Describing and Localizing Multiple Changes with Transformers

Yue Qiu[1]*, Shintaro Yamamoto[1,2]*, Kodai Nakashima[1], Ryota Suzuki[1],
Kenji Iwata[1], Hirokatsu Kataoka[1], Yutaka Satoh[1]
[1]National Institute of Advanced Industrial Science and Technology (AIST), [2]Waseda University
{qiu.yue, yamamoto.shintaro, nakashima.kodai, ryota.suzuki,
kenji.iwata, hirokatsu.kataoka, yu.satou}@aist.go.jp

Abstract

Change captioning tasks aim to detect changes in image pairs observed before and after a scene change and generate a natural language description of the changes. Existing change captioning studies have mainly focused on a single change. However, detecting and describing multiple changed parts in image pairs is essential for enhancing adaptability to complex scenarios. We solve the above issues from three aspects: (i) We propose a simulation-based multi-change captioning dataset; (ii) We benchmark existing state-of-the-art methods of single change captioning on multi-change captioning; (iii) We further propose Multi-Change Captioning transformers (MCCFormers) that identify change regions by densely correlating different regions in image pairs and dynamically determines the related change regions with words in sentences. The proposed method obtained the highest scores on four conventional change captioning evaluation metrics for multi-change captioning. Additionally, our proposed method can separate attention maps for each change and performs well with respect to change localization. Moreover, the proposed framework outperformed the previous state-of-the-art methods on an existing change captioning benchmark, CLEVR-Change, by a large margin (+6.1 on BLEU-4 and +9.7 on CIDEr scores), indicating its general ability in change captioning tasks. The code and dataset are available at the project page [1].

1. Introduction

Detecting and describing the changed parts in scenes at different times is essential in various scenarios, such as urbanization analysis [1, 2, 3], resource management [4, 5, 6, 7, 8], updating street-view maps for navigation [9, 10], damage detection [11, 12], video surveillance [13], and robotic applications [14, 15]. Recently, Jhamtani and Berg-Kirkpatrick [13] proposed the change captioning task to describe changes from image pairs of before and after scene changes. Describing change is useful for extracting semantic contents and conveying information to humans.

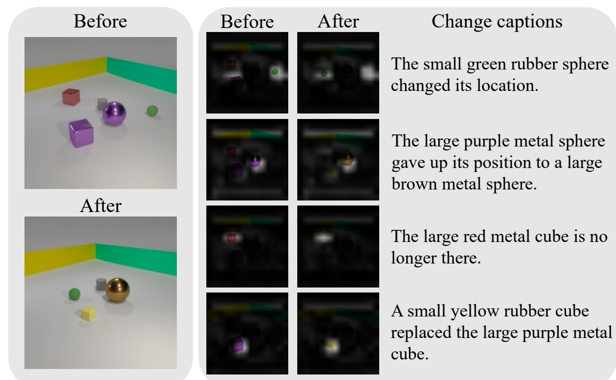

Figure 1. Given two images of a scene observed before and after multiple changes (the first column), we generate change captioning for each scene change along with attention maps (the second and third columns) indicating the region of changed objects.

Several methods have been proposed for the change captioning [13, 16, 17, 18, 19]. Most existing works focus on describing a single change. As a practical matter, multiple changes could be manifested within an image pair. Jhamtani and Berg-Kirkpatrick [13] studied scene change captioning with multiple changes; however, they addressed the problem with a known number of changes which is not provided in real-world problems. Practically, detecting and describing scene changes without prior information regarding changes is more useful in terms of providing information to users. We address multi-change captioning, where changed regions are localized and language descriptions of scene changes are generated from pair of images with an unknown number of changes, as shown in Figure 1.

Change captioning requires capturing relationships between image pairs, localizing changed regions, and gen-

*equal contribution
[1]https://cvpaperchallenge.github.io/Describing-and-Localizing-Multiple-Change-with-Transformers

erating language descriptions. In this work, we introduce a simple but effective framework Multi-Change Captioning transformers (MCCFormers) based on encoder-decoder transformer [20] which performs well in natural language processing. The encoder transformer captures relationship between local regions in two images to detect scene changes. Then, the decoder transformer attends over changed regions and generates language descriptions of the changes. In contrast to existing methods that generate a static attention map [16, 17], the decoder transformer changes the spatial attention for each generated word. Consequently, the decoder transformer can distinguish between different changes and avoid confusing them for one another.

To evaluate multi-change captioning and localization ability, we build a novel CLEVR-Multi-Change dataset consisting of image pairs containing multiple changes, change captions, and bounding boxes of the changed region. We compare the proposed MCCFormers model with several state-of-the-art methods under the multi-change setup. The experimental results show that the proposed method performed well in both change captioning and localization.

The contributions of our work are three-fold: (i) We address a novel task of multi-change captioning and propose a dataset for this task where multiple changes exist in before- and after-change images and the number of changes is unknown; (ii) We propose MCCFormers, which consists of encoder-decoder transformers that capture relationships between images and densely correlates image regions with words; (iii) The proposed MCCFormers outperforms existing methods in terms of four conventional image captioning evaluation metrics and shows promising ability on localization for multi-sentence change captioning.

2. Related Work

2.1. Change Detection

Change detection from scenes captured from different moments has been studied in various research fields. [14, 15, 21] discussed change detection from indoor scenes. There are also existing studies which discuss change detection for disaster management [11, 12], resource monitoring [6, 7], and vehicle navigation [9]. Among existing studies, [14, 15, 21] proposed rule-based methods for detecting changed parts from a set of 3-D maps. [6, 7, 9, 11, 12] discuss generating pixel-level maps for indicating the changed region between image pairs. Instead of change detection, we address localizing and describing changes.

2.2. Image Captioning

Image captioning is a well studied topic at the intersection of computer vision and natural language processing. Vinyals *et al*. [22] proposed encoder-decoder architecture where an encoder extracts image features and a decoder generates a description of an image. Xu *et al*. [23] introduced the attention mechanism to align each word and relevant region in an image. Inspired by the human visual system, Anderson *et al*. proposed a combined bottom-up and top-down attention mechanism [24]. Following the success of transformers [20] in natural language processing, transformer-based approaches have been introduced for image captioning [25, 26, 27]. Different from image captioning for a single image, we address change captioning which requires capturing the relationship between two images.

2.3. Change Captioning

Several studies have focused on change captioning which describes a change between two images from different moments. The Spot-the-Diff dataset which consists of 13,192 scene change image pairs was constructed by Jhamtani and Berg-Kirkpatrick [13]. Each image pair has 1.86 change description sentences on average. However, they addressed the problem with a known number of changes. By contrast, we study the multi-change captioning task in which the number of scene changes is not given. CLEVR-Change dataset was introduced by Park *et al*. [16] to overcome several limitations of the Spot-the-Diff dataset including lack of viewpoint change and localization ground truth. The authors of [28] and [29] discussed change captioning from image pairs observed from multiple viewpoints. This work addresses multi-change captioning with an unknown number of changes and we develop CLEVR-Multi-Change dataset to evaluate localization ability as well as captioning.

Jhamtani and Berg-Kirkpatrick [13] proposed DDLA which computes a pixel-level difference between image pairs, limiting the ability for situations with viewpoint changes. By contrast, DUDA [16] utilizes feature-level differences to enhance the robustness to viewpoint change. Similarly, the Siamese difference captioning model was proposed by Oluwasanmi *et al*. [18, 19]. M-VAM [17] separates viewpoint changes from semantic changes by evaluating the similarity of different patches of image pairs. Spatial attentions used in DUDA and M-VAM are static, which limits their ability to distinguish different changes.

In this work, we build transformer-based encoder-decoder models for multi-change captioning. The encoder transformer computes patch-level similarity with multi-head attention which captures different types of changes between image pairs. The decoder transformer performs multi-head attention over image patches from the encoder, which captures the relation between generated words and image regions and thus can distinguish different changes.

3. CLEVR-Multi-Change Dataset

Existing change captioning studies mainly focus on single changes. However, identifying and distinguishing multiple change regions simultaneously manifested in image

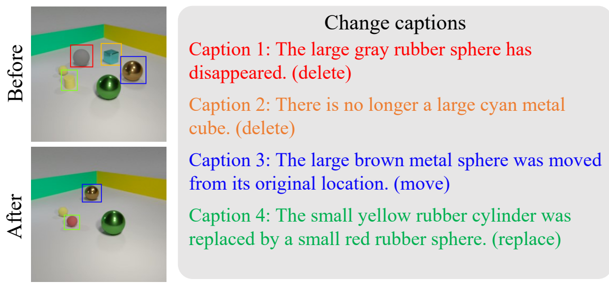

Figure 2. Example from the CLEVR-Multi-Change dataset. The changed objects are highlighted by bounding boxes with the same color as the associated change captions. More dataset examples can be found in the supplementary material.

Dataset	Multi-change	Viewpoint change	Localization	Total image pairs
Spot-the-Diff [13]	✓			13,192
CLEVR-Change [16]		✓	✓	79,606
CLEVR-Multi-Change	✓	✓	✓	60,000

Table 1. Change captioning dataset comparison.

	1 Change	2 Changes	3 Changes	4 Changes
Image Pairs	15,137	14,873	14,988	15,002
Captions	75,685	74,365	74,940	75,010
Bboxes	22,775	44,495	67,329	89,744

Table 2. CLEVR-Multi-Change dataset statistics.

pairs is necessary due to frequent human activities. Moreover, change region localization is also critical in a variety of applications. For example, localizing target objects is essential for robot manipulation applications. To address these issues, we propose the CLEVR-Multi-Change dataset for diagnosing the ability of change localization and captioning in image pairs involving multiple changes based on the CLEVR engine [30] and CLEVR-Change dataset [16].

Image Pairs Generation. To generate a variety of scenes, we place objects with random shapes (cube, sphere, cylinder), colors (red, blue, yellow, green, brown, cyan, gray, purple), sizes (large, small), and materials (metal, rubber) into a simulated environment. We considered four atomic change types, namely "add", "delete", "move", and "replace" an object. We set a virtual camera to create image pairs by observing a scene before and after scene change operations. We also add a random position change to cameras. We generate each scene consisting of one to four changes within image pairs. We also record bounding boxes of changed objects for localization evaluation.

Eliminating Scene Change Ambiguity. Unlike the CLEVR-Change dataset, we added two "walls" with solid colors as background to reduce ambiguous correspondences between images due to the lack of camera information. We deleted descriptions of object relationships within an image (*e.g.* to the left of) to prevent the ill-posed problem. To eliminate the ambiguity of change combinations (*e.g.* "replace a red cube with a blue cylinder" equals "delete a red cube" and then "add a blue cylinder"), we restrict the maximum change number to one for every object and region.

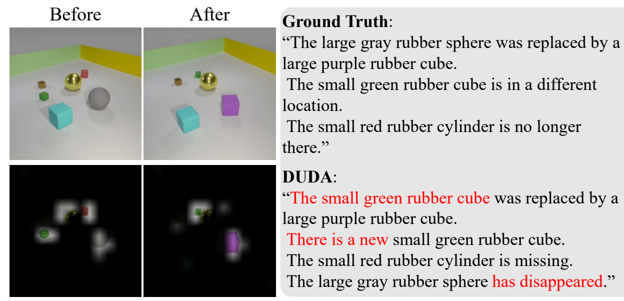

Figure 3. Preliminary study on DUDA for describing multiple changes. Incorrect captions are in red font.

Caption Generation. The change captions are automatically generated based on recorded scene change information and pre-defined change sentence templates. We create five captions for each image pair with different sentence templates. The sentence order is randomly determined for two-, three-, and four-change image pairs.

We show the comparison with two extant datasets in Table 1. Statistics of our dataset and example are provided in Table 2 and Figure 2, respectively. We split the dataset into 2/3, 1/6, and 1/6 for training, validation, and testing, respectively. We refer to the supplementary material for more details about the dataset generation process.

Task Definition. Given images of before and after changes I_{bef} and I_{aft}, there are N scene changes between them. Now, we define $S^i (i \in [1, \cdots, N])$ as a description of the i^{th} change consisting of a sequence of words $(w_1^i, \cdots, w_M^i)$ with a maximum length of M. The multi-change captioning task aims to generate all S^i from I_{bef} and I_{aft} with an unknown number of changes N. We consider predicting all sentences as a single sequence such as $(w_1^1, \cdots, w_M^1, < SEP >, w_1^2 \cdots)$. In our proposed framework, spatial attention is dynamically associated with each word and therefore localization of each change can be computed by averaging the attention maps of each word.

Preliminary Study. We evaluated a previous state-of-the-art method, DUDA, on this dataset and one example result is shown in Figure 3. Although DUDA can determine the changed region, it is confused by multiple changes and generates change captions which are only partially correct. For example, DUDA generated "The small green rubber cube was replaced by a large purple rubber cube", but the ground truth is "The large gray rubber sphere was replaced by a large purple rubber cube", which suggests that DUDA attended wrong objects mentioned in different ground truth.

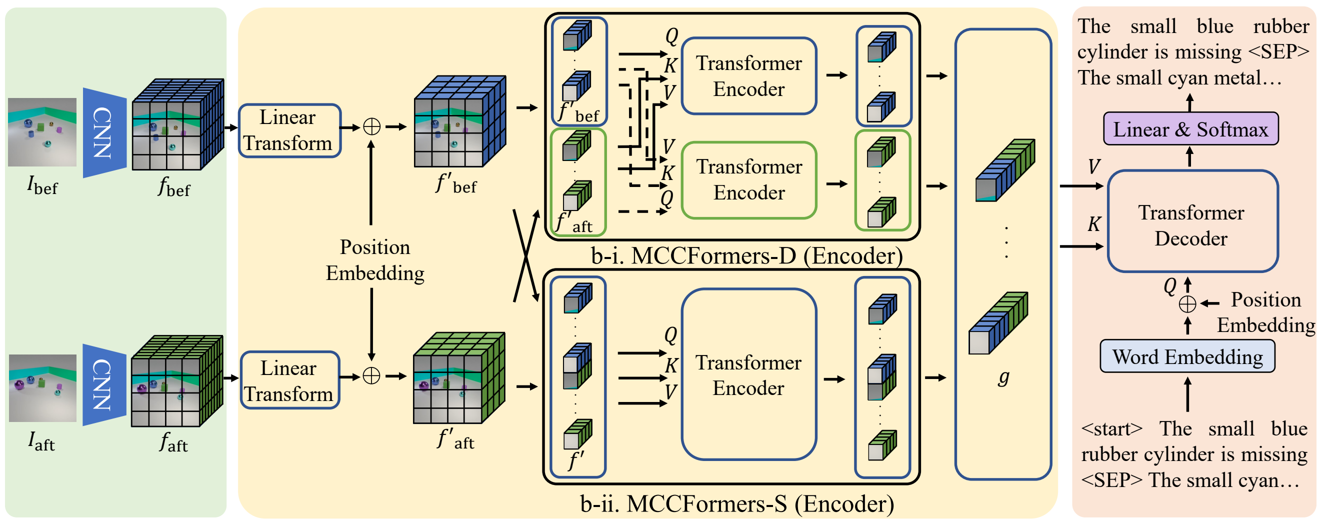

Figure 4. Overall framework of MCCFormers: (a) Image features are extracted using CNN. We experimented with two encoders: (b-i) MCCFormers-D (Encoder) and (b-ii) MCCFormers-S (Encoder). We then feed representations from the encoder to the decoder (c) for caption generation. Best viewed in color.

A dense correlation between different image regions of before- and after-change images is necessary for identifying changes. Furthermore, to distinguish and generate captions for each change, the correlation between change regions and sentences is critical.

4. Approach

Figure 4 shows the proposed Multi-Change Captioning transformers (MCCFormers). Given two images (I_{bef} and I_{aft}) of before and after multiple changes, MCCFormers generates a paragraph of descriptions of changes in image pairs. Following existing methods like DUDA [16] and M-VAM [17], we first extract image features f_{bef} and f_{aft} using the CNN structure.We then feed the features to the transformer [20]-based encoder-decoder model. A transformer encoder densely correlates each image patch of before- and after-change image pairs and a decoder further correlates each word with image patches for generating descriptions of multiple changes.

4.1. Change Encoder

It is necessary to distinguish and separate different change regions in scenes involving multiple changes, requiring a dense correlation of different regions between image pairs. To obtain the relationships of different image patches in image pairs, a mechanism to compare and correlate each image patch between image pairs is required. M-VAM correlates feature pairs by introducing an inner production operation of features. Compared with inner-production, the multi-head attention mechanism introduced in transformer-based encoders computes multiple types of attentions to correlate different patches. Thus, we consider adopting transformer-based encoders.

Different from recent transformer-based models for computer vision tasks such as DETR [31] which takes a single image as its input, change captioning tasks feed two images. Given image feature pairs f_{bef} and f_{aft} with dimension $\mathbb{R}^{W\times H\times D}$ (where W, H, and D are, respectively, the width, height, and channel of features), we consider two variants of encoder: Multi-Change Captioning transformers-Dual (MCCFormers-D) and Multi-Change Captioning transformers-Single (MCCFormers-S) (Figure 4 (b-i) and (b-ii), respectively). For both variants, we first transform f_{bef} and f_{aft} to f'_{bef} and f'_{aft} with dimension $\mathbb{R}^{W\times H\times d_{encoder}}$. To accomplish this, we use linear transformation and add a position embedding as follows:

$$f'(x,y) = W_{lt}f(x,y) + b_{lt} + pos(x,y) \tag{1}$$

where W_{lt} and b_{lt} are learnable parameters of linear transformation and $pos(x,y)$ is learnable position embedding.

MCCFormers-D. In this variant, we use two transformer encoders with shared weights. To capture relevance between local regions of two images, we employ the co-attention mechanism [32]. In contrast to the original co-attention mechanism which takes linguistic tokens and object proposals from an image as input, we consider a set of patches of before and after images as input. Given two feature maps from before and after images f'_{bef} and f'_{aft}, we consider that the query feature is either from before- or after-change images and the key and value feature is from the other one. After processing with the encoder with N_e layers, we concatenate the output from features g_{bef} and

g_{aft} over feature dimension as $g \in \mathbb{R}^{W \times H \times 2d_{encoder}}$.

MCCFormers-S. Different from MCCFormers-D, we use MCCFormers-S to capture image patch relationships among both inter- and intra-image pairs. We first concatenate f'_{bef} and f'_{aft} to $f' \in \mathbb{R}^{2W \times H \times d_{encoder}}$. We then pass f' to the standard transformer structure, which is similar to the BERT model [33] that takes a sequence of two sentences as input. Following MCCFormers-D, the feature maps are converted to $g \in \mathbb{R}^{W \times H \times 2d_{encoder}}$. Compared with MCCFormers-D, which only considers relevance between image pairs, this structure also captures image patches' relevance within both before- and after-change images.

4.2. Paragraph Decoder

In the multi-change captioning task, due to the coexistence of multiple changes, it is critical to distinguish different change regions and dynamically attend to different regions during the generation of different sentences. Transformer decoders accomplish this by attending to information from different patches during the generation process.

Therefore, we adopt a standard transformer decoder for generating captions. We first use a word embedding layer to transfer input sentences and add a learnable position embedding. Next, the sentence features are processed through a masked self-attention and feed-forward network. The cross-attention between a sentence and the encoder's output features is then computed and further processed by a feed-forward layer. The decoder layer is iterated for N_d layers.

The transformer decoder computes attention over image features for each word during the sentence generation process. Thus, image attention can be computed for every sentence by averaging the attention maps of each word in a sentence. In contrast, DUDA and M-VAM compute a single spatial attention map for an entire paragraph.

4.3. Learning Process

From the input of images (I_{bef} and I_{aft}) observed before and after scene changes, the decoder generates a word sequence with length T. We denote the target sequence as $(w_1^*, ..., w_T^*)$. We adopt a cross-entropy loss for network training, where θ indicates the learnable parameter:

$$L_{XE} = \sum_{t=1}^{T} -\log(p_\theta(w_t^* | (w_1^*, ..., w_{t-1}^*), I_{\text{bef}}, I_{\text{aft}})) \quad (2)$$

5. Experiments

5.1. Experimental Setup

Experiments and Datasets. We conducted experiments on both multi-change and single change setups. We also implemented DUDA and M-VAM without models modification compared to the models proposed in original papers and evaluated the models performance on paragraph generation. We also report the performance of the methods on two previous datasets i.e., Spot-the-Diff (containing multi- and single change setups, where we sampled instances within four changes for the multi-change setup) and the CLEVR-Change dataset (single change).

Methods	BLEU-4
DUDA [16]	76.1
DUDA Encoder + Transformer Decoder	79.1
M-VAM [17]	62.9
M-VAM Encoder + Transformer Decoder	65.8
MCCFormers-D (concat. over patches)	80.1
MCCFormers-D (concat. over feature dimension)	82.3
MCCFormers-S (concat. over patches)	80.6
MCCFormers-S (concat. over feature dimension)	**83.3**

Table 3. BLEU-4 evaluation of different methods applied to the CLEVR-Multi-Change dataset. (concat. : concatenation)

Evaluation Metrics. We adopt conventional image captioning and change captioning evaluation metrics for performance comparison: BLEU-4 [34], CIDEr [35], METEOR [36], and SPICE [37]. These metrics evaluate the similarity of generated sentences with ground truth from different aspects. We also evaluated the accuracy in terms of the number of sentences in multi-sentence generation. We compute accuracy which measures whether the number of sentences of the generated sequence is correct and mean absolute error (MAE) to evaluate the difference in the number of sentences between ground truth and generated sequence. We prepared five ground truth paragraphs with different sentence orders for each image pair. Therefore, the sentence order has less influence on evaluation results. To assess the localization ability in multi-change captioning, we introduce an evaluation metric based on the Pointing Game [38]. We record the bounding boxes for changed objects and transfer the obtained attention map to the original image size with bilinear interpolation. We then select the top-K pixels with the largest values in attention maps and compute the detected change region (where the top-K pixels are inside the bounding box of a changed region) over all of the changed regions in the ground truth. The overall accuracy is averaged over all changes contained in the test data. We set K to 1 for add and delete and 2 for move and replace where the bounding boxes of objects might be different due to the scene change.

Implementation Details. Similar to [16, 17], we use ResNet-101 [39] pretrained on the ImageNet dataset [40] to extract image features from images with a 224×224 resolution. The obtained feature maps dimension is $14 \times 14 \times 1024$. We implemented transformers with two layers and four heads for both encoders and decoders. The dimensions

Methods	BLEU-4 [34]					CIDEr [35]	METEOR [36]	SPICE [37]
	Overall	1 Change	2 Changes	3 Changes	4 Changes			
DUDA [16]	76.1	94.7	76.3	73.1	70.9	480.1	47.4	66.6
M-VAM [17]	62.9	79.0	61.7	59.9	57.9	338.1	41.3	55.9
MCCFormers-D	82.3	**98.4**	**82.9**	79.2	**80.6**	**539.3**	**52.1**	**71.7**
MCCFormers-S	**83.3**	96.5	82.6	**81.8**	80.0	523.3	51.5	70.0

Table 4. Results on the CLEVR-Multi-Change dataset.

Methods	Accuracy					MAE
	All	1 change	2 change	3 change	4 change	
DUDA [16]	83.3	96.8	84.8	74.4	76.0	0.169
M-VAM [17]	69.9	91.2	68.4	56.5	62.0	0.317
MCCFormers-D	**92.5**	**99.4**	94.8	81.9	**93.9**	**0.075**
MCCFormers-S	92.4	97.9	**95.9**	**91.9**	83.7	**0.075**

Table 5. Sentence number accuracy (%) and mean absolute error (MAE) on the CLEVR-Multi-Change dataset.

of input features to encoder $d_{encoder}$ and decoder $d_{decoder}$ are 512 and 1024, respectively. For the feedforward network, the dimensions are $4d_{encoder}$ and $4d_{decoder}$ for encoders and decoders, respectively. We set the learning rate to 0.0001 and trained models for 40 epochs with the Adam optimizer [41] during all implementations.

Baselines. We use DUDA and M-VAM for comparison during experiments and two variants of the transformer network. We set the hidden state dimension for all LSTM structures to 512 in both DUDA and M-VAM (see the supplementary material for details of the implementation of DUDA and M-VAM).

5.2. CLEVR-Multi-Change dataset

Ablation Study. We evaluated different structure choices of change encoders and decoders (Table 3). DUDA and M-VAM scored 76.1 and 62.9 for BLEU-4, respectively. We removed the sum operation over the spatial region of features (DUDA) and average pooling operations (M-VAM) which are applied before feeding features to decoders. We then replaced the decoders with transformer decoders. The use of transformer decoders improved the performance of both two methods.

Next, we conducted experiments regarding MCCFormers. Before feeding features to the decoder, we concatenate features of before- and after-change images in two ways: concatenation operation over patches (input to decoder: $g \in \mathbb{R}^{2W \times H \times d_{encoder}}$) and over feature dimension (input to decoder: $g \in \mathbb{R}^{W \times H \times 2d_{encoder}}$). All methods outperformed previous methods in terms of BLEU-4. MCCFormers-D and MCCFormers-S with concatenation over feature dimension performed the best. In the case of relatively small viewpoint change, patches of the same region in two images are concatenated, which could improve the effectiveness of the concatenation over feature dimension. We will investigate the robustness to viewpoint change (especially larger change) for future work. In the remaining experiments, we use MCCFormers-D and MCCFormers-S with concatenation over feature dimension.

Sentence Generation. Experimental results of the different evaluation metrics are shown in Table 4. Both MCCFormers-D and MCCFormers-S outperformed previous methods with respect to all metrics. MCCFormers-S obtained the highest BLEU-4 and outperformed previous methods by +7.2.

For one-change instances, the differences between the proposed methods and previous methods are relatively small. For instances with multiple changes, the two proposed methods exhibited better robustness. The transformer encoder learns a dense correlation among all local regions of change image pairs, and the decoder model further correlates each word with image regions, making the models better at distinguishing different changes.

Evaluation of Sentence Number Accuracy. Results concerning sentence number accuracy and MAE are shown in Table 5. Similar to the results in Table 4, the proposed methods obtained higher accuracy for sentence numbers compared to previous methods and achieved promising results for distinguishing changes, with an accuracy of 92.5% for the MCCFormers-D method. The MAE results show that the average errors of the number of sentences generated by all methods are less than 1.

All methods achieved the highest scores for one-change sentence and the MCCFormers-D model obtained 99.4% accuracy. For two-, three- and four-change instances, the accuracy of previous methods was degraded while the two proposed methods show promising stability for scenes with multiple changes.

Qualitative Results. We show one example result in Figure 5. This example contains four changes. DUDA predicted three changes and both variants of MCCFormers generated correct sentences in terms of both change number and change contents. In addition, two captions generated by DUDA contain incorrect change types whilst two proposed methods generated correct captions for each change.

DUDA generates a single attention map for each given image pair. Therefore, the network could be strug-

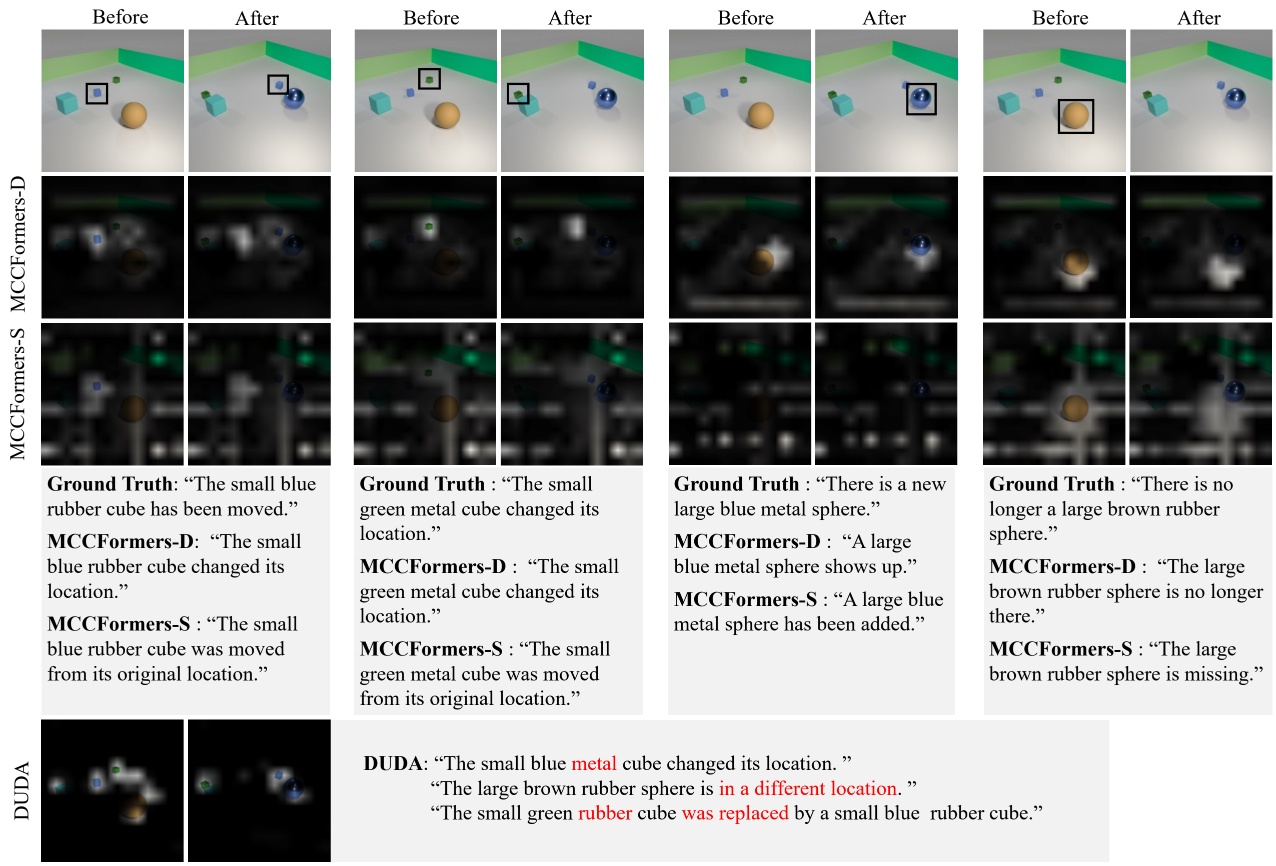

Figure 5. Visualization of an example from the CLEVR-Multi-Change dataset. We show the attention maps for the proposed methods and DUDA and generated sentences. Incorrect captions are in red font. We highlighted changed regions in black bounding boxes.

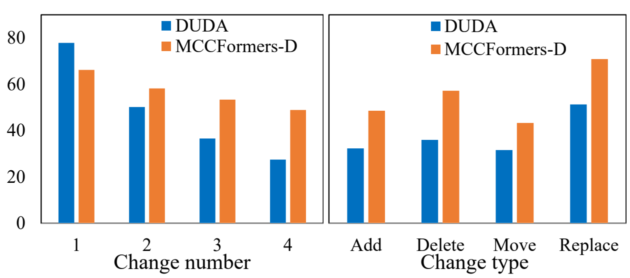

Figure 6. Pointing Game accuracy (%) over different change numbers (left) and types (right) on the CLEVR-Multi-Change dataset.

gling to distinguish each changed part from the attention map, limiting its ability in multi-change understanding. MCCFormers-S generated attention maps that tend to focus on unrelated regions as well as changed object regions. This structure refers to the patches of inter- and intra-image pairs, which might weaken the interpretability of attention maps. MCCFormers-D obtained separate attention maps for each sentence and attended to change regions. We refer to the supplementary material for more example results.

Pointing Game Evaluation for Attention Maps. We evaluated the localization accuracy of MCCFormers-D and DUDA (Figure 6). Since DUDA generates a single pair of attention maps for each image pair, we use the same attention map to evaluate each change. MCCFormers-D obtained an overall accuracy of 53.9% for change localization and 40.0% for DUDA. In both methods, localization performance degrades with the increase in change number. DUDA obtained higher localization accuracy for one change and our methods outperformed DUDA for two, three, and four changes, indicating the effectiveness of MCCFormers-D for detecting multiple changes.

Among different change types, both methods obtained the highest accuracy for replace change and the lowest accuracy for move change. The move change relates to two image positions, making it challenging for localization.

5.3. Spot-the-Diff Dataset

Spot-the-Diff contains multiple changes within image pairs. We first extract all instances containing one to four changes and report the results of DUDA and MCCFormers in Table 6 (top three rows). For this dataset, MCCFormers obtained comparable results with DUDA. We further show one example in Figure 7. For the example containing two changes, the two methods correctly generated two related

Methods	BLEU-4	CIDEr	METEOR	SPICE
Multiple change (one to four changes)				
DUDA [16]	5.4	24.8	**10.6**	12.9
MCCFormers-D	**6.2**	**28.8**	10.2	**17.8**
MCCFormers-S	5.8	18.2	10.5	10.1
Single change				
DUDA [16]	8.1	34.0	11.5	-
FCC [18]	9.9	36.8	**12.9**	-
SDCM [19]	9.8	36.3	12.7	-
DDLA [13]	8.5	32.8	12.0	-
M-VAM [17]	10.1	38.1	12.4	14.0
M-VAM + RAF [17]	**11.1**	42.5	**12.9**	17.1
MCCFormers-D	10.0	**43.1**	12.4	**18.3**
MCCFormers-S	9.8	41.6	12.3	16.3

Table 6. Results on the Spot-the-Diff dataset.

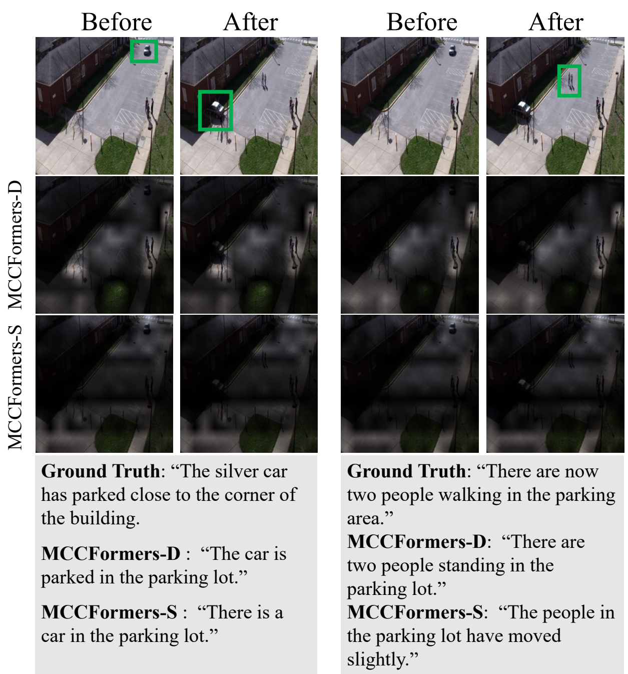

Figure 7. Visualization of an example of proposed methods on the Spot-the-Diff dataset (multiple change). We highlighted changed regions in green bounding boxes.

sentences. However, compared to ground truth sentences, the generated captions lack detailed attribute information, such as the color of the car is "silver" and the detailed location is the "corner of the building". We also found that for MCCFormers-D, the generated attention maps plausibly attended related regions but highlighted many unrelated image regions. MCCFormers-S failed to point out the related regions and struggled at recognizing detailed changes.

We show the results of the single change setup in Table 6 (bottom eight rows). The proposed methods obtained scores comparable to the state-of-the-art method M-VAM.

Methods	BLEU-4	CIDEr	METEOR	SPICE
DUDA [16]	47.3	112.3	33.9	24.5
M-VAM [17]	50.3	114.9	37.0	30.5
M-VAM + RAF [17]	51.3	115.8	37.8	30.7
MCCFormers-D	52.4	121.6	38.3	26.8
MCCFormers-S	**57.4**	**125.5**	**41.2**	**32.4**

Table 7. Results on the CLEVR-Change dataset.

Compared to the proposed dataset, the Spot-the-Diff dataset contains fewer images, which might limit the performance of the transformer-based methods because they tend to require a large amount of training data. Further exploration of change captioning on large-scale real-world imagery is warranted in future research.

5.4. CLEVR-Change Dataset

We compare the different methods on the previous single change dataset CLEVR-Change in Table 7. The CLEVR-Change dataset requires understanding object relationships inside each image (*e.g.* in front of) which are not included in the proposed dataset. Therefore, compared with MCCFormers-D, MCCFormers-S obtained the highest scores as MCCFormers-S can capture relationships between image patches within the same image. MCCFormers outperformed the previous methods on this dataset in terms of most evaluation metrics, indicating the ability of proposed structures to correlate different regions in change image pairs and further connect the change region information with words in sentences.

6. Conclusion

In this paper, we propose a novel multi-change captioning task and CLEVR-Multi-Change dataset for this task. To address the novel task, we proposed a transformer-based framework MCCFormers that densely correlates different image regions in image pairs and words. MCCFormers achieved state-of-the-art performance on both multi- and single-change captioning datasets, indicating the effectiveness of MCCFormers for change captioning tasks.

Acknowledgements

We want to thank Yoshitaka Ushiku, Seito Kasai, Hikaru Ishitsuka, and Tomomi Satoh for their helpful comments during research discussions. This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

References

[1] Qiaofeng Zhang, J Wang, X Peng, P Gong, and P Shi. Urban built-up land change detection with road density and spectral information from multi-temporal landsat tm data. *International Journal of Remote Sensing*, 23(15):3057–3078, 2002. 1

[2] Limin Yang, George Xian, Jacqueline M Klaver, and Brian Deal. Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, 69(9):1003–1010, 2003. 1

[3] Ibrahim Rizk Hegazy and Mosbeh Rashed Kaloop. Monitoring urban growth and land use change detection with gis and remote sensing techniques in daqahlia governorate egypt. *International Journal of Sustainable Built Environment*, 4(1):117–124, 2015. 1

[4] Pol R Coppin and Marvin E Bauer. Digital change detection in forest ecosystems with remote sensing imagery. *Remote Sensing Reviews*, 13(3-4):207–234, 1996. 1

[5] Robert E Kennedy, Philip A Townsend, John E Gross, Warren B Cohen, Paul Bolstad, YQ Wang, and Phyllis Adams. Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote Sensing of Environment*, 113(7):1382–1396, 2009. 1

[6] Salman H Khan, Xuming He, Fatih Porikli, and Mohammed Bennamoun. Forest change detection in incomplete satellite images with deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5407–5423, 2017. 1, 2

[7] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep change vector analysis for multiple-change detection in vhr images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3677–3693, 2019. 1, 2

[8] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067, 2018. 1

[9] Pablo F Alcantarilla, Simon Stent, German Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42(7):1301–1322, 2018. 1, 2

[10] Ken Sakurada, Takayuki Okatani, and Koichiro Deguchi. Detecting changes in 3d structure of a scene from multi-view images captured by a vehicle-mounted camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 137–144, 2013. 1

[11] Ken Sakurada and Takayuki Okatani. Change detection from a street image pair using cnn features and superpixel segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–12, 2015. 1, 2

[12] Aito Fujita, Ken Sakurada, Tomoyuki Imaizumi, Riho Ito, Shuhei Hikosaka, and Ryosuke Nakamura. Damage detection from aerial images via convolutional neural networks. In *2017 International Conference on Machine Vision Applications (MVA)*, pages 5–8, 2017. 1, 2

[13] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4024–4034, 2018. 1, 2, 3, 8

[14] Evan Herbst, Peter Henry, Xiaofeng Ren, and Dieter Fox. Toward object discovery and modeling via 3-d scene comparison. In *2011 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2623–2629, 2011. 1, 2

[15] Rareş Ambruş, Nils Bore, John Folkesson, and Patric Jensfelt. Meta-rooms: Building and maintaining long term spatial models in a dynamic world. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1854–1861, 2014. 1, 2

[16] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4624–4633, 2019. 1, 2, 3, 4, 5, 6, 8

[17] Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 574–590, 2020. 1, 2, 4, 5, 6, 8

[18] Ariyo Oluwasanmi, Enoch Frimpong, Muhammad Umar Aftab, Edward Y Baagyere, Zhiguang Qin, and Kifayat Ullah. Fully convolutional captionnet: Siamese difference captioning attention model. *IEEE Access*, 7:175929–175939, 2019. 1, 2, 8

[19] Ariyo Oluwasanmi, Muhammad Umar Aftab, Eatedal Alabdulkreem, Bulbula Kumeda, Edward Y Baagyere, and Zhiquang Qin. Captionnet: Automatic end-to-end siamese difference captioning model with attention. *IEEE Access*, 7:106773–106783, 2019. 1, 2, 8

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6000–6010, 2017. 2, 4

[21] Maciej Halber, Yifei Shi, Kai Xu, and Thomas Funkhouser. Rescan: Inductive instance segmentation for indoor rgbd scans. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2541–2550, 2019. 2

[22] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015. 2

[23] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015. 2

[24] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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 (CVPR)*, pages 6077–6086, 2018. 2

[25] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10578–10587, 2020. 2

[26] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8928–8937, 2019. 2

[27] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 2

[28] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. 3d-aware scene change captioning from multiview images. *IEEE Robotics and Automation Letters*, 5(3):4743–4750, 2020. 2

[29] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. Indoor scene change captioning based on multimodality data. *Sensors*, 20(17):4761, 2020. 2

[30] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017. 3

[31] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 4

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 4

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019. 5

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002. 5, 6

[35] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 5, 6

[36] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. 5, 6

[37] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 382–398, 2016. 5, 6

[38] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 5

[39] 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 (CVPR)*, pages 770–778, 2016. 5

[40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 6