Interacting Objects: A dataset of object-object interactions for richer dynamic scene representations

Asim Unmesh[1], Rahul Jain[1], Jingyu Shi[1], V. K. Chaithanya Manam[1], Hyung-Gun Chi[1], Subramanian Chidambaram[1], Alexander Quinn[1], Karthik Ramani[1,2]

***Abstract*—Dynamic environments in factories, surgical robotics, and warehouses increasingly involve humans, machines, robots, and various other objects such as tools, fixtures, conveyors, and assemblies. In these environments, numerous interactions occur not just between humans and objects but also between objects themselves. However, current scene-graph datasets predominantly focus on human-object interactions (HOI) and overlook object-object interactions (OOIs) despite the necessity of OOIs in effectively representing dynamic environments. This oversight creates a significant gap in the coverage of interactive elements in dynamic scenes. We address this gap by proposing, to the best of our knowledge, the first dataset* annotating for OOI categories in dynamic scenes. To model OOIs, we establish a classification taxonomy for spatio-temporal interactions. We use our taxonomy to annotate OOIs in video clips of dynamic scenes. Then, we introduce a spatio-temporal OOI classification task which aims to identify interaction categories between two given objects in a video clip. Further, we benchmark our dataset for the spatio-temporal OOI classification task by adopting state-of-the-art approaches from related areas of Human-Object Interaction Classification, Visual Relationship Classification, and Scene-Graph Generation. Additionally, we utilize our dataset to examine the effectiveness of OOI and HOI-based features in the context of Action Recognition. Notably, our experimental results show that OOI-based features outperform HOI-based features for the task of Action Recognition.**

I. Introduction

Scene graphs have been proposed to capture semantic representation of dynamic physical environments, for example, in factories, surgery rooms, and warehouses, using RGB videos/images. They have been used for various robotics and computer vision applications such as imitation learning [1], task planning [2], [3], human-robot collaboration [4], [5], human activity understanding [6], [7], [8], and embodied AI [9]. However, while existing scene graph datasets predominantly focus on spatio-temporal human-object interactions (HOI) they overlook the crucial compositional element of spatio-temporal object-object interactions (OOI). This is striking, as dynamic scenes not only comprise of HOIs, but also OOIs. Leaving out OOIs creates a significant gap in the coverage of interactive elements of the scene (see Fig. 1). Since OOIs are critical compositional elements of any dynamic scene, it is important to address this gap. Also, OOIs have tremendous potential to be used for robotics and computer vision tasks. OOIs often define the start and end of actions (e.g. the action of 'prying a nail from a wooden board' begins with the OOI of the hammer coming in contact with the nail and ends with the OOI of the nail coming out of the wooden board). During collaborative tasks (such as in Fig. 1), humans actively monitor both HOI and OOI occurring in the scene. This real-time monitoring enables humans to make informed decisions and adjust their actions accordingly. Thus, OOIs can be potentially used for coordinating actions in real-time, ensuring smooth human-robot and robot-robot collaborations. OOIs also have the potential to be utilized for learning tasks from human demonstrations, as human demonstrations not only involve HOIs but also encompass OOIs.

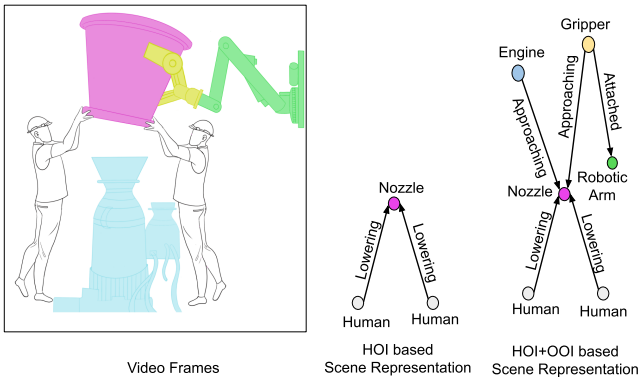

Fig. 1: Only using HOI in scene graphs (middle) restricts their scope. However by including OOI (right) we can make them rich and representative of dynamic environments (left).

Despite the importance of OOIs as a compositional element of dynamic scenes, and its numerous potential applications, existing datasets in the field of scene graphs and related areas (see Section II) have largely overlooked OOIs, leaving a major gap in the coverage of interactive elements of dynamic scenes. Motivated by this gap, and potential of OOIs for robotics and computer vision tasks, we introduce, to the best of our knowledge, the first dataset annotating semantic OOI categories in dynamic scenes. OOIs are defined as the category labels to spatio-temporal interactions between two objects present in the scene. When referring to objects, we denote objects such as tools (e.g., hammers, screwdrivers), furniture and their components (e.g., chairs, tables, legs), vehicles and

Manuscript received: June 20, 2023; Revised: September 16, 2023; Accepted: October 22, 2023

This paper was recommended for publication by Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by US National Science Foundation (FW-HTF 1839971) and Feddersen Chair Funds for Professor Karthik Ramani.

*Our dataset will be made available at https://engineering.purdue.edu/cdesign/wp/interacting-objects.

[1]Purdue University

[2]Corresponding Author: ramani@purdue.edu

Digital Object Identifier (DOI): see top of this page.

their components (e.g., cars, engines, wheels), and numerous other objects that encompass our surroundings.

To model OOIs, we establish a taxonomy of categories of OOIs. It comprises three super-categories: contact relations (cr), location relations (lr), and motion relations (mr). We provide a comprehensive description of our taxonomy in Section III-A. Subsequently, we employ this taxonomy to annotate interaction categories within micro video-clips (short duration video-clips) featuring objects. These clips are generated from the COIN dataset [10], which consists of instructional videos of various activities and tasks. Detailed information about the dataset and its construction is given in Section III-B.

Further, we propose the task of OOI classification, which aims to predict the interaction categories between two given objects in a video-clip. We include state-of-the-art (SoTA) methods adopted from domains of HOI classification, Scene Graph Generation, and Visual Relationship Classification in our benchmark. We present the formal definition of OOI classification task, details of the adopted SoTA methods from related areas, and our base features in Section IV-A. The results of benchmarking along with ablation studies can be found in Section V-A.

We also explore the application of OOIs, by using it for the task of Action Recognition. Our Action Recognition experiments aim at exploring the potential of OOI based features, and comparing their effectiveness with the HOI based ones. Mathematical definitions of OOI and HOI based features with other experimental details are presented in Section IV-B. The results of action recognition experiments are presented in Section V-B. Experimental results of action recognition show that OOI based features outperform HOI based ones significantly. Also, the ablation study for Action Recognition validates the structure of our taxonomy, showing that all 3 relation super-categories are important for performance. We present limitations and future work in Section VI, and conclude our paper in Section VII. In summary, while current scene-graph datasets annotate for spatio-temporal HOIs, they leave out the compositional element of spatio-temporal OOIs. We address this gap by:

- Proposing a novel dataset (atop COIN [10] dataset) focusing on spatio-temporal OOIs in dynamic scenes.
- We propose OOI classification task, and adopt and benchmark SoTA methods from related areas for this task.
- We explore the application of OOIs for the task of Action Recognition. Our results show that OOI based features significantly outperform HOI based features, thus providing strong evidence for the criticality of OOIs in effectively representing dynamic environments.

II. Related Works

A. Scene Graph Datasets

Chao et al. [11] introduced HICO dataset annotating relations between various objects and their interactions with humans in images. Krishna et al. [12] created an extensive scene graph dataset based on images. Recently, datasets such as [6], [13], [14] have focused on constructing scene graphs for dynamic scenes by annotating spatio-temporal HOIs. However, these datasets do not concentrate on OOIs. While certain datasets, such as [12], do annotate spatial relations like 'behind', 'near', and 'next to' between objects, solely annotating spatial relations falls short in capturing the temporal aspect of OOIs. To capture interactions, which inherently involve temporality, it is crucial to annotate spatio-temporal relations rather than solely focusing on spatial relations. We annotate motion relations between objects to capture temporality.

B. Visual Relationship Detection

Visual Relationship Detection is aimed at detecting relationships between entities in the scene. Thus, the OOI classification task is closely connected to the Visual Relationship Detection (VRD) task. However, existing VRD datasets like ImageNet-VidVRD (Shang et al., 2017) and VidOR (Shang et al., 2019) do not specifically emphasize OOI and instead include relation classes (e.g., 'chase', 'feed', 'kiss', 'throw' and 'kick' in VidOR and 'run-behind', 'move-behind', and 'jump-behind' in ImageNet-VidVRD) that are not relevant to interactions between objects. Thus current VRD datasets do not emphasize on OOIs.

C. Human Object Interaction classification

HOI classification and OOI classification are closely related problems, that are crucial for constructing detailed scene graphs. In order to benchmark our dataset, we surveyed recent graph neural network (GNN) based approaches for HOI classification, and then adapted these methods accordingly. Qi et al. [15] proposed a graph parsing neural network that predicts a parse graph with edge weights for a fully connected graph. These edge weights are utilized to modulate message passing between nodes, ensuring that only relevant neighboring nodes of the human and object contribute to HOI classification. We incorporate this approach into our benchmarking process.

D. Object-Object Interaction Affordance

While our focus is on categorization of OOIs that occur in dynamic scenes, existing research has explored the classification of Object-Object Interaction Affordances. Sun et al. [16] propose classifying OOI Affordances through human-object-object interactions modelling, leveraging hand motion and object state change based features. [17] propose learning affordance of tools and objects based on observation from RGB-D videos. More recently, Mo et al. [18] classify object-object interaction affordances by modeling 3D shapes and performing convolution over mesh models. In contrast to these works, which focus on learning Object-Object Interaction Affordances, we aim at classification of OOIs at a time instant, thus enabling richer dynamic scene graphs.

III. Dataset

A. Taxonomy of OOR categories

The relationship between two objects encompasses two dimensions: spatial and temporal. The spatial aspect pertains to the relative positioning of the objects in relation to each other. Spatial aspects include location relations (describing how an

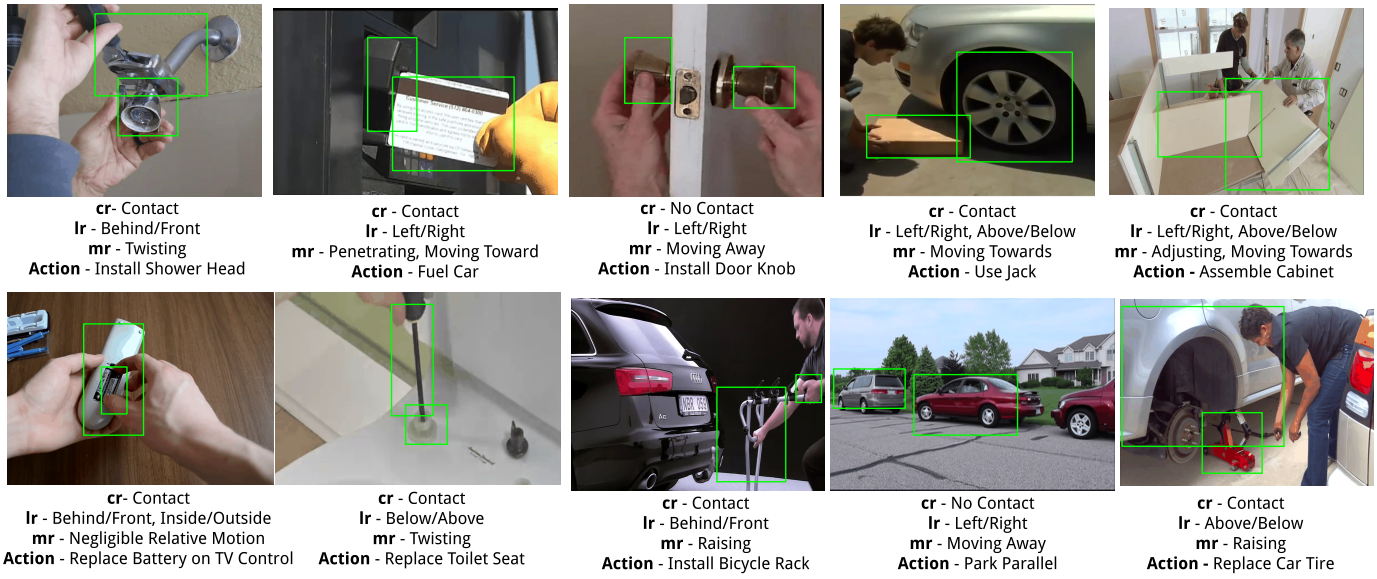

Fig. 2: Samples of annotated Object-Object Interactions in our dataset with bounding boxes, relation labels and action category labels. lr (Location Relations), mr (Motion Relations), and cr (Contact Relations). Only one interaction is shown in each frame for clarity.

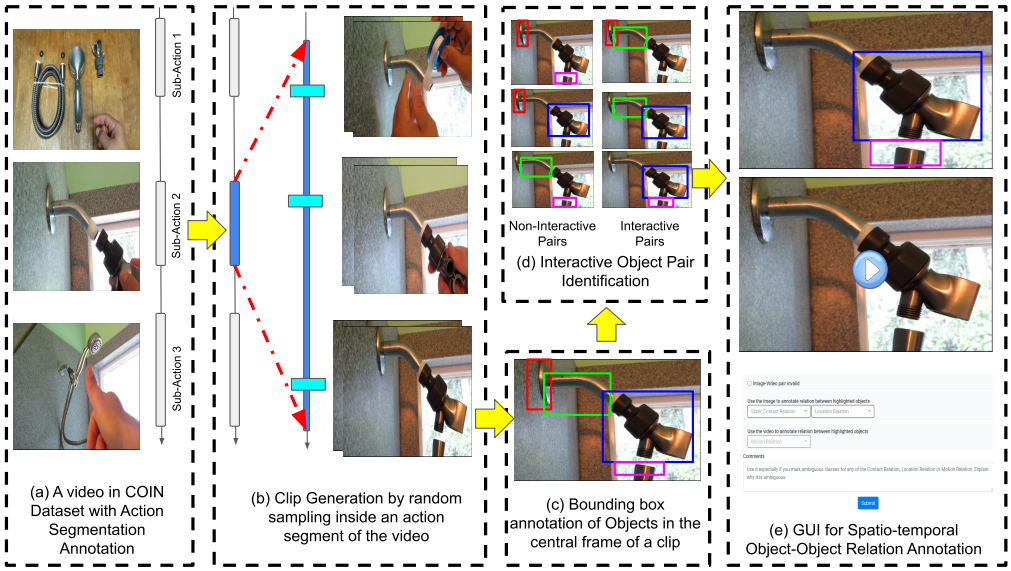

Fig. 3: Steps to generate clips (a, b), identify interactive object pairs (c, d), and annotate their interaction categories (e).

object is positioned relative to another) and contact relations (determining whether the objects are in contact). Temporal aspects relate to motion relations (depicting how the objects are moving relative to each other). We base our taxonomy on this framework to capture the spatial and temporal aspects of OOIs (see Table I for details).

While identifying sub-categories within contact and location relations is straightforward, the same task for motion relations is more complex. Employing principles of common sense physics, we classified motion relations into translational and rotatory categories and established common subcategories for each. Following this, we conducted an analysis of 500 pairs of objects from YouTube videos to identify any overlooked categories. Any newly identified relation categories were seamlessly integrated into the taxonomy.

Our taxonomy provides finely-grained categories, enabling precise annotation of interaction categories even in short-duration video clips.

B. Dataset Construction

Our dataset is built on top of COIN [10], a human activity understanding dataset consisting of YouTube videos. COIN includes various types of tasks from twelve activity domains, such as cooking and furniture assembling. We chose action

TABLE I: Taxonomy of spatio-temporal OOI categories

Contact relation	Location Relation	Motion Relation	
Contact	Right/Left	Holding	Rubbing
No contact	Behind/Front	Raising	Lowering
	Above/Below	Carrying	Rotating
	Inside	Twisting	Adjusting
		Sliding	Penetrating
		Moving Towards	Moving Away
		Negligible Relative Motion	

categories (Fig. 5) from COIN dataset, and generated clips from videos corresponding to those categories.

1) Clip Generation: To ensure the validity of the OOI classification task, we needed to select a clip duration in which the interaction category between two objects does not change. For COIN dataset, we determined through inspection that a duration greater than one-third of a second leads to a change of interaction categories between the objects. Thus, all the video clips have a duration of one-third of a second. Generation of micro video-clips was done by uniform sampling within action segments annotated by COIN dataset (see b in Fig. 3). We generated two clips for action segments less than two seconds long and five clips for longer ones. We also removed micro-clips having static objects.

2) Hand and Object Annotation: After selecting the clips, we annotate the central frame of each clip for the presence of human hands and objects. The annotation of objects is done manually. For annotation of human hands, we use pre-trained model [19] along with manual inspection and correction.

3) Relation Annotation: From the annotated objects in a clip, we select interactive pairs of objects for interaction category annotation. Annotators are shown the central frame with the object pairs indicated using overlaid bounding boxes alongside the video clip that displays object motion (see Fig. 3 (e)). Interactive object pairs were identified using a distance and mIoU based threshold as in [20]. To obtain high quality interaction category annotation, we selected and trained annotators using a four step process. In the first step, potential annotators received visual examples of interaction categories along with rationale. In second step, they were shown a tutorial

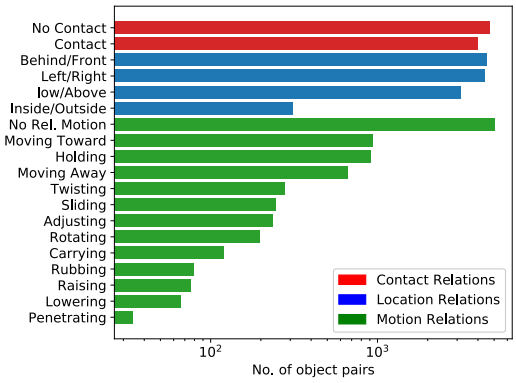

Fig. 4: Bar chart of number of annotated samples of each label in the interaction classification taxonomy.

video of the annotation interface. In third step, annotators did few sample annotations for further familiarisation. Finally,

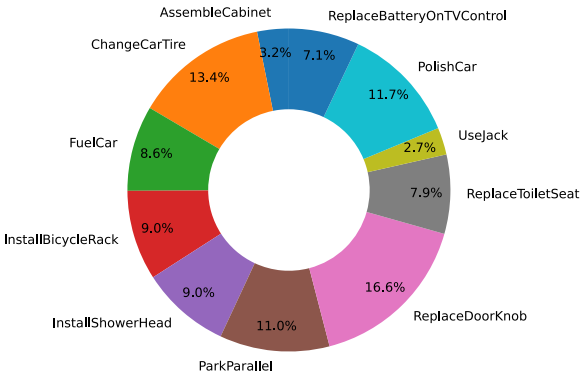

Fig. 5: Action Categories in our dataset

a quiz was conducted to check their understanding of the interaction categories and the annotation interface. Selected annotators initially received real-time feedback and query resolution. All of the steps were designed to avoid inconsistent understanding of the OOI categories by the different annotators. Our annotation system ensured that an object pair was annotated for relations by two different annotators. In case of conflicting annotations, a third annotator was used, followed by the majority voting rule to determine final annotations. We show samples of annotated interactions in Fig. 2.

4) Statistics: We annotate 9,155 object pairs across 2,200 scenes, resulting in 29,939 different OOI labels. We present the distribution of labels through a bar chart in Fig. 4.

IV. Method

We define the OOI classification task and describe the SoTA methods adopted from related areas for this task. Further, we introduce HOI and OOI based features and provide details for the Action Recognition task.

A. Object-Object Interaction Classification

1) Problem Definition: Given a clip M with frames $I_1, ..., I_K$ and objects with bounding box annotations $B_1, ..., B_N$ for the central frame $I_{\lceil K/2 \rceil}$, OOI classification task is to predict the relation labels R_{cr}, R_{lr}, and R_{mr}, for two given objects i and j. R_{cr} is addressed as a single-label multi-class and R_{lr} and R_{mr} is addressed as a multi-label multi-class classification problem. Our pipeline for OOI classification involves extracting various base features from the clip M followed by training and inference using SoTA methods. Base features and adopted methods are detailed below:

2) Base Features: Features are divided as: Object-Centric and Interaction-Centric features. While former captures information about the objects only, the latter captures information about the interaction between two given objects.

I3D based Object-Centric Feature:

$$O_i^{i3d} = RoIAlign\left(f_{res}(x_i, y_i, w_i, h_i)\right) \quad (1)$$

x_i, y_i, w_i, h_i are the x and y co-ordinate, width and height of bounding box of i^{th} object. f_{res} is the feature map of the central

frame generated by I3D backbone.

Vision-Transformer (ViT) based Object-Centric Feature:

$$O_i^{vit} = MOA\left(g_{res}(x_i, y_i, w_i, h_i)\right) \tag{2}$$

Masked object attention (MOA) [21] is used to extract region of interest (RoI) features from ViT. Existing RoI operators such as RoIAlign are not suited for extracting RoI features from ViT backbone due to its coarse output. g_{res} is feature map of central frame generated by ViT.

Vision-Transformer (ViT) based Interaction-Centric Feature:

$$I_{ij}^{vit} = MOA\left(g_{res}(x_{ij}, y_{ij}, w_{ij}, h_{ij})\right) \tag{3}$$

$x_{ij}, y_{ij}, w_{ij}, h_{ij}$ are the x co-ordinate, y co-ordinate, width and height of smallest bounding box containing bounding boxes of objects i and j.

Bounding-Box based Object-Centric Feature:

$$O_i^{box} = \left[\frac{x_i}{w}, \frac{y_i}{h}\frac{w_i}{w}, \frac{h_i}{h}, \frac{A_i}{A}\right], \tag{4}$$

A_i is the bounding box (in central frame) area, and w, h, A are the width, height, and the area of the entire image. This features informs about the geometry of the object.[22]

Bounding-Box based Interaction-Centric Feature:

$$I_{(i,j)}^{box} = \left\{I_{t,(i,j)}^{box}\right\}_{t=1}^{T} \tag{5}$$

where

$$I_{t,(i,j)}^{box} = \begin{bmatrix} \Delta(b_{i,t}, b_{j,t}), \Delta(b_{i,t}, b_{ij,t}), \Delta(b_{j,t}, b_{ij,t}), \\ \text{IoU}(b_{i,t}, b_{j,t}), \text{dis}(b_{i,t}, b_{j,t}) \end{bmatrix} \tag{6}$$

T represents total number of frames. $b_{i,t}$ represents i^{th} bounding box at time t. $b_{ij,t}$ represents the union box of $b_{i,t}$ and $b_{j,t}$. $IoU(b_{i,t}, b_{j,t})$ and $dis(b_{i,t}, b_{j,t})$ denote the IoU and normalized-distance between the i^{th} and j^{th} bounding boxes in t-th frame. $\Delta(b_{i,t}, b_{j,t})$ are the box deltas [23].

Word2Vec based Semantic features:

$$O_i^{w2v} = \sum_{i=1}^{n} p_k \cdot e_k \tag{7}$$

where $\sum_{i=1}^{n} p_k = 1$ and p_k is the probability score of the k^{th} class from the object detector. e_k represents the Word2Vec embedding of the k^{th} object class name. Inspired by [24], O_i^{w2v} provides semantic prior of the object category without explicit ground truth category annotations.

Segmentation Mask based Object-Centric shape features:

While we haven't provided ground-truth segmentation mask, we use the Segment-Anything model [25] to extract segmentation masks using ground truth bounding boxes. The predicted masks are used to extract object shape features.

$$O_i^{shape} = \left[A_i^s, C_i^x, C_i^y, E_i, S_i, H_i, P_i, L_i^{major}, L_i^{minor}\right], \tag{8}$$

Where:

- A_i^s: Normalized area of the object i. Computed as the ratio of the mask area to the total image area, $\frac{A_{\text{mask, i}}}{H \times W}$.
- C_i^x and C_i^y: Normalized x and y coordinates of the centroid of object i. Computed as $\frac{C_{x,\text{mask, i}}}{H}$ and $\frac{C_{y,\text{mask, i}}}{W}$ respectively.
- E_i: Eccentricity of object i, describing the shape of the mask. It is the ratio of the distance between the foci of the ellipse equivalent to the mask, to its major axis length.
- S_i: Solidity of object i, which is the ratio of the object's area to its convex hull's area, $\frac{A_{\text{mask, i}}}{A_{\text{convex\_hull, i}}}$.
- H_i: Extent of object i, which is the fraction of the pixels in the object's bounding box that are also in the region, $\frac{A_{\text{mask, i}}}{A_{\text{bounding\_box, i}}}$.
- P_i: Normalized perimeter of object i, computed as the ratio of the object's perimeter to the perimeter of the image, $\frac{P_{\text{mask, i}}}{2(H+W)}$.
- L_i^{major} and L_i^{minor}: Normalized major and minor axis lengths of the ellipse equivalent to object i. They are computed as $\frac{L_{\text{major, mask, i}}}{\max(H,W)}$ and $\frac{L_{\text{minor, mask, i}}}{\max(H,W)}$ respectively.

Segmentation Mask based Interaction-Centric shape features:

$$I_{(i,j)}^{shape} = \left\{O_{i,t}^{shape} - O_{j,t}^{shape}\right\}_{t=1}^{T} \tag{9}$$

3) Methods Adopted for OOI classification: In the context of OOI classification, our features are specifically crafted to capture both object-centric details and interaction-centric dynamics. However, it's crucial to integrate broader context of the objects as well. To do so, we model the scene as a graph where individual objects act as nodes, and their interactions form the edges. For effective feature aggregation within this graph structure, we leverage SoTA Graph Neural Networks (GNNs) from related fields. The node features are constructed by concatenating all object-centric features. Similarly edge features are formed by concatenating interaction-centric features. Afterwards, these features undergo aggregation by the GNN. Finally, for the object pair whose interaction category we need to predict, we concatenate the aggregated node and edge features and pass it to three classification heads to predict motion, location, and contact relations. From the HOI field, we adopt GPNN [15]. Graph-RCNN [26], Iterative Message Passing [27], Quad Attention Transformer [28] are adopted from Scene Graph generation field. Hierarchical Graph Attention Network [20] is adopted from Visual Relationship detection field. We also adopt the message passing frameworks Node-Edge Neural Net [29] and Graph Transformer [30] for context aggregation. All the architectures except HGAT assume a fully connected graph. HGAT only considers edges between two objects if they satisfy a IoU and distance based measure [20].

4) Implementation Details: We use a multi-object tracker to track objects across the video. I3D network [31] pre-trained on kinetics dataset [32] is used to extract O_i^{i3d}. Pretrained ViT [33] is used to extract ViT based features. We set the learning rate of benchmarked models at $1e-3$. Adam optimizer is used for all the networks. We report 5 times repeated 5 fold cross validation results averaged across different train-val splits. We implement our models in PyTorch Deep Learning framework. All experiments were conducted on NVIDIA RTX A6000 GPU.

B. Action Recognition

Interaction categories occuring during the course of an action can represent that action. Consider the action: "Fueling a car". This action involves interaction category of 'penetrating' and "moving away", occurring with fuel pump nozzle entering and leaving the tank. We define F_V, which captures normalized occurrence counts for interaction categories observed during the course of the action.

$$F_V = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{Q_m}\sum_{j=1}^{Q_m}\mathbf{q_j}\right) \tag{10}$$

Here, $\mathbf{q_j}$ is 19-dimensional label vector for the j-th interaction (HOI or OOI) within the m-th micro-clip of the video V. Q_m is the number of entity pairs (human-object or object-object) in the micro-clip m. M is the total number of micro-clips in the video V. Our Action Recognition experiment encompasses four distinct settings.

- HOI: F_V only considers HOI labels in each micro-clip.
- OOI: F_V only considers OOI labels in each micro-clip.
- HOI+OOI: F_V considers both HOI and OOI labels in each micro-clip.
- HOI‖OOI: Concatenates the vectors from HOI and OOI setting above.

We aim to (1.) compare effectiveness of HOI and OOI based feature vectors for action recognition (2.) investigate effect of distinguishing OOI with HOI. While 'HOI+OOI' setting eliminates any distinctiveness between HOI and OOI, HOI‖OOI setting preserves their distinctiveness. In total, our dataset has 11 action categories. We use Decision Trees to perform classification and report our results. We split our dataset into a training dataset of 350 and validation set of 88 videos. Results presented are the average of 1000 train-val runs, with randomized train-val splits for each new run. Results are reported in Table IV.

V. Experiments

TABLE II: Benchmarking Results. Cells with the highest scores in each column are highlighted. motion relations (mr), location relations (lr) and contact relations (cr).

Method	# Param (in M)	mAP_{all}	mAP_{mr}	mAP_{lr}	mAP_{cr}
GPNN [15]	1.56	86.72	71.35	93.08	95.73
GraphRCNN[26]	5.12	81.04	60.66	88.76	93.70
NENN[29]	1.64	83.47	65.05	90.70	94.67
HGAT[20]	3.26	84.23	65.50	91.60	95.58
IMP[27]	1.89	84.32	65.40	91.70	95.87
SQUAT[28]	9.89	80.97	61.98	87.43	93.51
GraphTrans[30]	4.30	85.16	67.32	93.22	94.93

A. OOI Classification

1) Benchmarking: We perform a thorough benchmarking of our dataset in Table II using SoTA methods described in previous section. We observe lower performances for mr classification (max 71.35%), as compared to lr and cr classification, thus indicating the challenge in classifying motion relations. Among all the tested approaches, GPNN shows the best performance overall as well as on motion relations. IMP achieves best performance for contact relations. Both of these architectures iteratively refine attention over edges and nodes for the fully connected graph. From their superior performance, and also looking at the nature of the OOIC problem, that it has many noisy and irrelevant edges, we posit that GNN architectures which iteratively perform attention are well suited for OOI classification task. We also show the Average Precision (AP) for each label in Fig. 6. We observe a significant variation in the AP values for labels in motion relations. Motion relations exhibit a strong long tailed nature (as show in Fig. 4) and poor performance of the labels in the far end of the tail can be caused by lesser number of training samples. However, we also see trends against the long tail (such as AP for rubbing is greater than AP for carrying). We suspect that some of the relation-labels (such as carrying or adjusting) has large intra-class variance, thus hampering their performance.

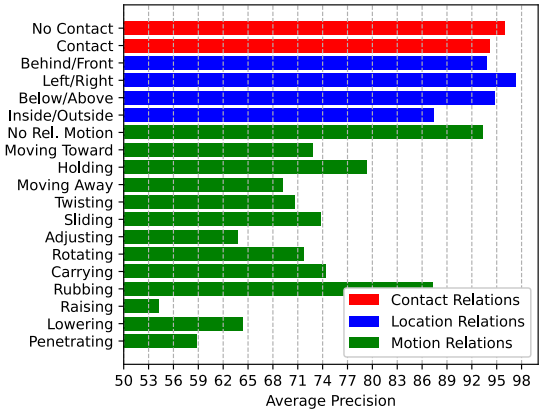

Fig. 6: Average Precision (AP) values for each label (using GPNN model). Predicting motion relations are more challenging compared to location and contact relations.

2) Base Feature Ablation for OOI classification: We perform feature ablation study (see Table III) to determine the importance of different features for the OOI classification task. Each ablated feature contributes significantly in final performance. However, some features like I3D-ResNet50 and ViT-based object-centric features appear to have a more significant impact on performance than others. Our observations are: (1) Contribution of I3D features - I3D features contribute exceptionally well to the performance of all the categories. Our observation indicates the critical nature of deep 3d convolutional features for the task of OOI classification. (2) ViT - While ViT contributes significantly, it's contribution is lower than expected. We posit that this is due to ViT's coarse 16x16 feature map which may cause a loss of object details, as noted in [21]. While MOA attempts to address this, the limitations persist because of the initial coarse map. Using ViT has benefits, but there's potential for improvement with a larger resolution feature map. (3) Success of Semantic Features - O_i^{w2v} plays a pivotal role in all relation categories. This is promising since this feature attempts to address the

open vocabulary nature of OOIs, catering to interactions with objects from unknown categories. The success of this feature encourages approaches in zero/few shot learning to be applied to OOI classification problem. (4) Contribution of Non-Deep features - Non-deep features contribute significantly to the performance and even outperform deep features (O_i^{w2v}, O_i^{vit} and $I_{i,j}^{vit}$ in many instances. The non-deep features are strikingly low-dimensional as compared to deep features, and are designed to capture the geometrical aspects of OOIs. Their success is noteworthy and indicates their crucial nature for achieving higher performance without much increase in computational cost. (5) Contribution of Interaction Centric Features - Interaction centric features have quadratic memory complexity with respect to number of objects, while object centric features have linear memory complexity with respect to number of objects. However, the crucial role of interaction-centric features in pushing performance, demands their inclusion. It's also noteworthy that for both bounding-box and shape based features, the interaction centric versions contribute more than object centric versions for all the relation categories.

TABLE III: Ablation study results. Drops are shown in blue parentheses, with the largest drop in each category boldfaced.

Ablated	mAP_{all}	mAP_{mr}	mAP_{lr}	mAP_{cr}
None	86.7	71.4	93.1	95.7
O_i^{vit}	81.7 (-5.0)	63.5 (-7.9)	87.7 (-5.4)	93.5 (-2.2)
O_i^{i3d}	69.1 **(-17.6)**	45.5 **(-25.9)**	80.2 **(-12.9)**	81.7 **(-14.0)**
O_i^{box}	81.3 (-5.4)	62.6 (-8.8)	87.8 (-5.3)	93.5 (-2.2)
O_i^{w2v}	82.6 (-4.1)	63.6 (-7.8)	90.3 (-2.8)	94.0 (-1.7)
O_i^{shape}	83.7 (-3.0)	65.7 (-5.7)	91.5 (-1.6)	93.9 (-1.8)
$I_{i,j}^{vit}$	83.4 (-3.3)	64.2 (-7.2)	90.6 (-2.5)	95.5 (-0.2)
$I_{i,j}^{box}$	80.55 (-6.2)	62.5 (-8.9)	87.3 (-5.8)	91.9 (-3.8)
$I_{i,j}^{shape}$	83.3 (-3.4)	64.2 (-7.2)	91.2 (-1.9)	94.5 (-1.2)

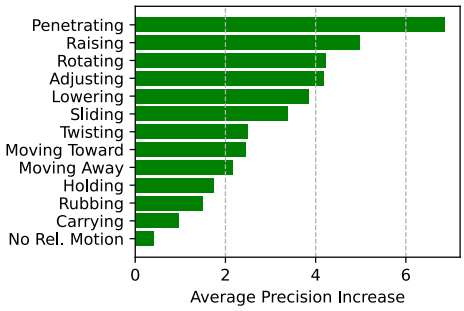

Fig. 7: Increase in AP values (going from 0% to 100% temporal observation) for motion relations labels.

3) Number of frames ablation for OOI classification: To validate micro-clip usage, we conducted an ablation study (Fig. 8), varying the frame count from one (central frame only) to eleven (five frames before/after the central frame) in steps of two (adding one frame each side of central frame). Performance, evaluated using GPNN network—best in benchmarking—improved with increased frame count across all relation super-categories (lr, mr, cr), affirming micro-clips' efficacy in predicting interactions. For motion relations prediction using GPNN, Fig. 7 indicates benefit of including temporal information for each label in motion relations. Categories having less motion (holding, negligible relative motion) benefit less from temporal information than categories having more motion (penetrating, rotating).

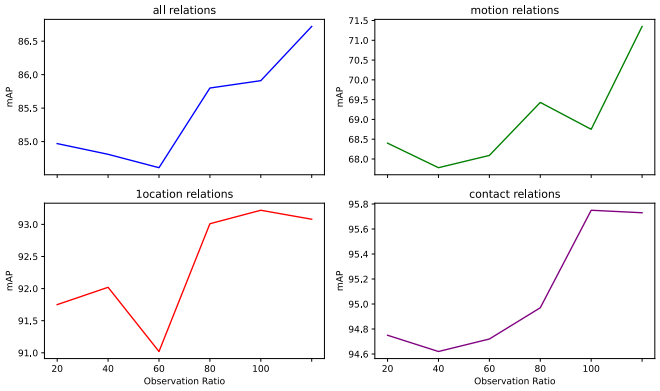

Fig. 8: Increasing the video observation ratio from 0% (using central frame) to 100% (using all 11 frames) positively affect mAP for lr (location relation), mr (motion relation) and cr (contact relation) due to the added temporal information.

B. Action Recognition on COIN

TABLE IV: Action Recognition accuracy using OOI and HOI based features measured across different combinations of relation categories. OOI consistently outperforms HOI based features. (Cells with highest score for a column are highlighted with light blue.)

Features	cr	lr	mr	cr+lr	cr+mr	lr+mr	cr+lr+mr
HOI	23.07	14.24	25.30	23.41	26.41	27.24	31.23
OOI	18.77	25.37	34.38	30.61	34.26	36.69	41.20
HOI+OOI	18.85	20.94	30.58	27.16	31.41	37.03	37.98
HOI‖OOI	24.88	25.72	34.07	32.30	36.06	39.33	43.14

OOI features prove to be significantly more performant and discriminative for Action Recognition, as compared to HOI features (9.97% increase from HOI to OOI when all 3 relations are used). Additionally, the performance of the fused feature, HOI‖OOI shows a significant increase of 5.15% compared to the naive fusion of HOI+OOI, suggesting that it is beneficial to consider HOI and OOI as distinct sources of information. The maximum performance for cr+lr+mr, indicates that the three relation categories (cr, lr and mr) are important for Action Recognition. High performance of OOI features, seen together with the high performance of motion relations, validates the inclusion of temporal aspects in our taxonomy.

VI. Limitations and Future Work

A larger dataset can enhance the generalizability of trained algorithms, enabling their application to challenging real world scenarios. Future OOI research can explore larger datasets and investigate segmentation settings using longer clips, allowing OOI categories between object pairs to change with time. Also,

OOIs may be applied to tasks which require understanding of compositional elements of dynamic environments such as human-robot or robot-robot collaboration or learning from human demonstrations.

VII. Conclusion

In this letter, we introduce a novel dataset focusing on OOI categories in dynamic scenes. Further, we propose the OOI classification task and benchmark our dataset. We also compare OOI and HOI-based features for Action Recognition, where OOI-based features outperform HOI-based ones thus highlighting potential of OOI for robotics and computer vision tasks.

References

[1] M. Sieb, Z. Xian, A. Huang, O. Kroemer, and K. Fragkiadaki, "Graph-structured visual imitation," in *Conference on Robot Learning*. PMLR, 2020, pp. 979–989.

[2] C. Agia, K. M. Jatavallabhula, M. Khodeir, O. Miksik, V. Vineet, M. Mukadam, L. Paull, and F. Shkurti, "Taskography: Evaluating robot task planning over large 3d scene graphs," in *Conference on Robot Learning*. PMLR, 2022, pp. 46–58.

[3] S. Amiri, K. Chandan, and S. Zhang, "Reasoning with scene graphs for robot planning under partial observability," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5560–5567, 2022.

[4] R. Inam, K. Raizer, A. Hata, R. Souza, E. Forsman, E. Cao, and S. Wang, "Risk assessment for human-robot collaboration in an automated warehouse scenario," in *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA)*, vol. 1. IEEE, 2018, pp. 743–751.

[5] S. Li, P. Zheng, Z. Wang, J. Fan, and L. Wang, "Dynamic scene graph for mutual-cognition generation in proactive human-robot collaboration," *Procedia CIRP*, vol. 107, pp. 943–948, 2022.

[6] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 236–10 247.

[7] Y. Ou, L. Mi, and Z. Chen, "Object-relation reasoning graph for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 133–20 142.

[8] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 399–417.

[9] X. Li, D. Guo, H. Liu, and F. Sun, "Embodied semantic scene graph generation," in *Conference on Robot Learning*. PMLR, 2022, pp. 1585–1594.

[10] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "Coin: A large-scale dataset for comprehensive instructional video analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1207–1216.

[11] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1017–1025.

[12] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[13] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J. C. Niebles, "Home action genome: Cooperative compositional action understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 184–11 193.

[14] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone, "3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans," *arXiv preprint arXiv:2002.06289*, 2020.

[15] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[16] Y. Sun, S. Ren, and Y. Lin, "Object–object interaction affordance learning," *Robotics and Autonomous Systems*, vol. 62, no. 4, pp. 487–496, 2014.

[17] Y. Zhu, Y. Zhao, and S.-C. Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," in *CVPR*, 2015.

[18] K. Mo, Y. Qin, F. Xiang, H. Su, and L. Guibas, "O2o-afford: Annotation-free large-scale object-object affordance learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 1666–1677.

[19] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding human hands in contact at internet scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[20] L. Mi and Z. Chen, "Hierarchical graph attention network for visual relationship detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 886–13 895.

[21] J. Park, J.-W. Park, and J.-S. Lee, "Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 152–17 162.

[22] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 792–807.

[23] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3588–3597.

[24] P. Huang, J. Han, D. Cheng, and D. Zhang, "Robust region feature synthesizer for zero-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 7622–7631.

[25] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[26] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–685.

[27] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.

[28] D. Jung, S. Kim, W. H. Kim, and M. Cho, "Devil's on the edges: Selective quad attention for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 664–18 674.

[29] Y. Yang and D. Li, "Nenn: Incorporate node and edge features in graph neural networks," in *Asian conference on machine learning*. PMLR, 2020, pp. 593–608.

[30] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. Sun, "Masked label prediction: Unified message passing model for semi-supervised classification," *arXiv preprint arXiv:2009.03509*, 2020.

[31] 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*, 2017, pp. 6299–6308.

[32] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.