I3N: Intra- and Inter-representation Interaction Network for Change Captioning

Shengbin Yue, Yunbin Tu, Liang Li, Ying Yang, Shengxiang Gao, Zhengtao Yu

***Abstract*—Change captioning aims to describe the disagreement of image pairs with a linguistic sentence. Compared with single image captioning, change captioning requires not only understanding the fine-grained information of each image, but also determining whether change occurs and further representing the differences of image pairs. Although much progress has been made, it remains a severe challenge of the precise difference representation in the distraction of viewpoint change, especially that of tiny difference. In this paper, we propose a novel Intra- and Inter-representation Interaction Network (I3N) to learn the fine difference representation and be immune to viewpoint change. In the Intra-representation Interaction stage, we design Geometry-Semantic Interaction Refining (GSIR) to explore the positional and semantic interactions of intra-image, which can be a prior knowledge of enduring viewpoint change and reinforce the cognition of semantic change. In the Inter-representation Interaction stage, to endow the model with the capability of pinpointing the latent difference in viewpoint change, Hierarchical Representation Interaction (HRI) models difference from coarse to fine representations through the Semantic Matcher and Change Amplifier module. The proposed approach outperforms the state-of-the-art methods with an encouraging performance on the existing change captioning benchmarks. Our code is available at https://github.com/yueshengbin/I3N.**

***Index Terms*—Change Captioning, Intra- and Inter- representation Interaction, Geometry-Semantic Interaction Refining, Hierarchical Representation Interaction.**

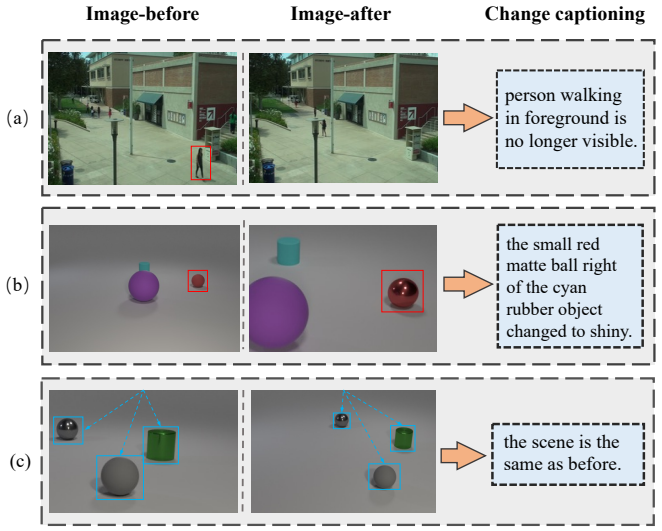


Fig. 1. Examples of the change captioning on CLEVR-Change and Spot-the-Diff datasets. This task aims to generate a natural language sentence to detail what has changed in a pair of images. (a) Disappearance of a person in a full-aligned image pair; (b) Slight movement of the tiny sphere in the presence of viewpoint change; (c) None-scene change of the object in the presence of viewpoint change.

I. Introduction

CHange captioning is a novel and challenging task since it requires comprehending the contents of the image pairs and further using a natural language sentence to summarize their disagreement. Unlike single image captioning [1]–[5], change captioning emphasizes the semantic change of pairs (*e.g.*, moving an object) and the greater challenge is to judge and further represent the semantic change in the distraction of illumination or viewpoint change. It has widespread applications, such as automatically generating reports about aerial imagery [6], [7], fault detection [8], [9], video surveillance [10], [11], and medical diagnosis [12], [13].

Thanks to the rapid development of image captioning [14]–[17], the existing change captioning methods have made significant inspiration. Pioneer works [18], [19] captioned the change on full-aligned image pairs without the distraction of illumination or viewpoint change as shown in Fig. 1 (a). However, as shown in the Fig. 1 (b), (c), there are numerous image pairs shot from different angles of viewpoint in a dynamic environment, where two images are not fully aligned. In this situation, the above works are unable to learn a difference representation explicitly [20], [21]. To address this limitation, Park *et al*. [20] first proposed to localize the change of the unaligned image pairs with consideration of viewpoint change, where they directly subtracted the images features to construct the difference representation. Nevertheless, this coarse way brings much noise to the obtained difference representation. Some recent endeavors [21]–[28] propose methods to handle the irrelevant change actively, which have boosted the performance to a certain extent. Specifically, on the one hand,

The work was supported by the National Natural Science Foundation of China (Grant Nos. 61972186, 61732005, U21B2027), Yunnan high-tech industry development project (Grant No. 201606), Yunnan provincial major science and technology special plan projects (Grant No. 202103AA080015, 202002AD080001-5), Yunnan Basic Research Project (Grant Nos. 202001AS070014), and Reserve Talents for Academic and Technological Leaders in Yunnan Province (Grant No. 202105AC160018). (*Corresponding author: Shengxiang Gao.*)

Shengbin Yue, Shengxiang Gao and Zhengtao Yu are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China, and also with the Yunnan Provincial Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China (e-mail: yueshengbin@foxmail.com; gaoshengxiang.yn@foxmail.com; ztyu@hotmail.com).

Yunbin Tu is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: tuyunbin22@mails.ucas.ac.cn).

Liang Li is with Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: liang.li@ict.ac.cn)

Ying Yang is with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China (e-mail: yangying98@foxmail.com)

some works [22], [24], [28] introduce a number of auxiliary strategies to improve the performance of the primary network. On the other hand, these works [21]–[23], [25]–[27] focus on improvements of the model structure.

Despite the advances, there are still two limitations in 1) the fine-grained interaction of intra-image in pairs; 2) the explicit difference representation learning of inter-images in the condition of viewpoint change. First, for the representation of intra-image: most approaches directly employ the object information extracted from CNN to learn change, ignoring the exploration of semantic and positional interactions. As shown in Fig. 1 (c), there is only viewpoint change in the scene and it appears that the locations of three objects have changed. This often makes the model confused in distinguishing semantic and viewpoint change. In fact, the semantic interaction and relative location of these objects do not change. Therefore, learning an intra-representation that integrates the position and semantic interaction is beneficial to differentiate the viewpoint and semantic change. In addition, embedding the absolute position of each feature improves the model's perception of position change (*e.g.*, movement). Second, the reliable difference representation in the pairs depends on a fine-grained interaction of inter-image. In the pairs with viewpoint shifting, due to most areas being unchanged, the semantic change would be overwhelmed by the majority of unchanged parts. Especially, this situation is obvious for the small object or tiny change as shown in Fig. 1 (b). The model will misidentify that the two images are well-matched. To capture this semantic change in the interacting process, an effective amplifier is needed to reveal the latent semantic change from the common parts to help the model describe accurate and fine differences.

In this paper, we propose an Intra- and Inter-representation Interaction Network (I3N), which explicitly models a reliable difference representation via exploiting the intra- and inter-representation interaction of the image pair. Specifically, we first design a novel Geometry-Semantic Interaction Refining (GSIR) to tailor the features of the "before" and "after" images. GSIR models the semantics and position interaction between the object features, and this comprehensive interaction is regarded as a priori knowledge to handle viewpoint change. Besides, GSIR integrates absolute geometric information to strengthen the cognition of position semantic change. Further, we propose the ingenious Hierarchical Representation Interaction (HRI) consisting of a Semantic Matcher and a Change Amplifier module, which aims to accurately learn difference representation from coarse to fine. First, the Semantic Matcher module roughly matches common features between the "before (after)" and "after (before)" images, then the Change Amplifier regards the "after (before)" image as the source of amplification and makes the latent change salient by refining the common features. Note that the entire process of HRI is based on the corresponding features and interactions (priors) in the "before (after)" and "after (before)" images. Next, we gain the final difference representation by fusing the bi-direction difference features. Finally, Dual Change Navigator locates the difference in the "before" and "after" images and feeds its output to Language Decoder to generate natural language that can describe the change.

The contributions of our work are three-fold:

- We design Intra- and Inter-representation Interaction Network (I3N) to learn the reliable difference representation and generate accurate and detailed captions. Extensive experiments show that our approach can achieve state-of-the-art performances on CLEVR-Change [20] and Spot-the-Diff [18] datasets.
- To the best of our knowledge, we present the first attempt to explore the comprehensive interaction of intra-representation for change captioning. By integrating positional and semantic interaction, GSIR not only enhances the model's perception of fine-grained differences, but also explores a new way of handling viewpoint change at the level of intra-image.
- We propose the Hierarchical Representation Interaction to explicitly learn change by matching multi-level correspondence (feature, position-interaction, semantic-interaction) of inter-image from coarse to fine. This ingenious hierarchical mechanism empowers the model the ability to obtain the fine difference representation from a large number of clutters and resist viewpoint change.

II. Related Work

A. Change Captioning

In recent years, change captioning [18]–[29] is a new branch of vision-language understanding task based on neural encoder-decoder framework [30]. Compared to image captioning, change captioning is more challenging as it requires locating and revealing local semantic features of change from the global semantic space. Specifically, DUDA [20], the pioneer work of change captioning, adopts subtraction between image pairs to learn their semantic difference. Nevertheless, direct removal yields an incorrect change in the presence of distractors. To make up for it, SRDRL [19], VACC [29] and R^3Net [25] measure the cross-semantic relation between image pairs to distinguish the semantic change from the illumination or viewpoint change. Furthermore, IFDC [23] designs the fine-grained feature extraction module to explore the semantic relationship, and SGCC [27] constructs the 3D scene graphs of image pairs to locate change. Noteworthily, some strategies have also been used to improve performance. Hosseinzadeh *et al*. [24] involve using an auxiliary task to improve the performance of the DUDA; Shi *et al*. [22] improve the decoding ability of the M-VAM by reinforcement learning; Yao *et al*. [28] design pre-training tasks and contrastive learning strategies to align visual differences and text descriptions. In addition, Qiu *et al*. [26] design two transformer-based models to describe the change. Different from the above state-of-the-art methods, this paper focuses on constructing interaction reasoning of intra- and inter-representation to learn the robust difference representation with distractors.

B. Geometric embedding in visual understanding

Geometric relation and geometric information are so important that fine-gained visual understanding tasks [31]–[36] often

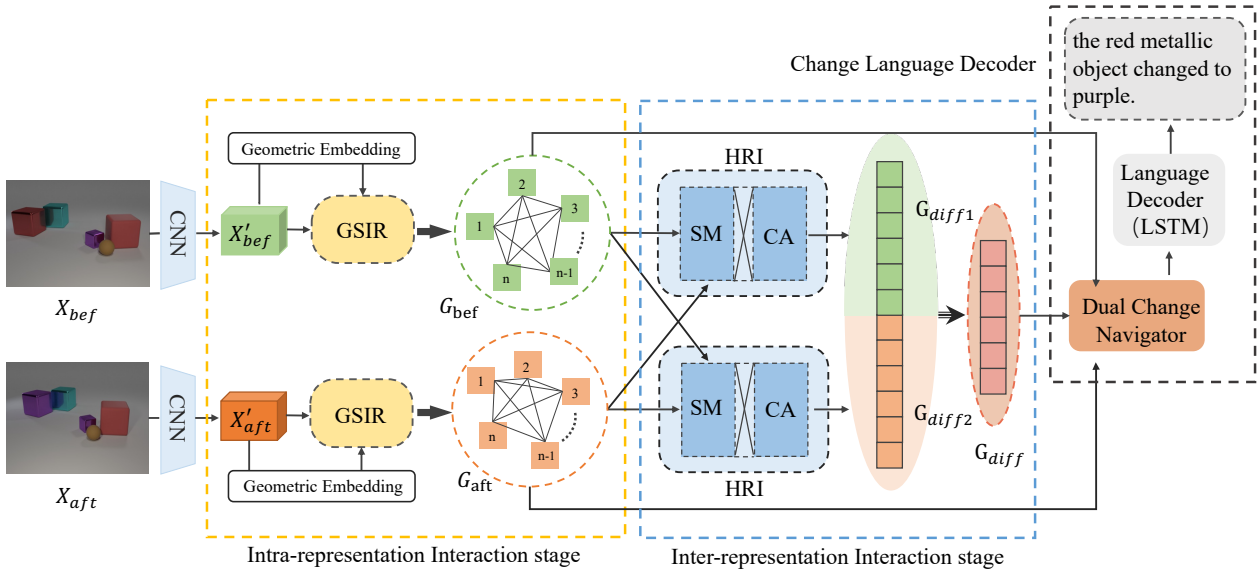


Fig. 2. Overview of the proposed Intra- and inter-representation Interaction Network (I3N), which consists of three key modules. The Geometry-Semantic Interaction Refining (GSIR) integrates semantic and positional interactions of intra-image, which is a special prior knowledge of handling viewpoint change; The Hierarchical Representation Interaction (HRI) constructs the reliable difference representation by means of Semantic Matcher (SM) and Change Amplifier (CA). Our I3N smartly tackles the distraction of irrelevant change by novel designs of interaction learning in the intra- and inter-representation.

emphasize them. Great efforts [2], [15], [17], [37], [38] have been dedicated to tackling this work over the past few years. The earliest work originated from the positional embedding of words [39], [40] in natural language processing. Inspired by this, some works [41], [42] consider embedding geometric information into visual information to perform downstream tasks better. However, most of them only consider a single and rigid geometric embedding for visual information. For the change captioning task, position relation and position information of visual features are critical to locate scene change and resist distractors, so we propose a Geometry-Semantic Interaction Refining (GSIR) to verify this idea.

III. Methodology

In this section, we elaborate each component of our model, as illustrated in Fig. 2- 4. Concretely, we first introduce the Intra-representation Interaction in Section III-A. Next, we present the implementation of Inter-representation Interaction in Sec III-B. Afterward, we perform the process of the Change Language Decoder to generate the captions in Section III-C. We ultimately detail the objective function utilized to optimize the network for change captioning in Section III-D.

A. Intra-representation Interaction

Unlike most methods that directly exploit image pairs to learn change, based on the intra-image perspective, the Intra-representation Interaction stage considers learning the comprehensive interaction to construct a priori knowledge for handling viewpoint change. Specifically, we design the novel Geometry-Semantic Interaction Refining (GSIR) to implement this stage by flexible token interaction learning of position and semantic. We first propose a method to powerfully model geometric interaction and geometric information of intra-image.

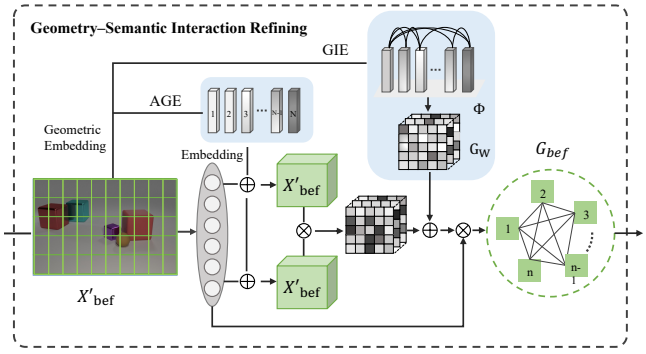


Fig. 3. The details of the Geometry-Semantic Interaction Refining (GSIR). Using X'_{bef} as the example, we present the uni-direction modeling process.

Geometric Interaction Embedding: the position between adjacent objects does not change as a result of viewpoint shift, which can be regarded as the spatial priors to distinguish the semantic and viewpoint change. we design the Geometric Interaction Embedding (GIE) to model the position interaction of features. Specifically, GIE designs the learnable geometric interaction weight $G_W^{i,j} \in \mathbb{R}^{N\times N}$ between feature i and j as:

$$G_W^{i,j} = \mathrm{ReLU}\left(\left(\max\left(0, W_r^1 \Phi(i,j) + b_r\right)\right) W_r^2\right), \quad (1)$$

where b_r, W_r^1, W_r^2 are bias and two learnable weight parameters, respectively. the $\Phi(i,j) \in \mathbb{R}^{N\times N\times 4}$ is a 4-d vector to denote the relative position between features as:

$$\Phi(i,j) = \left(\ln\left(\frac{x_i - x_j}{w_i}\right), \ln\left(\frac{y_i - y_j}{h_i}\right), \ln\left(\frac{w_i}{w_j}\right), \ln\left(\frac{h_i}{h_j}\right)\right)^T, \quad (2)$$

where $\ln$ is the natural logarithm. $(x_i,\ y_i)$, w_i and h_i are relative center coordinate, relative width and height of i. We calculate a pair of the 2D relative coordinate of the feature

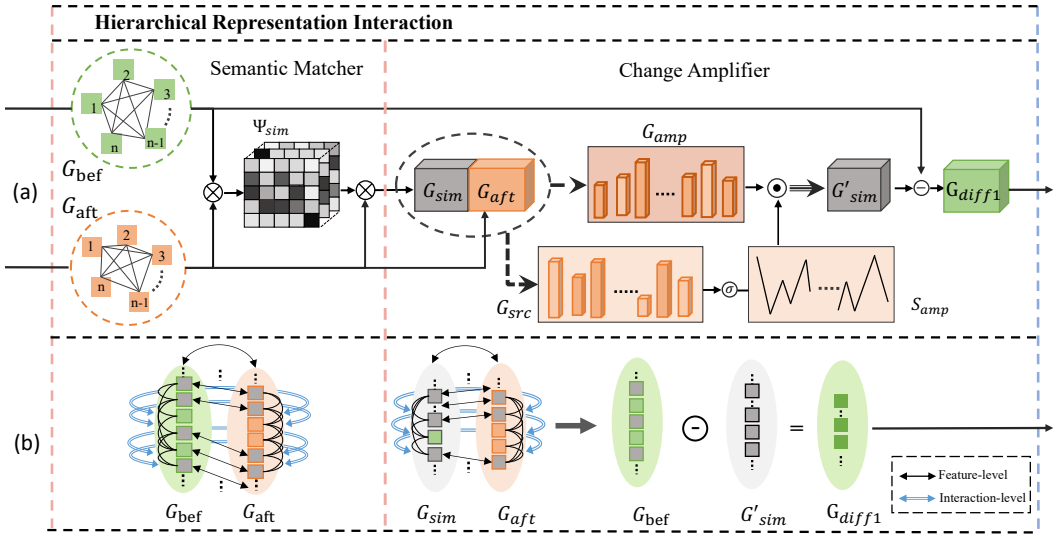


Fig. 4. Detail of proposed Hierarchical Representation Interaction (HRI), which consists of Semantic Matcher and Change Amplifier module. Using G_{bef} as an example, (a) presents the uni-direction modeling process; (b) vividly shows the hierarchical interaction process of the multi-levels (feature, semantic-interaction and position-interaction) in the presence of viewpoint change. The Semantic Matcher first roughly explores the unchanged features between the pair. Then, the Change Amplifier further reveals the fine-grained features of change. This particular architectural design gives the model the ability to locate semantic change and be immune to viewpoint change.

i: $\left[\left(x_i^{\min}, y_i^{\min}\right), \left(x_i^{\max}, y_i^{\max}\right)\right]$ and use it to calculate the x_i, y_i, w_i, h_i above:

$$\begin{aligned}(x_i, y_i) &= \left(\frac{x_i^{\min} + x_i^{\max}}{2}, \frac{y_i^{\min} + y_i^{\max}}{2}\right), \\ w_i &= \left(x_i^{\max} - x_i^{\min}\right) + 1, \\ h_i &= \left(y_i^{\max} - y_i^{\min}\right) + 1,\end{aligned} \tag{3}$$

where $\left(x_i^{\min}, y_i^{\min}\right)$ and $\left(x_i^{\max}, y_i^{\max}\right)$ are the relative coordinate of the upper left and the lower right corner, respectively.

Absolute Geometric Embedding: the spatial and geometric information of each image can be augmented by injecting absolute geometric information into the original image features, which can significantly improve the model's perception ability of position change. Absolute Geometric Embedding (AGE) assigns a constant token with an order to each feature. Specifically, the AGE employs sine and cosine functions of different frequencies to embed position as follow:

$$AGE\left(r, c\right) = \left[GE_r; GE_c\right], \tag{4}$$

where $[;]$, r, c indicate concatenation, row and column index, respectively. And $GE_r, GE_c \in \mathbb{R}^{N/2} (N = HW)$ as follow:

$$\begin{aligned}GE(pos, 2d) &= \sin(pos/10000^{2d/N}), \\ GE(pos, 2d+1) &= \cos(pos/10000^{2d/N}),\end{aligned} \tag{5}$$

where pos, d denote the position and dimension.

Geometry-Semantic Interactions Refining: in our work, we first extract the features X_{bef} and $X_{aft} \in \mathbb{R}^{C \times H \times W}$ of the image pairs from the pre-trained CNN [43], which C, H, W are the number of channels, height and width. Then, we reshape X_{bef}, X_{aft} into $X_{bef}^{'}$, $X_{aft}^{'} \in \mathbb{R}^{N \times C}$, where $N = HW$. The GSIR tailors the features of $X_{bef}^{'}$ and $X_{aft}^{'}$ to model the intra-representation interaction. Specifically, the geometry-semantic interactions are injected into individual features of each image based on scaled dot-product attention [39]:

$$\text{GSIR} = \text{softmax}\left(\Upsilon_W\left(Q, K, V\right)\right) V, \tag{6}$$

where Υ_W is the comprehensive interaction measure, which combines the geometric and semantic interaction measure:

$$\Upsilon_W = \frac{\left(Q + AGE_q\right)\left(K + AGE_k\right)^T}{\sqrt{d_k}} + \ln\left(G_W\right), \tag{7}$$

where AGE_q, AGE_k are Absolute Geometric Embedding of Q and K. G_W is geometric interaction wight, which can calibrate semantic interaction learning. In detail, before the operation of GSIR, Q, K and V are embedded in a same-dimensional embedding by the Emb method:

$$\left(Q, K, V\right) = \left(X_i W_i^Q, X_i W_i^K, X_i W_i^V\right), \tag{8}$$

where W_i^Q, W_i^K, W_i^V are learned parameter matrixes and $i \in (bef, aft)$. In this stage, $X_i^{'}$, $i \in (bef, aft)$, serve as the source field for GSIR:

$$G_i = \text{GSIR}\left(X_i^{'}, X_i^{'}, X_i^{'}\right). \tag{9}$$

When the model deeply grasps the comprehensive interactions of intra-image, it is better able to judge semantic change and irrelevant change. In other words, GSIR effectively constructs the multi-level correspondence (feature, position-interaction, semantic-interaction) as inputs to the Inter-representation Interaction stage, which is the basis for the Inter-representation Interaction stage to learn the reliable difference representation.

B. Inter-representation Interaction

The Inter-representation Interaction stage highlights the explicit learning of difference representation. It is implemented via the proposed Hierarchical Representation Interaction (HRI), which is an ingenious hierarchical mechanism

from coarse to fine to pinpoint the disagreement of inter-image in the presence of irrelevant change. The HRI contains two modules, Semantic Matcher (coarse) and Change Amplifier (fine) as shown in Fig. 4 (a).

Semantic Matcher: The Semantic Matcher obtains the common features of the before and after image by matching the multi-level correspondence, as shown in Fig. 4 (b). Specifically, with $G_{bef}, G_{aft} \in \mathbb{R}^{N\times C}$ as inputs, we first calculate the corresponding measure between the image pairs:

$$\Psi_{\mathrm{sim}}(G_{bef}, G_{aft}) = \mathrm{softmax}\left(\frac{G_{bef}G_{aft}^{T}}{\sqrt{ch_{aft}}}\right), \tag{10}$$

where ch_{aft} is the number of channel and softmax is applied to normalize measure scores. Then we employ Ψ_{sim} to compute common features over G_{bef}, which can be formulated as:

$$G_{sim}^{i} = \sum_{j} \Psi_{i,j} \otimes G_{aft}^{j}, \tag{11}$$

where i is the i^{th} G_{bef} and j is the j^{th} G_{aft}. $\Psi_{i,j}$ is the correspondence measure between the i^{th} G_{bef} and the j^{th} G_{aft}.

Change Amplifier: as shown in Fig. 11 (b), the change process is the local features change in the image pair, and we need to distinguish these tiny local features from a large number of clutter (global space) and generate corresponding semantic information based on the surrounding unchanged features. However, this surrounding unchanged information is disrupted by viewpoint shift, which makes the surrounding features appear to be pseudo-different. This makes these local features often overwhelmed by the mostly unchanged features, especially when the object is too small or changes slightly, such change features are easily missed during the first coarse interaction process. Hence, the Change Amplifier (CA) reveals the signal of latent change in the acquired common features G_{sim} by referring to source features G_{aft}. Concretely, we first construct the amplifier G_{amp} and amplification source G_{amp} by concatenating G_{aft} and $G_{sim}$to $\mathbb{R}^{2C\times N}$:

$$\begin{aligned} G_{src} &= W_s\left[G_{aft}; G_{sim}\right] + b_s, \\ G_{amp} &= W_a\left[G_{aft}; G_{sim}\right] + b_a, \end{aligned} \tag{12}$$

where $W_s, W_a \in \mathbb{R}^{2C\times N}$ and b_s, b_s are bias. Next, the fine common features $G_{sim}^{'}$ is constructed under the guidance of the amplifying signals $\mathcal{S}_{amp}$:

$$\begin{aligned} G_{sim}^{'} &= \mathrm{FC}\left(G_{src} \odot \mathcal{S}_{amp}\right), \\ \mathcal{S}_{amp} &= \mathrm{sigmoid}\left(G_{amp}\right), \end{aligned} \tag{13}$$

where $\odot$ and FC denote element-wise and a fully-connected layer. The amplifying signals $\mathcal{S}_{amp}$ measure the multi-level similarities between corresponding locations in G_{aft} and G_{sim}. Finally, the uni-direction difference G_{diff}^{b} is captured by purifying $G_{sim}^{'}$ from G_{bef}:

$$G_{diff}^{b} = G_{bef} - G_{sim}^{'}. \tag{14}$$

For the entire HRI, we cross-utilize G_{bef} and G_{aft} to match and amplify against each other, and then project them to obtain the bi-direction difference representation:

$$\begin{aligned} G_{diff}^{b} &= \mathrm{HRI}\left(G_{bef}, G_{aft}\right), \\ G_{diff}^{a} &= \mathrm{HRI}\left(G_{aft}, G_{bef}\right), \\ G_{diff} &= \mathrm{ReLU}\left(\mathrm{FC}\left[G_{diff}^{b}; G_{diff}^{a}\right]\right), \end{aligned} \tag{15}$$

where FC is a fully-connected layer. With this coarse-to-fine mechanism, the model can distinguish semantic change from irrelevant change and obtain the fine difference representation.

C. Change Language Decoder

The Change Language Decoder aims to generate a sentence that describes the change, which consists of two modules, namely Dual Change Navigator and Language Decoder.

Dual Change Navigator: the difference representation G_{diff} is constructed by the Equation 15. We exploit it as the query to tell the model where the change is in each representation (G_{bef}, G_{aft}). Specifically, the Dual Change Navigator first calculates two separate attention maps S_{nav}^{i}. Then the changed features d_i are localized via applying S_{nav}^{i} to G_i, $i \in (bef, aft)$:

$$\begin{aligned} S_{\mathrm{nav}}^{i} &= \mathrm{sigmoid}\left(f_2\left(\mathrm{ReLU}\left(f_1\left[G_i; G_{diff}\right]\right)\right)\right), \\ d_{\mathrm{i}} &= \sum_{H,W} S_{\mathrm{nav}}^{i} \odot G_{\mathrm{i}}\,, d_{\mathrm{i}} \in \mathbb{R}^{C}, \end{aligned} \tag{16}$$

where $[;]$ and f_1 f_2 are concatenation and convolution operations. This design navigates the model to focus on visual representation differently depending on the type of a semantic change and the amount of a viewpoint change.

Language Decoder: we first exploit attention LSTM_s to look for key semantic feature that is most relevant with the word (x_{w}) from three kinds of features d_{bef}, d_{aft} and d_{diff} $(d_{\mathrm{bef}} - d_{\mathrm{aft}})$:

$$d_{\mathrm{s}}^{(t)} = \textstyle\sum_j \mho_j^{(t)} d_j, \tag{17}$$

where $j \in (bef, aft, diff)$. The attention map $\mho_i^{(t)}$ is computed by the attention $LSTM_s$, which take as input the previous hidden state $x_w^{(*)}$ of LSTM_w and the projection d_{all} of three visual features:

$$\begin{aligned} d_{all} &= \mathrm{ReLU}\left(\mathrm{FC}\left[d_{bef}; d_{diff}; d_{aft}\right]\right), \\ x_s^{(t)} &= \mathrm{LSTM}_s\left(x_s^{(t)} \mid \left[d_{all}; x_w^{(t-1)}\right], x_s^{(0:t-1)}\right), \\ \mho^{(t)} &\sim \mathrm{softmax}\left(W_1 x_s^{(t)} + b_1\right), \end{aligned} \tag{18}$$

where FC, $x_s^{(*)}$, W_1 and b_1 are the fully-connected operation, hidden states of the module LSTM_s, learnable parameters and bias, respectively.

The language generation process is guided by the above operation. We exploit attended visual feature $d_{\mathrm{s}}^{(t)}$ and the previous word w_{t-1} (ground-truth word during training, predicted word during inference) to the LSTM_w to predict a series of distributions over the next word:

$$\begin{aligned} c_{t-1} &= Embed\left(w_{t-1}\right), \\ x_w^{(t)} &= \mathrm{LSTM}_w\left(x_w^{(t)} \mid \left[d_{\mathrm{s}}^{(t)}; c_{(t-1)}\right], x_s^{(0:t-1)}\right), \\ w^{(t)} &\sim \mathrm{softmax}\left(W_2 x_w^{(t)} + b_2\right), \end{aligned} \tag{19}$$

TABLE I
TOTAL PERFORMANCE COMPARED WITH STATE-OF-THE-ART METHODS ON CLEVR-CHANGE, WHERE B-4, C, M, R-L, AND S ARE SHORTHAND FOR BLEU-4, CIDER, ROUGE-L, METEOR, AND SPICE, RESPECTIVELY. RL AND PT ARE THE SHORTHAND OF REINFORCEMENT LEARNING TRAINING AND PRE-TRAINING STRATEGIES, RESPECTIVELY. "-" DENOTES NO RESULTS REPORTED AND "*" DENOTES USING TRANSFORMER DECODER(TD) AS DECODER. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLDFACE.

Method	RL	PT	Total				
			B-4	C	M	R-L	S
Capt-Dual (ICCV2019) [20]	×	×	43.5	108.5	32.7	-	23.4
DUDA (ICCV2019) [20]	×	×	47.3	112.3	33.9	-	24.5
M-VAM (ECCV2020) [22]	×	×	50.3	114.9	37.0	69.7	30.5
M-VAM+RAF (ECCV2020) [22]	√	×	51.3	115.8	37.8	70.4	30.7
DUDA+AT (CVPR2021) [24]	×	×	51.2	115.4	37.7	70.5	31.1
VACC (ICCV2021) [29]	×	×	52.4	114.2	37.5	-	31.0
IFDC (TMM2021) [23]	×	×	49.2	118.7	32.5	69.1	-
SGCC (MM2021) [27]	×	×	51.1	121.8	40.6	73.9	32.2
R^3Net+SSP (EMNLP2021) [25]	×	×	54.7	123.0	39.8	73.1	32.6
SRDRL (ACL2021) [21]	×	×	54.8	121.0	40.1	73.2	32.6
SRDRL+AVS (ACL2021) [21]	×	×	54.9	122.2	40.2	73.3	32.9
I3N (OURS)	×	×	**55.5**	**123.0**	**41.0**	**74.2**	**33.4**
*MCCFormers-D (ICCV2021) [26]	×	×	52.4	121.6	37.0	-	32.6
*MCCFormers-S (ICCV2021) [26]	×	×	**57.4**	125.5	**41.2**	-	32.4
*IDC (AAAI2022) [28]	×	×	32.7	89.8	27.2	57.2	-
*IDC (AAAI2022) [28]	×	√	51.2	**128.9**	32.5	71.7	-
*I3N-TD (OURS)	×	×	55.8	**125.6**	40.6	**73.9**	**32.8**

where W_2, b_2 are learnable parameters. Embed is a one-hot encoding of the word w_{t-1}.

D. Training

The proposed network is trained end-to-end by minimizing the negative likelihood of the observed word sequence. Given the target ground-truth words $(w_1, ..., w_m)$ and a captioning model with parameters θ, we minimize the following cross-entropy loss:

$$\mathfrak{L}_{lan}(\theta) = -\sum_{t=1}^{m} \log\left(p_\theta\left(w_t^* \mid w_{1:t-1}^*\right)\right), \tag{20}$$

where m is the length of the caption. The Intra-representation Interaction and Inter-representation Interaction receive no direct supervision for difference representation learning. The only available supervision is obtained through the language decoder, which then guides the Intra- and Inter-representation Interaction to learn the difference representation.

IV. EXPERIMENTS

A. Datasets

CLEVR-Change dataset [20] is an enormous dataset generated by the CLEVR engine [44], which contains 79,606 complex scenarios and 493,735 captions. The dataset is divided into two main categories: One (*i.e.*, none-scene change) has only illumination/ viewpoint change (distractors). The other (*i.e.*, scene change) has both scene change (color/ material change, adding/ dropping/ moving an object) and distractors. Based on the official split, we use the split with 67,660 for training, 3,976 for validation, and 7,970 for testing.

Spot-the-Diff dataset [18] consists of 13,192 image pairs without distractors, which are extracted from surveillance videos of different time periods. This dataset is set up as a change for each image pair with no distractors. The dataset is split into training, validation, and testing with an official ratio of 8:1:1.

B. Evaluation Metrics

We use metrics commonly used in captioning to evaluate the quality of the generated sentence. Including BIEU-4 [45], METEOR [46], CIDEr [47], ROUGE-L [48] and SPICE [49]. Generally, those scores are higher if the semantic content and grammatical structure are more accurate. In addition, to evaluate the accuracy of change localization, we use the Pointing Game [50] to evaluate our method. Concretely, we up-sampled the attention map to the size of the original image and check if the points with the highest activation are within the bounding box of ground truth.

C. Implementation Details

Following the approaches of change captioning, we employ the pre-trained ResNet-101 [43] on the ImageNet [51] to extract visual features, with the dimension of $1024 \times 14 \times 14$. Then, the visual features are embedded into a low-dimensional embedding of 512. The hidden size of the whole model is set to 512 and the number of attention heads is set to 4. Besides, each word is represented by a 300-dim vector. For the 38-epoch training stage, the model is trained by the Adam Optimizer [52] with a learning rate of 0.001/0.0002 and a batch size of 128/64 on the CLEVR-Change/ Spot-the-Diff dataset. Both training and inference are implemented with PyTorch [53] on the Titan XP GPU. Moreover, the training time and memory cost for the convergence of our model are about 15/10 GB and 3/2 hours on CLEVR-Change/Spot-the-Diff dataset.

D. Experimental Results

1) Results on CLEVR-Change Dataset

To demonstrate the superiority of our model, we compare it with state-of-the-art methods [20]–[29] on CLEVR-Change [20]. We present the conclusion from five perspectives: Total change in Table I; Scene change in Table II; None-scene change in Table II; Representative scene change in Table III; Measuring robustness to viewpoint change in Fig 5.

TABLE II
SCENE/ NONE-SCENE CHANGE PERFORMANCE COMPARED WITH STATE-OF-THE-ART METHODS ON CLEVR-CHANGE. RL IS THE SHORTHAND OF REINFORCEMENT LEARNING TRAINING STRATEGY. "-" DENOTES NO RESULTS REPORTED AND THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLDFACE.

Method	RL	Scene Change					None-scene Change				
		B-4	C	M	R-L	S	B-4	C	M	R-L	S
Capt-Dual (ICCV2019) [20]	×	38.5	89.8	28.5	-	18.2	56.3	108.9	44.0	-	28.7
DUDA (ICCV2019) [20]	×	42.9	94.6	29.7	-	19.9	59.8	110.8	45.2	-	29.1
M-VAM+RAF (ECCV2020) [22]	√	-	-	-	-	-	-	**122.6**	**66.4**	-	33.4
DUDA+AT (CVPR2021) [24]	×	49.9	101.3	34.3	65.4	27.9	**62.4**	116.3	50.5	53.9	35.0
SGCC (MM2021) [27]	×	-	-	-	-	-	-	116.5	52.2	-	35.0
IFDC (TMM2021) [23]	×	47.2	105.4	29.3	63.7	-	52.5	114.2	40.1	74.4	-
R^3Net+SSP (EMNLP2021) [25]	×	52.7	116.6	36.2	69.8	30.3	61.9	116.4	50.5	76.4	34.8
SRDRL+AVS (ACL2021) [21]	×	52.7	114.2	36.4	69.7	30.8	62.2	117.0	51.3	76.9	34.9
I3N (OURS)	×	**53.1**	**117.0**	**37.0**	**70.8**	**32.1**	62.2	116.7	52.0	**77.5**	**35.0**

Total change performance. We simultaneously evaluate for scene change and none-scene change in Table I. To keep the comprehensiveness and fairness of the comparison, we compare SOTA methods from two perspectives based on the structure of the decoder. One is based on the LSTM decoder. We observe that our model outperforms the above SOTA methods on all metrics with an encouraging performance. Compared with the previous best SRDRL+AVS [21], the main motivation of I3N is how to learn an effective difference representation while being immune to the interference of viewpoint change, which is similar to the SRDRL. In contrast to SRDRL, I3N circumvents the distraction of irrelevant change by learning the interaction of intra- and inter-representation. Compared to SRDRL, I3N achieves improvements of 1.3%, 1.7%, 2.2%, 1.4% and 2.5% on BLEU-4, CIDEr, METEOR, ROUGE-L and SPICE metrics, respectively. This indicates that our model can construct more reliable difference representation. When generating the words, SRDRL introduces Part-of-Speech knowledge and designs an AVS module to calibrate the cross-modal alignment, which further improves its performance. In this case, our I3N still outperforms SRDRL+AVS on all metrics. Since changed objects are subtle and easy to ignore, their features would be weak. This makes the model difficult to align them with target words. Further exploration of cross-modal alignment is necessary in future research.

The other is based on the Transformer decoder. Our model achieves the best performance on SPICE, CIDEr and ROUGE-L without the aid of the pre-training strategy. Although the IDC [28] with the pre-training strategy improves the performance, its computational cost is too expensive and almost all metrics are lower than ours, which indicates that I3N can achieve the best performance while saving computational costs. Note that compared with MCCFormers-D [26], I3N-TD outperforms it on most metrics, and even I3N without the Transformer decoder outperforms it on ROUGE-L and SPICE. This indicates that our intra- and inter-representation interaction mechanism can construct a more reliable difference representation than the common Transformer encoder structure. To conclude, the generalization ability of the proposed method is superior to that of the current SOTA methods, which benefits from the intra- and inter-representation interaction mechanism.

None-scene/ Scene change performance. As the M-VAM [22], VACC [29], MCCFormers-D [26], MCCFormers-S [26] and IDF [28] do not report the results on none-scene/scene change, we only compare with the methods that provide official results. In the case of scene change, the image pairs have both scene change and illumination/viewpoint change. From Table II, our method surpasses all SOTA methods by a large margin, especially SPICE (from 30.8 to 32.1). In the case of none-scene change, the image pairs have only illumination or viewpoint change. From Table II, we can observe that the METEOR and CIDEr of M-VAM+RAF are higher than ours, and we consider that this lies in the introduction of reinforcement learning. This shows that reinforcement learning does have an improvement on performance in this setting, but it dramatically increases the training time and computational complexity simultaneously. Nevertheless, I3N far surpasses M-VAM+RAF and M-VAM in total performance (Table I), so our approach is more robust while reducing training complexity.

Representative scene change performance. We evaluate five typical change scenes: color/texture change (static change) and adding/dropping/moving an object (dynamic change) in Table III. We compared with the SOTA approaches that reported the results on CLDEr, METEOR and SPICE. we can observe that our model achieves competitive results for all kinds of change types and obtains impressive results for "COLOR", "ADD", "DROP" and "TEXTURE" change types. In particular, "TEXTURE" and "MOVE" are challenging in this task, because the viewpoint shifting makes locating these types of changes more difficult. Obviously, benefiting from the intra- and inter-representation interaction, we achieve an excellent effect in the change of the object's texture and disappearance. In general, this experiment indicates that our model can accurately capture the dynamic and static change of object in the distraction of viewpoint change.

Measuring robustness to viewpoint change. To explore the robustness of our model against viewpoint change, we measure the amount of viewpoint shift for an image pair by IoU and sort the test examples based on the mean of these IoUs. Specifically, we calculate the IoU of the object's bounding boxes in the scene across the image pair. When the viewpoint change more, the less the bounding boxes of image pairs overlap. As shown in Fig 5, we plot the graph of CIDEr scores (on the left) and Locating Accuracy (on the right) under different degrees of viewpoint shift (measured by IoU), and compare with some SOTA methods that provided official code: SRDRL+AVS [21] and R^3Net+SSP [25]. For the

TABLE III
A DETAILED BREAKDOWN OF CHANGE CAPTIONING EVALUATION ON CLEVR-CHANGE BY DIFFERENT CHANGE TYPES: "COLOR" (C), "TEXTURE" (T), "ADD" (A), "DROP" (D), AND "MOVE" (M). RL IS THE SHORTHAND OF REINFORCEMENT LEARNING TRAINING STRATEGY. "-" DENOTES NO RESULTS REPORTED AND THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLDFACE.

Method	RL	PT	CIDEr					METEOR					SPICE				
			C	T	A	D	M	C	T	A	D	M	C	T	A	D	M
Capt-Dual (ICCV2019) [20]	×	×	120.4	86.7	108.2	103.4	56.4	32.8	27.3	33.4	31.4	23.5	21.1	18.3	22.4	22.2	15.4
DUDA (ICCV2019) [20]	×	×	120.4	86.7	108.2	103.4	56.4	32.8	27.3	33.4	31.4	23.5	21.1	18.3	22.4	22.2	15.4
M-VAM+RAF (ECCV2020) [22]	√	×	122.1	98.7	126.3	115.8	82.0	35.8	32.3	37.8	36.2	27.9	28.0	26.7	30.8	32.3	22.5
DUDA+AT (CVPR2021) [24]	×	×	120.8	89.9	119.8	123.4	62.1	36.1	30.4	37.8	36.7	27.0	29.7	27.4	31.4	30.8	23.5
SGCC (MM2021) [27]	×	×	128.0	122.9	117.1	116.9	77.1	37.8	36.1	38.9	36.7	**32.8**	30.0	31.1	30.8	30.1	25.3
IFDC (TMM2021) [23]	×	×	133.2	99.1	128.2	118.5	82.1	33.1	27.9	36.2	31.4	31.2	-	-	-	-	-
R^3Net+SSP (EMNLP2021) [25]	×	×	139.2	123.5	122.7	121.9	**88.1**	38.9	35.5	38.0	37.5	30.9	31.6	30.8	32.3	31.7	25.4
SRDRL+AVS (ACL2021) [21]	×	×	136.1	122.7	121.0	126.0	78.9	39.0	35.6	38.9	38.0	30.1	32.4	30.9	33.0	32.4	25.4
IDC (AAAI2022) [28]	×	√	131.2	101.1	**133.3**	116.5	81.7	-	-	-	-	-	-	-	-	-	-
I3N (OURS)	×	×	**139.2**	**127.6**	125.5	**131.8**	78.5	**39.9**	**36.7**	**39.9**	**38.1**	30.6	**33.6**	**33.1**	**33.9**	**33.8**	**25.8**

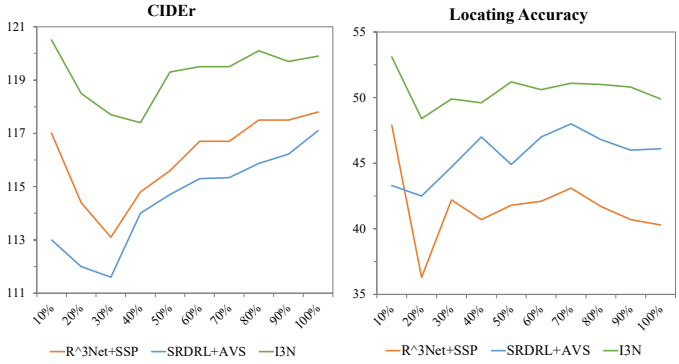


Fig. 5. CIDEr and Locating Accuracy (%) by IoU difficulty. The more significant the viewpoint shifting, the more difficult the IoU. Values are acquired for every 10th percentile.

CIDEr scores, our method outperforms the SOTA methods, including the more significant viewpoint shift. In the Locating Accuracy graph, all methods trend downward as IoU difficulty increases, as accurate localization becomes more difficult with the drastic viewpoint shift, while our method achieves the highest accuracy rate. These results that the captioning and locating performance of our method is more robust to varying degrees of viewpoint change compared to existing SOTA methods.

2) Results on Spot-the-Diff Dataset

We evaluate the performance of our model in the real environment. We compare with the state-of-the-art methods [18], [20]–[27], [29], [54], [55] that reported the results on the Spot-the-Diff dataset [18]. From Table IV, we can observe that our I3N attains the best results on BLUE-4, METEOR and SPICE when training without reinforcement learning. Even when compared to M-VAM+RAF [22] with reinforcement learning, our method still beats it on METEOR and SPICE. As this dataset does not consider the distractions of viewpoint/illumination change and each pair is set to a change, the model does not have to judge whether a change has occurred. the superiority mainly results from the fact that the GSIR module can enhance the model's understanding of the fine-grained information and interaction. We note that the CIDEr and ROUGE-L of SGCC [27] are higher than ours, and the reason is that they introduce additional 3D information of

TABLE IV
PERFORMANCE OF OUR METHOD AGAINST STATE-OF-THE-ART METHODS ON SPOT-THE-DIFF DATASET. RL IS THE SHORTHAND OF REINFORCEMENT LEARNING STRATEGY. "-" DENOTES NO RESULTS REPORTED AND THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLDFACE.

Method	RL	B-4	C	M	R-L	S
DDLA (EMNLP2018) [18]	×	8.5	32.8	12.0	28.6	-
DUDA (ICCV2019) [20]	×	8.1	32.5	11.8	29.1	-
FCC [54]	×	9.9	36.8	12.9	29.9	-
SDCM [55]	×	9.8	36.3	12.7	29.7	-
M-VAM (ECCV2020) [22]	×	10.1	38.1	12.4	31.3	14.0
M-VAM+RAF (ECCV2020) [22]	√	**11.1**	42.5	12.9	33.2	17.1
DUDA+AT (CVPR2021) [24]	×	8.1	34.5	12.5	29.9	-
VACC (ICCV2021) [29]	×	9.7	41.5	12.6	32.1	-
R^3Net+SSP (EMNLP2021) [25]	×	-	36.6	13.1	32.6	18.8
SGCC (MM2021) [27]	×	8.6	**42.9**	13.1	**38.2**	17.2
IFDC (TMM2021) [23]	×	8.73	37.0	11.7	30.2	17.2
SRDRL+AVS (ACl2021) [21]	×	-	35.3	13.0	31.0	18.0
I3N (OURS)	×	**10.1**	38.3	**13.2**	33.1	**18.9**
*MCCFormers-D (ICCV2021) [26]	×	10.0	**43.1**	12.4	-	18.3
*MCCFormers-S (ICCV2021) [26]	×	9.8	41.6	12.3	-	16.3
*I3N-TD (OURS)	×	**10.3**	42.7	**13.0**	**31.5**	**18.6**

images and semantic attributes of objects. Nevertheless, our I3N still outperforms it on most metrics, which proves the strength of our GSIR. For I3N with Transformer Decoder (I3N-TD), our method achieves the best performances on BLUE-4, METEOR, ROUGE-L and SPICE compared with MCCFormers-S [26] and MCCFormers-D [26]. These results prove that the proposed method is more competitive than the current SOTA method.

In conclusion, the Spot-the-Diff dataset ignores the existence of distractors and the image pairs are mostly perfectly aligned. Our performance on Spot-the-Diff demonstrates the ability to accurately caption change in noisy real-life and generic scenarios without distractors.

E. Ablation Studies

To figure out contributions from different components of the proposed model, we provide extensive ablation studies on CLEVR-Change. 1) We begin with the baseline based on DUDA [20]; 2) The SIR is the semantic interaction refining without geometric embedding; 3) The SIR+AGE is the semantic interaction refining with absolute geometric embedding; 4) The SIR+GIE is the semantic interaction refining with geometric interaction embedding; 5) The GSIR is the

TABLE V
ABLATION STUDIES IN TERMS OF TOTAL PERFORMANCE TO DEMONSTRATE CONTRIBUTIONS FROM DIFFERENT PROPOSED COMPONENTS ON CLEVR-CHANGE, WHERE B-4, C, M, R-L, AND S ARE SHORTHAND FOR BLEU-4, CIDER, ROUGE-L, METEOR, AND SPICE, RESPECTIVELY. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLDFACE

	Total					Scene Change					None-scene Change				
Method	B-4	C	M	R-L	S	B-4	C	M	R-L	S	B-4	C	M	R-L	S
Baseline	53.1	115.6	37.6	70.8	30.2	50.9	102.4	33.3	65.7	27.3	61.0	114.3	49.9	75.8	34.5
SIR	53.1	116.3	38.3	71.5	31.1	51.8	103.6	33.5	65.8	28.3	61.8	115.6	51.4	77.0	34.7
SIR+AGE	54.4	119.0	39.4	72.4	31.5	52.4	109.6	35.5	68.6	30.3	61.5	114.4	50.6	76.3	34.6
SIR+GIE	54.1	118.2	38.8	72.0	31.8	51.2	106.9	34.2	67.0	28.8	62.4	115.8	51.6	77.0	34.8
GSIR	54.7	119.4	39.3	72.5	31.9	52.0	108.9	34.9	67.8	28.9	**62.5**	116.2	51.8	77.1	34.9
HRI w/o CA	54.5	120.0	40.2	73.1	32.6	52.1	111.7	36.2	69.8	30.6	60.8	114.9	51.6	77.2	34.6
HRI	54.4	121.6	40.4	73.6	33.2	52.3	112.8	36.3	69.8	31.1	61.2	116.1	50.7	77.2	34.7
I3N	**55.5**	**123.0**	**41.0**	**74.2**	**33.4**	**53.1**	**117.0**	**37.0**	**70.8**	**32.1**	62.2	**116.7**	**52.0**	**77.5**	**35.0**

TABLE VI
ABLATION STUDIES ON CLEVR-CHANGE BY DIFFERENT CHANGE TYPES: "COLOR" (C), "TEXTURE" (T), "ADD" (A), "DROP" (D), AND "MOVE" (M). THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLDFACE.

Metrics	Method	C	T	A	D	M
BLUE-4	GSIR	52.2	48.9	56.9	54.3	48.9
	HRI w/o CA	50.8	49.1	56.3	52.7	48.7
	HRI	51.9	49.3	57.5	53.4	49.7
	I3N	**52.2**	**49.9**	**57.8**	**54.3**	**50.5**
CIDEr	GSIR	134.0	115.3	122.4	125.6	75.1
	HRI w/o CA	131.8	120.4	121.6	126.0	73.7
	HRI	133.4	121.0	123.1	125.6	**81.8**
	I3N	**139.2**	**127.6**	**125.5**	**131.8**	78.5
METEOR	GSIR	38.5	34.0	39.0	37.5	28.4
	HRI w/o CA	38.9	35.4	38.8	37.0	29.3
	HRI	39.0	35.7	39.0	37.2	30.4
	I3N	**39.9**	**36.7**	**39.9**	**38.1**	**30.6**
SPICE	GSIR	32.1	30.5	32.9	32.3	23.4
	HRI w/o CA	32.7	33.0	33.1	32.3	25.9
	HRI	32.3	31.1	33.5	32.6	**26.0**
	I3N	**33.6**	**33.1**	**33.9**	**33.8**	25.8

TABLE VII
ABLATION STUDIES OF THE G_i AND G_{sim},$i \in (bef, aft)$ CONNECTIONS ON CHANGE AMPLIFIER. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLDFACE.

Method	B-4	C	M	R-L	S
Scene Change					
Sum	**53.2**	117.0	36.5	70.2	31.1
Concat(ours)	53.1	**117.0**	**37.0**	**70.8**	**32.1**
None-scene Change					
Sum	**62.3**	116.0	50.6	76.4	34.8
Concat(ours)	62.2	**116.7**	**52.0**	**77.5**	**35.0**

geometry-semantic interactions refining combining two types of geometric embedding; 6) The HRI w/o CA module is hierarchical representation interaction without change amplifier; 7) HRI is the whole hierarchical representation interaction; 8) I3N is the full model that incorporated all the proposed components.

From total performance in Table V, we can easily observe that the metrics for each of the added modules are higher than the baseline. The GSIR is better than the SR, where the GSIR is improved by 3%, 2.7%, 2.6%, 1.4% and 2.6% on metrics BLEU4, CIDEr, METEOR, ROUGE-L and SPICE. This proves the necessity of positional interaction and position information to locate change. Among them, we find that the addition of the GIE is a little worse than the AGE, and we argue that the learning of the geometric interaction information is disturbed by the noise of the direct subtraction [20]. The HRI is better than the HRI w/o CA, where the HRI is improved by 1.3%, 0.5%, 2.6%, 0.7% and on CIDEr, METEOR, ROUGE-L metrics. This proves the Change Amplifier capture those latent semantic change to compensate for the Semantic Matcher in viewpoint change. Note that I3N is the best of them and improves 4.5%, 6.4%, 9.0%, 4.8% and 10.6% over baseline on BLEU-4, CIDEr, METEOR, ROUGE-L and SPICE metrics.

From scene change performance in Table V, we obtain the fact that 1) all the modules have a substantial performance improvement over the baseline. 2) I3N is the best of them and I3N improves 4.3%, 14.2%, 11.1%, 7.8% and 17.6% over baseline on BLEU-4, CIDEr, METEOR, ROUGE-L and SPICE metrics. 3) Note that the augmenting of HRI shows the indispensability of this hierarchical mechanism. The above observations indicate that 1) understanding the semantics-positional interaction of intra-representation is a prerequisite for learning fine difference under the viewpoint change. 2) HRI is an effective way to distinguish semantic change from irrelevant change, which also proves the superiority of our method compared to the direct subtraction approach.

From none-scene change performance in Table V, we can observe that all values exceed the baseline. Note that the BLUE-4 of GSIR is higher than I3N. We consider the reason is that the model with GSIR treats the unchanged and some changed regions as unchanged in the baseline setting of using direct subtraction, so the performance of total and scene change is not well. The experimental results prove that our model can effectively determine whether semantic change occurs in the distraction of viewpoint change.

Different change types performance. To further reveal the contribution of GSIR (Intra-representation Interaction), HRI (Inter-representation Interaction) and HRI w/o CA in dealing with viewpoint change. We study five representative types of change in the presence of distractors. From Table VI, we get the fact that 1) I3N is better than GSIR, HRI and HRI w/o CA; 2) Note that most of the HRI metrics are higher than the GSIR and HRI w/o CA. We can draw the following inferences: 1) GSIR is the basis of HRI, and the semantic-position interactions by GSIR are of great importance in handling viewpoint change and significantly perceiving position and attribute change; 2) As the key to highlight change, this

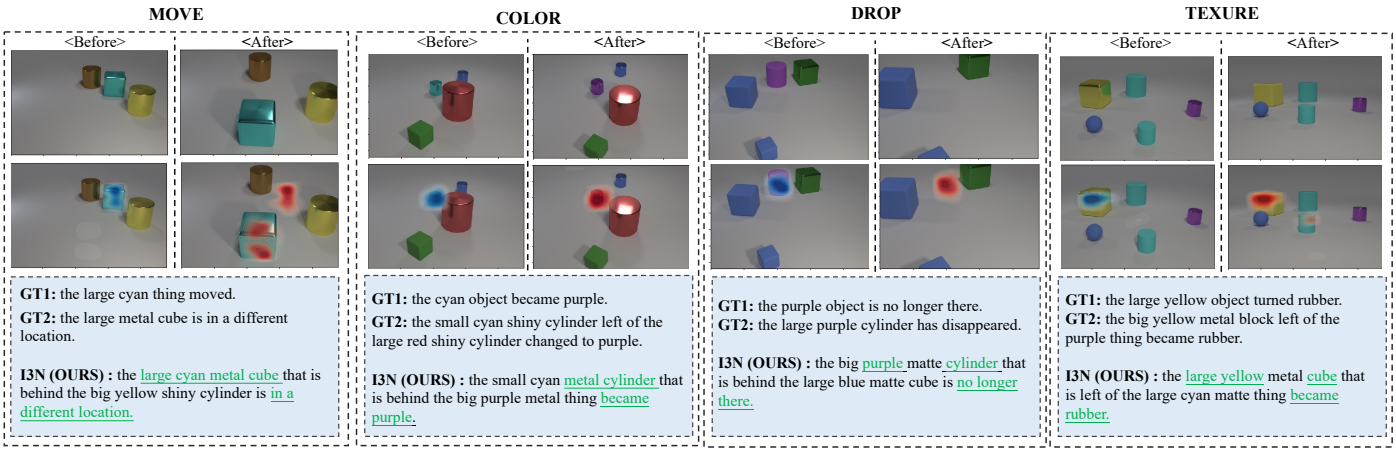


Fig. 6. Four visualizations of our I3N model under viewpoint change on CLEVR-Change dataset, namely "Move", "COLOR", "DROP", and "TEXTURE". The visualizations of localization results are shown on "Before" (blue) and "After"(red). Two Ground Truths (*i.e.*, GT1, GT2) are listed for each case and green font indicates highlights of the change description.

hierarchical mechanism makes the model's performance shine in the location of different scene change; 3) The learning of accurate difference representation is dependent on the entire coarse-to-fine mechanism and not just on the Semantic Matcher.

In addition, we investigate the way of the G_i and G_{sim},$i \in (bef, aft)$ connection on the Change amplifier. From Table VII, it is obtained that the Concat for G_i and G_{sim} is better than the Sum.

F. Qualitative Results

To present the performance generated by the proposed I3N, we conduct qualitative studies from different perspectives on both CLEVR-Change [20] and Spot-the-Diff [18] datasets.

Fig. 6 illustrates several examples of I3N from the test set of the CLEVR-Change dataset. In the setting of different types of scene change, our model not only accurately describes the change process, but emphasizes some characteristics such as relative location and object attribute. For example, "large cyan metal", "small cyan metal", "big purple matte", and "large yellow matte" emphasize the size and attributes of the changed object in these examples. "is behind ..." and "to the left ..." highlight the position information. As indicated by these examples, with intra- and inter-representation interaction, our model can pinpoint the different scene changes to generate accurate and detailed captions. In Fig. 7, we show several examples of I3N from the test set of the Spot-the-diff dataset. We can see that the sentences generated by our model can capture the difference in the full-aligned image pairs. This demonstrates the ability of our model to describe real-life change.

In Fig. 8, we present some examples of our model compared with DUDA [20], SRDRL+AVS [21], and R^{3}Net+SSP [25], respectively. For none-scene change: in the first example, SRDRL+AVS and R^{3}Net+SSP mistakenly consider that the "purple ball" moved under the viewpoint change; In the second example, DUDA, SRDRL+AVS, and R^{3}Net+SSP incorrectly describe as texture change in the scene, whereas I3N successfully describes that the scene has not changed, even though most of the "purple cylinder" is out of sight. For scene change: in the first example, DUDA and SRDRL+AVS both confuse scene change and viewpoint change, and think no change in the scene. R^{3}Net+SSP fails to locate and describe the changed "green cylinder"; In the second example, when a large cylinder obscures the change object, DUDA, SRDRL+AVS, and R^{3}Net+SSP are unable to describe the change of "gray sphere". Instead, our model captions the change with an accurate sentence. In addition, we compare the accuracy of locating change with some SOTA methods by evaluating their attention maps using the Pointing Game. As shown in Fig. 9, our method achieves the highest localization accuracy in each type of change, especially for COLOR (from 43.2% to 50.8%). Compared to DUDA/ R^{3}Net+SSP/ SRDRL+SSP, our average localization accuracy improved by 28.7%/ 30.6%/ 21.6%. These examples indicate that our model locates and captions change more accurately than SOTA methods.

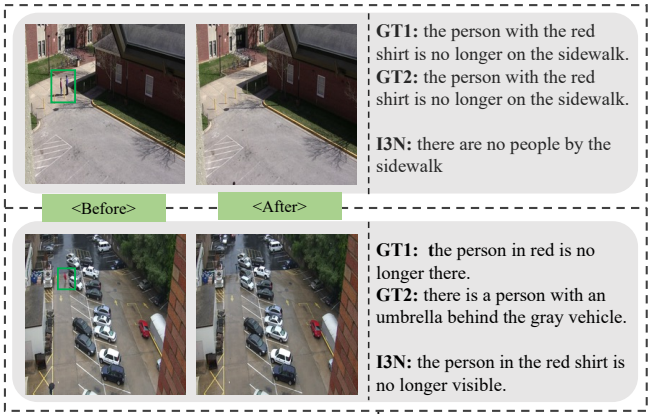


Fig. 7. Qualitative results on the Spot-the-Diff dataset. the Ground Truths (i.e, GT1, GT2) are listed here for each case.

Fig. 10 illustrates several examples of ablation studies on the CLEVR-Change dataset. In the setting of none-scene change with viewpoint change, GSIR and HRI both think that the

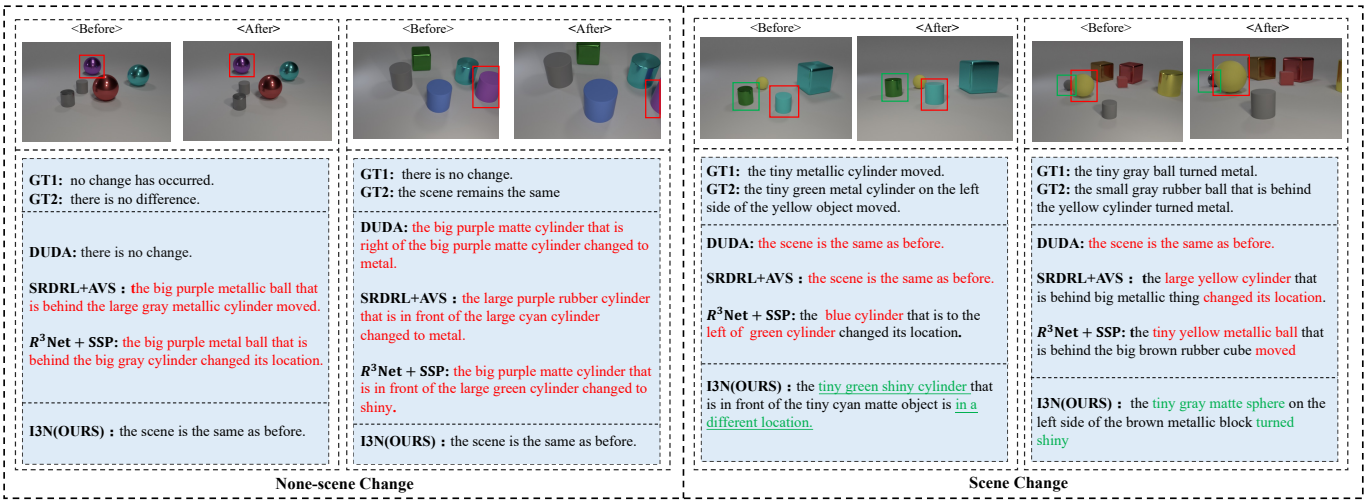


Fig. 8. Qualitative results of our method against state-of-the-art methods on CLEVR-Change dataset. The left is the none-scene change with viewpoint change and the right is scene change with viewpoint change. The underlined words and bounding boxes in green denote highlight captions, while the words and boxes in red denote incomplete or incorrect captions. Two Ground Truths (i.e, GT1, GT2) are listed here for each case.

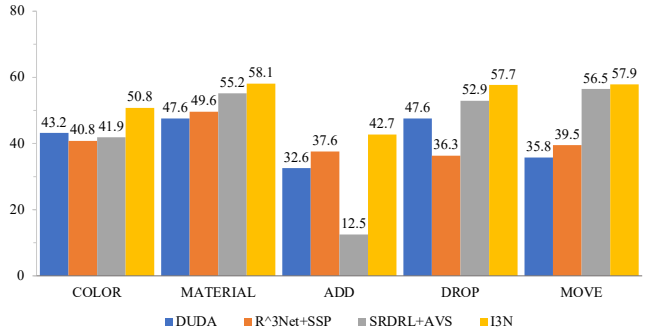


Fig. 9. Locating Accuracy (%) over different change types on the CLEVR-Change dataset.

tiny block has changed. In the setting of scene change with viewpoint change, "blue metal" and "in front of" show that GSIR can get more details about the location and properties. In addition, as shown in Fig. 11, we provide some examples to illustrate the effect of the Change Amplifier. In the first example, when both the small object and viewpoint change slightly, the I3N w/o CA is confused to distinguish between visual signals of viewpoint change and semantic change, resulting in incorrectly locating the changed object. In the second example, I3N w/o CA localizes an incorrect region on the "after" image and therefore misidentifies "Move" as "Drop". Instead, I3N correctly locates and describes the "Move" of the yellow cylinder. In the third example, the I3N w/o CA locates an extra wrong region on both "before" and "after" images, and the reason is that when viewpoint change overwhelms texture change, the I3N w/o CA tends to mistake the signal of viewpoint changes as semantic changes. In contrast, the I3N successfully locates the region of semantic change and obtains the correct caption. In the fourth example, when the large red cube blocks the tiny changed object, I3N w/o CA ignores the signal of the semantic change from the tiny object and thus fails to locate the changed region on the "after" image. These examples indicate that the Semantic Matcher is not sufficient to learn an accurate difference representation, and this coarse-to-fine mechanism does help the model to reveal semantic change and obtain a more accurate caption.

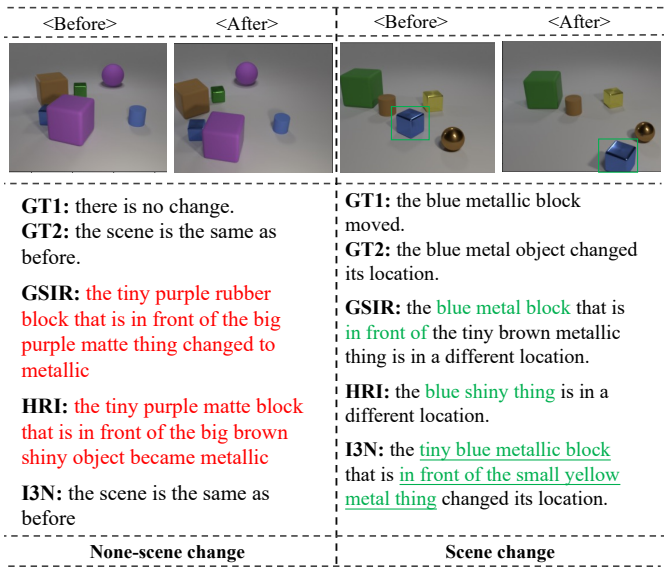


Fig. 10. Qualitative results of ablation study. The left is the none-scene change with viewpoint change and the right is movement (i.e., scene change) with viewpoint change. The underlined words and boxes in green denote highlight captions, while the words in red denote incomplete or incorrect captions. Two Ground Truths (i.e, GT1, GT2) are listed here for each case.

From these examples, we can draw the following conclusions: 1) By the mechanism of intra- and inter-representation interaction, the model can accurately locate and describe the change of different scenarios in the presence or absence of distractors; 2) The proposed method is more competitive compared to current state-of-the-art methods; 3) GSIR and HRI play an indispensable role in the proposed model; 4) HRI empowers the model to accurately locate semantic change and be immune to viewpoint change, which indicates the

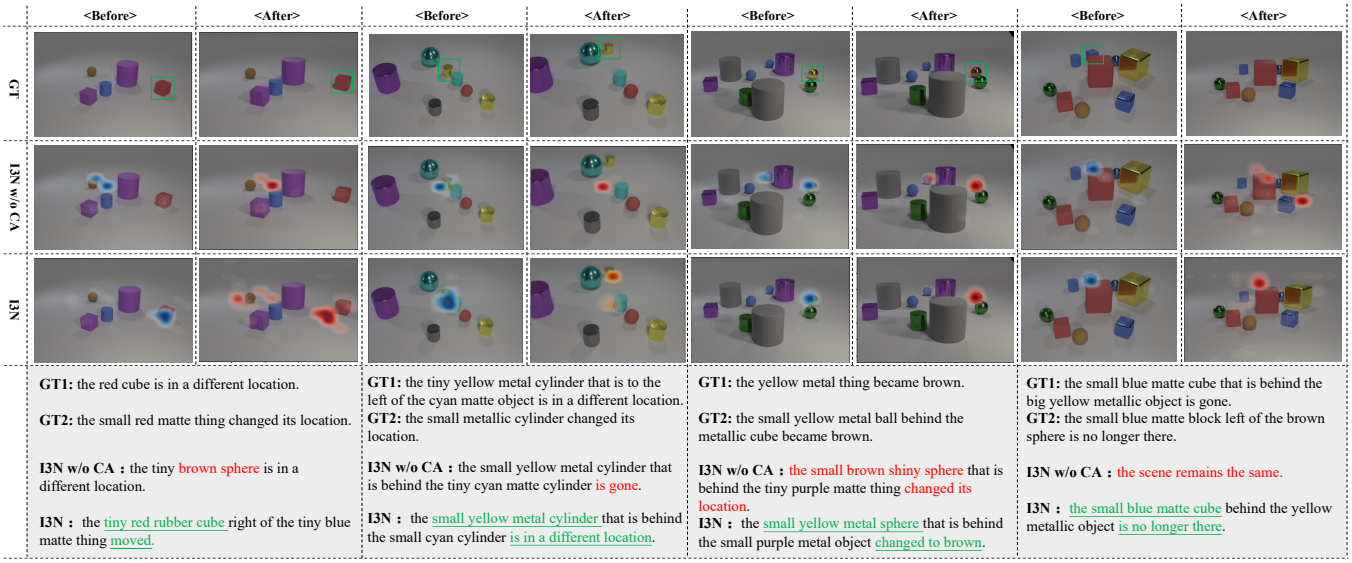


Fig. 11. Qualitative results of I3N and I3N w/o CA. The underlined words in green denote highlight captions and words in red denote incorrect captions. Two Ground Truths (i.e, GT1, GT2) are listed for each case.

superiority of this particular architectural design.

V. Conclusion

In this paper, we propose a novel and generic encoder-decoder architecture for change captioning. Specifically, the proposed Intra- and Inter-representation Interaction Network (I3N) learns the reliable difference representation and generates accurate captions in the presence of viewpoint change from different angles. In the Intra-representation Interaction stage, our model represents the comprehensive interactions of semantic and geometric by applying Geometry-Semantic Interactions Refining, and these interactions serve as priori knowledge for immunizing viewpoint change. In the Inter-representation Interaction stage, we propose the Hierarchical Representation Interaction to resist distractors and pinpoint the semantic change of inter-representation, which consists of the Semantic Matcher and Change Amplifier. Our I3N smartly tackles the distraction of irrelevant change by novel designs of flexible interaction learning in the intra- and inter-representation. Extensive experiments and encouraging performances demonstrate the superiority of our approach and achieve new state-of-the-art results on the two public datasets, CLEVR-Change [20] and Spot-the-Diff [18]. In the future, we will attempt to exploit reinforcement learning strategies to further boost the performance of the proposed method. In addition, we will optimize the decoder to better align the change features with text features to obtain more accurate captions.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. 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*, 2018, pp. 6077–6086.

[2] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[3] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep hierarchical encoder–decoder network for image captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2942–2956, 2019.

[4] L. Yu, J. Zhang, and Q. Wu, "Dual attention on pyramid feature maps for image captioning," *IEEE Transactions on Multimedia*, 2021.

[5] L. Yang, H. Wang, P. Tang, and Q. Li, "Captionnet: A tailor-made recurrent neural network for generating image descriptions," *IEEE Transactions on Multimedia*, vol. 23, pp. 835–845, 2021.

[6] Z. Liu, G. Li, G. Mercier, Y. He, and Q. Pan, "Change detection in heterogenous remote sensing images via homogeneous pixel transformation," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 27, no. 4, p. 1822, 2018.

[7] M. Zanetti and L. Bruzzone, "A generalized statistical model for binary change detection in multispectral images," in *IGARSS 2016 - 2016 IEEE International Geoscience and Remote Sensing Symposium*, 2016.

[8] B. Liao, Y. Du, and X. Yin, "Fusion of infrared-visible images in ue-iot for fault point detection based on gan," *IEEE Access*, vol. 8, pp. 79 754–79 763, 2020.

[9] B. N. Subudhi, T. Veerakumar, S. Esakkirajan, and A. Ghosh, "Kernelized fuzzy modal variation for local change detection from video scenes," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 912–920, 2019.

[10] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[11] R. T. Collins, A. J. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 745–746, 2000.

[12] J. Patriarche and B. Erickson, "A review of the automated detection of change in serial imaging studies of the brain," *Journal of digital imaging*, vol. 17, no. 3, pp. 158–174, 2004.

[13] M. Bosc, F. Heitz, J.-P. Armspach, I. Namer, D. Gounot, and L. Rumbach, "Automatic change detection in multimodal serial mri: application to multiple sclerosis lesion evolution," *NeuroImage*, vol. 20, no. 2, pp. 643–656, 2003.

[14] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[15] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," *arXiv preprint arXiv:2101.06462*, 2021.

[16] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4634–4643.

[17] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "Rstnet: Captioning with adaptive attention on visual and non-

visual words," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 465–15 474.

[18] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," in *EMNLP*, 2018.

[19] H. Tan, F. Dernoncourt, Z. Lin, T. Bui, and M. Bansal, "Expressing visual relationships via language," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1873–1883.

[20] D. H. Park, T. Darrell, and A. Rohrbach, "Robust change captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4624–4633.

[21] Y. Tu, T. Yao, L. Li, J. Lou, S. Gao, Z. Yu, and C. Yan, "Semantic relation-aware difference representation learning for change captioning," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 63–73.

[22] X. Shi, X. Yang, J. Gu, S. Joty, and J. Cai, "Finding it at another side: A viewpoint-adapted matching encoder for change captioning," in *European Conference on Computer Vision*. Springer, 2020, pp. 574–590.

[23] Q. Huang, Y. Liang, J. Wei, C. Yi, H. Liang, H.-f. Leung, and Q. Li, "Image difference captioning with instance-level fine-grained feature representation," *IEEE Transactions on Multimedia*, 2021.

[24] M. Hosseinzadeh and Y. Wang, "Image change captioning by learning from an auxiliary task," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2725–2734.

[25] Y. Tu, L. Li, C. Yan, S. Gao, and Z. Yu, "Rˆ3net: Relation-embedded representation reconstruction network for change captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9319–9329.

[26] Y. Qiu, S. Yamamoto, K. Nakashima, R. Suzuki, K. Iwata, H. Kataoka, and Y. Satoh, "Describing and localizing multiple changes with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1971–1980.

[27] Z. Liao, Q. Huang, Y. Liang, M. Fu, Y. Cai, and Q. Li, "Scene graph with 3d information for change captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5074–5082.

[28] L. Yao, W. Wang, and Q. Jin, "Image difference captioning with pre-training and contrastive learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[29] H. Kim, J. Kim, H. Lee, H. Park, and G. Kim, "Agnostic change captioning with cycle consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2095–2104.

[30] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[31] Y. Tu, C. Zhou, J. Guo, S. Gao, and Z. Yu, "Enhancing the alignment between target words and corresponding frames for video captioning," *Pattern Recognition*, vol. 111, p. 107702, 2021.

[32] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: Spatial-temporal attention mechanism for video captioning," *IEEE transactions on multimedia*, vol. 22, no. 1, pp. 229–241, 2019.

[33] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, and X. Gao, "Task-adaptive attention for image captioning," *IEEE Transactions on Circuits and Systems for Video technology*, vol. 32, no. 1, pp. 43–51, 2021.

[34] L. Qu, M. Liu, J. Wu, Z. Gao, and L. Nie, "Dynamic modality interaction modeling for image-text retrieval," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1104–1113.

[35] Y. Tu, L. Li, L. Su, S. Gao, C. Yan, Z.-J. Zha, Z. Yu, and Q. Huang, "I2transformer: Intra-and inter-relation embedding transformer for tv show captioning," *IEEE Transactions on Image Processing*, 2022.

[36] L. Li, X. Gao, J. Deng, Y. Tu, Z.-J. Zha, and Q. Huang, "Long short-term relation transformer with global gating for video captioning," *IEEE Transactions on Image Processing*, 2022.

[37] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *arXiv preprint arXiv:1906.05963*, 2019.

[38] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 327–10 336.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[40] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[41] Y. Jung, D. Cho, S. Woo, and I. S. Kweon, "Global-and-local relative position embedding for unsupervised video summarization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 2020, pp. 167–183.

[42] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, and X. Fan, "Position focused attention network for image-text matching," *arXiv preprint arXiv:1907.09748*, 2019.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[44] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[46] S. Banerjee and A. 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*, 2005, pp. 65–72.

[47] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[48] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[49] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*. Springer, 2016, pp. 382–398.

[50] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

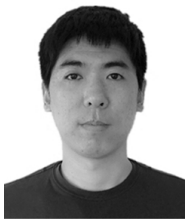
[53] P. Adam, G. Sam, C. Soumith, C. Gregory, Y. Edward, D. Zachary, L. Zeming, D. Alban, A. Luca, L. Adam *et al.*, "Automatic differentiation in pytorch," in *Proceedings of Neural Information Processing Systems*, 2017.

[54] A. Oluwasanmi, E. Frimpong, M. U. Aftab, E. Y. Baagyere, Z. Qin, and K. Ullah, "Fully convolutional captionnet: Siamese difference captioning attention model," *IEEE Access*, vol. 7, pp. 175 929–175 939, 2019.

[55] A. Oluwasanmi, M. U. Aftab, E. Alabdulkreem, B. Kumeda, E. Y. Baagyere, and Z. Qin, "Captionnet: Automatic end-to-end siamese difference captioning model with attention," *IEEE Access*, vol. 7, pp. 106 773–106 783, 2019.

Shengbin Yue received the B.S. degree in Communication from Kunming University of Science and Technology. He is currently pursuing a M.S. degrees in Communication and Information System at Kunming University of Science and Technology. His research interests include multimedia content analysis, especially for change captioning.

Yunbin Tu received the B.S. degree in Automation from Hangzhou Dianzi University, and the M.S. degree in Pattern Recognition and Intelligent System from Kunming University of Science and Technology. He is currently pursuing the Ph.D. degree from the School of Computer Science and Technology, University of Chinese Academy of Sciences. His research interests include multimedia content analysis, especially for video and change captioning.



Liang Li received his B.S. degree from Xi'an Jiaotong University in 2008, and Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China in 2013. From 2013 to 2015, he held a post-doc position with the Department of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China. Currently he is serving as the associate professor at Institute of Computing Technology, Chinese Academy of Sciences. He has also served on a number of committees of international journals and conferences. Dr. Li has published over 60 refereed journal/conference papers. His research interests include multimedia content analysis, computer vision, and pattern recognition.

Ying Yang received her B.S. degree in Electrical and Information Engineering from Wenhua College. She is currently pursuing a M.S. degree in Electronics and Communication Engineering from Kunming University of Science and Technology. Her research interests are image semantic segmentation.

Shengxiang Gao received the M.S. degree in Pattern Recognition and Intelligent System and the Ph.D. degree in Control Engineering from Kunming University of Science and Technology in 2005 and 2016, respectively. She is currently associate professor in School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. Her main research interests include machine learning, nature language processing and machine translation.

Zhengtao Yu received his Ph.D. degree in computer application technology from Beijing Institute of Technology, Beijing, China, in 2005. He is currently a professor in the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His main research interests include natural language processing, information retrieval and machine learning.