Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

RESEARCH ARTICLE

Bi-directional Difference Locating and Semantic Consistency Reasoning for Change Captioning

Yaoqi Sun[1] | Liang Li*[2] | Tingting Yao[1] | Tongyv Lu[1] | Bolun Zheng*[1] | Chenggang Yan[1] | Hua Zhang[1] | Yongjun Bao[3] | Guiguang Ding[4] | Gregory Slabaugh[5]

[1]Hangzhou Dianzi University, Hangzhou, China
[2]Institute of Computing Technology, CAS, Beijing, China
[3]JD.com, Beijing, China
[4]Tsinghua University, School of Software, Beijing, China
[5]Queen Mary University of London, Digital Environment Research Institute (DERI), London, UK

Correspondence
*Liang Li, Institute of Computing Technology, CAS, Beijing 100086, China.
Email: liang.li@ict.ac.cn

*Bolun Zheng, School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China.
Email: blzheng@hdu.edu.cn

Abstract

Change captioning is an emerging task to describe the changes between a pair of images. The difficulty in this task is to discover the differences between the two images. Recently, some methods have been proposed to address this problem. However, they all employ unidirectional difference localization to identify the changes. This can lead to ambiguity about the nature of the changes. Instead, we propose a framework with bi-directional difference localization and semantic consistency reasoning to describe the image changes. First, we locate the changes in the two images by capturing bi-directional differences. Then we design a decoder with spatial-channel attention to generate the change caption. Finally, we introduce semantic consistency reasoning to constrain our bi-directional difference localization module and spatial-channel attention module. Extensive experiments on three public datasets show that the performance of our proposed model outperforms the state-of-the-art change captioning models by a large margin.

KEYWORDS:
Change Captioning; Semantic Consistency Reasoning; Spatial-Channel Attention

1 | INTRODUCTION

Change captioning [1,2,3] is an emerging subject to describe the differences between paired images. Change captioning is important in many practical applications including anomaly detection, infrastructure inspection, and disaster prevention using images from satellites or surveillance cameras.

Different from image captioning which characterizes the complete image content of a single image, change captioning only focuses on the differences between two images. Often the changes are localized to regions in the images, but the images contain plenty of other unchanged visual information. Furthermore, there may be global distractors, e.g. viewpoint shift and illumination variations. Both will increase the difficulty of locating changes. Another difficulty is that change captioning must comprehensively analyze the correspondences and disagreements between paired images to generate a fine-grained textual description.

Change captioning methods tried to model the mapping between two input images and output sentences describing the changes. For example, Tan et al. [2] utilized a linear fusion of the paired image features to generate the change caption. But the fused image features can not provide enough contrastive information, which produced generated captions with limited ability to describe the changes. Later, Oluwasanmi et al. [1,4] directly calculated the weighted L1 distance of paired images features, then generated the change caption simply using the obtained contrastive features. Li et al. [5] proposes a simple and effective robust

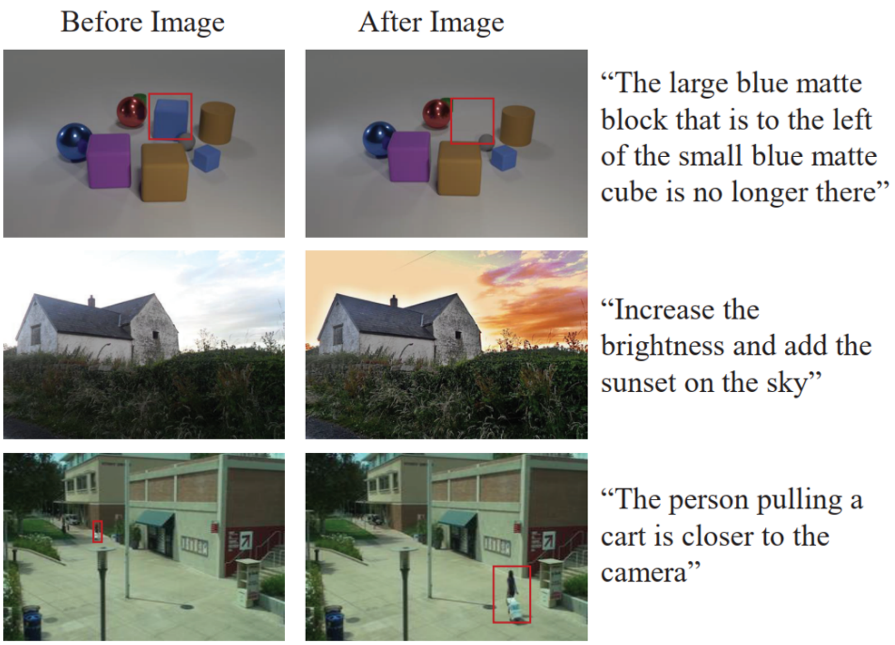
Before Image
After Image
“The large blue matte block that is to the left of the small blue matte cube is no longer there”
“Increase the brightness and add the sunset on the sky”
“The person pulling a cart is closer to the camera”

FIGURE 1 Overview of change captioning on three datasets: CLEVR-Change Dataset (top), Image Editing Request Dataset (middle), Spot-The-Diff Dataset (bottom). The annotation on the right side is the corresponding referenced caption. Given a pair of images, this task is to locate the changes between them to generate captions.

subspace learning method by incorporating the feature learning and visual understanding, which achieves surprising improvement. However, these features lacked expressiveness to represent unchanged objects and the positional relationship between unchanged objects and those that change. Recently, Dong Huk Park et al.[3] leveraged the features of images before and after the change as well as the unidirectional change between them to generate the change caption. However, there are two apparent limitations in this approach: (1) due to the the unidirectional differencing, the method can be easily confused about changes such as ‘add’, ‘drop’ and ‘move’, which require more sophisticated modeling to understanding to the semantics of the change. For example, the localization of a ‘move’ change is similar to a ‘drop’ change in the ‘before’ image, but also similar to an ‘add’ change in ‘after’ image. (2) the method ignores the attributes of unchanged objects and the positional relationship between unchanged objects and changed objects which results in inaccurate descriptions in the final output.

To mitigate the above-mentioned limitations, we propose a framework with bi-directional difference localization and semantic consistency reasoning to describe the image changes. First, we compute and localize the bi-directional differences between the ‘before’ images and the ‘after’ images. Such bi-directional difference localization greatly reduces the confusion about certain change types, especially ‘add’, ‘drop’ and ‘move’. Second, we design the decoder with spatial-channel attention to generate the change captions. This attention module enables our model to not only focus on key feature maps but also further attend to the core positions of each feature map at each decoding stage. This generates a more accurate description of the positional relationship between changed unchanged objects and significantly facilitates downstream tasks that require more precise visual content understanding[6]. Finally, we introduce semantic consistency reasoning to enhance the ability of our model to locate changed regions and systematically analyze both features by reasoning about the relationship between captions and images. The reasoning mechanism measures the relationship between captions and images by computing a consistency score which is enforced as a loss. Extensive experiments on three public datasets show that the performance of our proposed model outperforms the state-of-the-art change captioning models by a large margin.

The contributions of our work are summarized as follows:

- We introduce a novel change captioning method, where bi-directional difference localization reduces the confusion about the change types and spatial-channel attention attends to the core positions of the key image.
- We propose a semantic consistency reasoning mechanism to match the textual information to the visual information by calculating a consistency score between them.

- Extensive experiments on three public datasets show that our model outperforms the state-of-the-art methods by a large margin. In addition, we will release our code on Github.

2 | RELATED WORK

Our proposed method is inspired by existing work in image captioning, change captioning and text-to-image mapping. This section discusses the related work in these three areas.

Image Captioning. Image captioning[7] provides textual a description of visual content in an image and has been extensively studied. The most popular method is the neural-network-based encoder-decoder architecture[8,9,10,11,12]. Vinyals et al.[11] first proposed an image captioning model based on the encoder-decoder framework[13]. This method had difficulty in exploiting all the necessary visual information to produce captions. Xu et al.[14] introduced a spatial attention mechanism to help the model attend to the corresponding image region at each decoding step. Li et al.[15] proposes a very interesting and novel end-to-end learning framework based on CNN for visual understanding, which can seamlessly and simultaneously integrate multi-matrix factorization to significantly improve the performance. Yao et al.[16] proposed a graph convolutional neural network in the image encoding phase to integrate semantic and spatial relationships between objects. Wang et al.[17] presented a Long Short-Term Memory plus Relation-aware pointer network architecture (LSTM-R) which explored geometrical relationship between OCR tokens. Chen et al.[18] proposed a new control signal for Controllable Image Captioning (CIC): Verb-specific Semantic Roles (VSR), and it is the first control signal to consider both event-compatible and sample-suitable requirements. Ji et al.[19] leverage global representation to address the issue of missing objects and relationship bias. In recent work, attention plays a vital role not only in image captioning but also video captioning[20,21,22,23,24,25,26,27,28]. These methods apply visual attention to different spatial regions at each text generation time step. For example, You et al.[23] employed a semantic attention model to combine the visual features with visual concepts in a recurrent neural network that generates the image caption. Pan et al.[21] presented X-Linear Attention Networks to leverage higher order intra and inter-modal interactions. Guan et al.[22] proposed a particular spatial-channel noise attention unit to separate fixed pattern noise and real scene related features, to maximize the noise reduction and preserve details. Chen et al.[24] presented a video captioning framework named Motion Guided Spatial Attention (MGSA), which utilizes optical flow to guide spatial attention. Lei et al.[25] proposed a video captioning model based on channel soft attention and a semantic reconstructor, which considers the global information for each channel. Therefore, to generate fine-grained captions, we introduce spatial-channel attention mechanisms into the decoder in our framework.

Change Captioning. Change captioning combines the subjects of image captioning (discussed above) and change detection. Each of these subjects has been extensively studied in isolation[29,21,20,22]. In change detection, Doi et al.[29] utilized deep graph matching which can estimate pixel-wise changes with object-wise change annotation to detect changed regions. Daudt et al.[30] presented a network architecture to perform land cover mapping, which is then used to predict changes. Such methods however lack the ability to describe, with text, the changes between images. Therefore, change captioning has been proposed to characterise changes between iamges with textual output. Jhamtani et al.[31] proposed a framework named Difference Description with Latent Alignment (DDLA) and an accompanying dataset. The dataset consists of images collected from two video frames at different time steps of a given scene, along with a human textual annotation. In DDLA, the pixel-level difference is calculated between the paired images and then input to the decoder for caption generation. Dong Huk Park et al.[3] presented a Dual Dynamic Attention model (DUDA) and a CLEVR-Change Dataset. DUDA utilized dynamic attention structures to locate changed regions. Then the images before and after the change as well as information capturing their differences are input into the decoder. Qiu et al.[32,33] described changes based on multi-view image information. Hosseinzadeh et al.[34] formulated a training scheme that uses an auxiliary task to improve the training of the change captioning network.

Text-To-Image. Text-to-image is a task to generate images from a linguistic description. Goodfellow et al.[35] introduced the adversarial process into generative models. Generative adversarial networks (GANs) have been widely studied due to their excellent performance and applied to the text-to-image problem. However, a large domain gap exists between images and linguistic descriptions. To maintain the semantic consistency when modeling each modality, most text-to-image methods attempt to constrain the semantics of these two modalities. Qiao et al.[36] proposed a framework named MirrorGAN in which a mirror structure has been embodied. Briefly, this method addresses not only a text-to-image task but also a text-to-image-to-text task. In IR-GAN[37], an adversarial mechanism is used in which the discriminator incrementally determines the consistency between visual and semantic information. Inspired by these two methods, we design a semantic consistency reasoning mechanism to align the visual and linguistic information.

3 | APPROACH

As shown in Figure 2 , we propose a framework for change captioning that consists of bi-directional difference localization, spatial-channel attention, and semantic consistency reasoning. We locate the changed region between the 'before' and the 'after' images (I_{bef} and I_{aft}) with the bi-directional difference localization module. Then we leverage the spatial-channel attention mechanism to select helpful information to generate change captions. Finally, we utilize semantic consistency reasoning to measure the consistency between visual and linguistic information to direct the decoder to generate better captions.

3.1 | Bi-directional Difference Localization

The bi-directional difference localization module takes f_{bef} and f_{aft} as input. f_{bef} and f_{aft} are the feature maps of I_{bef} and I_{aft}. Bidirectional modeling can reduce the confusion of different changes by locating the changes on both the 'before' and 'after' images.

First, in order to capture the semantic difference in the input pairs, we use a bi-directional subtraction to obtain two contrastive feature maps $F_{forward}$ and $F_{backward}$, whereas conventional methods only subtract f_{bef} from f_{aft}:

$$F_{forward} = f_{aft} - f_{bef}, \tag{1}$$

$$F_{backward} = f_{bef} - f_{aft}. \tag{2}$$

Second, to pay more attention to the changed region, we calculate two separate attention maps a_{bef} and a_{aft} by contrastive feature maps $F_{forward}$ and $F_{backward}$. As shown in the following formulas, we not only concatenate visual features f_i and contrastive feature map $F_{forward}$, but we also concatenate visual features f_i and contrastive feature maps $F_{forward}$ and $F_{backward}$ to incorporate the comprehensive change information into visual features. This produces the fused features $F_{i_{one}}$ and $F_{i_{two}}$:

$$F_{i_{one}} = ReLU(conv_{i_{one}}([f_i; F_{forward}]), \tag{3}$$

$$F_{i_{two}} = ReLU(conv_{i_{two}}([f_i; F_{forward}; F_{backward}]). \tag{4}$$

In the fused features $F_{i_{one}}$ and $F_{i_{two}}$, the changing region has stronger semantic information than other regions. We can concatenate $F_{i_{one}}$ and $F_{i_{two}}$ to obtain the corresponding change information contained in these fused features which represent spatial attention maps which have a greater weights in the changing region:

$$F_i = [F_{i_{one}}; F_{i_{two}}], \tag{5}$$

$$a_i = \sigma(conv(F_i)), \tag{6}$$

where $i \in (bef, aft)$. The symbols [;], conv, σ indicate concatenation, convolutional layer, and sigmoid function. Finally, the two attention maps are applied to f_{bef} and f_{aft} to locate the changing region in the images based on the visual features. This produces a weighted visual feature map r_i

$$r_i = a_i \odot f_i, \tag{7}$$

where $\odot$ indicates element-wise multiplication.

Compared to unidirectional difference localization, our bi-directional difference localization module increases the accuracy in change localization, since this module not only provides the changes information on the 'before' images, it also provides the changes on the 'after' images producing comprehensive change information at the spatial level. Furthermore, our localization module greatly reduces confusion of change types, especially 'add', 'move' and 'drop'. The reason is that bi-directional difference localization can increase the representational power of features for all types of change.

3.2 | Spatial-Channel Attention

After the changing regions are located, we further employ a spatial-channel attention module to focus on regions which contribute to word generation. The spatial-channel attention module contains two attention mechanisms: channel attention and spatial attention.

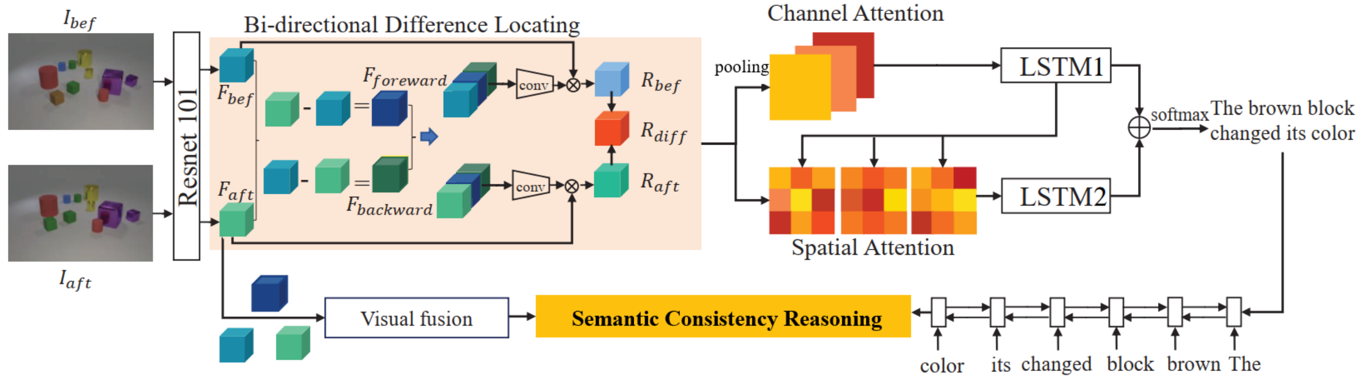

FIGURE 2 Our model consists of three modules: bi-directional difference localization, spatial-channel attention, and semantic consistency reasoning. We firstly locate changes with the bi-directional difference localization module. Then, we employ the decoder with spatial-channel attention to describe the changes. Finally, semantic consistency reasoning is used to measure the consistency between visual and linguistic information.

The purpose of channel attention is to help our model pay more attention to the key feature maps obtained by the bi-directional difference localization module at each decoding stage t. For example, when generating a word to describe the change type "add", the channel attention part should force the spatial part to focus more on the $R_{diff}(R_{diff} = R_{aft} - R_{bef})$ and R_{aft}.

Equally importantly, spatial attention assists our model to focus on different regions in each feature map at each decoding step. When describing the attributes of unchanged objects, our model should attend to the unchanged object rather than pay attention to the changed region all the time.

3.2.1 | Channel Attention

At each decoding stage t, we obtain the channel attention weights for the three feature maps R_{bef}, R_{diff} and R_{aft}. According to Equation (8), first we calculate the sum pooling of each R_i so that the channels' degree of importance can be retained. This helps in the subsequent calculation of the key channel at current decoding stage t,

$$S_i = \sum_{H,W} R_i. \tag{8}$$

As shown in Equation (9), we then fuse S_i to get the latent projection v, with which we can encode the channel attention information.

$$v = ReLU(W_{d_1}[S_{bef}; S_{diff}; S_{aft}] + b_{d_1}). \tag{9}$$

After we get the latent projection v, we take v and sum of the hidden states $h_1^{(t-1)}$, $h_2^{(t-1)}$ as input u^t to an LSTM to get the hidden state $h_0^{(t)}$,

$$u^{(t)} = [v; (h_2^{(t-1)} + h_1^{(t-1)})], \tag{10}$$

$$h_0^{(t)} = LSTM(u^{(t)}, h_0^{(t-1)}), \tag{11}$$

where $i \in (bef, diff, aft)$, $h_1^{(t-1)}$ and $h_2^{(t-1)}$ are the hidden states of LSTM1 and LSTM2 at the decoding stage $t-1$. As shown in Figure 2 , LSTM1 is used to decode the channel attention information. LSTM2 acts as a spatial decoder to decode the visual information which has been enhanced by spatial attention. The information provided by $h_1^{(t-1)}$ and $h_2^{(t-1)}$ enables the LSTM to focus on the key feature maps at current decoding stage t, thus we could obtain the hidden state $h_0^{(t)}$ which contains the channel attention information. Further we obtain the comprehensive weighted feature $L_{com}^{(t)}$ at step t using the channel attention weights predicted from $h_0^{(t)}$ according to Equations (12) and (13):

$$\alpha_c^{(t)} \sim softmax(W_{d_2} h_0^{(t)} + b_{d_1}), \tag{12}$$

$$L_{com}^{(t)} = \alpha_{bef}^t S_{bef} + \alpha_{aft}^t S_{aft} + \alpha_{diff}^t S_{diff}. \tag{13}$$

Finally, to help compute the spatial weights via channel attention, we utilize LSTM1 to decode the comprehensive feature and the embedding of the previous word $w_{(t-1)}$ (ground-truth word during training, predicted word during inference):

$$x^{(t-1)} = E\mathbf{1}_{w_{(t-1)}}, \tag{14}$$

$$c^{(t)} = [x^{(t-1)}; L_{com}^{(t)}], \tag{15}$$

$$h_1^{(t)} = LSTM1(c^{(t)}, h_1^{(t-1)}), \tag{16}$$

where $h_1^{(t)}$ represents the hidden states of LSTM1 at decoding stage t, used in part to generate the distributions over the next word. W_{d_1}, b_{d_1}, W_{d_2} and b_{d_2} are learnable parameters. The $\mathbf{1}_{w_{(t-1)}}$ is the one-hot encoding of the word $w_{(t-1)}$ and E is an embedding layer.

3.2.2 | Spatial Attention

We calculate the spatial attention with the output of channel attention $h_1^{(t)}$ and the previous hidden state of LSTM2 $h_2^{(t-1)}$. First, we concatenate $h_1^{(t)}$ and $h_2^{(t-1)}$ as well as the embedding word and get the spatial attention tensor A_{SF}^t:

$$A_{SF}^t = [h_1^{(t)}; w_{(t-1)}; h_2^{(t-1)}], \tag{17}$$

where the hidden state $h_1^{(t)}$ plays the role of providing the channel attention information for calculating the spatial attention. Therefore, A_{SF}^t brings the channel information into the calculation of spatial attention. Second, we calculate each changed visual feature r's spatial attention map A_i using A_{SF}^t to predict the spatial attention weight $\alpha_{S_i}^{(t)}$ via a softmax:

$$A_i = \frac{A_{SF}^t r_i}{||A_{SF}^t|| \cdot ||r_i||}, \tag{18}$$

$$\alpha_{S_i}^{(t)} \sim softmax(A_i), \tag{19}$$

hence the weighted feature maps can be obtained by the sum pooling of element-wise multiplication between the spatial attention weight and the corresponding visual feature:

$$A_{I_i}^{(t)} = \sum_{H,W} \alpha_{S_i}^{(t)} \odot r_i, \tag{20}$$

where $i \in (bef, diff, aft)$. Next, all the weighted feature maps are concatenated together which will act as one part of LSTM2's input for getting the representation of spatial attention in hidden space, to ensure that the computation of spatial attention has the time continuous property as word generation. The hidden state of LSTM2 is also used in part to generate the distributions over the next word,

$$S_I = ReLU(W_{d_S}[A_{I_{bef}}^{(t)}; A_{I_{diff}}^{(t)}; A_{I_{aft}}^{(t)}] + b_{d_S}), \tag{21}$$

$$S^{(t)} = [w_{(t-1)}; S_I], \tag{22}$$

$$h_2^{(t)} = LSTM2(s^{(t)}, h_2^{(t-1)}), \tag{23}$$

where W_{d_S}, b_{d_S} are learnable parameters.

As above mentioned, the word generation at each step includes the comprehensive features from channel attention and spatial attention. Therefore the probability distribution of the next word can be calculated via softmax:

$$w^{(t)} \sim Softmax(h_1^{(t)} + h_2^{(t)}), \tag{24}$$

where W_w and b_w are learnable parameters.

3.3 | Semantic Consistency Reasoning

To ensure the consistency of textual and visual information, we propose a semantic consistency reasoning mechanism. This mechanism calculates a relevance score between the visual information and the text description, thereby generating more accurate captions. Further, due to the constraints among before and after the visual information change , generated captions will be more robust to describe changes. The relevance score can be calculated as:

$$r = R(f_i, D_I), \tag{25}$$

where $f_i \in (f_{bef}, f_{aft})$. D_I is the caption to describe the changes between the input pair I_{bef} and I_{aft}. First, we fuse the three feature maps, which are the feature maps used in spatial-channel attention module, to generate captions:

$$C_I = W_{C_I}[\sum_{H,W} f_{bef}; \sum_{H,W} F_{forward}; \sum_{H,W} f_{aft}] + b_{C_I}, \tag{26}$$

where W_{C_I} and b_{C_I} are learnable parameters, $F_{forward}$ is predicted from Equation (1). For the captioning information, we employ a bi-directional GRU to encode the caption:

$$d_I = BiGRU(w_t^1, w_t^2, ..., w_t^n), \tag{27}$$

where w_t^1, w_t^2,..., w_t^n denote the embedding of each word in D_I, hence we obtain d_I which represents the semantic information of the corresponding caption. Finally, to measure relevance between the visual information and the textual description, we leverage semantic consistency reasoning to align the fused feature and caption:

$$r = R(f_i, D_I) = d_I C_I + \phi(C_I). \tag{28}$$

The first term in Equation (28) is the inner product of d_I and C_I which represents the degree of relevance, in accordance with our intuition that similar features' inner product will be larger. ϕ is a fully connected layer which projects C_I into a scalar.

3.4 | Optimization

Similar to the image captioning, we train our model end-to-end with a word-level cross-entropy loss (XE) which minimizes the distance L between the generated sequence W_I and the ground truth sequence W_I^* :

$$L_{XE}(\theta) = log(p_\theta(W_I^*|W_I)), \tag{29}$$

where θ denotes all the parameters in our model. Besides the cross-entropy loss, we apply a regularization to the attention maps generated by the bi-directional difference localization module to minimize unnecessary activations. In addition, we use a consistency loss calculated by the relevance score to maximize the consistency between visual information and generated captions. The final loss function is as follows:

$$L(\theta) = L_{XE} + \lambda L_1 + \mu L_{consistency}, \tag{30}$$

where

$$L_{consistency} = L_{consistency}^{fake} + L_{consistency}^{real}, \tag{31}$$

$$L_{consistency}^{fake} = -\mathbb{E}[min(0, -R(F_i, D_I)) - 1], \tag{32}$$

$$L_{consistency}^{real} = -\mathbb{E}[min(0, R(F_i, D_I^{GT})) - 1]. \tag{33}$$

4 | EXPERIMENTS

4.1 | Experimental settings

Evaluation Metrics. We evaluate the performance of our model using CIDEr[38], BLEU[39], ROUGE[40], METEOR[41] and SPICE[42]. CIDEr is voting-based metric which penalizes the often-seen but uninformative n-grams in the dataset. BLEU is the metric for computing the n-gram based precision between the candidate sentence and reference sentences. In BLEU metrics, we mainly calculate the popular BLEU-4 metric to evaluate the matching degree of 4-word sub-sequences between the generated

TABLE 1 Total performance comparison of our model on the CLEVR-Change Dataset. Our model outperforms all other methods.

Approach	CIDEr	BLEU-4	METEOR	SPICE
Capt-Pix-Diff	75.9	30.2	23.7	17.1
Capt-Rep-Diff	87.9	33.5	26.7	19
Capt-Att	106.4	42.7	32.1	23.2
Capt-Dual-Att	108.5	43.5	32.7	23.4
DUDA	112.3	47.3	33.9	24.5
Ours	**118.1**	**54.2**	**38.3**	**31.7**

caption and the ground truth caption. ROUGE is also n-gram based, which computes recall between candidate sentence and reference sentences. In our experiments, we adopt the ROUGEL metric which computes F-measure with a recall bias according to longest common subsequence between candidate sentence and reference sentences. Similar to ROUGEL, METEOR also computes the F-measure based on matches, and returns the maximum score among references. SPICE parses each caption to derive a scene graph and computes the F-SCORE to measure the scene graph's similarity. Further, we utilize the Pointing Game evaluation[43] to validate the ability of our model to accurately locate the changed regions.

Implementation Details. The spatial resolution of visual features f is 1024*14*14. The features are obtained from the convolutional layer before the global average pooling in ResNet101[44] pre-trained on ImageNet[45]. The hidden state dimension of the LSTM in our spatial-channel attention module is 512. The word embedding layer is trained from scratch and each word is represented by a 300 dimensional vector. The hidden state dimension of Bi-GRU which is used to encode the captions is 1024. We train our model for 80 epochs using the Adam Optimizer[46] with a learning rate of 0.0007 and a batch size of 128. The hyperparameter for our consistency loss is 0.025 and for regularization terms is 0.0025. Our model is implemented using PyTorch[47], and our code will be made publicly available.

4.2 | Results on CLEVR-Change Dataset

The CLEVR-Change Dataset[48] consists of 39803 images before the change, 39803 images for the scene change (add, move, drop, color, texture) and 39803 for the distractors (viewpoint change, illumination change) without scene change. The dataset is split into 33830*3, 1988*3, and 3985*3 for training, validation, and testing.

To evaluate the effectiveness of our model on the CLEVR-Change Dataset, we compare it with several models.The Capt-Pix-Diff[3] first utilizes a pixel-level difference image which is obtained by calculating the difference of the corresponding pixels in the two input images. Then we downsample the spatial resolution of the difference image to the same size as the features of the original images. Finally, the downsampled difference image is concatenated with the input image features, and then is input to the LSTM for caption generation. Capt-Rep-Diff[3] is similar to Capt-Pix-Diff, the difference is that Capt-Rep-Diff utilizes a feature-level difference image to describe changes. Capt-Att[3] introduces a single spatial attention into Capt-Rep-Diff, and applies the learned single spatial attention to the two images. Compared to Capt-Att, Capt-Dual-Att[3] learns two separate spatial attentions for the paired images. Dual Dynamic Attention Model (DUDA)[3] is currently the state-of-the-art model for change captioning and includes a dual attention module to locate the change regions and a dynamic speaker for caption generation.

We compare the performance of our model with the above approaches in several respects: (1) total performance (2) scene change and distractor and (3) change types.

Total performance. We evaluate the total performance of our model for three aspects: the accuracy of generated captions(Table 1), robustness for viewpoint shift (Figure 4 (a)) and accuracy of change localization in different viewpoints (Figure 4 (b)). We observe that our model outperforms DUDA.

As for the accuracy of change localization, we upsample the attention maps to the original image size and then check whether the point with the highest activation is in the ground truth bounding box[43]. We observe good performance of our model, as shown in Figure 4 (b). We owe the performance improvement to the bi-directional difference localization module which identifies differences for all change types.

For the robustness to viewpoint shift, in Figure 4 (a), we can observe that the performance of our model not only outperforms DUDA, it is also more stable than DUDA. We speculate the reason is that the spatial-channel attention module helps our model

TABLE 2 Total performance comparison of our model on the CLEVR-Change Dataset. Our model outperforms all other methods.

Approach	Scene Change				Distrator			
	CIDEr	BLEU-4	METEOR	SPICE	CIDEr	BLEU-4	METEOR	SPICE
Capt-Pix-Diff	36.2	21.9	17.7	7.9	98.2	43.4	38.9	26.3
Capt-Rep-Diff	51.8	26.0	21.1	10.1	105.3	49.4	41.7	27.8
Capt-Att	87.2	38.3	27.9	18.0	106.6	53.5	43.2	28.4
Capt-Dual-Att	89.8	38.5	28.5	18.2	108.9	56.3	44.0	28.7
DUDA	94.6	42.9	29.7	19.9	110.8	59.8	45.2	29.1
Ours	**105.8**	**52.1**	**34.3**	**28.5**	**116.1**	**62.7**	**50.3**	**34.8**

TABLE 3 Evaluation for every change type on the CLVER-Change Dataset.

Approach	Color			Material			Drop			Add			Move		
	C	M	S	C	M	S	C	M	S	C	M	S	C	M	S
Capt-Pix-Diff	4.2	7.4	1.3	16.1	16.0	6.8	27.1	20.9	10.6	30.1	24.4	11.4	18.0	18.2	9.2
Capt-Rep-Diff	44.5	19.2	8.2	21.9	18.2	8.8	49.7	23.5	12.0	50.1	25.7	12.1	26.5	18.9	9.6
Capt-Att	112.1	30.5	17.9	75.9	25.4	16.3	98.4	31.2	19.0	91.5	30.2	19.0	49.6	22.2	14.5
Capt-Dual-Att	115.8	32.1	19.8	82.7	26.7	17.6	103.0	31.7	16.9	85.7	29.5	16.9	52.6	22.4	14.7
DUDA	**120.4**	32.8	21.2	86.7	27.3	18.3	103.4	31.4	22.4	108.2	33.4	22.4	56.4	23.5	15.4
Ours	115.9	**36.1**	**28.9**	**106.8**	**31.6**	**26.2**	**124.9**	**36.8**	**32.0**	**121.3**	**38.7**	**32.0**	**71.8**	**28.1**	**23.5**

utilize more useful information to for word decoding at each step, and reduces the influence of location error generated in localization module to some extent.

Finally, the semantic consistency reasoning mechanism further aligns the visual information and text captions which helps our model generate a more accurate output. Therefore, our method outperforms competing techniques as shown in Table 1 .

Scene Change and Distrators. In Table 2 , we observe a larger performance improvement for scene change than distractors. This is due to bi-directional difference localization, which is more effective at distinguishing different scene change types. Compared to the unidirectional difference, the bi-directional difference localization increases the information difference for all change types.

Change Types. The performance in Table 3 demonstrates that our model surpasses DUDA in almost all change types. C, M and S in the table refer to CIDEr, METEOR and SPICE respectively. In particular, the relative improvement of the 'move' change is surprising. The reason is that for the 'move' change, we need to overall analyze the different change localization regions in the 'before' and 'after' images. Further, the 'move' change is more easily confused with the 'add' and 'drop' changes. In this case, if the model lacks the ability to systematically analyze the information in the 'before' image and the 'after' image, it will be difficult to identify the 'move' change. However, the semantic consistency reasoning in our model can enhance the ability to systematically analyze visual information by ensuring the consistency between captions and images. Additionally, the bidirectional difference localization module reduces the confusion by increasing the information difference for all change types.

We show the qualitative comparison of the 'move' change in Figure 3 . We observe that our model can well locate the different change regions in the 'before' and 'after' images compared to DUDA. Less attention to the position of change object in 'before' images in DUDA accounts for the confusion between the 'move' and 'add' change. In contrast, our model pays attention to the two change regions equally through bi-directional difference localization which proves the ability of our model to recognize various change types.

4.3 | Results on Image Editing Request Dataset

To validate the usefulness of our model for different tasks, we conduct further experiments on the Image Editing Request Dataset[2]. This dataset consists of 3939 images which are split into 3061, 383, and 495 for training, validation, and testing. The changes in this dataset are more diverse including scene changes and more comprehensive change. Therefore, it is more difficult to accurately caption the changes in this dataset.

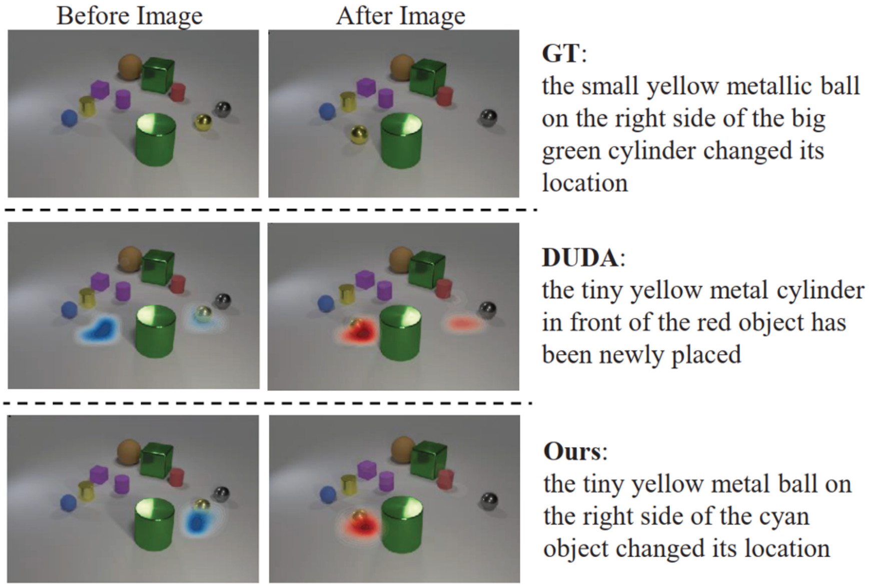
Before Image
After Image
GT: the small yellow metallic ball on the right side of the big green cylinder changed its location
DUDA: the tiny yellow metal cylinder in front of the red object has been newly placed
Ours: the tiny yellow metal ball on the right side of the cyan object changed its location

FIGURE 3 Qualitative comparison of the 'move' change

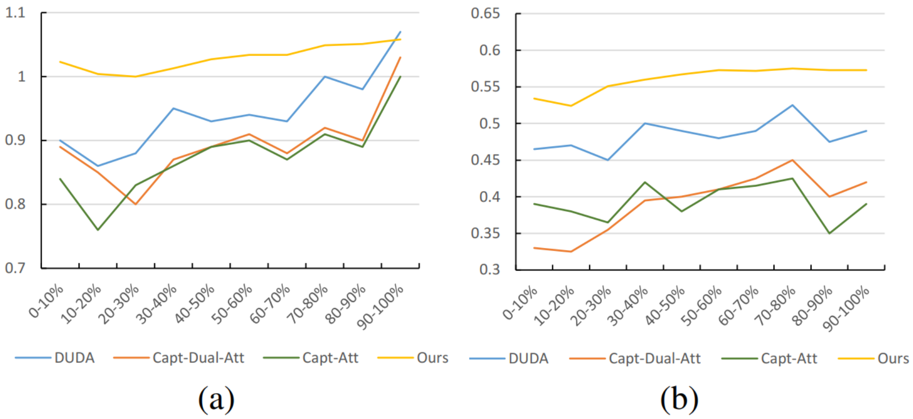

FIGURE 4 a: Evaluation for robustness to viewpoint shift. Horizontal axis is IOU and the normal axis is CIDEr score. b: Location accuracy evaluation in different viewpoints.

TABLE 4 Total performance comparison on Image Editing Request Dataset.

Approach	CIDEr	BLEU-4	ROUGEL	METEOR
static+dynamic rel-att	26.36	6.72	37.35	12.80
DUDA	22.82	6.54	37.29	12.40
Ours	**27.68**	**6.89**	**38.51**	**14.62**

The performance comparison in Table 4 indicates that although the task in this dataset is different from scene change, our model also is effective. In Figure 5 , compared to DUDA which locates the wrong region, our model can locate the background, which changes in these pictures, and then describes the visual change precisely. We speculate that since DUDA lacks the bi-directional difference localization of our model, it does not accurately identify the change. Further, the spatial channel attention also helps our model learn to focus on the other information related to changes, and accounts for the accurate description. Also, the semantic consistency reasoning's constraint between the visual information and textual description also contributes to a description that accurately matches the visual content.

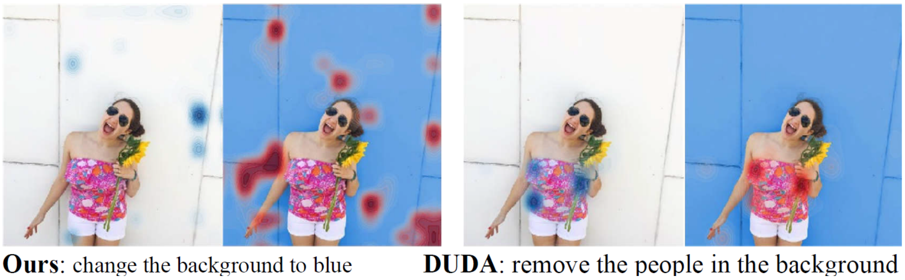
Ours: change the background to blue
DUDA: remove the people in the background

FIGURE 5 Qualitative comparison on Image Editing Request Dataset.

TABLE 5 Total performance comparison on the Spot-the-Diff Dataset.

Approach	CIDEr	BLEU-4	ROUGEL	METEOR
DDLA	32.8	**8.5**	28.6	**12.0**
DUDA	32.5	8.1	29.1	11.8
Ours	**42.2**	6.6	**29.5**	10.6

4.4 | Results on Spot-the-Diff Dataset

The Spot-the-Diff dataset[31] possesses one or more scene changes with real images and human-provided descriptions. We evaluate our model in a single change setting by picking out the images with one caption in the dataset. The images are split into 3852, 653, and 789 for training, validation, and testing.

We present the comparison on the Spot-the-Diff dataset in Table 5 . We notice that our model achieves a performance gain of +9.7 compared to the DUDA and +9.4 compared to the DDLA on the CIDEr metric. CIDEr is a widely recognized evaluation metric in the task of image captioning. On the BLEU-4 metric, it is important to give each caption four reference captions to reduce the impact of language diversity. However, we select the images with a single description to evaluate our model which results in the low performance on the BLEU-4 score to some extent. Besides, as Dong Huk Park et al.[3] mentioned, the Spot-the-Diff dataset shouldn't be the definitive test for the change captioning task because the dataset doesn't consider the presence of distractors, which means more experiments need to be set up for verifying captioning effectiveness.

4.5 | Ablation Study

We conduct experiments on different variants of our model to evaluate the effectiveness of each of the model's components including the bi-directional difference localization, spatial-channel attention, and semantic consistency reasoning, which are abbreviated as bidiff, attention and reasoning in Table 6 . To evaluate the performance improvement from bi-directional difference localization module, we implement a degraded model which includes spatial-channel attention, semantic consistency reasoning and unidirectional difference localization. All ablation studies are performed on the CLEVR-Change dataset.

As seen in Table 6 , thanks to the bi-directional difference localization, our model achieves a performance improvement of +2.1 in the CIDEr score. The CIDEr scores in the third and fourth rows prove that our spatial-channel attention contributes a performance improvement with +0.4. In addition, compared to DUDA, we observe that our semantic consistency reasoning module boosts the CIDEr score by +3.6, which makes the descriptions generated by our model more consistent with the images. We also observed a performance decline of 0.3 on the SPICE metric with spatial-channel attention module, and we speculate that the combination of the bi-directional difference localization and spatial-channel attention is the reason. When the bi-directional difference localization and spatial-channel attention work together, a more precise description will be generated so the arbitrariness of potential description will decrease. This means that under the F-score based metric, such as the SPICE metric, a more accurate description may result in a lower score. Finally, the model with three modules outperforms the model only with spatial-channel attention and semantic consistency reasoning.

TABLE 6 Ablation experimental results of different variants conducted on CLEVR-Change Dataset.

Approach	CIDEr	METEOR	SPICE
DUDA	112.3	33.9	24.5
bidiff	114.4	37.9	**32.0**
reasoning	115.9	37.7	31.5
attention+reasoning	116.3	38.0	31.6
bidiff+reasoning	117.0	38.2	**32.0**
bidiff+reasoning+attention	**118.1**	**38.3**	31.7

5 | CONCLUSION

In this paper, we propose the bi-directional difference localization and semantic consistency reasoning network to address the change captioning task. We first employ the bi-directional difference localization module to locate changes which can well distinguish all change types. Second, we design a decoder with spatial-channel attention to generate the change caption. This attention module enables our model to not only focus on key feature maps but also attend to the core positions of each feature map at each decoding stage. Finally, we introduce semantic consistency reasoning to enhance the ability of our model to locate the change regions and systematically analyze both features by reasoning the relationship between text captions and images. Extensive experiments on three public datasets show the effectiveness of our approach.

References

1. Oluwasanmi A, Aftab MU, Alabdulkreem E, Kumeda B, Baagyere EY, Qin Z. *CaptionNet: Automatic end-to-end siamese difference captioning model with attention*. *IEEE Access* **2019**; *7*: *106773–106783*.

2. Tan H, Dernoncourt F, Lin Z, Bui T, Bansal M. *Expressing visual relationships via language*. *arXiv preprint arXiv:1906.07689* **2019**.

3. Park DH, Darrell T, Rohrbach A. *Robust Change Captioning*. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ; **2019**: *4623–4632*.

4. Oluwasanmi A, Frimpong E, Aftab MU, Baagyere EY, Qin Z, Ullah K. *Fully convolutional captionnet: Siamese difference captioning attention model*. *IEEE Access* **2019**; *7*: *175929–175939*.

5. Zechao , Li , Jing , et al. *Robust Structured Subspace Learning for Data Representation.*. *IEEE transactions on pattern analysis and machine intelligence* **2015**.

6. Yu J, Yang Y, Murtagh F, Gao X. *Fine-grained visual understanding and reasoning*. *Neurocomputing* **2020**; *398*: *408–410*. doi: 10.1016/j.neucom.2019.07.055

7. Ben H, Pan Y, Li Y, et al. *Unpaired Image Captioning with Semantic-Constrained Self-Learning*. *IEEE Transactions on Multimedia* **2021**: 1-1. doi: 10.1109/TMM.2021.3060948

8. Liu D, Zha ZJ, Zhang H, Zhang Y, Wu F. *Context-Aware Visual Policy Network for Sequence-Level Image Captioning*. In: MM '18. Association for Computing Machinery; **2018**; New York, NY, USA: *1416–1424*

9. Donahue J, Anne Hendricks L, Guadarrama S, et al. *Long-term recurrent convolutional networks for visual recognition and description*. In: **2015** Proceedings of the IEEE conference on computer vision and pattern recognition: *2625–2634*.

10. Bin Y, Yang Y, Shen F, Xu X, Shen HT. *Bidirectional long-short term memory for video description*. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Proceedings of the 24th ACM international conference on Multimedia; **2016**: *436–440*.

11. Vinyals O, Toshev A, Bengio S, Erhan D. *Show and tell: A neural image caption generator*. In: **2015** Proceedings of the IEEE conference on computer vision and pattern recognition: *3156–3164*.

12. Wang J, Cao Z, Xiao Y, Qi X. *Supervised guiding long-short term memory for image caption generation based on object classes*. In: . *10609*. International Society for Optics and Photonics. ; **2018**: 106090P.

13. Cho K, Van Merriënboer B, Gulcehre C, et al. *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. *arXiv preprint arXiv:1406.1078* **2014**.

14. Xu K, Ba J, Kiros R, et al. *Show, attend and tell: Neural image caption generation with visual attention*. In: PMLR. *International conference on machine learning*; **2015**: *2048–2057*.

15. Li Z, Tang J, Mei T. *Deep Collaborative Embedding for Social Image Understanding*. *IEEE Transactions on Pattern Analysis Machine Intelligence* **2018**: *1-1*.

16. Yao T, Pan Y, Li Y, Mei T. *Exploring visual relationship for image captioning*. In: **2018** Proceedings of the European conference on computer vision (ECCV): *684–699*.

17. Wang J, Tang J, Yang M, Bai X, Luo J. *Improving OCR-Based Image Captioning by Incorporating Geometrical Relationship*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; **2021**: *1306–1315*.

18. Chen L, Jiang Z, Xiao J, Liu W. *Human-like Controllable Image Captioning with Verb-specific Semantic Roles*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; **2021**: *16846–16856*.

19. Ji J, Luo Y, Sun X, et al. *Improving image captioning by leveraging intra-and inter-layer global representation in transformer network*. In: . *35*. Proceedings of the AAAI Conference on Artificial Intelligence; **2021**: *1655–1663*.

20. Wang J, Wang W, Wang L, Wang Z, Feng DD, Tan T. *Learning visual relationship and context-aware attention for image captioning*. *Pattern Recognition* **2020**; *98*: *107075*.

21. Pan Y, Yao T, Li Y, Mei T. *X-linear attention networks for image captioning*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; **2020**: *10971–10980*.

22. Guan J, Lai R, Xiong A, Liu Z, Gu L. *Fixed pattern noise reduction for infrared images based on cascade residual attention CNN*. *Neurocomputing* **2020**; *377*: *301–313*.

23. You Q, Jin H, Wang Z, Fang C, Luo J. *Image captioning with semantic attention*. In: Proceedings of the IEEE conference on computer vision and pattern recognition; **2016**: *4651–4659*.

24. Chen S, Jiang YG. *Motion guided spatial attention for video captioning*. In: . *33*. Proceedings of the AAAI conference on artificial intelligence; **2019**: *8191–8198*.

25. Lei Z, Huang Y. *Video Captioning Based on Channel Soft Attention and Semantic Reconstructor*. *Future Internet* **2021**; *13*(2): *55*.

26. Cong R, Lei J, Fu H, et al. *An iterative co-saliency framework for RGBD images* . *IEEE Transactions on Cybernetics* **2019**; *49*(1): *233–246*.

27. Zhang Q, Cong R, Li C, et al. *Dense attention fluid network for salient object detection in optical remote sensing images* . *IEEE Transactions on Image Processing* **2021**; *30*: *1305-1317*.

28. Cong R, Lei J, Fu H, Huang Q, Cao X, Hou C. *Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation* . *IEEE Transactions on Image Processing* **2018**; *27*(2): *568–579*.

29. Doi K, Hamaguchi R, Iwase S, et al. *Epipolar-Guided Deep Object Matching for Scene Change Detection*. *arXiv preprint arXiv:2007.15540* **2020**.

30. Daudt RC, Le Saux B, Boulch A, Gousseau Y. *Multitask learning for large-scale semantic change detection*. *Computer Vision and Image Understanding* **2019**; *187*: *102783*.

31. Jhamtani H, Berg-Kirkpatrick T. *Learning to describe differences between pairs of similar images*. *arXiv preprint arXiv:1808.10584* **2018**.

32. Qiu Y, Satoh Y, Suzuki R, Iwata K, Kataoka H. *3D-Aware Scene Change Captioning From Multiview Images*. *IEEE Robotics and Automation Letters* **2020**; *5*(3): *4743–4750*.

33. Qiu Y, Satoh Y, Suzuki R, Iwata K, Kataoka H. *Indoor Scene Change Captioning Based on Multimodality Data*. *Sensors* **2020**; *20*(17): *4761*.

34. Hosseinzadeh M, Wang Y. *Image Change Captioning by Learning From an Auxiliary Task*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; **2021**: *2725–2734*.

35. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. *Generative adversarial networks*. *arXiv preprint arXiv:1406.2661* **2014**.

36. Qiao T, Zhang J, Xu D, Tao D. *Mirrorgan: Learning text-to-image generation by redescription*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; **2019**: *1505–1514*.

37. Liu Z, Deng J, Li L, et al. *IR-GAN: Image Manipulation with Linguistic Instruction by Increment Reasoning*. In: Proceedings of the 28th ACM International Conference on Multimedia; **2020**: *322–330*.

38. Vedantam R, Lawrence Zitnick C, Parikh D. *Cider: Consensus-based image description evaluation*. In: Proceedings of the IEEE conference on computer vision and pattern recognition; **2015**: *4566–4575*.

39. Papineni K, Roukos S, Ward T, Zhu WJ. *Bleu: a method for automatic evaluation of machine translation*. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics; **2002**: 311–318.

40. Lin CY. *Rouge: A package for automatic evaluation of summaries*. In: Text summarization branches out,**2004**: *74–81*.

41. Banerjee S, Lavie A. *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,**2005**: *65–72*.

42. Anderson P, Fernando B, Johnson M, Gould S. *Spice: Semantic propositional image caption evaluation*. In: Springer. European conference on computer vision; **2016**: *382–398*.

43. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S. *Top-down neural attention by excitation backprop*. *International Journal of Computer Vision,**2018***; *126*(10): *1084–1102*.

44. He K, Zhang X, Ren S, Sun J. *Deep residual learning for image recognition*. In: Proceedings of the IEEE conference on computer vision and pattern recognition,**2016**: *770–778*.

45. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. *Imagenet: A large-scale hierarchical image database*. In: Ieee. **2009** IEEE conference on computer vision and pattern recognition: *248–255*.

46. Da K. *A method for stochastic optimization*. *arXiv preprint arXiv:1412.6980* **2014**.

47. Paszke A, Gross S, Chintala S, et al. *Automatic differentiation in pytorch*. **2017**.

48. Johnson J, Hariharan B, Van Der Maaten L, Fei-Fei L, Lawrence Zitnick C, Girshick R. *Clevr: A diagnostic dataset for compositional language and elementary visual reasoning*. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,**2017**: *2901–2910*.

49. Crispell D, Mundy J, Taubin G. *A variable-resolution probabilistic three-dimensional model for change detection*. *IEEE Transactions on Geoscience and Remote Sensing,**2011***; *50*(2): *489–500*.

50. Huertas A, Nevatia R. *Detecting changes in aerial views of man-made structures*. *Image and Vision Computing,**2000***; *18*(8): *583–596*.

How to cite this article: Author One, Author One, Author One, Author One, and Author One (2021), Bi-directional Difference Locating and Semantic Consistency Reasoning for Change Captioning , *Q.J.R. Meteorol. Soc.*, *2017;00:1–6*.