Ordered Atomic Activity for Fine-grained Interactive Traffic Scenario Understanding

Nakul Agarwal[1] Yi-Ting Chen[2]
[1]Honda Research Institute USA [2]National Yang Ming Chiao Tung University
nakul_agarwal@honda-ri.com ychen@cs.nycu.edu.tw

Abstract

*We introduce a novel representation called Ordered Atomic Activity for interactive scenario understanding. The representation decomposes each scenario into a set of ordered atomic activities, where each activity consists of an action and the corresponding actors involved and the order denotes the temporal development of the scenario. This design also helps in identifying important interactive relationships, such as yielding. The action is a high-level semantic motion pattern that is grounded in the surrounding road topology, which we decompose into zones and corners with unique IDs. For example, a group of pedestrians crossing in front is denoted as **C1** → **C4**: **P+**, as depicted in Figure 1. We collect a new large-scale dataset called OATS*[1] *(**O**rdered **A**tomic Activities in interactive **T**raffic **S**cenarios), comprising 1026 video clips (∼ 20s) captured at intersections in San Francisco Bay Area. Each clip is labeled with the proposed language, resulting in 59 activity categories and 6512 annotated activity instances. We propose three fine-grained scenario understanding tasks, i.e., multilabel atomic activity recognition, activity order prediction, and interactive scenario retrieval. We also propose a Graph Convolutional Network based framework that models both appearance and motion of traffic participants to tackle the above tasks, that performs favorably against state-of-the-art methods. However, we find that the methods cannot achieve satisfactory performance, indicating rising opportunities for the community to develop new algorithms for these tasks towards better interactive scenario understanding.*

1. Introduction

Intelligent transportation systems (ITS) have made significant progress in addressing traffic fatalities, with advancements in perception [20, 55, 70, 26, 19], prediction [90, 48, 16], hazard identification [54, 71, 39, 31], and planning [5, 1]. To deploy ITS at scale, extensive research has been conducted on scenario-based simulation assessment [15, 5, 85, 1] to validate ITS in challenging scenarios and identify the causes of failure. Effective scenario understanding and retrieval of real-world data, therefore, can enhance scenario-based assessment [65, 21] and help improve ITS deployment.

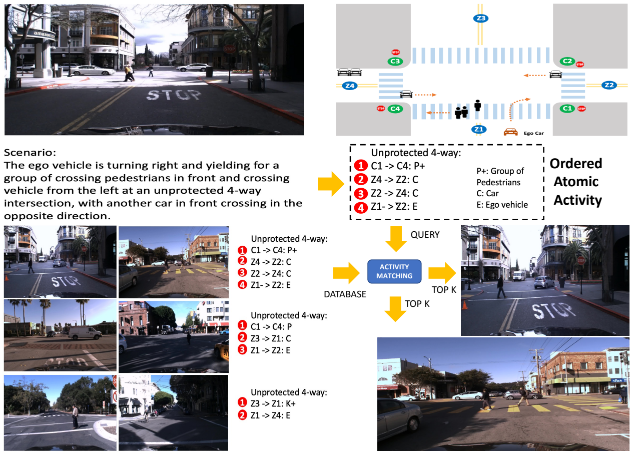
Scenario:
The ego vehicle is turning right and yielding for a group of crossing pedestrians in front and crossing vehicle from the left at an unprotected 4-way intersection, with another car in front crossing in the opposite direction.

Figure 1. **Ordered Atomic Activity.** Instead of using natural language to describe a scenario which is verbose and time-consuming, *Ordered Atomic Activity* decomposes each interactive scenario into a set of ordered atomic activities, where each activity consists of an action and the corresponding actors involved and the order represents the temporal development of the scenario. The activity of "a group of pedestrians crossing in front" is represented as **C1** → **C4: P+**, where **C1** and **C4** are the two corners, → denotes the moving direction and the actors **P+** (a group of pedestrians) perform the action. We denote the temporal development in a scenario by tagging the order of activities based on their occurrence. *Ordered Atomic Activity* also enables efficient scenario retrieval for applications such as scenario-based assessment.

In this work, we focus on intersection scenarios, which involve the highest number of interactions among traffic participants and account for approximately 40% of all crashes [53]. We specifically examine scenarios where traffic participants are crossing at the intersections, rather than

[1] https://usa.honda-ri.com/oats

static participants who do not *directly* interact with the ego car. This design choice is motivated by the practice of scenario-based safety assessment [77, 63, 21, 13], which focuses on activities relevant to the ego vehicle. Our primary research objective is to identify an effective way to describe an interactive scenario involving traffic participants, in order to achieve a better scenario understanding and retrieval.

According to [77, 13], an interactive scenario representation should include information about the activities of road users (including their location, moving direction, and goal), the static environment, and the temporal evolution of these activities. We can use natural language [43, 34, 33], road scene graphs [89, 74], and attributes [58, 42, 48, 65] to represent interactive scenarios. While it is intuitive and explainable for humans using natural language, it often lacks explicit information about the motion directions and goals of road users, leading to verbose descriptions that are difficult to label and control for quality, as shown in Figure 1. Additionally, it is challenging to retrieve videos efficiently with lengthy natural language descriptions. Road scene graphs have also become a popular representation of traffic scenes. Each road user is represented as a node with low-level states such as location, speed, and direction, whereas edges capture pair-wise relationships between pairs of road users such as ”following.” However, these approaches do not represent the high-level semantics between actors' actions and the underlying road topology, which enables efficient tagging for scenario-based assessment [21]. Attribute-based representations are promising options. However, existing works have not explored bridging the gap between activity and underlying road scene structures. Moreover, both road scene graphs and attribute-based approaches have not tackled the temporal development of a scenario and also lack activity order information.

To this end, we propose a novel representation of an interactive scenario called *Ordered Atomic Activity*. In Figure 1, we see an ego vehicle turning right and yielding to a group of crossing pedestrians in the front and a crossing vehicle from the left side, while another car in front is crossing in the opposite direction of the ego vehicle. We represent this detailed description in a concise yet interpretable manner using our proposed representation, where each activity comprises an action and the corresponding actors, such as a group of pedestrians or the ego vehicle. The action is a high-level semantic motion pattern grounded in the surrounding road topology. Specifically, we decompose a road scene, such as a 4-way intersection, into a set of regions that represent corners and zones. For example, a right-turn action is denoted as **Z1** → **Z2**, representing a motion pattern from an initial position to the destination. ”A group of pedestrians crossing in front” is denoted as **C1** → **C4: P+**, where **C1** and **C4** are the two corners, → is the motion direction and **P+** are the group of pedestrians performing the action. Moreover, we denote temporal development in a scenario by tagging the order in which the activities occur. This also helps in representing yielding relationships. In Figure 1, the order of the group of pedestrians, crossing vehicle in the same direction as ego vehicle, and ego vehicle activities represent *yielding*. *Ordered Atomic Activity* is designed at a video level, making it highly scalable and easing the burden of annotation.

We collect a large-scale dataset OATS and annotate monocular videos with *Ordered Atomic Activity*. The OATS dataset comprises 1026 clips, each approximately 20 seconds long, of real human driving in the San Francisco Bay Area captured using an instrumented vehicle. We propose three new and challenging fine-grained scenario understanding tasks, including multilabel atomic activity recognition, activity order prediction, and interactive scenario retrieval. We design a graph convolutional network-based algorithm that models both appearance and motion of traffic participants for the above tasks. We also implement multiple baselines (including recent state-of-the-art video understanding algorithms [6, 79, 87, 18, 76, 2, 83]) and evaluate them on the proposed dataset.

Our extensive experiments demonstrate that the methods perform inadequately on the proposed three tasks. Specifically, the best-performing algorithms for multilabel activity recognition, activity order prediction, and interactive scenario retrieval achieve 26.7% mAP, 16.1% matching score, and 16.6% Recall@top50. To successfully recognize multiple *Atomic Activities* in a scenario, a model should detect and track moving road users, associate their spatio-temporal action with respect to the underlying road topology, and capture the concept of groups. We believe the proposed *Ordered Atomic Activity* and the dataset introduce new challenges to the video scene understanding community. We hope to encourage the research community to work collectively on this important and challenging area.

This paper makes the following contributions. **First**, we introduce a novel representation called Ordered Atomic Activity for interactive scenario understanding. **Second**, we construct a large-scale dataset and propose three essential fine-grained scenario understanding tasks: multilabel atomic activity recognition, activity order prediction, and interactive scenario retrieval. **Third**, we propose a graph convolutional based network that models both appearance and motion of traffic participants and performs favorably against state-of-the-art methods on the above tasks. We also conduct extensive experiments on the three tasks and establish a comprehensive benchmark suite for future research.

2. Related Work

Traffic Scene Datasets. In recent years, many traffic scene datasets have been proposed to stimulate progress in detection [20, 84, 55, 70, 26], tracking [20, 84, 55, 70, 26, 60, 59],

	PIE [59]	JAAD [60]	STIP [46]	LOKI [22]	OATS
# Clips	-	346	556	644	1026
RGB Images	✓	✓	✓	✓	✓
LiDAR Point Cloud	✗	✗	✗	✓	✓
# Agent type	1	1	1	8	6
Type of Intersection	✗	✗	✗	✗	✓
Atomic Activity	✗	✗	✗	✗	✓
Activity Order	✗	✗	✗	✗	✓

Table 1. Comparison of OATS with other datasets.

semantic segmentation [11, 84, 88, 19], trajectory forecast [26, 90, 69], and long-term localization [47]. In addition to perception and prediction tasks, nuScenes [26], Argoverse [8], and Lyft [32] provide high definition maps that enable road topology understanding from monocular videos [64, 56]. Moreover, the datasets containing annotations for activity recognition and intention prediction of traffic participants [60, 58, 59, 46, 48, 22] are essential for the field. Some recent datasets [65] also provide HD maps, but lack a concept of spatial zones and only contain information in the form of semantic labels (road segments, lanes) or polylines/polygons. A dataset that jointly considers the relationship between behaviors of traffic participants, ego vehicle, and underlying road topology is under-explored.

To this end, we collect a large-scale dataset with diverse interactive scenarios labeled with the proposed representation *Ordered Atomic Activity*. Although *Ordered Atomic Activity* can be used to represent scenarios in existing datasets [16, 26, 19], these interactive scenarios are not predefined in the released annotations. Moreover, IP-related issues can make it challenging to augment and release data from such datasets. Therefore, we collect our own dataset to facilitate the new research direction. In addition, we propose and benchmark three novel fine-grained interactive scenario understanding tasks. We provide extensive discussion on the challenges posed by the proposed dataset and tasks. Table 1 shows a comparison of OATS with existing datasets.

Video Analysis and Understanding. Video analysis and understanding have been an active research topic in the community. A significant amount of temporal models have been proposed for action recognition [75, 78, 6, 79, 92, 82, 30, 23]. Recently, the community has exploited object proposals [2, 49, 25, 40] as an inducted bias to boost the performance of video understanding. Recently, Nagarajan et al., [51] propose a method that converts egocentric videos of daily kitchen activities into a topological map consisting of activity "zones" and their rough spatial proximity. In contrast, we investigate the connection between traffic scenes. Specifically, the proposed *Ordered Atomic Activity* offers a new challenge to the video understanding community. We believe there is a great synergy between the recent road topology understanding works [80, 56, 64, 74, 28] and the proposed tasks. Additionally, we study the link for all traffic participants instead of only the ego agent as in [51]. All these unique aspects mentioned above introduce new challenges for understanding traffic scenes from videos.

Activity Order Prediction. The task of generating an ordered list of actions is commonly referred to as transcript prediction within the action understanding community. While many studies utilize this activity order as weak supervision to address complex tasks like action segmentation [29, 61, 62, 14, 7, 41, 3, 67], only a few studies explicitly focus on generating this transcript using the ordered list of actions as the ground truth [67, 3]. Existing works have concentrated on cooking activities [37, 4, 17, 68] where object movements are limited or only a single action occurs at a given point in time. The community has not yet explored this aspect in traffic scenes, where the order of actions can help to identify critical interactions, such as yielding. In traffic scenes, individual actions begin at different times, and their progression happens simultaneously, making it extremely challenging to identify the multiple actions and corresponding orders. Note that this order cannot be automatically generated even if timestamps [16] are available. Timestamps alone cannot tell when a pedestrian starts crossing. Moreover, a car yielding to a crossing pedestrian might start crossing the zones before the pedestrian based on timestamps. The complicated interactive patterns make automatic generation infeasible. Thus, we label the activity orders manually and release the dataset. To the best of our knowledge, our work is the first to introduce and tackle this problem in driving scenes.

Traffic Scenario Retrieval. The focus of most existing work in the analysis is detection and tracking of traffic participants [12, 73, 52], recognition of actions in traffic scenes [59, 60, 58, 46, 22], and traffic anomaly detection [52]. To enable fine-grained scenario retrieval, we need to analyze human driving data at multiple levels, ranging from object detection to fine-grained activity orders. Existing traffic scenario retrieval solutions describe traffic scenes using natural language [43], traffic scene attributes [42, 65], and latent representation [27]. Our work is closely related to [42, 65] as the proposed description language can be treated as traffic scene attributes. The paper [42] proposes attributes such as driver behavior/intention/attention. For [65], they consider vehicles' speed and density, vehicles' actions (e.g., braking, keeping lane), and vehicle-to-vehicle interactions (e.g., braking for another vehicle). However, we observe the following differences. First, we predict traffic scene attributes using monocular videos instead of multimodal data (e.g., GPS/IMU, LiDAR, and HD maps). Second, existing scene attributes do not jointly consider the relationship between action and road structures. While the work [65] offers HD maps to link actions and road structures, their attributes only describe the locations of vehicles. Third, *Ordered Atomic Activity* enables the development of predicting the activity order in a scenario.

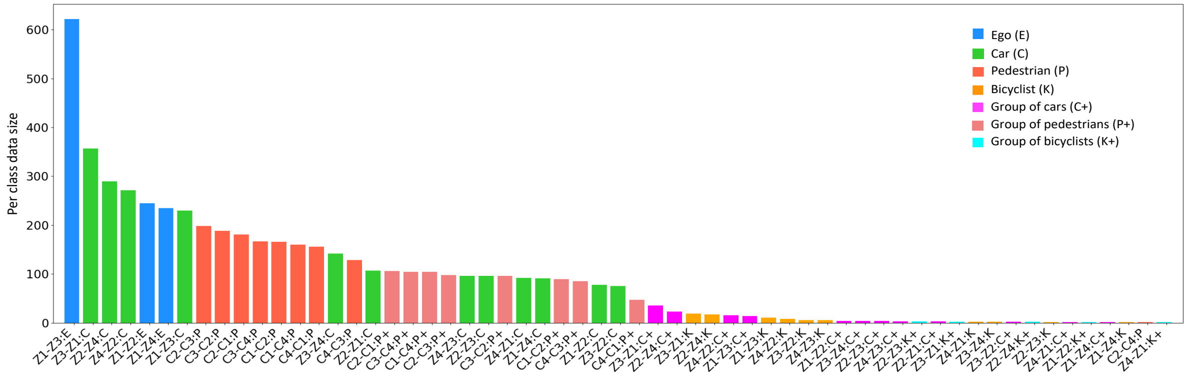

Figure 2. Sizes of each atomic activity class in the OATS dataset sorted by descending order, with colors indicating actor types.

3. The OATS Dataset

3.1. Data Collection Platform

The data is collected using an instrumented vehicle equipped with 3 Point Grey Grasshopper video cameras with a resolution of 1920 × 1200 pixels, a Velodyne HDL-64E S2 LiDAR sensor, and high precision GPS. All sensor data are synchronized and timestamped using ROS and customized hardware and software. The data is collected in the San Francisco Bay Area region and captures a diverse set of traffic scenarios at intersections, including different environments such as urban and suburban areas.

3.2. Annotation Methodology

Modeling both driver and traffic agent behavior is complex and involves different levels of cognitive processes, particularly in complicated interactive scenes, and is essential for the development of robust intelligent driving systems. Therefore the data selection and annotation protocol must be carefully designed. The first step involves the manual selection of short clips from hours of recording that include appropriate scenarios. Although in this work we only focus on 4-way intersections as mentioned in Section 1, we show the flexibility of *Ordered Atomic Activity* for different topologies in the supplementary material. We clip ∼20s short clips around intersections. Each clip contains both the entering and exit of the ego vehicle from the intersection. We discuss the different aspects of our annotation in the following.

Annotation Consistency. We conduct the following quality control strategy. Given a video, it is first annotated by 2 independent human annotators who are experienced drivers residing in the USA to ensure familiarity with the rules of the road, signs etc. Finally, we have an internal expert annotator to review and obtain the final version. To demonstrate the consistency and validity of the annotations, we compute the intra-class correlation coefficient (ICC) [66], which is widely used [58] for the assessment of consistency made by different observers measuring the same quantity. The ICC for our annotations is 0.91, which indicates "excellent" consistency (ICC = 1 for absolute agreement) according to [10].

Atomic Activity. An atomic activity consists of an action and the corresponding actors. The action is a high-level semantic motion pattern that is grounded in the surrounding road topology. We first decompose the surrounding road topology into four corners (C1, C2, C3, C4) and four zones (Z1, Z2, Z3, Z4) with respect to the ego vehicle in an anticlockwise manner, as shown in Figure 1. *Actor* consists of seven classes: *car* (C), *bicyclist* (K), *pedestrian* (P), *car* (E), *group of cars* (C+), *group of bicyclists* (K+), *group of pedestrians* (P+). Given a clipped video, the annotator watches the full video to understand both the type of agents in the scenario and their direction of motion. After analyzing the video, the annotator annotates the *Atomtic Activity*. For instance, we denote a group of pedestrians' crossing activity from the near-right corner to the near-left corner with respect to the ego vehicle as **C1** → **C4: P+**.

Activity Order. Besides labeling activities, we also assign labels to their corresponding order. The starting time of the activity, i.e. when the agent starts crossing, is the basis for annotating this order except for yielding actions. In such cases, the agent that yields is labeled later, regardless of when the action begins or ends. For instance, when a car yields to a pedestrian at a crossing, it may eventually cross the zones before the pedestrian, but the pedestrian should be labeled first. The strategy is challenging to be labeled automatically using timestamps This is why human observation is necessary when annotating the activity order, and it cannot be inferred heuristically or programmatically using timestamps.

Type of Intersection. We further categorize the scenario into the following four types of intersections: *Four-Way Unprotected Turn* (4WUT), *Four Way Protected Turn* (4WPT), *Four way Two Stop* (4W2S) and *Four Way Four Stop* (4W4S). These scenarios are differentiated based on the number of stop signs and traffic light types at intersections. This is particularly useful for traffic scene understanding and scenario retrieval based on a specific type of intersection.

3.3. Dataset Statistics

Our dataset comprises 1026 video clips and 59 activity categories. Figure 2 depict the label distribution of the different activities in our dataset, sorted in descending order. As shown in the figure, the label distribution is not uniform, which is typical of real-world data. For instance, activities comprising a group of actors (e.g., Z3-Z1:K+) are very rare compared to more frequent activities, such as **Z3** → **Z1**:**C** or **C2** → **C3**:**P**. More details regarding the frequency of individual activities, actors, actions and also the types of intersections can be found in the supplementary material.

4. Methodology

4.1. Model Description

In this section, we present the proposed model for multilabel atomic activity recognition. The framework is depicted in Figure 3. As atomic activity comprises both action and actor, our network considers both appearance and motion features. Given a video sequence, we begin by extracting tracklets of the actors in the scene using pre-trained Mask R-CNN [24] pre-trained on COCO dataset [44] and Deep SORT [81]. We then select a set of Z frames from the video and extract appearance features of N traffic agents using the Inception-v3 [72] backbone and RoIAlign [24]. We also extract motion features from tracklets using bounding boxes of agents. After completing the feature extraction process, we pass the motion and appearance features through separate graph convolution networks [35], with each node representing an actor. Finally, we fuse the learned features from both graphs for multilabel atomic activity recognition. We discuss the specifics of each module below.

Appearance Model. We utilize the graph structure to explicitly model pair-wise relations between different agents in the driving scene. Unlike prior graph-based algorithms [2, 83] that treat objects in the graph independently, we construct our graph by utilizing tracking. Given a set of N agents in the traffic scene with their corresponding tracklets, we construct a spatio-temporal graph $G_t^a = (V_t^a, A_t)$, where $V_t^a = \{v_t^i | \forall i \in \{1, ..., N\}\}$ is the set of vertices of graph G_t^a and $A_t = \{a_t^{ij} | \forall i, j \in \{1, ..., N\}\}$ is the adjacency matrix $\forall t \in \{1, ..., Z\}$. In our graph, a_t^{ij} models the appearance relation between two agents at time t and is formally defined as:

$$a_t^{ij} = \frac{f_p(v_t^i, v_t^j)\text{exp}(f_a(v_t^i, v_t^j))}{\sum_{j=1}^{N} f_p(v_t^i, v_t^j)\text{exp}(f_a(v_t^i, v_t^j))}, \tag{1}$$

where $f_a(v_t^i, v_t^j)$ indicates the appearance relation between agents i and j at time t, and $f_p(v_t^i, v_t^j)$ is an indicator function that determines the presence of a tracklet. The softmax function is used to normalize the influence on agent i from other objects. The appearance relation is calculated as below:

$$f_a(v_t^i, v_t^j) = \frac{\theta(v_t^i)^T \phi(v_t^j)}{\sqrt{D}}, \tag{2}$$

where $\theta(v_t^i) = \mathbf{w} v_t^i$ and $\phi(v_t^j) = \mathbf{w'} v_t^j$. Both $\mathbf{w} \in \mathbb{R}^{D \times D}$ and $\mathbf{w'} \in \mathbb{R}^{D \times D}$ are learnable parameters that map appearance features to a subspace and enable learning the correlation of two objects, and $\sqrt{D}$ is a normalization factor. While [2, 83] can fill the graph with any random node at time t, we need to take into account missing nodes due to both inconsistencies in tracking and agents entering and leaving the traffic scene at different times. To mitigate this issue, we set adjacency matrix values to zero when an object is missing using indicator function f_p as:

$$f_p(v_t^i, v_t^j) = \mathbb{I}(v_t^i = \text{present } \textbf{and } v_t^j = \text{present}) \tag{3}$$

Once the nodes and adjacency matrix values are defined, we reason over the Graph Convolutional Network (GCN) [35]. GCN takes a graph as input, performs computations over the structure, and returns a graph as output. For a target node i in the graph, it aggregates features from all neighbor nodes according to values in the adjacency matrix. Formally, one layer of GCN can be written as:

$$Z^{(l+1)} = \sigma A Z^{(l)} W^{(l)}, \tag{4}$$

where $A \in \mathbb{R}^{NZ \times NZ}$ is the adjacency matrix for appearance model and $Z^l \in \mathbb{R}^{NZ \times D}$ is the feature representations of nodes in the l^{th} layer. The matrix $W^l \in \mathbb{R}^{D \times D}$ is the layer-specific learnable weight matrix. The function $\sigma(\cdot)$ denotes an activation function. We adopt ReLU as the activation function. Note that this layer-wise propagation can be stacked into multi-layers.

Motion Model. An essential aspect of our problem is how we represent the motion of agents, as it is a crucial part of our annotations. In this work, we model the motion of agents in 2D. Our motion model is inspired by recent success in the motion-based action recognition [86] and trajectory prediction [50, 36, 91]. We use the relative motion between tracklets at different times t as input to the graph rather than absolute coordinates in the image space. We construct another spatio-temporal graph $G_t^m = (V_t^m, E_t)$, where $V_t = \{u_t^i | \forall i \in \{1, ..., N\}\}$ is the set of vertices of graph G_t^m and $E_t = \{e_t^{ij} | \forall i, j \in \{1, ..., N\}\}$ is the set of edges. We set $e_t^{ij} = 1$ if e_t^i and e_t^j are connected, and $e_t^{ij} = 0$ otherwise. We attach a value b_t^{ij} to model the relation between two nodes i and j, which is computed by some kernel function for each e_t^{ij}. The b_t^{ij} are organized into the weighted adjacency matrix B_t. We introduce $b_{sim,t}^{ij}$ as a kernel function to be used within the adjacency matrix B_t.

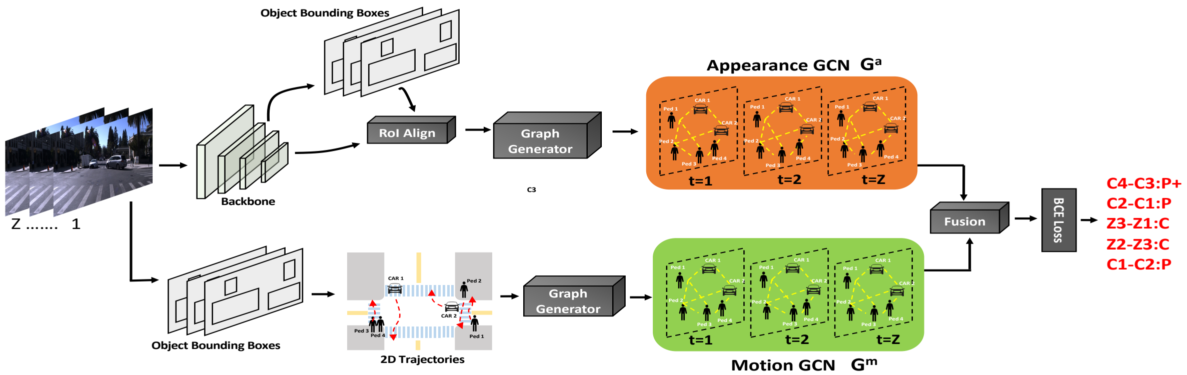

Figure 3. **An overview of our framework.** Given Z frames from an input video, we first extract tracklets of traffic agents and then construct two separate spatio-temporal GCNs G^a and G^m each to model both their appearance and motion in the scene respectively. Unlike previous methods, we utilize tracking to capture the position of the traffic agent in the graph. Finally, the extracted GCN features are then fused for multi-label classification.

Algorithm 1 Retrieval Algorithm

Input: L, P, V and model I
Output: recall@top$K$

```
$t_p = 0, f_p = 0, f_n = 0$
top $K$ = Hamming$(I(S_q), L)$
for $k \leftarrow 1$ to $K$ do:
    for $i \leftarrow 1$ to $q$ do:
        if $M_{S_q} \subseteq M_{S_f}$ then:
            $t_p += 1$
        else
            $f_p += 1$
for $k \leftarrow 1$ to $\{F - K\}$ do:
    for $i \leftarrow 1$ to $q$ do:
        if $M_{S_q} \subseteq M_{S_f}$ then:
            $f_n += 1$
return recall = $\frac{t_p}{t_p + f_n}$
```

The function $b_{sim,t}^{i,j}$ is defined as:

$$b_{sim,t}^{i,j} = \begin{cases} \frac{1}{||u_t^i - u_t^j||_2}, & ||u_t^i - u_t^j||_2 \neq 0 \\ 1, & u_t^i \text{ or } u_t^j \text{ missing} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where u_t^i is 2D position of agent i at time t and can be defined as:

$$u_t^i = [c_x, c_y], \tag{6}$$

where c_x and c_y are centers of bounding box for agent i at time t. Once the nodes, edges, and adjacency matrix for motion GCN are formed, we perform spatio-temporal graph convolution operation similar to [50, 86].

Loss Function. After constructing the motion and appearance graphs, we extract feature representations learned through the GCNs and combine them for recognition. As both our motion and appearance graphs are composed of tracklets, there are multiple ways to link information from the two graphs. We experimented with various approaches and discovered that late fusion works best. We conjecture that it is because the tracklet cues are noisy. We discuss the performance of different fusion schemes in the supplementary material. The entire model can be trained end-to-end with backpropagation. The final loss function is defined as:

$$\mathcal{L} = -\frac{1}{R}\sum_{i=1}^{R} y_i log(p(y_i)) + (1 - y_i) log(1 - p(y_i)), \tag{7}$$

where y is the label, $p(y)$ is the predicted probability and R is the batch size.

4.2. Activity Order Prediction

The appearance and motion features are obtained from our spatio-temporal graphs G_t^a and G_t^m, respectively. We fuse them to form a latent video representation $O \in \mathbb{R}^{T \times D}$ of an input video $X_t = \{x_t | \forall t \in \{1, ..., Z\}\}$. To model the activity order, we incorporate the frame classification and segment generation branches (denoted as FC and SG respectively) proposed in [67] for calculating the mutual consistency loss. The branch FC takes the shared latent video representation O as input and predicts the class probabilities $FC(O) = Y \in \mathbb{R}^{T \times P}$, where P is the number of activities in the dataset. The branch SG predicts the segments $SG(O) = S$, where each segment s_m consists of predicted ordered activity probabilities a_m and the estimated relative log length l_m of that segment. To calculate the mutual consistency loss for training the network, we first start with the estimated relative log length l_m for each segment s_m, which is then converted into the absolute length l'_m. For each segment, we compute its absolute starting position p'_m within the video. Given l'_m and p'_m, for each segment we generate a mask w_m using the differentiable mask generation module. For each segment, we use the ground-truth action label

Method	Backbone	Pretrain	Splits s_1	s_2	s_3	mAP
CSN [76]	Resnet152	IG65M	12.14	12.57	12.89	12.53
TPN [87]	Resnet50	ImageNet	11.57	13.26	12.94	12.59
SlowOnly [18]	Resnet50	ImageNet	11.20	14.73	12.92	12.95
SlowFast [18]	Resnet50	None	10.80	15.09	14.45	13.45
I3D (NL) [79]	Resnet50	ImageNet	11.87	15.48	14.00	13.78
I3D [6]	Resnet50	ImageNet	11.82	14.30	16.84	14.32
ORN [2]	Resnet50	ImageNet	16.83	13.39	18.14	16.12
ARG [83]	Inceptionv3	ImageNet	20.21	21.34	19.25	20.26
Ours	Inceptionv3	ImageNet	24.34	28.56	27.21	26.70

Table 2. Comparison of our framework with other state-of-the-art algorithms for multi-label atomic activity recognition.

$\hat{a}_m$ instead of the estimated class probabilities. The final loss function is given by:

$$L_\mu = \sum_{m=1}^{M} L_{\mu_m}(Y, w_m, \hat{a}_m) \qquad (8)$$

For more details regarding the loss function, please refer to [67].

4.3. Scenario Retrieval

We denote an intersection scenario as $S_q = (V, P)$ in the query set Q, where V is the set of agents in a scenario and P is the set of tracklets. Each activity l in video-level labels L has the format of Action:Actor where the former gives direction of motion and the latter denotes the type of agent. When retrieving scenarios for driver behavior understanding, not only is it important to match the configuration of the traffic scene based on activities, but also to independently focus on the type of agent and motion dynamics. This is because sometimes the focus for retrieving a certain scenario could be based on the number of type of agents in the scene, e.g. car or pedestrian or both, while other times the user might be more interested in the direction of motion of traffic participant, e.g. Z1-Z2 or C1-C2. Some users might be interested in both aspects. Therefore, we formulate a metric that takes into account all three factors: activities, action, and actor.

We denote the scenario database as F and our goal is to retrieve top K similar scenarios $S_f^i = (V, P) \forall i \in \{1, ..., K\}$ given S_q. We first retrieve top K scenarios using the hamming distance between predicted binary labels (through model I in Figure 3) for S_q and ground truth labels for $S_f \forall f \in \{1, ..., F\}\}$. Once the top K scenarios are retrieved, we convert predicted labels for S_q and L into dictionaries M_{S_q} and M_{S_f} respectively, where the key is noun, verb or action unit and values are their instances/frequencies in the output. We consider a retrieved scenario to be a match if M_{S_q} is a subset of M_{S_f}. The algorithm of the proposed retrieval system is discussed in Algorithm 1.

Method	Splits s_1	s_2	s_3	mAP
Motion	13.68	15.70	14.27	14.55
Appearance	18.25	21.48	20.25	19.99
Motion + Appearance	24.34	28.56	27.21	26.70

Table 3. Ablation studies showing the effect of motion and appearance modeling in our framework.

5. Experiments and Analysis

Datasets & Metrics. To account for the variability in the number of labels in each video, we divided our OATS dataset into three groups, namely s_1, s_2, s_3, with a sufficient number of samples for each class. The splits s_1 and s_2 consist of 350 video clips while split s_3 has 326 clips. These groups are permuted as splits for training and testing to ensure fair evaluation [38]. For further details on the dataset creation and experimental setup, refer to the supplementary material.

We also conduct action understanding experiments on the Collective Activity dataset [9], which contains 44 short video sequences from 5 group activities (crossing, waiting, queueing, walking, and talking) and 6 individual actions (NA, crossing, waiting, queueing, walking and talking). We follow the same evaluation scheme as in [83, 57]. We report mean average precision (mAP) for *multilabel atomic activity recognition*, matching score [67] for *activity order prediction*, and recall@topK for *scenario retrieval*.

Implementation Details. Our implementation is based on PyTorch and we use a single Nvidia Quadro RTX 6000 GPU. We use 32 frames as input each having an image size of 224 × 224, and a batch size of 1 for experiments. We adopt stochastic gradient descent with ADAM to learn the network parameters and train the model for 50 epochs using a learning rate ranging from 0.0002 to 0.0001. All feature layers are jointly updated during training. More implementation details and qualitative results can be found in the supplementary material.

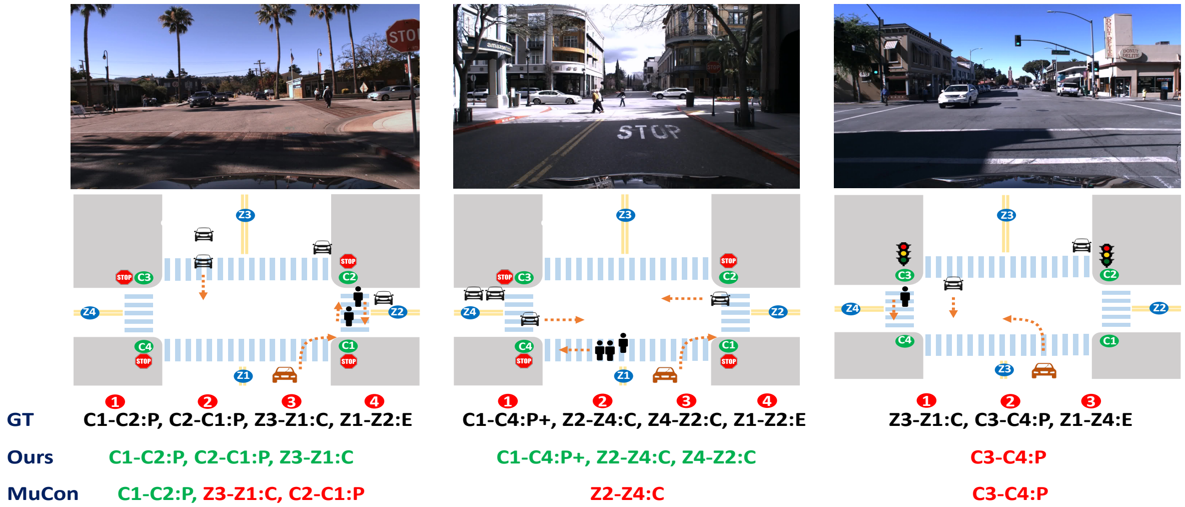
STOP
Z3
Z4
Z2
Z1
C1
C2
C3
C4
STOP
1 2 3 4
GT C1-C2:P, C2-C1:P, Z3-Z1:C, Z1-Z2:E
Ours C1-C2:P, C2-C1:P, Z3-Z1:C
MuCon C1-C2:P, Z3-Z1:C, C2-C1:P
1 2 3 4
C1-C4:P+, Z2-Z4:C, Z4-Z2:C, Z1-Z2:E
C1-C4:P+, Z2-Z4:C, Z4-Z2:C
Z2-Z4:C
1 2 3
Z3-Z1:C, C3-C4:P, Z1-Z4:E
C3-C4:P
C3-C4:P

Figure 4. Qualitative results for activity order prediction of our method against MuCon [67]. All ground truths contain ego vehicle action, i.e. activities starting with 'E', just for reference and it is not used for order prediction. The GT denotes ground truth, and green and red color denote true and false positives respectively.

Multilabel Atomic Activity Recognition. Given the labeled *Ordered Atomic Activity* in the OATS dataset, we formulate the problem as a multi-label recognition problem. We compare our method in Sec. 4 against the different state-of-the-art video understanding algorithms. We start with algorithms extracting features at the video-level [6, 79, 87, 18, 76]. As shown in Table 2, given that our scene deals with multiple agents in the scene spanning approximately 20s of video, implicitly focusing on agents through video-level features cannot recognize activities successfully. For algorithms [2, 83], which explicitly models objects in videos, the results improve. While these object-aware models perform better than purely video-based modeling, they focus on the appearance modeling of the agents, instead of motion modeling. As atomic activity comprises both action and actor, both appearance and motion should be modeled explicitly. The efficacy of the proposed modeling is presented in Table 2. We also conduct ablation studies to diagnose the effects of appearance and motion modeling of our framework in Table 3. It is worth noting that the unsatisfactory performance demonstrated in Table 2 indicates the challenging nature of the task. In addition, also highlight the scope for improvement before moving on to even more complicated scenarios and descriptions.

We also provide results on the Collective Activity dataset in Table 4 to show the efficacy and generalizability of our proposed framework. Since we design our framework keeping in mind video-level supervision, we avoid individual actions and only use group activities in our experiments. We compare our method against [83], which is the state-of-the-art algorithm on the Collective dataset. To have a fair comparison, we run their publicly available code[2] using their own fixed setting but without individual action supervision. We observe that even though [9] has much less object motion in the image space compared to OATS, our framework can perform favorably against the state-of-the-art.

Method	Group Action
ARG [83]	85.28%
Ours	86.58%

Table 4. Comparison with the state-of-the-art algorithm on the Collective dataset

Method	Matching Score s1	s2	s3	Avg
MuCon [67]	48.12/16.45	52.06/15.93	47.93/13.23	49.37/15.20
Ours	52.09/17.32	52.94/16.28	51.73/14.65	52.25/16.08

Table 5. Activity order prediction results on OATS w/ (left) and w/o (right) start and end tokens.

Activity Order Prediction. Very few works focus on explicitly predicting activity order using only the ordered list of activities as ground truth, and none do so in traffic scenes to the best of our knowledge. Therefore, we adapt state of the art method MuCon [67] to our task as a baseline for comparison. Table 5 demonstrates that our method outperforms MuCon [67] for activity order prediction on the OATS dataset. We report results with both w/ and w/o start and end tokens, which help in correctly identifying the number of activities in the predicted sequence [67]. The results indicate that while our method is much better when it comes to predicting the correct number of activities, it is only slightly better in predicting the actual activity or-

[2] https://github.com/wjchaoGit/Group-Activity-Recognition

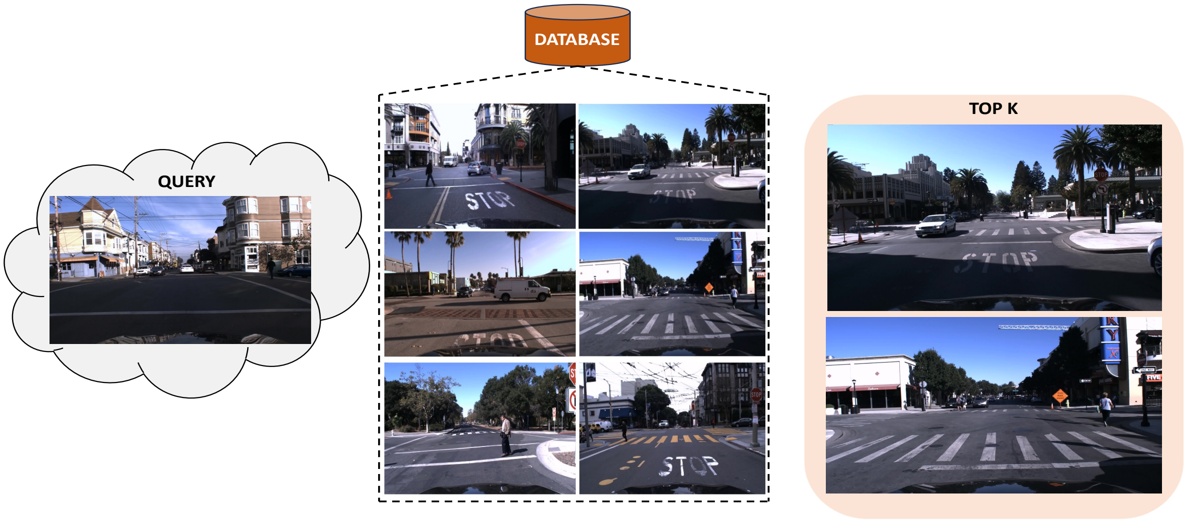

Figure 5. Qualitative results of our method for scenario retrieval. On the **left** is the query input, **middle** contains some examples from the scenario database, and **right** shows the top two retrieved scenarios from the database.

Method	Actor			Action			Activity		
	R@10	R@30	R@50	R@10	R@30	R@50	R@10	R@30	R@50
ARG [83]	0.60	1.91	3.22	0.32	1.49	2.31	0.32	1.49	2.31
Ours	6.89	12.92	18.87	1.53	6.33	16.56	1.53	6.33	16.56

Table 6. Scenario retrieval results on the OATS dataset

der itself, indicating that a more sophisticated approach is required to solve this challenging task. Some qualitative results are shown in Figure 4.

Scenario Retrieval. Given the low performance in Table 2 and Table 5, we select a subset of videos from the 3 splits s_1, s_2, and s_3, which comprises 13 activities and train the proposed model for multi-label atomic activity recognition. We then use those results for the task of scenario retrieval. More specifically, we use one split, e.g. s_1 as our query set and train our model using the database which comprises the other two splits, e.g. s_1 and s_2. We then use the trained model to predict binary labels for the query set which is then used for retrieval as described in Section 4.3. We calculate the recall@topK (for different Ks) for Actor, Action, and Activity individually. Due to the lack of existing methods in this space, we compare with the best-performing method in Table 2. Figure 5 and Table 6 show that although our method performs favorably against [83], the absolute numbers are quite low, indicating the challenging nature of the task and encouraging future research. This is also natural given the low performance of multi-label classification by state-of-the-art methods in Table 2. These low results are also at par with other retrieval applications [45] where the ground truth generally focuses on a single aspect, whereas we are trying to retrieve a complex traffic configuration that covers various aspects such as type of agent, motion information, and scene information.

Inference Speed. Although real-time inference is not the focus of this paper, our model is quite fast. While scenario retrieval has an offline use case, the model takes around 40 ms for multi-label classification in Table 2 and about 70 ms for Activity Order prediction in Table 5 using a single Nvidia Quadro RTX 6000 GPU, making it viable from an application point of view.

Limitations. One of the limitations of our work is that we do not explicitly model group actions in our framework, i.e. P+, C+, and K+. Distinguishing between individual and group actions can play an important role in downstream applications. Additionally, we primarily rely on [67] for Activity Order modeling whereas results in Table 5 show that a more sophisticated approach is required to solve this challenging task. We plan to address these issues in future work.

6. Conclusion

We introduce a novel representation called *Ordered Atomic Activity* for interactive scenario understanding. The representation decomposes each scenario into a set of ordered atomic activities, where each activity consists of an action and the corresponding actors involved and the order denotes the temporal development of the scenario. Given the lack of an appropriate dataset, we introduce a new large-scale dataset OATS along with three crucial fine-grained interactive scenario understanding tasks, i.e., multi-label atomic activity recognition, activity order prediction, and scenario retrieval. We also provide an algorithm to solve scenario retrieval and benchmark these tasks on our dataset with existing SOTA algorithms, as well as with our proposed framework that performs favorably against these methods, for enabling future research.

References

[1] CARLA Autonomous Driving Challenge. https://carlachallenge.org/, 2022.

[2] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object Level Visual Reasoning in Videos. In *ECCV*, 2018.

[3] Nadine Behrmann, S Alireza Golestaneh, Zico Kolter, Jürgen Gall, and Mehdi Noroozi. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In *European Conference on Computer Vision*, pages 52–68. Springer, 2022.

[4] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly Supervised Action Labeling in Videos under Ordering Constraints. In *ECCV*, 2014.

[5] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric M. Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *ArXiv*, abs/2106.11810, 2021.

[6] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017.

[7] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3TW: Discriminative Differentiable Dynamic Time Warping for Weakly Supervised Action Alignment and Segmentation. In *CVPR*, 2019.

[8] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019.

[9] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289. IEEE, 2009.

[10] Domenic V Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284, 1994.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 2016.

[12] Aleksa Ćorović, Velibor Ilić, Siniša urić, Mališa Marijan, and Bogdan Pavković. The real-time detection of traffic participants using yolo algorithm. In *2018 26th Telecommunications Forum (TELFOR)*, pages 1–4. IEEE, 2018.

[13] Erwin de Gelder, Jan-Pieter Paardekooper, Arash Khabbaz Saberi, Hala Elrofai, Olaf Op den Camp, Steven Kraines, Jeroen Ploeg, and Bart De Schutter. Towards an ontology for scenario definition for the assessment of automated vehicles: An object-oriented framework. *IEEE Transactions on Intelligent Vehicles*, 7(2):300–314, 2022.

[14] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6508–6516, 2018.

[15] Hala Elrofai, Jan-Pieter Paardekooper, Erwin de Gelder, Sytze Kalisvaart, and Olaf Op den Camp. Scenario-based safety validation of connected and automated driving. 2018.

[16] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles Qi, Yin Zhou, Zoey Yang, Aurelien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large Scale Interactive Motion Forecasting for Autonomous Driving: The Waymo Open Motion Dataset. *arXiv*, 2021.

[17] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011.

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[19] Xin Wang Wenqi Xian Yingying Chen Fangchen Liu Vashisht Madhavan Trevor Darrell Fisher Yu, Haofeng Chen. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *CVPR*, 2020.

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012.

[21] Erwin De Gelder, Jeroen Manders, Corrado Grappiolo, Jan-Pieter Paardekooper, Olaf Op Den Camp, and Bart De Schutter. Real-world scenario mining for the assessment of automated vehicles. In *ITSC*, 2021.

[22] Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, and Chiho Choi. Loki: Long term and key intentions for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021.

[23] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video Action Transformer Network. In *CVPR*, 2019.

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[25] Roei Herzig, Elad Levi, Huijuan Xu, Hang Gao, Eli Brosh, Xiaolong Wang, Amir Globerson, and Trevor Darrell. Spatio-Temporal Action Graph Networks. In *ICCVW*, 2019.

[26] Alex H. Lang Sourabh Vora Venice Erin Liong Qiang Xu Anush Krishnan Yu Pan Giancarlo Baldan Oscar Beijbom Holger Caesar, Varun Bankiti. nuScenes: A Multimodal Dataset for Autonomous Driving. In *CVPR*, 2020.

[27] Sascha Hornauer, Baladitya Yellapragada, Arian Ranjbar, and Stella X. Yu. Driving Scene-Retrieval by Example from Large-Scale Data. In *CVPRW*, 2019.

[28] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. FIERY: Future Instance Prediction in Bird's-Eye View from Surround Monocular Cameras. In *ICCV*, 2021.

[29] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist Temporal Modeling for Weakly Supervised Action Labeling. In *ECCV*, 2016.

[30] Noureldien Hussein, Efstratios Gavves, and Arnold W.M. Smeulders. Timeception for Complex Action Recognitions. In *CVPR*, 2019.

[31] Otniel-Bogdan Mercea A. Sophia Koepke Zeynep Akata Andreas Geiger Katrin Renz, Kashyap Chitta, Kashyap Chitta, Otniel-Bogdan Mercea, A. Sophia Koepke, Zeynep Akata, and year=2022 Andreas Geiger, booktitle=CoRl. PlanT: Explainable Planning Transformers via Object-Level Representations.

[32] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Level 5 perception dataset 2020. https://level-5.global/level5/data/, 2019.

[33] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John F. Canny. Grounding Human-To-Vehicle Advice for Self-Driving Vehicles. In *CVPR*, 2019.

[34] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John F. Canny, and Zeynep Akata. Textual Explanations for Self-Driving Vehicles. In *ECCV*, 2018.

[35] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[36] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, S Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *arXiv preprint arXiv:1907.03395*, 2019.

[37] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[38] Hilde Kuehne, Ali Arslan, and Thomas Serre. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In *CVPR*, 2014.

[39] Chengxi Li, Stanley H Chan, and Yi-Ting Chen. Who Make Drivers Stop? Towards Driver-centric Risk Assessment: Risk Object Identification via Causal Inference. *IROS*, 2020.

[40] Chengxi Li, Yue Meng, Stanley H. Chan, and Yi-Ting Chen. Learning 3D-aware Egocentric Spatial-Temporal Interaction via Graph Convolutional Networks. In *ICRA*, 2020.

[41] Jun Li, Peng Lei, and Sinisa Todorovic. Weakly Supervised Energy-Based Learning for Action Segmentation. In *ICCV*, 2019.

[42] Max Guangyu Li, Zhengping Che Bo Jiang, Xuefeng Shi, Mengyao Liu, Yiping Meng, Jieping Ye, and Yan Liu. DBUS: Human Driving Behavior Understanding System. In *ICCVW*, 2018.

[43] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *CVPR*, 2014.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[45] Yen-Liang Lin, Son Tran, and Larry S Davis. Fashion outfit complementary item retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3311–3319, 2020.

[46] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

[47] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017.

[48] Srikanth Malla, Behzad Dariush, and Chiho Choi. TITAN: Future Forecast using Action Priors. In *CVPR*, 2020.

[49] RJoanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks. In *CVPR*, 2019.

[50] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020.

[51] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. EGO-TOPO: Environment Affordances from Egocentric Video. In *CVPR*, 2020.

[52] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Liang Zheng, Anuj Sharma, Rama Chellappa, and Pranamesh Chakraborty. The 4th ai city challenge. In *CVPRW*, 2020.

[53] National Highway Traffic Safety Administration. Crash Factors in Intersection-Related Crashes: An On-Scene Perspective, 2010.

[54] Eshed Ohn-Bar and Mohan Manubhai Trivedi. Are all objects equal? Deep Spatio-temporal Importance Prediction in Driving Videos. *Pattern Recognition*, 64:425–436, 2017.

[55] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The H3D Dataset for Full-Surround 3D Multi-Object Detection and Tracking in Crowded Urban Scenes. In *ICRA*, 2019.

[56] Jonah Philion and Sanja Fidler. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In *ECCV*, 2020.

[57] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018.

[58] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for

learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018.

[59] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019.

[60] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017.

[61] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 754–763, 2017.

[62] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. NeuralNetwork-Viterbi: A Framework for Weakly Supervised Video Learning. In *CVPR*, 2018.

[63] Stefan Riedmaier, Thomas Ponn, Dieter Ludwig, Bernhard Schick, and Frank Diermeyer. Survey on Scenario-Based Safety Assessment of Automated Vehicles. *IEEE Access*, PP:1–1, 05 2020.

[64] Thomas Roddick and Roberto Cipolla. Predicting Semantic Map Representations from Images using Pyramid Occupancy Networks. In *CVPR*, 2020.

[65] Sean Segal, Eric Kee, Wenjie Luo, Abbas Sadat, Ersin Yumer, and Raquel Urtasun. Universal Embeddings for Spatio-Temporal Tagging of Self-Driving Logs. In *CoRL*, 2020.

[66] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[67] Yaser Souri, Mohsen Fayyaz, Luca Minciullo, Gianpiero Francesca, and Juergen Gall. Fast weakly supervised action segmentation using mutual consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[68] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013.

[69] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020.

[70] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset, 2019.

[71] Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. Anticipating Traffic Accidents with Adaptive Loss and Large-scale Incident DB. In *CVPR*, 2018.

[72] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[73] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *CVPR*, 2019.

[74] Yafu Tian, Alexander Carballo, Ruifeng Li, and Kazuya Takeda. Road Scene Graph: A Semantic Graph-Based Scene Representation Dataset for Intelligent Vehicles. In *ICRA*, 2021.

[75] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri1. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*, 2015.

[76] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

[77] Simon Ulbrich, Till Menzel, Andreas Reschka, Fabian Schuldt, and Markus Maurer. Defining and substantiating the terms scene, situation, and scenario for automated driving. In *ITSC*, pages 982–988, 2015.

[78] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, , and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[79] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local Neural Networks. In *CVPR*, 2017.

[80] Ziyan Wang, Buyu Liu, Samuel Schulter, and Manmohan Chandraker. A Parametric Top-View Representation of Complex Road Scenes. In *CVPR*, 2019.

[81] Nicolai Wojke and Alex Bewley. Deep Cosine Metric Learning for Person Re-identification. In *WACV*, 2018.

[82] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *CVPR*, 2019.

[83] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019.

[84] Xinjing Cheng Dingfu Zhou Qichuan Geng Ruigang Yang Xinyu Huang, Peng Wang. The ApolloScape Open Dataset for Autonomous Driving and its Application. In *CVPR*, 2018.

[85] Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. SafeBench: A Benchmarking Platform for Safety Evaluation of Autonomous Vehicles. In *NeurIPS Track on Datasets and Benchmarks*, 2022.

[86] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[87] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 591–600, 2020.

[88] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sumanth Chennupati, Sanjaya Nayak, Saquib Mansoor, Xavier Perroton, and Patrick Perez. WoodScape: A Multi-task, Multi-camera Fisheye Dataset for Autonomous Driving. In *ICCV*, 2019.

[89] Shih-Yuan Yu, Arnav Vaibhav Malawade, Deepan Muthirayan, Pramod P. Khargonekar, and Mohammad Abdullah Al Faruque. Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–11, 2021.

[90] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kümmerle, Hendrik Königshof, Christoph Stiller, Arnaud de La Fortelle, and Masayoshi Tomizuka. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. *arXiv:1910.03088 [cs, eess]*, 2019.

[91] Lidan Zhang, Qi She, and Ping Guo. Stochastic trajectory prediction with social graph network. *arXiv preprint arXiv:1907.10233*, 2019.

[92] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018.