IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

3D-Aware Scene Change Captioning from Multiview Images

Yue Qiu[1,2], Yutaka Satoh[1,2], Ryota Suzuki[1], Kenji Iwata[1] and Hirokatsu Kataoka[1]

***Abstract*— In this paper, we propose a framework that recognizes and describes changes that occur in a scene observed from multiple viewpoints in natural language text. The ability to recognize and describe changes that occurred in a 3D scene plays an essential role in a variety of human-robot interaction applications. However, most current 3D vision studies have focused on understanding the static 3D scene. Existing scene change captioning approaches recognize and generate change captions from single-view images. Those methods have limited ability to deal with camera movement, object occlusion, which are common in real-world settings. To resolve these problems, we propose a framework that observes every scene from multiple viewpoints and describes the scene change based on an understanding of the underlying 3D structure of scenes. We build three synthetic datasets consisting of primitive 3D object and scanned real object models for evaluation. The results indicate that our method outperforms the previous state-of-the-art 2D-based method by a large margin in terms of sentence generation and change understanding correctness. In addition, our method is more robust to camera movements compared to the previous method and also performs better for scenes with occlusions. Moreover, our method also shows encouraging results in a realistic scene-setting, which indicates the possibility of adapting our framework to a more complicated and extensive scene-settings.**

Fig. 1. Illustration of 3D scene change captioning. The proposed approach observes a scene from different viewpoints before and after it changes and describes the change with natural language.

I. INTRODUCTION

The ability to understand changes occurring in 3D scenes is crucial for a variety of human-robot interaction (HRI) applications. For example, the arrangement and locations of objects usually change in a home, and home robots need the ability to understand the dynamics in home scenes to avoid the necessity of re-scanning and re-recognizing its 3D surroundings frequently. For video surveillance systems, the ability to automate the recognition and description of scene changes can help reduce the labor costs of manually checking every frame of a video.

The development in deep neural networks (DNNs) and graphic processing units, along with 3D sensing technologies, have brought massive success to 3D recognition related research. Many new tasks related to 3D recognition have emerged, such as 3D object detection [1], 3D semantic segmentation [2], and shape completion [3]. In contrast, despite its significance in robotics applications, recognizing 3D scenes with dynamics and changes remains less studied.

Recently, several scene change captioning approaches have emerged, which aim to understand scene changes and describe changes with natural language text. Scene change captioning is highly practical in HRI applications as it transfers scene understanding to natural language. Current scene change captioning methods [4], [5] are mostly based on 2D scene images. These works predict a natural language description (caption) to describe the change that occurred between two 2D images of a scene. These methods report results on 2D scene image datasets with slight camera jittering. However, those 2D-based methods have limited abilities to handle scenes with heavy occlusions from single-view scene observation. In addition, those methods do not establish 3D-based scene recognition that reasons about the underlying 3D structure of scenes, which makes them not practical for huge camera movements. Furthermore, these situations are common in real-world applications. For example, it is difficult for robots to always photograph a scene from the same viewpoints.

To solve these problems, as shown in Figure 1, we propose a framework that establishes a scene understanding containing underlying 3D information of the scenes and describes scene changes based on the comprehensive information of before- and after-change scenes. Based on the above, after observing the original scene, our framework can identify and describe scene changes by observing the changing scene from different viewpoints. In addition, based

[1]Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata and Hirokatsu Kataoka are with National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8560 Japan {yu,ryota,kenji,hirokatsu}.{satou, suzuki, iwata, kataoka}@aist.go.jp

[2]Yue Qiu and Yutaka Satoh are with the Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577 Japan s1830151@s.tsukuba.ac.jp

on a continuous scene representation, our framework can achieve scene change captioning directly from input multiview images without any registration and tracking stages. We build three synthetic scene change captioning datasets consisting of simple geometric primitives and photorealistic daily supply object models. Experiments on these datasets showed that our framework outperformed the previous state-of-the-art 2D-based method by a large margin in terms of change captioning accuracy and sentence construction. In addition, our framework showed higher robustness for camera movements and scenes with occluded objects. Moreover, our framework also performs better on the dataset with photorealistic objects, which provides the possibility for our method to be adapted to real-world applications.

II. Related Work

A. 3D Scene Understanding

3D scene understanding plays an essential role in various robot applications. In [6], the authors proposed a method that detects objects from multiview images and applies a tree-structured inference strategy to determine the support relationships in object clutters. The authors of [7] proposed a method to choose observation viewpoints from multiple viewpoints, which is intentioned for support relationships understanding of scenes with occlusions. These non-neural methods usually require less runtime and memory costs in implementation. But these methods often have limited ability for large scale and complex scenes.

Neural 3D scene understanding methods use CNNs for learning latent representations. Among those methods, there are 3DCNN-based methods that utilize 3DCNN structure to learn representations for various 3D scene understanding tasks such as 3D object detection [1], 3D semantic segmentation [2], shape completion [3]. Due to the limitations in memory and computing power, the resolution of input data of 3DCNNs is restricted and 3DCNNs often requires a massive amount of training data, high memory, and execution time costs.

In contrast, recently, a series of 2DCNN-based neural 3D scene representation methods have been proposed [8], [9]. Those methods receive multiview 2D images of 3D scenes as input and learn representations containing the underlying 3D information. The generative query network (GQN) [8] is a conditional variational autoencoder (CVAE) [10]-based network that learns meaningful 3D representation from multiview images of scenes. Scene representation networks (SRNs) [9] predict 3D range maps along with scene representation of the input scene, which makes their framework robust to unknown camera poses.

We applied 2DCNN-based scene representation method GQN [8] to our framework because the model encodes semantic scene information into latent vectors, and has lower memory- and runtime- costs compared with 3DCNN-based methods. Compared with 2DCNN-based methods PointNet [11], PointNet++ [12] (handles coordinates only), GQN handles 2D images containing both geometry and color information. Moreover, the GQN network requires less labeled training data, reducing annotation costs. The GQN also has limitations. For example, it handles fixed camera positions. We believe that recent works such as SRNs, which is more robust to unknown camera positions, could enhance the model to unknown camera positions.

B. Active Perception

Active perception [13] is widely used in robot applications to select behaviors for increasing information from a series of data. In [7], the authors proposed a method to select viewpoints in a scene for lowering occlusion. The authors of [14] introduced the active perception to help attain autonomous robotics in unstructured environments by providing robust perception. In embodied question answering task [15], the active perception is introduced for choosing navigation routes for answering scene-related questions. In our work, we use a fixed number of observation viewpoints of scenes. It is an interesting topic to integrate active perception in the context of scene change captioning.

C. Change Detection

Change detection for street view images [16], [17] has been widely studied. The authors of [16] proposed a panoramic change detection dataset along with a method that integrates CNN features with superpixel segmentation for change detection. In [17], the authors proposed a dense optical flow-based network for change detection from multiview images.

[18], [19], [20] have discussed change detection in robotics. [18] proposed a method, which aligns two 3D maps and computes the movement of surfaces to determine the changed part. [19] proposed a method for recording dynamic scenes (point clouds) by detecting and separating dynamic and static elements in a scene. In [20], the authors proposed a method to speed up the detection of novel objects by calculating the regions of interest (changes) in a scene. These previous works handle point cloud data while our work directly deals with multiview images. In addition, our framework is end-to-end with learnable parameters, while previous works often require multiple processes, manually designed functions, and parameters. Additionally, our model describes detailed scene changes through language texts. Moreover, our model can recognize scene change based on partially observed scenes, while previous methods require full observation of scenes.

D. Image Captioning

Traditional image captioning methods [21], [22] generate a text description given an image. [21] proposed a model with spatial attention, which generates captions about highlighted image regions. In [22], the authors combined the bottom-up and top-down attention to better determine the highlighted region for caption generation.

DDLA [5] computes the pixel-level RGB difference between before- and after-change images for change captioning. DUDA [4] proposes an extracted image features-based approach. The authors of DUDA introduced a dual-attention

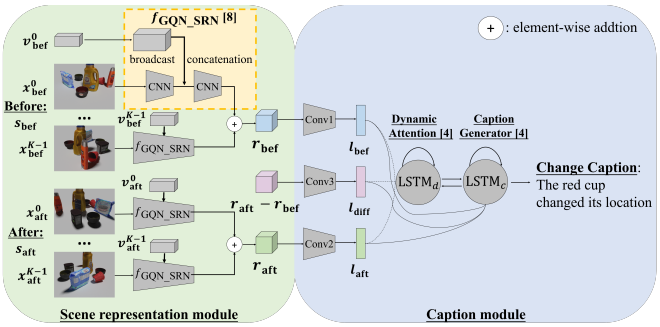

Fig. 2. Overall framework. Given an input of a scene before and after changes observed from multiple virtual cameras , our approach produces a change caption with models consisting of a scene representation module and a captioning module. The input images are observed from viewpoints (numbered 1 to K) sampled from the pre-defined viewpoint set.

mechanism to highlight the important region of scene images and a dynamic attention mechanism for weighting the image features during the caption generation process. DUDA achieves state-of-the-art performance among 2D-based scene change captioning methods. However, both DDLA and DUDA do not recognize the underlying 3D structure of the scene; thus, these works have only limited ability for camera movements between scene changes and scenes with occlusions.

In contrast to DDLA and DUDA, our work is based on a scene representation containing the underlying 3D structure and semantic information. Therefore, our work establishes 3D correlations between before- and after-change scenes and could handle scene observations from different viewpoints, and scenes with occlusions. These characteristics make our method more practicable for real-world applications with constant object occlusions and changing observation viewpoints.

III. Approach

We show the overall framework in Figure 2. We denote the scene observed before scene changes as s_{bef}, and the scene observed after scene changes as s_{aft}. Our overall framework generates a text caption indicating the scene changes based on images of s_{bef} and s_{aft} from multiple viewpoints. Our framework consists of two modules: the scene representation module, which extracts scene representation from the input scenes s_{bef} and s_{aft}; the caption module, which predicts a change caption based on the extracted scene representation features.

A. Scene Representation Module

We use the scene representation network proposed by Eslami et al. [8] to obtain an integrated scene representation from multiview images. For scene s_j observed from K viewpoints ($K \geq 1$), the observation o_j is defined via (1). x_j^k indicates image observed from viewpoint v_j^k (five-dimensional vector (**w**, **y**, **p**), where **w**, **y**, **p** indicate the three-dimensional position, yaw, and pitch of the camera):

$$o_j = \{(x_j^k, v_j^k)\}_{k=0,\dots,K-1} \tag{1}$$

The scene representation network $r_j = f_{\text{GQN\_SRN}}(o_j)$ and renderer network $g_{\text{GQN\_RN}}(x^m|v^m, r_j)$ are jointly trained for image rendering from arbitrary viewpoint m to maximize the likelihood between the predicted image x^m and the ground truth image. $f_{\text{GQN\_SRN}}$ integrates multiview information into a compact scene representation. We train the overall framework. The renderer network $g_{\text{GQN\_RN}}$ is discarded after training. We use the latent scene features extracted via pre-trained $f_{\text{GQN\_SRN}}$ to represent scenes.

B. Caption Module

The caption module predicts a text change caption given input scene representations r_{bef} and r_{aft}. Inspired by [4], we first compute the difference between r_{bef} and r_{aft} via:

$$r_{\text{diff}} = r_{\text{aft}} - r_{\text{bef}} \tag{2}$$

In [4], the authors compute the difference between image features using equation (2). We compute the scene representation difference instead.

Next, as shown in (3), we modify the original dynamic attention structure proposed in [4] by applying convolution operations to r_{bef}, r_{aft}, and r_{diff} respectively.

$$\begin{cases} l_{\text{bef}} = Conv1(r_{\text{bef}}) \\ l_{\text{aft}} = Conv2(r_{\text{aft}}) \\ l_{\text{diff}} = Conv3(r_{\text{diff}}) \end{cases} \tag{3}$$

The remaining processes are adapted from [4]. We use long short-term memory (LSTM)-structured networks LSTM$_d$ (dynamic attention) and LSTM$_c$ (caption generator) proposed in [4] to generate captions from l_{bef}, l_{aft}, l_{diff}. More details of LSTM$_d$ and LSTM$_c$ can be found in [4].

IV. Experiments

In this section, we describe the evaluation of our model against the previous state-of-the-art 2D-based method [4] through several experimental setups. In detail, we first conducted an experiment to test the effectiveness of models for situations with different observation viewpoints of before- and after-change scenes. We also conducted ablation experiments on different sub-structures, including image feature attention, ensembles of image features, attention mechanism for decoder steps. Next, we conducted an experiment on the robustness to camera movements brought by rotating cameras. Finally, we tested the validity under a dataset setting with photorealistic objects.

A. Experimental Setups

We built three datasets for the experiments mentioned above. The statistics of these three datasets are shown in Table I. Each dataset includes scene pairs (before and after changes) observed from multiple viewpoints and related change captions. The scene generation process is based on CLEVR [23].

P_change_occlusion dataset. This dataset is composed of scenes with geometric **P**rimitives. We observe each scene from four virtual cameras (Figure 3 left). Each scene consists

TABLE I

DATASET STATISTICS OF P_CHANGE_OCCLUSION, P_CHANGE_CAMERAS (GEOMETRIC PRIMITIVES), AND ROM_CHANGE (REALISTIC OBJECT MODELS).

Dataset	Scenes (train/test)	Captions (train/test)	Change types	Viewpoints	Object classes	Unique objects	Objects / scene
P_change_occlusion	(24,000 / 6,000)	(720,000 / 180,000)	5	4	-	30	2-6
P_change_cameras	(24,000 / 6,000)	(720,000 / 180,000)	5	7	-	30	2-7
ROM_change	(24,000 / 6,000)	(720,000 / 180,000)	5	4	13	72	3-8

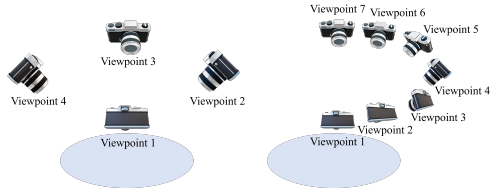

Fig. 3. Virtual camera setup for P_change_occlusion, ROM_change (left), and P_change_cameras (right).

TABLE II

OBJECT MODELS OF ROM_CHANGE DATASET.

Classes	Object instances	Source dataset
snack	12	NEDO:2; YCB:5; Bigbird:5
cup	11	NEDO:6; YCB:4; Bigbird:1
shampoo	10	NEDO:1; Bigbird:9
dishwasher	9	NEDO:8; Bigbird:1
minicar	7	NEDO:7
bar soap	6	NEDO:3; Bigbird:3
bowl	4	NEDO:3; YCB:1
sponge	4	NEDO:3; YCB:1
soft drink	3	Bigbird:3
weight	2	NEDO:2
water bottle	2	NEDO:2
soccerball	1	YCB:1
teddy bear	1	NEDO:1

of geometric primitives numbered between 2 and 6 with random colors (five colors), shapes (three shapes), and sizes (two sizes). To evaluate occluded scene understanding ability and make the scene change understanding infeasible to be resolved through information from the same viewpoint, we added an occlusion setting for before-change scenes, which ensures that there are occluded objects (less than 100 pixels, original image size: 320×240 pixels) from at least two viewpoints. This dataset is used to evaluate the effectiveness of models to integrate multiview information for change captioning, and the robustness for occluded scenes.

P_change_cameras dataset. We created this dataset to evaluate the models' robustness to camera movements which are achieved by rotating cameras between before- and after-change scenes. This dataset is constructed from geometric primitives with virtual cameras setup shown in Figure 3 right. To exclude the effect of occlusion, we added a restriction that each object can be observed from all viewpoints.

ROM_change dataset. This dataset is composed of scenes with **R**ealistic **O**bject **M**odels. To evaluate the validity under more realistic dataset settings, we built ROM_change with photorealistic object models collecting from three previously published object model datasets: NEDO item [24], YCB [25], and Bigbird dataset [26]. We annotated each model with color (14 colors) and class label (13 classes). We show the object classes, number of object instances per class, and source datasets in Table II. The diversity in object shapes and sizes makes it more difficult to include two occluded views comparing with the P_change_occlusion dataset. Therefore, we applied a restriction that at least one observation of each scene contains one or more occluded objects.

Captions generation. To generate before- and after-change scene pairs, we rendered twice for each scene and added a change operation in between. We defined five change types: add - adding an object to the original scene; delete - removing an object; move - moving an object; swap - swapping the locations of two existing objects; and replace - replacing an object with a new object. We also added a distractor type that does not contain any changes. We recorded scene information through scene graphs [27], which records object attributes and spatial relationships between objects. The ground truth change captions were automatically generated from recorded scene graphs and predefined templates that record the sentence structures. We defined five sentence templates for each change type. Through the above processes, each scene is paired with thirty captions (five captions per change type (or the distractor)).

Although we currently generate change captions directly from scene observations, the framework also could be expanded to simultaneously generate detailed change information (e.g., change type) by integrating a multi-task structure or a two-stage structure which first generates change information and creates captions using pre-defined templates.

Implementation details. In all experiments, we pre-trained the scene representation network for 200 epochs. For both our and the previous method DUDA [4], we trained the caption models for 40 epochs. We set the learning rate to 10^{-4} and used Adam optimizer for all experiments.

The DUDA model generates captions from two images (before- and after-change images). The model first extracts image features using the ResNet model [28], which is trained on the ImageNet dataset. We input before- and after-change image features into the DUDA model for caption generation during training and evaluation process.

B. Experiments on 3D Scenes with Geometric Primitives

In this section, we conducted experiments on the P_change_occlusion dataset to evaluate the validity of models to integrate multiview information for change captioning.

TABLE III

EVALUATION ON P_CHANGE_OCCLUSION (TOP TEN ROWS) AND ROM_CHANGE (BOTTOM SIX ROWS).

Approaches, (viewpoint pair)	ROUGE_L [29]	SPICE [30]	METEOR [32]	BLEU-4 [31]						
				overall	add	delete	move	swap	replace	distractor
no-diff/-/spa-att, (1,2-3,4)	73.4	40.5	45.3	58.8	66.9	84.5	64.5	48.6	79.8	11.7
no-diff/-/no-att, (1,2-3,4)	76.3	41.0	46.9	63.2	67.4	84.6	68.2	50.6	82.3	24.5
no-diff/conv/no-att, (1,2-3,4)	87.8	47.8	54.1	77.9	79.0	90.2	85.8	57.4	81.1	78.6
diff/conv/no-att, (1,2-3,4)	87.9	49.3	54.1	78.5	80.5	88.9	83.9	55.9	83.1	81.5
diff/conv/dyn-att, (1,2-3,4)	88.4	49.4	54.7	79.4	81.2	89.7	85.2	56.3	83.1	84.4
diff/conv/dyn-att, (1,2,3,4-1,2,3,4)	**96.4**	**56.9**	**64.3**	**92.3**	**99.8**	**99.8**	**97.6**	**70.3**	**97.1**	**100.0**
diff/conv/dyn-att, (1-3)	79.4	42.6	48.1	67.9	64.0	75.9	80.8	53.4	71.2	60.1
diff/conv/dyn-att, (1-1)	84.6	45.9	51.5	74.2	71.7	74.5	75.8	50.8	72.8	99.7
DUDA, (1-3)	66.6	34.3	37.6	50.1	50.7	58.1	52.0	36.1	61.2	33.4
DUDA, (1-1)	80.7	43.6	48.7	68.6	63.8	71.1	62.2	47.5	67.5	99.4
Ours, (1-1)	73.2	33.1	40.8	58.6	51.0	57.0	58.2	28.1	52.4	**100.0**
Ours, (1-3)	64.0	25.7	33.7	47.6	36.9	54.7	54.3	25.5	48.9	56.3
Ours, (1,2-3,4)	73.4	32.8	40.0	58.1	54.5	68.2	59.4	30.1	63.3	65.6
Ours, (1,2,3,4-1,2,3,4)	**94.6**	**46.9**	**59.5**	**87.1**	**96.1**	**97.3**	**89.4**	**50.2**	**91.9**	**100.0**
DUDA, (1-1)	66.7	27.9	35.4	51.9	47.0	52.2	39.8	24.1	47.8	**100.0**
DUDA, (1-3)	46.3	14.6	21.9	26.1	22.4	45.3	32.2	12.4	27.9	10.4

TABLE IV

CHANGE CAPTION CORRECTNESS EVALUATION ON P_CHANGE_OCCLUSION DATASET.

Approaches	Accuracy				
	change type	Object	Color	Shape	Size
Ours (1-1)	87.1	74.6	89.4	84.9	91.7
Ours (1-3)	77.8	58.6	78.0	69.5	85.4
Ours (1,2-3,4)	90.4	70.1	87.4	78.8	92.3
Ours (1,2,3,4-1,2,3,4)	**99.4**	**91.5**	**98.0**	**96.6**	**99.2**
DUDA (1-1)	76.9	59.8	78.8	74.9	84.0
DUDA (1-3)	60.1	43.4	60.0	59.0	73.8

We also performed experiments on sub-structure ablations, viewpoint settings of before- and after-change scenes.

Refer to the virtual camera setting in Figure 3 left, we first created scene pairs by observing the before-change scene s_{bef} from viewpoints 1, 2, and after-change scene s_{aft} from viewpoint 3, 4. This setting requires models to integrate information from multiview images for scene understanding, and associate before- and after-change scenes observed from different viewpoints. We tested several structural ablations. We represent different ablations as $E/F/A$, where E denotes the type of ensembled features, F represents the feature extraction structure, and A represents the decoder attention mechanism. We introduce the details of each model below.

no-diff/-/spa-att: This structure is based on [21]. This structure takes the concatenation of scene representation r_{bef} and r_{aft} as input. At each decoding step, we apply spatial attention to concatenated scene representation. The spatial attention mechanism is adopted by most current image captioning methods. It is firstly introduced in the context of image captioning in [21].

no-diff/-/no-att: This structure is created by removing the spatial attention structure of no-diff/-/spa-att.

no-diff/conv/no-att: Compared with no-diff/-/no-att, this structure applies separated convolution operations for r_{bef} and r_{aft} before the concatenation operation.

diff/conv/no-att: Compared with no-diff/conv/no-att, this structure takes r_{bef}, r_{aft}, and the difference between these two representations, $r_{\mathrm{aft}} - r_{\mathrm{bef}}$ as input and applies separated convolution operations on the three representations.

diff/conv/dyn-att: Compared with diff/conv/no-att, this structure further uses a dynamic attention operation [4] on scene representations processed by separated convolution layers during decoding steps.

Quantitative Results. We show the evaluation results for different evaluation metrics in Table III (top ten rows). These metrics evaluate the similarity between generated captions with ground truth caption sets. In detail, ROUGE_L [29] measures the recall rate of ground truth caption sets in generated captions. SPICE [30] extracts semantic structures, such as objects, attributes, relationships from captions, and evaluates the similarity between generated and ground truth captions based on semantic structures. BLEU-k [31] evaluates the precision of consequent words (length 1 to k) in generated captions compared with ground truth captions. METEOR [32] further introduces the semantic similarity of words in addition to evaluation process of BLEU. We also report BLEU-4 scores for different change types.

We first analyze the effects of different ablations. The results on first, second and third row show that spatial attention dramatically degraded the performance since the no-diff/-/spa-att structure obtained the worst result and the performance elevated by removing the spatial attention operation, which provides a perspective that the scene representation feature might not be suitable for directly adopting spatial attention operations. By comparing the second and third row, we found that the use of the convolution layers improved the performance in a large margin. From the results of the third and fourth row, we found that concatenating the difference feature $r_{\mathrm{aft}} - r_{\mathrm{bef}}$ helped improve the performance. From the comparison of the fourth and fifth row, we noticed that the dynamic attention mechanism also helped improve the performance slightly. Based on the above, we used the diff/conv/dyn-att structure for all the following experiments.

We also implemented the diff/conv/dyn-att model with

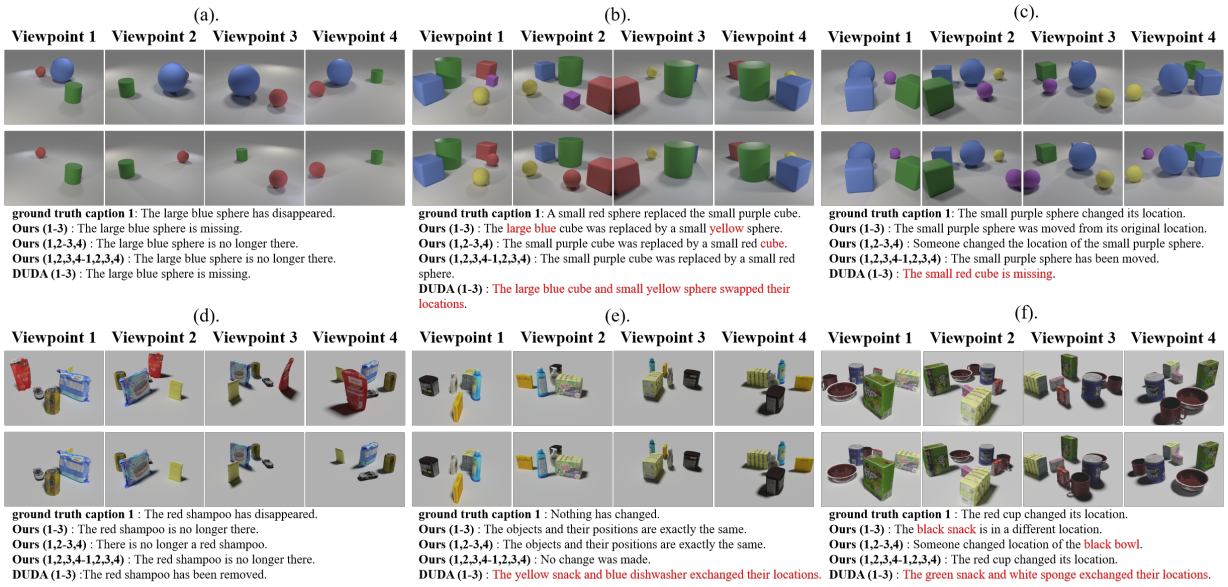

Fig. 4. Example results on P_change_occlusion ((a-c)) and ROM_change ((d-f)). The images in the first line show scene observed before changes and the second line for after changes. The incorrect caption predictions are shown in red.

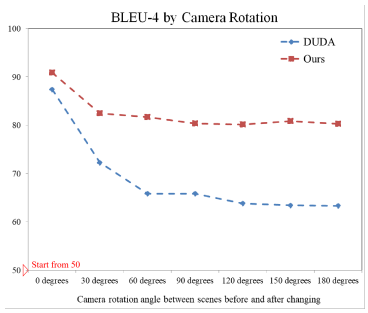

Fig. 5. BLEU-4 evaluation results on P_change_cameras.

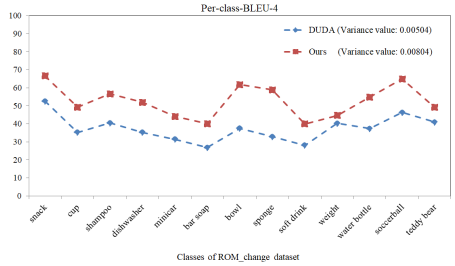

Fig. 6. Per-class-BLEU-4 evaluation results on ROM_change.

viewpoint pair setting (1,2,3,4-1,2,3,4) (before-after) for before- and after-change scenes. For comparison, we implemented viewpoint pair settings (1-3) and (1-1) for our and the state-of-the-art 2D-based model DUDA [4], which handles single view images for before- and after-change scenes.

Our model obtained higher performance when more viewpoints were used except the (1-1) viewpoint setting. Our models outperformed the DUDA under the viewpoint settings (1-1) and (1-3) for all evaluation metrics on all change types.

These results show that, compared with the previously reported 2D-based method DUDA, our model can effectively integrate scene information observed from different viewpoints and correlate before- and after-change scene information. The significant performance gap between our and the 2D-based method indicates that the understanding of underlying 3D structure of scenes is essential for the 3D scene change captioning task.

For both our and the DUDA model, the performance of viewpoint setting (1-3) degraded compared with (1-1), although the former setting brings more information. Comparing with the viewpoint setting (1-3), which requires alignment of multiview information, viewpoint setting (1-1) makes it less challenging to point out the scene difference.

TABLE V

CHANGE CAPTION CORRECTNESS EVALUATION ON ROM_CHANGE DATASET.

Approaches	Accuracy			
	change type	Object	Color	Class
Ours (1-1)	74.1	42.6	56.4	54.3
Ours (1-3)	63.2	31.7	45.5	43.0
Ours (1,2-3,4)	75.4	40.1	53.0	50.9
Ours (1,2,3,4-1,2,3,4)	**98.2**	**81.6**	**88.1**	**86.9**
DUDA (1-1)	64.2	37.8	50.9	46.8
DUDA (1-3)	37.0	19.8	31.5	28.2

Caption correctness evaluation. Conventional evaluation metrics, such as BLEU-4, cannot precisely evaluate the correctness of generated captions. Therefore, we added a caption correctness evaluation in addition to the conventional evaluation metrics. In this evaluation process, we ignored the validity of sentence structures and extracted "change type," "objects," "colors," "shapes," and "sizes" from each generated caption based on relevant words appeared in captions.

Then, we calculated the overall accuracy for change type, object, and attributes among the entire test set. The results are shown in Table IV.

We implemented our best model diff/conv/dyn-att for before- and after-change scenes with viewpoint pairs (1-1), (1-3), (1,2-3,4), (1,2,3,4-1,2,3,4) and (1-1), (1-3) for the DUDA model.

Compared with the DUDA model, our models with the same viewpoint settings obtained higher accuracy in terms of change type, object, and object attribute prediction, especially for change type prediction. Our model with four viewpoints achieved nearly perfect correctness, which means that alongside with generating sentences with correct structure, our model can give detailed and correct predictions for change captioning. Also, there is still room for our model to improve the object information related accuracy, especially for our model with viewpoint pair (1-3).

Qualitative results. We show three example results in Figure 4 ((a-c)). For the scene of example (a) with relatively less occlusion, both the DUDA model and our models correctly predicted the caption. However, for scene examples (b) and (c) involving severe occlusions, our models predicted correct caption types, while the object attributes are slightly incorrect. In contrast, the DUDA method failed to provide related captions in terms of both caption type and object attributes.

Compared with the DUDA method, our framework can identify the underlying 3D structure of scenes and associate before- and after-change scenes based on 3D correspondence. Thus, our model is more practical for interpreting scenes with occlusions. This ability plays an essential role in potential robot applications, which always require observing the surroundings with occlusions from multiple viewpoints.

C. Robustness to Camera Movement

Here, we discuss the effect on the performance of camera movement between the before- and after-change scenes. We used P_change_cameras dataset for evaluation. We observed the original scene from viewpoint 1 (Figure 3 right) and changed the observation viewpoint for the after-change scene from viewpoint 1 to 7 (0 degrees to 180 degrees). We trained the DUDA and our model under these seven viewpoint pairs setting. The BLEU-4 results are shown in Figure 5.

For viewpoint (1-1) pairs (same observation viewpoint for the before- and after-change scenes), both the DUDA method and our method obtained the highest BLEU-4 scores. After rotating the virtual cameras for the after-change scenes, the performance of the DUDA method was degraded significantly, while our method achieved relatively stable performance for different viewpoint pairs. This result indicates that compared with DUDA, our method is more robust against camera rotations, which makes our model more practical for real-world applications, where observing surroundings constantly from the same viewpoint is difficult.

D. Experiments on 3D Scenes with Realistic Object Models

In this experiment, we used ROM_change, which consists of photorealistic object models to evaluate the performance.

Quantitative results. Here, we used the viewpoint settings adopted in experiments on 3D scenes with geometric primitives. The overall and per-change type performance across different evaluation metrics are shown in Table III (bottom six rows). For our models and the DUDA model, performance with ROM_change dataset was degraded compared with the performance with the CLEVR-based datasets. Compared with CLEVR-based datasets, the ROM_change dataset has more complicated object models, which makes it more challenging in terms of image content understanding. Our models outperformed DUDA on all change types, especially for distractors except the (1-1) viewpoint pair. Although there is still room for our model with (1-3), (1,2-3,4) viewpoint pairs to improve, the model with four viewpoints obtained 87.1 for BLEU-4, which shows a possibility of adapting the model to more complex scenes.

Both models obtained higher performance with viewpoint setting (1-1) compared with (1-3). Alignment from the same viewpoint brings advantages to this task.

Imbalances sensitiveness evaluation. The ROM_change dataset has an imbalanced distribution over the object instances. To measure the sensitiveness to imbalances, we recorded test samples involving single-object change (add, delete, move) to the changed object, and recorded samples to the two changed objects for swap and replace. Then, we computed the BLEU-4 score for each class for our and the DUDA models with observation viewpoint pair (1-3). As shown in Figure 6, for both two models, the BLEU-4 score varies for different classes. Our model obtained a higher variance value than the DUDA model. Our model tends to obtain higher BLEU-4 scores for classes with large objects (e.g., soccerball) and lower scores for classes with small objects (e.g., bar soap). Increasing input image resolution could help to improve performance for small objects.

Caption correctness evalutation. We show the correctness evaluation result of generated captions in Table V. Although compared to CLEVR-based datasets, all models obtained a degraded accuracy, our model with four viewpoints still achieved high accuracy for generating correct captions, especially for change types. Our model with the same viewpoint pair to DUDA obtained a significantly higher accuracy on change type, which indicates that despite the correctness on detailed object information identifying, our model can effectively learn what kind of change has occurred in the input scene.

Qualitative results. We show three example results in Figure 4 ((d-f)). For example (d), which is relatively less occluded, all models produced the correct caption. For distractor shown in (e), all of our models recognized distractor and gave correct captions. In contrast, the DUDA model tended to be confused for correctly identifying distractors. For example (f), the DUDA model failed to identify change types while all our models correctly recognized change types but failed to predict the detailed object attributes.

V. CONCLUSION

We propose a framework that identifies and describes changes that occur in scenes observed from multiple viewpoints through natural language. The ability to understand the changes involved in 3D scenes plays a vital role in various robotic applications. Existing approaches focus on 2D images and have limited ability to handle occlusion, camera movements, which is critical in real-world applications. Therefore, we propose a framework that establishes scene representations containing underlying 3D structures of scenes and describes the changes through associating the before- and after-change scene representations. We created three synthetic datasets. The experimental results indicate that our method outperforms the previous state-of-the-art 2D-based method by a large margin in both sentence construction and captioning correctness. In addition, our method performs better for scenes with occlusions and shows higher robustness for camera movements. Our framework also shows encouraging results in a realistic dataset setting, which indicates the possibility of adapting our framework to a more complicated and broader scene-setting.

We conducted experiments on simulated datasets containing solid color scenes and object models. To provide a more realistic environment for scene change understanding, we will expand dataset setting by introducing more complex scenes, object models with higher diversity, and placing object models at various locations in the scenes.

Currently, our model builds scene representation from images observed from viewpoints sampled from pre-defined camera positions. Incorporating explicit 3D structures, such as depth maps, could enable our model to handle flexible camera positions and increase practicability in real-robot applications. In addition, integrating semantic understanding structures, such as semantic segmentation networks, could help improve the performance of our model in real-robot environments with higher complexity and diversity.

REFERENCES

[1] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3d object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[2] Graham B, Engelcke M, van der Maaten L. 3d semantic segmentation with submanifold sparse convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[3] Dai A, Ruizhongtai Qi C, Nießner M. Shape completion using 3d-encoder-predictor cnns and shape synthesis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[4] Park D H, Darrell T, Rohrbach A. Robust change captioning. Proceedings of the IEEE International Conference on Computer Vision. 2019.

[5] Jhamtani H, Berg-Kirkpatrick T. Learning to Describe Differences Between Pairs of Similar Images. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[6] Panda, Swagatika, AH Abdul Hafez, and C. V. Jawahar. "Single and multiple view support order prediction in clutter for manipulation." Journal of Intelligent and Robotic Systems 83.2. 2016.

[7] Grotz, Markus, David Sippel, and Tamim Asfour. "Active Vision for Extraction of Physically Plausible Support Relations."IEEE-RAS 19th International Conference on Humanoid Robots. 2019.

[8] Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., ... and Reichert, D. P. Neural scene representation and rendering[J]. Science, 2018, 360(6394).

[9] Sitzmann V, Zollhöfer M, Wetzstein G. Scene representation networks: Continuous 3D-structure-aware neural scene representations. Advances in Neural Information Processing Systems. 2019.

[10] Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. Advances in neural information processing systems. 2014.

[11] Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[12] Qi, C. R., Yi, L., Su, H., and Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Advances in neural information processing systems. 2017.

[13] Bajcsy, R. Active perception. Proceedings of the IEEE, 76(8), 1988.

[14] Lepora, N. F., Martinez-Hernandez, U., and Prescott, T. J. Active touch for robust perception under position uncertainty. In 2013 IEEE International Conference on Robotics and Automation. 2013.

[15] Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., and Batra, D. Embodied question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018.

[16] Sakurada, K., and Okatani, T. Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation. In The British Machine Vision Association. 2015.

[17] Sakurada, K., Wang, W., Kawaguchi, N., and Nakamura, R. Dense optical flow based change detection network robust to difference of camera viewpoints. arXiv preprint arXiv:1712.02941.

[18] Herbst, E., Henry, P., Ren, X., and Fox, D. Toward object discovery and modeling via 3-d scene comparison. In 2011 IEEE International Conference on Robotics and Automation. 2011.

[19] Ambruş, R., Bore, N., Folkesson, J., and Jensfelt, P. Meta-rooms: Building and maintaining long term spatial models in a dynamic world. In IEEE International Conference on Intelligent Robots and Systems. 2014.

[20] Langer, E., Ridder, B., Cashmore, M., Magazzeni, D., Zillich, M., and Vincze, M. On-the-fly detection of novel objects in indoor environments. In 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO). 2017.

[21] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. Show, attend and tell: Neural image caption generation with visual attention. International conference on machine learning. 2015.

[22] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[23] Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[24] Araki, R., Yamashita, T., and Fujiyoshi, H. Arc2017 rgb-d dataset for object detection and segmentation. In Late Breaking Results Poster on International Conference on Robotics and Automation, 2018.

[25] Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., and Dollar, A. M. The ycb object and model set: Towards common benchmarks for manipulation research. 2015 international conference on advanced robotics (ICAR). IEEE, 2015.

[26] Singh, A., Sha, J., Narayan, K. S., Achim, T., and Abbeel, P. Bigbird: A large-scale 3d database of object instances. 2014 IEEE international conference on robotics and automation (ICRA). IEEE, 2014.

[27] Johnson, J., Krishna, R., Stark, M., Li, L. J., Shamma, D., Bernstein, M., Fei-Fei, L. Image retrieval using scene graphs. In Proceedings of conference on computer vision and pattern recognition. 2015.

[28] He, K., Zhang, X., Ren, S.,and Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[29] Lin, Chin-Yew. ROUGE: a Package for Automatic Evaluation of Summaries. Association for Computational Linguistics, 2004.

[30] Anderson, P., Fernando, B., Johnson, M., and Gould, S. Spice: Semantic propositional image caption evaluation. European Conference on Computer Vision. Springer, Cham, 2016.

[31] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[32] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. Proceedings of the ninth workshop on statistical machine translation. 2014.