete set of predictions covering all the activities defined by each evaluation, we generate predictions using the presented TRN baseline for a set of activities that we do not yet have SAFER formulation for.

4.4. Interpretable Diagnosis of Detections

Another aspect of SAFER is that it is interpretable by design. SAFER enables more explicit reasoning about model predictions under a coarse-to-fine design paradigm during both implementation and diagnosis stages. During our implementation of SAFER, we exploited the interpretable aspect of our predictions to initially design coarse detectors and gradually converged towards more fine-grained detectors by composing additional slots and fillers. For example, consider the set of slots, {human, box}, and a filler representing the co-occurrence of two slots. Under such definitions, activities *Pulling* and *Transport Heavy Carry* are entirely ambiguous under SAFER. Resulting ambiguities are illustrated in Figure 4. Additional filler modeling spatial relations of the two slots enables discrimination at a finer scale, leading to more accurate detection results as shown in Table 4.

The visualization in Figure 4c represents the incorrectly detected *Pulling* instance even with the spatial constraint. We note that the model's incorrect prediction is inter-

	P_m@ 0.15 *rfa*		P_m@ 1 *rfa*	
	with GT	w/o GT	with GT	w/o GT
Entering	0.394	0.667	0.127	0.183
Exiting	0.677	0.789	0.200	0.231

Table 5: Activity detection performance of Entering and Exiting on the DIVA validation data given ground truth object locations and identities. GT: SAFER with ground truth information.

pretable given the relatively lower position of the carried object which is a common spatial location of boxes in *Pulling* instances. We believe that modeling human pose as a Slot will resolve this ambiguity but we reserve that for future work.

Representational power of SAFER grows effectively as detectors for slots continue to improve. At the same time, we identify errors in slot detectors as a common source of missed/false detections in DIVA. Consider *Entering* and *Exiting* from a vehicle. They are defined as a human appearing or disappearing near a stationary vehicle. Figure 5 illustrates false positive detections for both cases found in the predictions. Due to heavy occlusion by a vehicle, human is barely visible and at which point, detector fails to locate the human. Such cases may be removed using a visual tracker which may re-identify the object. Table 5 demonstrates the upper-bound of SAFER for the example activities where we perform activity detection with object identifications provided by ground truth.

4.5. Synthesis of Slots: Vehicles Parts

Interesting aspect of SAFER is how slots naturally define dimensions for data synthesis. When more detailed slots describing a particular object configuration is required for more granular analysis, simulation can effectively produce labeled data to train such slot detectors. We demonstrate slot simulation by learning to detect *opened-door* and *opened-trunk* detectors through simulation using UnrealCV [25]. The dataset includes 19,944 rgb crops of vehicles with opened/closed trunks and 74,576 rgb crops of opened/closed doors. Examples of simulated crops are shown in Figure 6.

Trunks (19944 crops)		Doors (74576 crops)	
# Real Crops	Acc.	# Real Crops	Acc.
0	0.655	0	0.599
2	0.675	2	0.631
50	0.721	150	0.686
100	0.718	300	0.706
150	0.781	600	0.772
330	0.805	1234	**0.806**
660	**0.838**	2468	0.796
660, No Sim.	0.615	2468, No Sim.	0.710

Table 6: Classification accuracy of models trained on a simulated dataset. Having a mix of real and simulated images greatly improves the model's generalization capability. The amount of real training samples is varied from 0 to approximately 3 % of the generated simulation data.

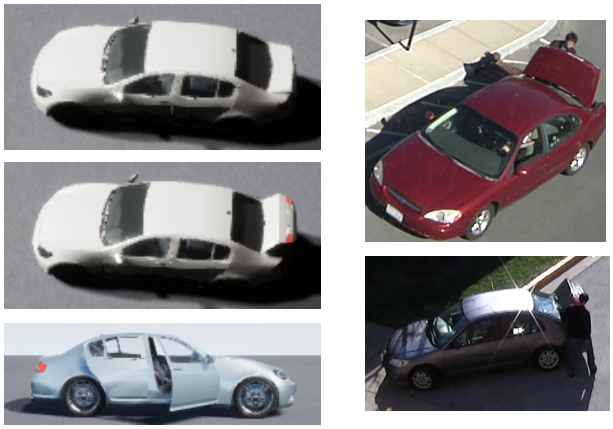


Figure 6: The simulated dataset of vehicles with opened or closed vehicle parts is shown on the left side. The two crops on the right of the figure represent samples from the DIVA dataset. Best viewed in color.

We fix the model presented in [18] as our detection model as it has successfully demonstrated training models using simulated data for inference tasks on real images. As shown in Table 6, by introducing simulation data for training, the opened-trunk and opened-door classification results improve by 22.3 % and 9.6 % respectively. Obtaining detectors for very fine-grained slots greatly increases fine-grained activity detection potential of SAFER and we wish to investigate further of bringing more simulation into activity detection systems through SAFER.

5. Conclusion

SAFER is a framework where complex activities are modeled as explicit structured time-series of objects, object parts and relationships among them. We demonstrated on multiple evaluation settings using the DIVA dataset that compositional construction of activities through slots and fillers leads to detection models that generalizes more effectively to unseen data at test time. As available detectors and means of slot synthesis mature more as the field progresses, we believe SAFER is a powerful framework for activity analysis under structured environments and zero-shot applications where generalization is a significant measure of success.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 2

[3] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning*, 2017. 3

[4] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[5] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron. https://github.com/facebookresearch/detectron, 2018. 3

[6] G. Gkioxari and J. Malik. Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 759–768, 2015. 2, 3

[7] J. Gleason, R. Ranjan, S. Schwarcz, C. Castillo, J.-C. Chen, and R. Chellappa. A proposal-based solution to spatio-temporal action detection in untrimmed videos. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 141–150. IEEE, 2019. 2, 6

[8] H. Hattori, V. Naresh Boddeti, K. M. Kitani, and T. Kanade. Learning scene-specific pedestrian detectors without real data. In *CVPR*, 2015. 3

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[10] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2

[11] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 3

[12] R. Hu, J. Andreas, T. Darrell, and K. Saenko. Explainable neural computation via stack neural module networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[13] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017. 2

[14] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2014. 2

[15] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998, 2017. 2

[16] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017. 2

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 1, 2

[18] C. Li, M. Z. Zia, Q.-H. Tran, X. Yu, G. D. Hager, and M. Chandraker. Deep supervision with intermediate concepts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 2, 3, 8

[19] W. Li and N. Vasconcelos. Recognizing activities by attribute dynamics. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012. 2

[20] B. Liu, S. Yeung, E. Chou, D.-A. Huang, L. Fei-Fei, and J. Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–568, 2018. 3

[21] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR 2011*, pages 3337–3344. IEEE, 2011. 2

[22] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4942–4950, 2018. 2

[23] NIST. Trecvid 2017 evaluation for surveillance event detection. https://www.nist.gov/itl/iad/mig/trecvid-2017-evaluation-surveillance-event-detection, 2017. 6

[24] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011. 5

[25] W. Qiu, F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, T. S. Kim, and Y. Wang. Unrealcv: Virtual worlds for computer vision. In *ACM MM*, 2017. 3, 8

[26] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 3

[27] C. Roberto de Souza, A. Gaidon, Y. Cabon, and A. Manuel Lopez. Procedural generation of videos to train deep action recognition networks. In *CVPR*, 2017. 3

[28] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 3

[29] G. A. Sigurdsson, O. Russakovsky, and A. Gupta. What actions are needed for understanding human actions in videos? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2137–2146, 2017. 2

[30] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 1, 2

[31] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield. Training deep networks with synthetic data: Bridging

the reality gap by domain randomization. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 969–977, 2018. 2

[32] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017. 3

[33] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 2

[34] X. Wang and A. Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018. 2

[35] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018. 1, 2

[36] H. Xu, A. Das, and K. Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 2, 3, 6, 7

[37] J. Zheng, Z. Jiang, and R. Chellappa. Submodular attribute selection for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2242–2255, 2017. 2

[38] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 1, 2, 6

[39] M. Zolfaghari, K. Singh, and T. Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 695–712, 2018. 1, 2