RealGraph: A Multiview Dataset for 4D Real-world Context Graph Generation

Haozhe Lin[*], Zequn Chen[*], Jinzhi Zhang[*], Bing Bai, Yu Wang, Ruqi Huang, Lu Fang[†]
Tsinghua University
fanglu@tsinghua.edu.cn

Abstract

Understanding 4D scene context in real world has become urgently critical for deploying sophisticated AI systems. In this paper, we propose a brand new scene understanding paradigm called "Context Graph Generation (CGG)", aiming at abstracting holistic semantic information in the complicated 4D world. The CGG task capitalizes on the calibrated multiview videos of a dynamic scene, and targets at recovering semantic information (coordination, trajectories and relationships) of the presented objects in the form of spatio-temporal context graph in 4D space. We also present a benchmark 4D video dataset "RealGraph", the first dataset tailored for the proposed CGG task. The raw data of RealGraph is composed of calibrated and synchronized multiview videos. We exclusively provide manual annotations including object 2D&3D bounding boxes, category labels and semantic relationships. We also make sure the annotated ID for every single object is temporally and spatially consistent. We propose the first CGG baseline algorithm, Multiview-based Context Graph Generation Network (MCGNet), to empirically investigate the legitimacy of CGG task on RealGraph dataset. We nevertheless reveal the great challenges behind this task and encourage the community to explore beyond our solution. Our project page is at https://github.com/THU-luvision/RealGraph .

1. Introduction

Understanding our natural world in 4D space-time is the fundamental challenge in building sophisticated AI systems. Recently, significant progress has been made in this area, including 3D object detection [12, 44] and 3D multi-object tracking [5, 30, 47]. In addition to detecting and tracking 3D objects, understanding the interaction and relationships between humans and objects in dynamic environments are also essential to intellectual perception and cognition science [59, 25]. Take for instance, deploying embodied AI [24, 17] in a café, as shown in Fig. 1, requires scene understanding in complicated 4D space-time in order to make prompt decisions: service robots need to estimate objects' 3D trajectories and to understand their semantic relationships (e.g. $\langle man - sit\ at - table \rangle$) to infer human intention and to offer help. A legitimate solution to approach 4D scene understanding is to interpret the scenes in the form of a semantic graph, which encompasses 4D

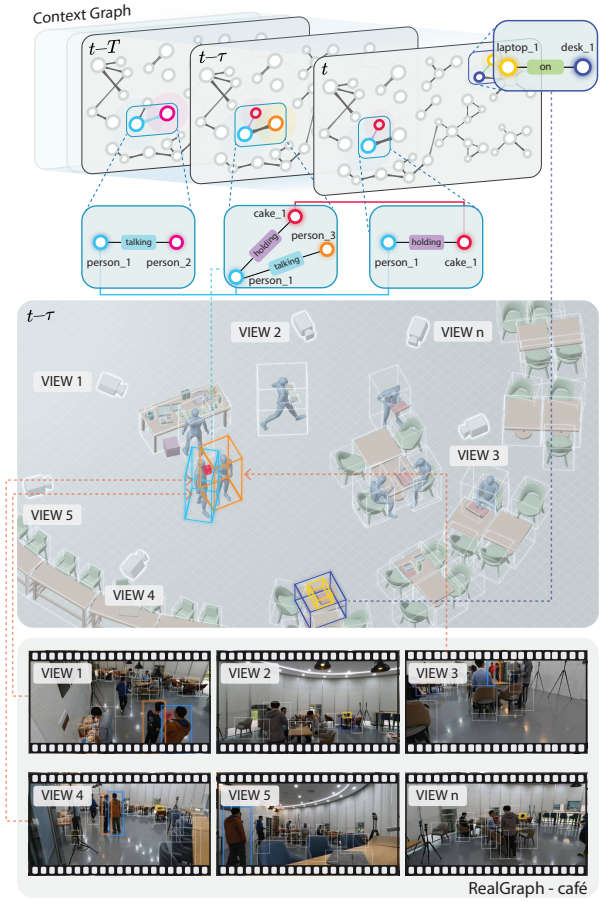


Figure 1. From multiview videos (bottom) to context graph (top). The holistic understanding of 4D scenes empowers the noise-robust, context-aware, and scene-adaptive visual analysis.

[1]These authors contributed equally to this work.
[2]Lu Fang is the corresponding author (www.luvision.net).

information and is denoted as a **context graph**[1].

Existing approaches may not even work to generate context graph. Current scan-based method for 3D scene graph generation [2, 52, 66] can only deal with static indoor scenes, given spatial relationship annotation like $\langle sofa - close\,to - table \rangle$ and static attribute like geometry and color. Owing to the time consuming and complicated 3D scan (e.g. with RGB-D camera or LiDAR) processing procedure, it's hard to capture semantic information of the complete scene in real-time. Therefore, there is a high demand for a novel approach to representing 4D semantics as a context graph, which is currently unexplored and presents a significant challenge.

Given the numerous inherent benefits of multiview videography, such as easy accessibility, comprehensive coverage, robustness, and high reliability, we believe it is practically valuable to consider a brand new modeling paradigm, named as **context graph generation (CGG)** from multiview videos. Specifically, CGG takes multiple synchronized 2D videos of the same dynamic scene as input, and the goal is to estimate 3D coordination and trajectory of targeted objects, and the inference on semantic relationships between them within the 4D space-time, and finally generate a context graph that is consistent across temporal and spatial contexts. With the information from context graph, questions subject to temporal clues like "when did the man sit down on the chair" or "what did the man do after eating a cake" becomes easier to address, which is crucial for successful deployment of embodied AI system. However, without explicit 3D information, it's non-trivial for annotators to accurately annotate 3D labels, and it is even more challenging for algorithms to process these information in 4D space. In general, CGG requires both large scale annotated data and tailored model to succeed.

By taking into account the above considerations, we propose the first multiview video (4D) dataset tailored for context graph generation (CGG), named RealGraph. In general, RealGraph dataset captures 13 real-world scenes, with more than 2.4M video frames, and provides various human annotated labels, including 2.3M 2D bounding boxes, 760K 2D relationships, 420K 3D bounding boxes and 130K 3D relationships out of 37 object categories and 18 relationship categories in total, and each object has a unique identifier across different views and frames. Apart from CGG, Realgraph supports the deployment of several traditional tasks like 2D scene graph generation, 3D detection, 3D Multi-object tracking. The multiview cameras in RealGraph are synchronized and uniformly distributed in each scene. We demonstrate one example frame of café in Fig. 1 with annotations, certain objects are occluded or missing in a single view, while the missing information can be complemented from the other views, depending on the spatial-temporal context. This could compensate for the limitation of inference from single view, but also remains as a challenge in terms of information fusion across different views in dynamic scenes.

To solve the CGG task, models face two major challenges: tracking objects of various scales from multiview with random occlusion and out-of-view problems; and inference of semantic relationships from multiple views with large perspective disparity. We propose a tailored learning-based model to address these problems. For multi-object tracking, we adopt a multi-scale feature fusion module to better detect small-scale objects, and a double association scheme to improve tracking performance. For relationship prediction, we reuse the 3D feature volume from the detector and apply a sequential network to extract context information. We nevertheless encourage the community to explore alternative solutions on the new task.

To summarize, our contributions are three-folds:

- Inspired by 3D scene graph [2], we propose a new task: context graph generation (CGG) from multiview 2D videos, which aims to describe dynamic 3D objects and their relationships as an abstract of the 4D real world in the form of graph. We hope the paradigm of CGG could further benefit downstream applications like VQA, robotics and augmented reality.

- We propose a new benchmark dataset RealGraph, the first dataset tailored for CGG task. This dataset consists of multiview synchronized videos for various scenes along with CGG annotations. RealGraph also provides benchmarks on basic 3D tasks like 3D object detection, 3D multi-object tracking and 3D scene graph generation.

- We propose the first baseline method to solve the CGG task based on RealGrpah dataset. Extensive experiments demonstrate that CGG task is non-trivial, and remains as an intriguing and open problem that is of potential interests of the community.

2. Related work

Dataset	Video	#Views	#Frames	#ObjClass	#RelClass	Calibration
3DSG[2]	✗	-	-	80	7	✓
3DSSG[52]	✗	-	-	187	40	-
SGNet[66]	✗	-	-	51	6	-
KITTI[16]	✓	4	15K	8	-	✗
Waymo[47]	✓	5	230K	4	-	✓
nuScenes[5]	✓	6	40K	23	-	✓
STCrowd[9]	✓	1	11K	1	-	-
HOI4D[32]	✓	1	2.4M	16	54	-
LEMMA[22]	✓	4	4.6M	14	16	✗
BEHAVE[4]	✓	4	15.2K	20	1	✗
HAKE-3D[29]	✓	1	9K	40	48	-
RealGraph (ours)	✓	8∼15	2.4M	37	18	✓

Table 1. Comparisons between RealGraph and relevant datasets. #ObjClass: number of object classes. #RelClass: number of relationship classes.

[1]Formal definition is presented in Sec. 3.1.

2.1. Object detection and tracking

The collection of 2D object detection datasets [15, 42, 31, 43, 21, 62, 63, 53] has contributed to remarkable progress of the object detection in 2D images. Recently, there are also growing interests in 3D perception tasks, along with an increasing need for large-scale 3D datasets. For instance, SUN RGB-D [44] provides labeled 2D polygons and 3D cuboids for various indoor scenes. ScanNet [12] is an RGB-D video dataset annotated with camera poses, surface reconstruction, and semantic segmentation. These annotations are generally obtained through 3D-2D projection techniques based on the dense depth information collected from RGB-D cameras. More advanced and elaborated labels such as geometry primitives and instance segmentations can be generated from a photo-realistic simulations, such as SceneNet RGB-D [35] and Structured3D [65]. However, these datasets only consider data collection in static indoor rooms, making it hard to apply to 3D perception in dynamic environments.

Another line of research focuses on robust 3D analysis for autonomous driving [34]. Datasets in this domain include [16, 30, 47, 5, 39, 8]. These works often rely on data collected from mixed multi-modal sensors, including synchronized cameras, radar and LiDAR point cloud. Though these datasets usually contain temporal information, the annotations are always in the form of 3D bounding boxes and trajectories of objects, neglecting semantic relationships and interactions within the scene.

Nevertheless, we still believe the tasks of 3D object detection and tracking form the very foundation of 3D scene understanding. Therefore, our proposed RealGraph dataset also provides rich annotations for these tasks. The main difference is that RealGraph is not confined to specific application like autonomous driving, and contains various scenes both indoor and outdoor.

2.2. Scene understanding

The perception of 3D scenes plays a critical role for human vision system. Early works investigate room layout estimation with cuboids [13, 20, 27] or primitives [11, 67, 46, 40]. With the advance of deep learning, many works start to estimate 3D bounding boxes and object poses [7, 50, 51, 14]. Some works [61, 37] explore further to jointly solve multiple scene understanding tasks, including estimating the room layout, object poses and shapes. Human-Object Interaction (HOI) in 3D [32, 22, 29] has made great progress these days, it provides elaborate understanding on how human interact with objects in daily life. However, most of HOI works focus on local actions of human during specific interaction, neglecting various semantic relationships within the whole scene in long-term observation.

Among existing scene understanding techniques, scene graph [25, 26, 23] provides a compact yet comprehensive structure to abstracting semantic from the scene. Apart from object class and location, it includes object attributes and semantic relationships between objects. Video Scene Graph [10, 48, 57, 28], extends this structure to temporal dimension for dynamic scene understanding. 3D scene graph [60, 2, 52] borrows the idea of scene graph and apply it to the realm of 3D vision. Some works[7] introduces physical commonsense to help improve the 3D relationship estimation. Beyond that, context priors can also be employed to give a more holistic understanding with larger respective field in a 3D context network [1]. However, the related datasets [3, 6, 65, 60, 2, 52] usually lack pose annotations or human interactions in dynamic environments, which hinders the algorithms from general scene understanding.

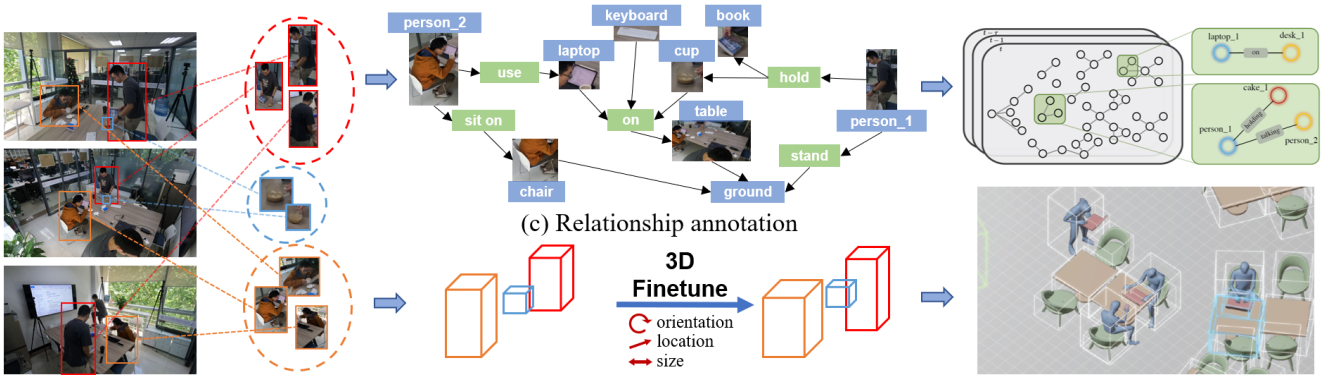
Compared to the previous paradigm, our proposed 4D context graph generation (CGG) task leverages multiview videos to parse semantic relationships in dynamic 3D environments. CGG waives the need of 3D signals from RGD-D or LiDAR equipment, and provides more robust and compact representation for 3D scene understanding. Relatively, our proposed RealGraph dataset is the first dataset exclusively designed for CGG task. It's most suitable for real-world 4D scene understanding, especially for those with human activities and interaction with objects. We believe the proposed RealGraph dataset can benefit the community in better understanding dynamic 3D scenes, and encourage the development of real-world perception algorithms.

3. Dataset and problem formulation

We always prioritize the privacy issue during the collection of the dataset: we ensure that all participants appeared in the dataset are aware of the data collection process and provide their formal consent for the usage of their information.

3.1. Formal definition

The CGG task aims to construct a 4D *context graph*, a structured representation of context information of a dynamic scene, from multiview calibrated videos. Context graph requires scene graph generation process respect additional 4D "context" consistency constraint imposed both temporarily and spatially. Beyond the definition of conventional 3D scene graph [60], the new context graph contains four parts: 1) a set of 3D bounding boxes $\mathcal{B} = \{B_1, ..., B_n\}$, $B_n = (x, y, z, l, w, h, \theta)$ indicates the 3D location, size, rotation angle of the 3D box; 2) a corresponding set of object labels $\mathcal{O} = \{o_1, ..., o_n\}$, $o_n \in \mathcal{C}$, where $\mathcal{C}$ is all object categories; 3) a set of object "*global*" tracking ID $\mathcal{I} = \{i_1, ..., i_n\}$, $i_n \in \mathbb{N}$ (by "*global*", we mean the annotated ID is spatially and temporally consistent across all the views, 3D space and timestamps); 4) a set $\mathcal{R} = \{r_1, ..., r_m\}$ of relationships between those ob-



(a) 2D annotation (b) ID annotation (c) Relationship annotation (d) 3D initialization (e) 3D annotation (f) Context graph

Figure 2. The annotation pipeline of one timestamp in RealGraph. (a) The 3D regions of objects are firstly labeled on each 2D frame separately. (b) Each object is manually assigned with a global ID across different views and timestamps. (c) Then we annotate the semantic relationships between the objects on 2D images. (d)(e) The 3D location of objects are firstly initialized from multiview triangulation with 2D boxes, then manually adjusted for better quality. (f) Relationships in 3D are automatically gathered from 2D annotation and objects' global ID. We continue this procedure along all the views and frames to annotate the whole scene.

jects. Each relationship $r_m \in \mathcal{R}$ is a triplet of a "head" node $(B_h, o_h, i_h) \in \mathcal{B} \times \mathcal{O} \times \mathcal{I}$, a "tail" node $(B_t, o_t, i_t) \in \mathcal{B} \times \mathcal{O} \times \mathcal{I}$, and a predicate label $x_{h \to t} \in \mathcal{P}$, where $\mathcal{P}$ is all predicate categories. A demonstration of the context graph is shown in Fig. 1. Major categories of objects $o_n \in \mathcal{C}$ and predicates $x_{h \to t} \in \mathcal{P}$ can be found in Fig. 3. Example scenes and extra statistics are in supplementary materials.

3.2. Data collection

The raw data in RealGraph dataset is a set of multiview RGB videos captrued with GoPro Hero 10 cameras. The capture mode is cinematic, with the resolution of 5312×2988, the frame rate of 30 FPS, and the horizontal field of view (FOV) of $87°$. For each scene, we set up several synchronized GoPro cameras and fix their position. The principle of arranging camera position is to ensure the union of camera views covers as much 3D space as possible. This multiview setup aims to provide abundant visual information of the scene. Before the start of recording, we carefully calibrate all the fixed cameras. Example scenes of camera distribution and detailed calibration process can be found in supplementary materials.

Currently, RealGraph dataset covers 13 dynamic scenes[2] with human activities. Each scene is captured with 8 to 15 cameras with 3 to 20 minutes' 30Hz HD video, the number of camera is determined by the scale and complexity of the scene. We manually synchronize all the camera recordings in post-processing. Compared with single-view recording, the multiview setup helps to capture the missing information due to occlusion and out-of-view problems in dynamic scenes.

[2]We will continue to update the dataset by collecting new scenes and providing more annotations.

3.3. Data annotation

We provide fully manually annotated labels on RealGraph dataset, including 2.3M 2D bounding boxes, 760K 2D relationships, 420K 3D bounding boxes and 130K 3D relationships out of 37 object categories and 18 relationship categories in total, each object has a unique identifier across different views and frames. Note that the raw data of RealGraph is 30 FPS, we only annotate semantic labels at 1 FPS. The general statistics of semantic annotations are illustrated in Fig. 3.

In the following sections, we introduce the annotation procedure of each component. A demonstration of the annotation pipeline is shown in Fig. 2.

3.3.1 2D annotations

We specify the bounding boxes of humans and objects for 2D detection in each view, in the form of $b_n = (x, y, w, h)$. Besides, we carefully assign each object a global identifier i_n, which is shared across different views and frames. The identifiers will be used in 3D initialization during annotation, and as a part of ground truth in 3D tracking and CGG tasks. We omit the objects having severe occlusions, only annotate 2D objects obscured for less than 30% in areas in an image.

3.3.2 3D annotations

The 3D spatial annotation refers to specifying objects' 3D bounding boxes $B_n = (x, y, z, l, w, h, \theta)$, where (x, y, z) are the coordinate of the center, (l, w, h) are length, width and height, and θ are the rotation angle around x-axis. Given annotated 2D bounding boxes (with label and id) and camera poses, an initial 3D bounding box is calculated by multiview triangulation [58]. By back projection, the 2D

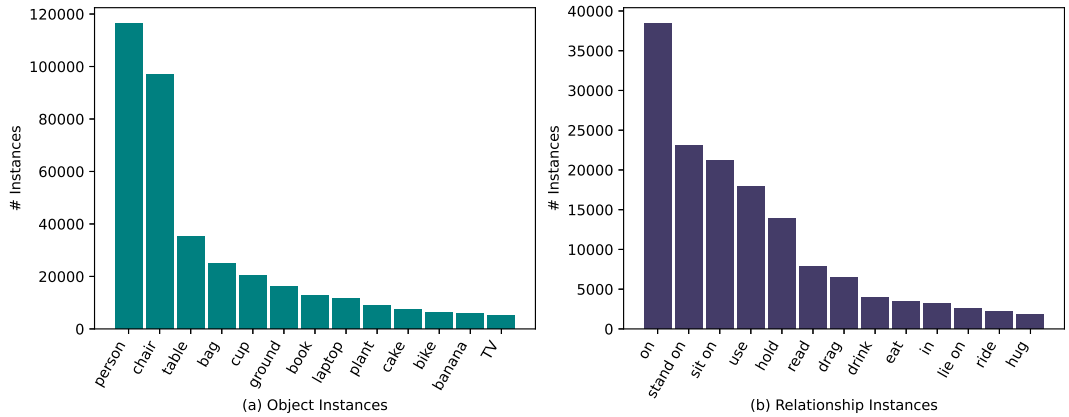


Figure 3. Number of 3D instances of different semantic categories in RealGraph dataset.

bounding boxes of the same ID in each view forms a view frustum in 3D space, and the initial 3D bounding box lies in the intersection of these frustums. Detailed illustration can be found in supplementary materials. After initialization, we manually adjust the parameters of each 3D bounding box so that its projection into all 2D views fits well with original 2D boxes.

3.3.3 Relationship annotations in 4D

After object annotation, we proceed to annotate the relationships between annotated objects from 2D images. The majority of object-object relationships defined in existing datasets [25, 2] tends to simply reflect view-dependent spatial relationships such as "in front of" and "near". In contrast, we constrain the object-object relationships in RealGraph dataset to be "supportive" (on) or "containing" (in). For example, these relationships may include $\langle box - in - bag \rangle$ or $\langle book - on - table \rangle$, where objects need to demonstrate actual physical contact with each other. Since context graph is constructed in 4D space, most view-dependent spatial relationships mentioned above can be directly inferred from relative 3D location of the objects.

The human-object[3] relationships include actions like "hold" and "drag", and activities like "drink", "read" and "ride". To minimize the ambiguity in the start and end moments of the action, the relationships are only labeled between objects that are in physical contact . We connect pairs of objects (including human) with relationships in 2D views, where relationships can be clearly observed. The annotated relationships will be displayed in all other views simultaneously based on objects' identifier across different views. Once the annotation in 2D is complete, the relationships in between 3D objects can be directly inferred based on objects' global ID.

[3]We slightly abuse the term of "object" in this paragraph to only indicate non-human objects in the dataset

4. Method

Intuitively, context graph generation from multiview videos has to tackle three major points: 1) detecting objects of various scales; 2) predicting semantic relationships from multiple views with great perspective disparity; 3) dealing with unpredictable occlusion in dynamic scenes and fuse information effectively. To tackle these challenges, we introduce Multi-view Context Graph inference Network, **MCGNet**, a multi-stage model to extract the context multiview videos in 4D space. The specific design of MCGNet is explained as follows.

Detection. Let $I_n \in \mathbb{R}^{3\times H\times W}$ be one of the images among all N multiview RGB images $\{I_n\}_{n=1}^{N}$ in timestamp τ, we omit notation of τ in this timestamp without loss of generality. We firstly extract features $F_n = F(I_n) \in \mathbb{R}^{C_1\times H\times W}$ from each view separately with a pretrained 2D backbone and a multi-scale Feature Pyramid Network (FPN), C_1 is the feature dimension. The extracted 2D features are then back-projected into the 3D space with camera calibrations, and then aggregated by an element-wise averaging to construct the 3D feature volume V, with the size of $H_V \times W_V \times D_V$. A detailed explanation of the back-projection procedure can be found in supplementary materials. Following [41, 58], we apply a Feature Fusion (FF) network to fuse 3D features in different levels. The FF network has three down-sampling residual blocks, each with three 3D convolutional layers, and three up sampling blocks, each with a transposed 3D convolutional layer followed by another 3D convolutional layer. The network outputs 3 feature maps $\mathbf{P} = [P_1, P_2, P_3]$ with the size of $\frac{H_V}{4} \times \frac{W_V}{4} \times \frac{D_V}{4}$, $\frac{H_V}{2} \times \frac{W_V}{2} \times \frac{D_V}{2}$, $H_V \times W_V \times D_V$, and of the same channel size C_2. In comparison, a baseline model without feature fusion applies a simple 3D CNN to refine 3D feature volume.

Following [41, 49], the 3D detection head takes the 3-level feature maps from the 3D feature decoder, and predicts a set of class distribution p_n, centerness c_n and 3D bounding boxes B_n. We adopt the loss function L_{det} as

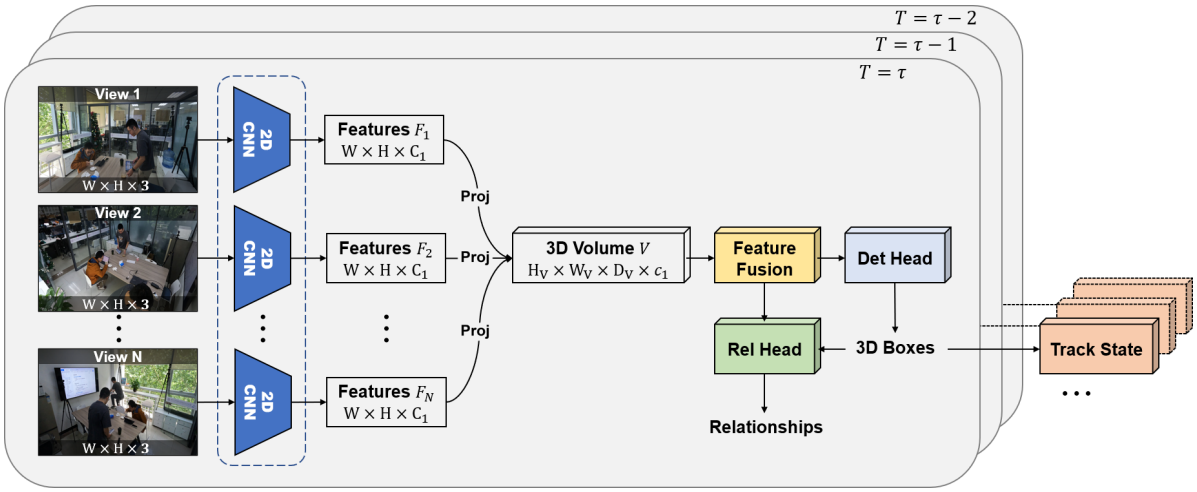


Figure 4. The network architecture of our proposed MCGNet for the CGG task. At the timestamp $T = \tau$, a 2D backbone extracts features from each image independently. The features are then back-projected into a 3D feature volume. A Feature Fusion network refines the accumulated 3D features. The detection head predicts 3D bounding boxes and their class labels, and the relationship head infers the relationships between the detected objects. The tracker uses detected 3D boxes to match the tracklets and update their states. A context graph is generated and updated frame by frame in this pipeline. Detailed demonstrations are explained in Sec.4.

in [41], L_{det} sums up a focal loss for classification L_{cls}, a cross-entropy loss for centerness L_{cntr}, and an IoU loss for 3D location L_{loc}. The detection head produces a set of 3D bounding boxes $\mathcal{B}$ with class labels $\mathcal{O}$ and scores.

Relationship prediction. In terms of the relationship inference, we follow [60] by applying a bidirectional LSTM network to refine and propagate the context feature across the object proposals from the detection head. Inspired by [18], we design a 3D ROI align based layer $\mathrm{F}_{\mathrm{extr}}$ to explicitly encode the object feature $\mathbf{E} = [\mathbf{e}_n]_{n=1}^{n_{pos}}$ and the context feature $\mathbf{C} = [\mathbf{c}_n]_{n=1}^{n_{pos}}$ as: $\mathbf{E} = \mathrm{F}_{\mathrm{extr}}(\mathbf{P}, \mathcal{B}), \mathbf{C} = \mathrm{BiLSTM}(\mathbf{E}, \mathcal{O})$, where n_{pos} is the number of positive proposals from the detection head. Then the n_{pos}^2 edge features are modeled as the combination of "head" and "tail" object context: $\mathbf{f}_{n,m} = (\mathbf{c}_n \oplus \mathbf{c}_m) \odot \hat{\mathbf{e}}_{n \cup m}$. We use the union box, i.e., the convex hull of the union of two bounding boxes, to pool their joint volume feature $\hat{\mathbf{e}}_{n \cup m}$. The $\oplus$ denotes the element-wise sum, and $\odot$ denotes the element-wise product. Finally, a softmax layer produces the probability distribution of the relationships. Unlike some previous works, we do not require bounding box overlap to recognize the relationship between objects.

Tracking. The tracking model of MCGNet mainly consists of two parts: the motion model and the association model. Following [55, 38], we adopt Kalman filter as the motion model for it fits better for high-frequency cases because of more predictable motions; and we adopt Intersection over Union (IoU) as association metric and Hungarian algorithm as matching strategy [55]. Besides, to better associate low-score detection results due to occlusion and scale variation, we introduce a double association approach [38, 64] by splitting detection results with two score thresholds thr_1 and thr_2. In the first association we follow the conventional procedure [55, 54] with boxes whose scores are higher than thr_1. Then we match the left boxes with scores higher than thr_2 to the unmatched tracklets. However, we do not use low-score boxes to update motion states. Instead, we use KF predictions as the latest tracklet states as in [38].

5. Experiments

RealGraph dataset supports a broad spectrum of CV applications, including 3D Detection (3D Det), 3D Multi-Object Tracking (3D MOT), 3D Scene Graph Generation (3D SGG) and 4D Context Graph Generation (CGG). In this section, we first provide detailed experiment setups, and then evaluate our method on RealGraph dataset. The experimental results highlight the challenges behind the proposed CGG problem. In RealGraph dataset, there are 13 scenes in total, we split the dataset in the unit of scenes: 10 scenes for training and 3 scenes for testing. For 3D MOT and CGG tasks which consider temporal consistency, we split each scene into 30-second clips, training and evaluating of the model will base those clips.

5.1. Tasks & Metrics

3D Object Detection (3D Det). We follow the existing standard metric Average Precision (AP) and mean Average Precision (mAP) for 3D detection as in [15, 16]. We adopt 3D intersection over union (IoU), requiring IoU of a detection over 25% to be true positive, redundant detection of the same object is identified as false positive as mentioned

in [12, 44]. As an extra evaluation, we examine the performance of the model when using only half of the 2D views of each scene.

3D Multi-Object Tracking (3D MOT). We evaluate standard MOT metrcis as in [30, 5, 55]: multi-object tracking accuracy (MOTA), multi-object tracking precision (MOTP) in centimeters, identity switches (IDs), False Positive (FP) and False Negative (FN).

3D Scene Graph Generation (3D SGG). Similar to previous works [33, 56], we adopt criteria recall @ K (R@K) and mean recall @ K (mR@K) on 3D SGG task. We set up three graph generation experiments as in those works:

- **Scene Graph Detection (SGDet)**. The input is a set of multiview images with camera calibration, the task is to construct 3D scene graph of the scene, i.e., to detect 3D objects and predict possible pairwise predicates at the same time.
- **Scene Graph Classification (SGCls)**. Apart from images and camera calibration, the input includes ground truth 3D bounding boxes. The task is to predict the labels of these 3D objects and the relationships between each pair of them.
- **Predicate Classification (PredCls)**. Apart from images and camera calibration, the input includes ground truth 3D bounding boxes and labels. The task is to predict relationships between each pair of these annotated 3D objects.

There are two ways to evaluate SGG tasks: with/without constraint. The constraint forbids the model to predict multiple relationships between the same object pair. Note that the true positive relationships are those whose head object, tail object are true positive in detection and relationship is correctly classified.

Context Graph Generation (CGG). In comparison to previous 3D Scene Graph Generation task, CGG focuses on daily activity in dynamic scenes and takes temporal consistency into account. Based on that point, we propose a new metric, Context Graph Recall (CGR) to evaluate the performance of models on CGG task:

$$CGR = 1 - \frac{IDs + FN}{TP + FN}$$

Compared to traditional recall, we add an extra punishment term IDs (number of identity switches occurred in all the tracklets) to assess temporal consistency of the result. The idea is borrowed from MOTA [45]. We also evaluate CGG task with/without constraint.

5.2. Implementation details

3D Det. Following previous 3D volume-based networks [36, 60], we estimate the spacial limits of all the scenes in our dataset as $(8 \times 8 \times 2.4)$ meters, and set the voxel size 0.08m. We use ResNet-50 [19] as 2D backbone. The output feature channels of each layer are set to $C_1 = 256, C_2 = 128$. We use Adam optimizer with initial learning rate = 2×10^{-4} and weight decay = 10^{-4}. We train the detection model for 14 epochs on the training set of RealGraph. The learning rate is reduced by 10 times after the 6th and 8th epoch. In the benefit of subsequent tracking and relationship prediction tasks, we apply non-maximum suppression (NMS) to the detection results with IoU threshold 25%.

3D SGG. We use the feature volume generated in detection model and ground-truth bounding boxes to train the relationship prediction model in Sec. 4. During inference, the predicted boxes from detection model are used instead. To represent the feature of the relationship between boxes B_n, B_m, we resize the edge feature $\mathbf{f}_{n,m}$ to $7 \times 7 \times 7 \times 128$. We use SGD optimizer with initial learning rate 8×10^{-4} and reduce it by 10 times after 2 epochs. We train the detection and relationship prediction models on 8 Nvidia 3090 GPUs.

3D MOT. For 3D MOT, we adopt a 3D Kalman Filter as motion model as in [55]. We choose $thr_1 = 0.25$ and $thr_2 = 0.05$ for double association. Similar to 3D Det, a minimum 3D IoU over 0.25 with the ground truth is considered as a successful match.

CGG. The proposed MCGNet for CGG task is a multi-stage model composed of detection, relationship prediction and tracking models mentioned above. During inference, we run the whole pipeline on input multiview videos and eventually output the context graph of the scene.

5.3. Results

In this section, we evaluate the results of aforementioned tasks of MCGNet on RealGraph dataset.

FF	#views	chair	table	person	laptop	cup	box	whiteboard	mAP
	half	43.96	12.27	23.36	58.93	0.85	12.32	12.07	18.70
✓	half	53.38	20.32	38.57	64.49	2.10	16.28	19.52	23.33
	full	71.52	32.77	61.19	**80.98**	5.04	19.73	38.26	34.96
✓	full	**75.14**	**33.70**	**68.44**	80.29	**7.16**	**22.49**	**43.29**	**38.29**

Table 2. Performance of MCGNet on 3D detection on RealGraph. AP@0.25 on typical object classes is also shown. "FF" denotes feature fusion, "#views" denotes views used as input during training and inference.

FF	DA	MOTA↑	MOTP↓	IDs↓	FP↓	FN↓
		35.16	18.12	3379	8603	13246
✓		37.06	16.60	2658	6859	9933
	✓	35.20	18.08	1921	8096	11853
✓	✓	**37.17**	**16.54**	**1658**	**6269**	**8773**

Table 3. Performance of MCGNet on 3D MOT on RealGraph dataset. The tracker is based on AB3DMOT [55]. MOTP is in centimeters. "FF" denotes feature fusion module in the detector, "DA" denotes double association in the tracker.

Task	FF	With Constraint						No Constraint					
		R@20	R@50	R@100	mR@20	mR@50	mR@100	R@20	R@50	R@100	mR@20	mR@50	mR@100
SGDet		27.4	32.4	33.2	17.8	20.3	22.1	30.6	32.8	33.5	18.0	21.6	22.9
	✓	**28.8**	**35.4**	**36.3**	**19.0**	**22.1**	**22.9**	**32.4**	**33.6**	**34.2**	**21.7**	**23.5**	**25.7**
SGCls		36.1	38.9	40.4	22.3	25.6	28.6	38.0	40.1	**41.4**	25.7	31.4	33.2
	✓	**36.5**	**39.4**	**40.5**	**22.9**	**26.0**	**28.8**	**38.6**	**40.5**	**41.4**	**28.2**	**31.7**	**33.3**
PredCls		-	-	-	-	-	-	-	-	-	-	-	-
	✓	58.3	64.5	66.3	31.3	35.2	37.9	66.7	69.4	71.6	39.2	41.5	43.0

Table 4. Performance of MCGNet on RealGraph dataset in 3D SGG tasks. 3D SGG includes three sub-tasks: scene graph detection (SGDet), scene graph classification (SGCls) and predicate classification (PredCls). **With Constraint** forbids the model to output multiple results of the same object pair which could promote the recall. "FF" makes little difference in PredCls sub-task since groud-truth boxes are given, so we omit the comparison.

FF	DA	With Constraint						No Constraint					
		CGR@20	CGR@50	CGR@100	mCGR@20	mCGR@50	mCGR@100	CGR@20	CGR@50	CGR@100	mCGR@20	mCGR@50	mCGR@100
		25.9	30.9	31.7	16.3	18.8	20.6	29.1	31.3	32.0	16.5	20.1	21.4
✓		27.7	34.3	35.2	17.9	21.0	21.8	31.3	32.5	33.1	20.6	22.4	24.6
	✓	26.6	31.6	32.4	17.0	19.5	21.3	29.8	32.0	32.7	17.2	20.8	22.1
✓	✓	**28.2**	**34.8**	**35.7**	**18.4**	**21.5**	**22.3**	**31.8**	**33.0**	**33.6**	**21.1**	**22.9**	**25.0**

Table 5. Performance of MCGNet on CGG on RealGraph dataset. CGR@K and mCGR@K are new evaluation metrics proposed in this work and detailed in Sec. 5.1. "FF" denotes feature fusion module in the detector, "DA" denotes double association approach in the tracker.

3D Det. We compare the results of the 3D volume based MCGNet with/without feature fusion module on full/half view settings in Table 2. Apparently, MCGNet with feature fusion outperforms the vanilla version in most object categories, especially small-scale ones like "cup" and "box", showing improvement on AP by 2.12 and 2.76 respectively, and in general, FF model achieves higher mAP by 3.3. The primary reason is the random occlusion and largely diverse object scale that makes detection of small objects difficult. Another interesting observation is, the performance of the same model trained and tested with only half views drops significantly as shown in Table 2. This reveals that sufficiency and proper arrangement of cameras are necessary to provide adequate information to construct good context graphs.

3D MOT. As in Table 3, the model with feature fusion (FF) and double association (DA) outperforms other versions. Similar to 3D SGG, models with FF produces better detection results hence achieves higher performance in tracking task. Comparing two models with FF with/without double association, we see DA contributes to higher MOTA by 0.11 and lower MOTP by 0.06. As analyzed in Sec. 4, it's mainly due to the low-score detections could carry extra information of object motion.

3D SGG. Test results on 3D SGG task show similar trend to 3D Det, where the method with feature fusion module undoubtedly make better detection results with higher recall on SGG. Specifically, MCGNet with feature fusion outperforms the vanilla model on SGDet sub-task with constraint by 1.2, 1.8 and 0.8 of mR@20, mR@50 and mR@100 respectively as in Table 4.

CGG The overall performance shown in Table 5 indicates that the proposed 3D feature volume based MCGNet effectively learns the complex spatial and semantic information in a dynamic scene. Different from 3D SGG task, CGG task takes temporal perception into account. Therefore we can see a boost in CGR (0.5 on average) with double association, which lowers the ID switches just like in 3D MOT task. However, the proposed MCGNet is a multi-stage pipeline, and suffers from accumulated error through each stage. We encourage the community to further present joint optimization schemes for more robust, efficient and effective models.

6. Conclusion

The capability of understanding the semantics in the real world is critical for AI systems. In this paper, we rethink the way of 3D scene graph generation, and propose a new paradigm called 4D context graph generation (CGG) from multiview images to better parse dynamic scenes of human activities. We correspondingly propose a multiview video dataset "RealGraph" tailored for the proposed CGG task. The RealGraph dataset provides multiple synchronized videos for daily scenes along with various semantic annotations. This paper reveals the great challenges behind CGG task and we explore a feasible baseline for the task. Empirical results demonstrate that the proposed baseline partially addresses the CGG problem, yet we encourage future study to propose more effective and efficient solutions, e.g., besides tracking objects, the relationships can also be constrained with temporal consistency. We believe CGG could also benefit downstream applications like visual grounding and robotic intelligence, and will be of interests of the community.

7. Acknowledgement

This work is supported in part by Natural Science Foundation of China (NSFC) under contract No. 62125106, 61860206003 and 62088102, in part by Ministry of Science and Technology of China under contract No. 2021ZD0109901.

References

[1] Abdulaziz Alajaji, Walter Gerych, Kavin Chandrasekaran, Luke Buquicchio, Emmanuel O. Agu, and Elke A. Rundensteiner. Deepcontext: Parameterized compatibility-based attention cnn for human context recognition. *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 53–60, 2020. 3

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5664–5673, 2019. 2, 3, 5

[3] Iro Armeni, Sasha Sax, Amir Roshan Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *ArXiv*, abs/1702.01105, 2017. 3

[4] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. 1, 2, 3, 7

[6] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *2017 International Conference on 3D Vision (3DV)*, pages 667–676, 2017. 3

[7] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8647–8656, 2019. 3

[8] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19:934–948, 2018. 3

[9] Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19608–19617, 2022. 2

[10] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16372–16382, 2021. 3

[11] James M. Coughlan and Alan Loddon Yuille. Manhattan world: compass direction from a single image by bayesian inference. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2:941–947 vol.2, 1999. 3

[12] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017. 1, 3, 7

[13] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 616–624, 2016. 3

[14] Yilun Du, Zhijian Liu, Hector Basevi, Ale*vs*. Leonardis, Bill Freeman, Joshua B. Tenenbaum, and Jiajun Wu. Learning to exploit stability for 3d scene parsing. In *NeurIPS*, 2018. 3

[15] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009. 3, 6

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32:1231 – 1237, 2013. 2, 3, 6

[17] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):1–12, 2021. 1

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[20] Varsha Hedau, Derek Hoiem, and David A. Forsyth. Recovering the spatial layout of cluttered rooms. *2009 IEEE 12th International Conference on Computer Vision*, pages 1849–1856, 2009. 3

[21] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In-So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015. 3

[22] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *European Conference on Computer Vision*, pages 767–786. Springer, 2020. 2, 3

[23] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015. 3

[24] Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S Bernstein. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39):e2115730119, 2022. 1

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 1, 3, 5

[26] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 3

[27] David C. Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2136–2143, 2009. 3

[28] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13874–13883, 2022. 3

[29] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10166–10175, 2020. 2, 3

[30] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022. 1, 3, 7

[31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014. 3

[32] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2, 3

[33] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 7

[34] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*, 2019. 3

[35] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2697–2706, 2017. 3

[36] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *ECCV*, 2020. 7

[37] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 52–61, 2020. 3

[38] Ziqi Pang, Zhichao Li, and Naiyan Wang. Simpletrack: Understanding and rethinking 3d multi-object tracking. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 680–696. Springer, 2023. 6

[39] 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. *2019 International Conference on Robotics and Automation (ICRA)*, pages 9552–9557, 2019. 3

[40] Srikumar Ramalingam, Jaishanker K. Pillai, Arpit Jain, and Yuichi Taguchi. Manhattan junction catalogue for spatial reasoning of indoor scenes. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3065–3072, 2013. 3

[41] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022. 5, 6

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 3

[43] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019. 3

[44] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 1, 3, 7

[45] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. The clear 2006 evaluation. In *Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006, Southampton, UK, April 6-7, 2006, Revised Selected Papers 1*, pages 1–44. Springer, 2007. 7

[46] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1047–1056, 2019. 3

[47] 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 M. 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. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451, 2020. 1, 2, 3

[48] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13688–13697, 2021. 3

[49] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 5

[50] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *CoRL*, 2018. 3

[51] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7657–7666, 2019. 3

[52] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020. 2, 3

[53] Xueyan Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David J. Brady, Qionghai Dai, and Lu Fang. Panda: A gigapixel-level human-centric video dataset. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3265–3275, 2020. 3

[54] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366. IEEE, 2020. 6

[55] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics. *arXiv preprint arXiv:2008.08063*, 2020. 6, 7

[56] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 7

[57] Li Xu, Haoxuan Qu, Jason Kuen, Jiuxiang Gu, and Jun Liu. Meta spatio-temporal debiasing for video scene graph generation. In *European Conference on Computer Vision*, pages 374–390. Springer, 2022. 3

[58] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*, 2018. 4, 5

[59] Jeffrey M Zacks, Barbara Tversky, and Gowri Iyer. Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General*, 130(1):29, 2001. 1

[60] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Conference on Computer Vision and Pattern Recognition*, 2018. 3, 6, 7

[61] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12612–12621, 2021. 3

[62] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4457–4465, 2017. 3

[63] Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, S. Li, and Guodong Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE Transactions on Multimedia*, 22:380–393, 2020. 3

[64] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 1–21. Springer, 2022. 6

[65] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *ECCV*, 2020. 3

[66] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7384–7392, 2019. 2

[67] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2051–2059, 2018. 3