Neighborhood Contrastive Transformer for Change Captioning

Yunbin Tu, Liang Li, Li Su, Ke Lu, and Qingming Huang, *Fellow, IEEE*

arXiv:2303.03171v1 [cs.CV] 6 Mar 2023

Abstract—Change captioning is to describe the semantic change between a pair of similar images in natural language. It is more challenging than general image captioning, because it requires capturing fine-grained change information while being immune to irrelevant viewpoint changes, and solving syntax ambiguity in change descriptions. In this paper, we propose a neighborhood contrastive transformer to improve the model's perceiving ability for various changes under different scenes and cognition ability for complex syntax structure. Concretely, we first design a neighboring feature aggregating to integrate neighboring context into each feature, which helps quickly locate the inconspicuous changes under the guidance of conspicuous referents. Then, we devise a common feature distilling to compare two images at neighborhood level and extract common properties from each image, so as to learn effective contrastive information between them. Finally, we introduce the explicit dependencies between words to calibrate the transformer decoder, which helps better understand complex syntax structure during training. Extensive experimental results demonstrate that the proposed method achieves the state-of-the-art performance on three public datasets with different change scenarios. The code is available at https://github.com/tuyunbin/NCT.

Index Terms—Change captioning, Neighborhood contrastive transformer, Syntax dependencies.

I. INTRODUCTION

CHANGE captioning aims to describe what has changed between two semantically similar images, which is a novel task in the community of vision and language [1]–[3]. It extends the conventional image captioning [4], [5] further, *i.e.*, it needs to simultaneously deal with two images and describe their disagreement. This pushes forward the

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).

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

Ke Lu is with the School of Engineering Science, University of Chinese Academy of Sciences, Beijing, China, with Peng Cheng Laboratory, Nanshan, Shenzhen, Guangdong, China. (e-mail: luk@ucas.ac.cn).

Qingming Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China, with Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: qmhuang@ucas.ac.cn).
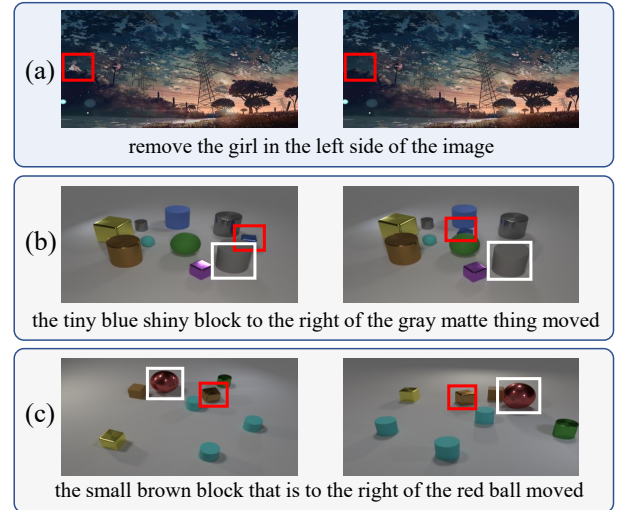
Fig. 1. The examples of change captioning. The first one is from image editing scene, where the removed object is inconspicuous. The second one shows that with both object move and moderate viewpoint change, and the changed object is partially occluded. The last one shows that with both object move and extreme viewpoint change, where the real movement is overwhelmed by pseudo movements. The changed objects and referents are shown in the red and white boxes, respectively

research of exploring the relationship and difference of image pair. In addition, it has wide applications, such as providing explanation of complex image editing effects for laypersons or visually-impaired users, outputting logs about monitored areas, and generating reports about pathological changes [6]–[8].

The key challenges are mainly embodied in the two aspects. First, the model should have the ability of fine-grained semantic comprehension, because change information is usually hard to pinpoint. For example, in Fig. 1 (a), the removed girl is easy to ignore, due to her inconspicuous position and vague shape. In Fig. 1 (b), The change is hard to be located, because the moved block is partially occluded. Second, the model should be immune to irrelevant distractors and only describe genuine semantic change. In a dynamic environment, it is nearly impossible to acquire two images under same viewpoint due to various factors, such as camera shaking, different shoot time, etc. In Fig. 1 (c), extreme viewpoint change leads to obvious pseudo movement for unchanged objects, which could overwhelm the real change and mislead the model into generating inaccurate sentences.

There have been previous endeavors for the above challenges. Despite progresses, these methods suffer from learning the effective change representation. Specifically, they compare

two images mainly at global or local level, where global refers to direct subtraction [9], [10] and local is to compute their similarity based on individual feature matching [8], [11], [12]. The former is too coarse to capture inconspicuous or occluded changes. The latter is more reasonable, but it is easily influenced by extreme viewpoint change, *e.g.,* in Fig. 1 (c) every object seems to move. In this case, such individual matching is unable to learn the stable features of change. We argue that to learn effective features across viewpoint changes, the model should compare details at neighborhood level. The reasons are that spatially neighboring objects are highly correlated in an image [13], where if an object changed, its relations with neighboring objects would change as well. Such relation change helps mine those inconspicuous changes. Besides, pseudo changes are actually the distortion of objects' scale and location, so the relations of these neighboring objects are not affected. Considering this, we try to dynamically integrate features at neighborhood level, thus helping the model resist viewpoint changes and locate the real change.

Besides, we observe that a change description usually consists of two parts: the semantic change and a referent, which makes it contain complex syntax structure. As shown in Fig. 1 (c), the main clause of this sentence is "the small brown block moved". However, the subject "block" and its predicate "moved" are separated by a subordinate clause describing the referent "ball". In this case, the word "moved" is closer to the word of "ball" than "block". During training, if a model does not understand syntax relations between words, it might learn wrong information from the ground-truth caption. To the best of our knowledge, this problem is disregarded by the existing methods. In fact, the above misunderstanding could be avoided if the model notices the direct dependency relation between "block" and "moved". Hence, it is necessary to introduce explicit dependency relations during training, which helps the model understand the syntax structure of captions.

In this paper, we propose a Neighborhood Contrastive Transformer to pinpoint change under different change scenarios, and endow the model with the syntax knowledge of dependency relation to address structural ambiguity. Concretely, given an image pair, a neighborhood feature aggregating is first designed to integrate neighboring context into features of each image. This helps the model resist viewpoint changes and perceive the fine-grained change under the guidance of neighboring referents. Then, based on similarity matching, a common feature distilling is customized to establish correspondences between the above two image features, so as to summarize their common features. Next, the stable features of change in each image are computed by removing common features, which are fused to learn contrastive features between the image pair. These contrastive features are subsequently fed into a transformer decoder to generate descriptions. During training, we provide the decoder with the prior knowledge of dependencies between words, which is beneficial to understand the complex syntax structure in ground-truth captions.

The contributions of this paper are summarized below: (1) A neighborhood contrastive transformer is proposed to pinpoint changes via performing neighborhood contrast between image pairs, where a neighborhood feature aggregating is designed to explore fine-grained changes and resist viewpoint change; a common feature distilling is devised to capture discriminative properties of each image and construct their contrastive features for sentence generation. (2) This work is the first attempt in this task to solve syntax structural ambiguity via introducing explicit dependencies between words. (3) Extensive experiments demonstrate that our method performs favorably against the state-of-the-art methods on three public datasets.

II. RELATED WORK

Image/Video Captioning. Before introducing the works of change captioning, we first review recently published works in conventional image/video captioning. TTA [14] detects visual tags from videos to bridge visual-textual gap, and presents a textual-temporal attention model to build alignment between words and frames. LSRT [15] proposes the long short-term relation transformer to fully mine objects' relations for caption generation. I^{2}Transformer [16] learns the intra- and inter-relation embedded representation from different modalities, which is fed into the standard transformer for caption generation. P+D attention [17] proposes a dual attention module on pyramid image feature maps, which can explore the visual-semantic correlations and refine generated captions. MSA [18] presents a multi-branch self-attention and duplicates it multiple times, in order to increase the expressive power of general self-attention model during caption generation. HTG+HMG [19] proposes a relation-aware attention by designing two kinds of graphs, namely linguistics-to-vision heterogeneous graph and vision-to-vision homogeneous graph.

Change Captioning. It is a new task in visual captioning, while it is more challenging. This is because it needs to understand the contents of two images, and further to describe their difference. The pioneer work [6] describes the change based on the surveillance scenarios. The work [7] elaborates the editing transformation between two images, as shown in Fig.1 (a). The common point of these two works is that they detect and describe changes between two well-aligned images. In fact, there exist viewpoint shifts as we shoot pictures, which poses a challenge to distinguish the real change from pseudo changes. Considering this, Park *et al.* [9] and Kim *et al.* [11] respectively release two datasets with moderate (Fig.1 (b)) and extreme viewpoint changes (Fig.1 (c)). To describe semantic change under viewpoint changes, Park *et al.* propose a DUDA model for localizing and describing changes, where they model the difference by subtracting two unaligned images, which might compute the difference features with noise [20].

To ease this problem, Hosseinzadeh *et al.* [10] leverage a retrieval model of TIRG [21] to regularize DUDA. Tu *et al.* [20] measure the relations between the subtracted change and image pair to judge if the change has actually happened. Instead of using direct subtraction, on the one hand, the works [8], [11], [22] first distill the common features between two images based on feature similarity. Then, they remove these features to explicitly capture the features of change. On the other hand, the works [12], [23] match the similar features between two images to implicitly infer the features of change. To enhance the visual-textual alignment, Kim *et al.*

[11] introduce a cycle consistency module to refine generated sentences. Yao *et al.* [12] model the fine-grained cross-modal alignment by the paradigm of pre-training to fine-tuning.

In addition, Liao *et al.* [24] introduce the 3D information of depths of objects to deal with viewpoint changes. They first input images into a pre-trained depth estimation model to obtain the depth maps. Meanwhile, they utilize a pre-trained Yolov4 to obtain the bounding boxes of the objects. Then, with these depth maps and bounding boxes of objects, they obtain the depths of objects. Since the accuracy of depths of objects heavily depends on the efficiency of two pre-trained models, the computed depth information is unreliable. Besides, the introducing of 3D information increases the complexity of model. Even so, leveraging 3D knowledge is another idea to overcome the influence of viewpoint changes. This inspires us to further explore this task in the future.

However, the aforementioned methods capture changed features between two images mainly based on the global (direct subtraction) or local (individual feature matching) level, while not trying to learn the features of change based on the neighborhood level. In addition, the problem of syntax ambiguous in ground-truth captions are disregarded. Instead, we propose a neighborhood contrastive transformer. It compares two images at neighborhood level to first capture differentiating properties from each image and then learn contrastive information between them. In addition, it employs dependency relations to solve the problem of structure ambiguity in change captions.

Contrastive Feature learning in Captioning. Learning contrastive features is to model similar/dissimilar image representations from similar/dissimilar image pairs [25]. This idea has been attempted by recent works in group captioning [26] and chest X-ray report generation [25]. On the one hand, given two groups of images, Li *et al.* [26] propose to use self-attention mechanism to capture common properties from each image group and then capture contrastive information between them. On the other hand, given a chest X-ray image and a set of norm images, Liu *et al.* [25] present a contrastive attention model to learn contrastive features between the input image with normal images. There are two major differences between our method and them. First, there exist irrelevant distractors (*e.g.,* viewpoint change) in our task, which brings the additional challenge to distinguish real change from pseudo change. Second, different from them matching feature individually, our method is first to aggregate neighboring features, and then perform feature matching at neighborhood level to construct contrastive features, which aims to identify fine-grained change while being immune to viewpoint change.

Syntax Knowledge Used in Captioning. There have been some attempts that use the syntax knowledges of Part-of-Speech (PoS) and syntax dependencies between words in captioning. On the one hand, Hou *et al.* [27] propose to model the syntactic structure and exploit the semantic primitive by learning the joint probability of the PoS sequence and words. Wang *et al.* [28] present a PoS generator to predict the global syntactic PoS information of sentences. Zhang *et al.* [29] and Deng *et al.* [30] propose to make the model adaptively generate each word based on its PoS, thus improving the cross-modal alignment. On the other hand, Zheng *et al.* [31] propose

to decode syntax components (subject, object and predicate) for targeting the action in video clips. Zhao *et al.* [32] devise a multi-modal dependency tree construction approach to capture the syntactic and semantic dependencies in long and complex video captions.

In change captioning, most works focus on learning an accurate change representation for caption generation, while ignoring the exploitation of syntax knowledge. Similar to Zhang *et al.* and Deng *et al.*, Tu *et al.* [20] introduce PoS information and propose an attention-based visual switch to dynamically use visual information. Different from this work, we aim to exploit explicit syntax dependencies between words to disambiguate syntax structure of change captions, which is beneficial to help the model differentiate changed object and its referent in ground-truth captions during training.

III. METHODOLOGY

As shown in Fig. 2, the architecture of our method consists of four parts: (1) a neighborhood feature aggregating module identifies the fine-grained change and resists irrelevant viewpoint changes; (2) a common feature distilling module extracts differentiating information from each image, and learns contrastive information between them; (3) a contrastive change localizer locates the specific change features on the two images; (4) a syntax-aware transformer decoder translates the learned features of change into a natural language sentence, and predicts the syntax dependencies between words.

A. Neighborhood Feature Aggregating

Formally, given two images of "before" I_{bef} and "after" I_{aft}, we exploit an off-the-shelf CNN to extract grid features for them, denoted as X_{bef} and X_{aft}, where $X \in \mathbb{R}^{C \times H \times W}$. C, H, W indicate the number of channels, height, and width. Although the CNN can capture local spatial context, these correlations are modeled based on single image without viewpoint changes and cannot be directly transferred to change captioning. Besides, the latest work [33] in semantic correspondence shows that local self-attention performs well in capturing relations between neighboring elements. Inspired by this, we design a neighborhood feature aggregating module to dynamically update each feature by integrating spatial context from the same neighborhood of two images.

Concretely, for $X_{bef(aft)} = \{x_1, \ldots, x_N\}$ ($N = HW$), where $x_i \in \mathbb{R}^C$, we first project the feature of i-th grid cell x_i into a low-dimensional embedding space of $\mathbb{R}^D$ by a shared linear transformation:

$$x'_i = M_v x_i + b_v + pos(\tilde{x}, \tilde{y}), \tag{1}$$

where $M_v \in \mathbb{R}^{D \times C}$ and $b_v \in \mathbb{R}^D$ are trainable parameters. $pos(\tilde{x}, \tilde{y}) \in \mathbb{R}^D$ is a learnable position embedding for i-th grid feature. Herein, $\tilde{x}$ and $\tilde{y}$ are the orders of each feature in the height and width dimensions of an image. Position embedding layers are two lookup tables of size (H, D/2) and size (W, D/2). Based on the orders of each feature, we can learn its position embeddings from height and width dimensions, and concatenate them as its position embedding. Then, for every grid feature x'_i, we pick out its $r \times r$ neighboring features
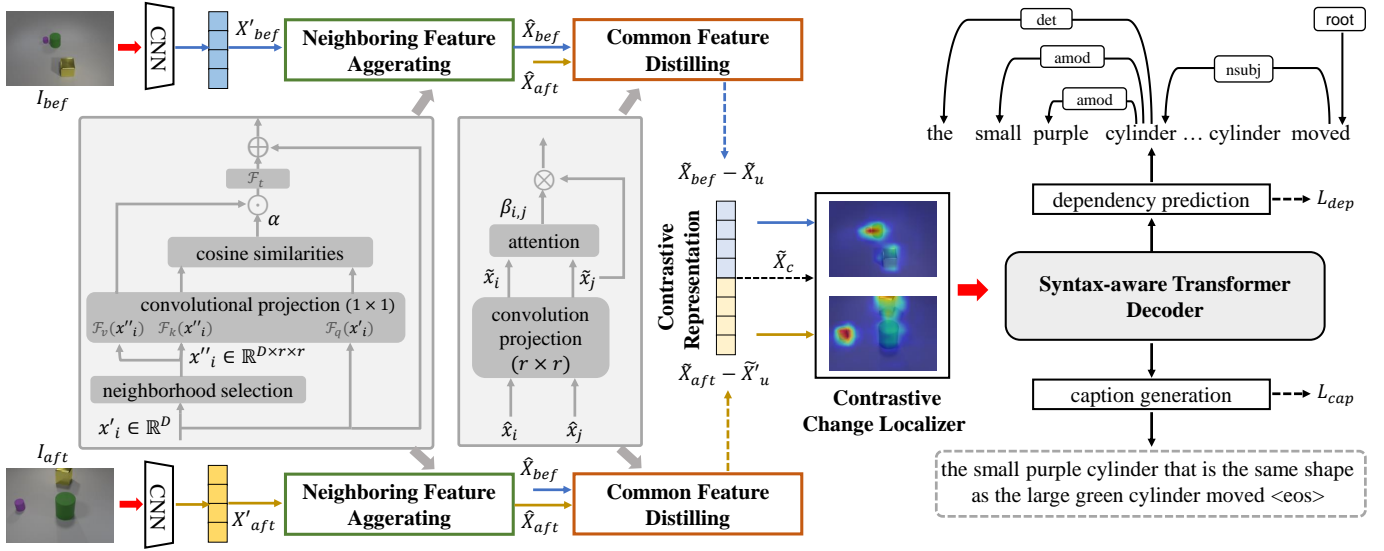
Fig. 2. The architecture of the proposed neighborhood contrastive transformer, including a neighborhood feature aggregating, a common feature distilling, a contrastive change localizer and a syntax-aware transformer decoder.

and acquire a neighborhood feature representation $X''_{bef(aft)} \in \mathbb{R}^{C \times H \times W \times r \times r}$. Next, we measure feature cosine similarities between x'_i and its $r \times r$ neighborhood:

$$
\begin{aligned}
e &= \Phi \left[\mathcal{F}_q \left(x'_i \right), \mathcal{F}_k \left(x''_i \right) \right], \\
\alpha &\sim \text{Softmax} \left(e \right),
\end{aligned}
\tag{2}
$$

where $\alpha \in \mathbb{R}^{r \times r}$ is the relation coefficient indicating how much message to obtain from the neighboring features; Φ is the cosine similarity function. $\mathcal{F}_q \in \mathbb{R}^D$ and $\mathcal{F}_k \in \mathbb{R}^{D \times r^2}$ are two convolution layers. Finally, x'_i is updated to $\hat{x}_i$ via aggregating related information from the neighboring features:

$$
\hat{x}_i = x'_i + \mathcal{F}_t(\sum_{r,r} \alpha \odot \mathcal{F}_v \left(x''_i \right)), \hat{x}_i \in \mathbb{R}^D,
\tag{3}
$$

where $\odot$ refers to element-wise multiplication. The above operation updates the original features into $\hat{X}_{bef}$ and $\hat{X}_{aft}$, which enables the model to identify inconspicuous and occluded changes, while being immune to viewpoint change.

B. Common Feature Distilling

As shown in Fig. 1, compared to the tiny change, most properties are identical between the image pair. Hence, it is natural to find and remove the common portion from the two images, and the remaining information can be treated as contrastive features. Motivated by this, a common feature distilling module is designed to compare two images and learn an effective contrasive representation.

Herein, we learn the change features in $\hat{X}_{bef}$ compared to $\hat{X}_{aft}$. In detail, we first exploit a shared transformation layer with depth-wise separate convolutions to project $\hat{X}_{bef}$ and $\hat{X}_{aft}$ into a common semantic space. This further captures spatial correlations in the same neighborhood:

$$
\tilde{X}_{bef(aft)} = \mathcal{F}_{depth} \left(\hat{X}_{bef(aft)}, s \right), \tilde{X} \in \mathbb{R}^{D \times H \times W},
\tag{4}
$$

where s is with the kernel size of $r \times r$, and we reshape $\tilde{X}_{bef(aft)}$ to $\tilde{X}_{bef(aft)} \in \mathbb{R}^{N \times D}$. Then, we measure the similarity between every feature $\tilde{x}^b_i$ in $\tilde{X}_{bef}$ and every feature $\tilde{x}^a_j$ in $\tilde{X}_{aft}$ by the dot-product attention:

$$
\beta_{i,j} = \frac{\exp \left(\beta'_{i,j} \right)}{\sum_j \exp \left(\beta'_{i,j} \right)}, \quad \beta'_{i,j} = \tilde{x}^{bT}_i \tilde{x}^a_j,
\tag{5}
$$

where $\beta_{i,j} \in B$ is a set of similarity scores to indicate which features are the common properties between $\tilde{X}_{bef}$ and $\tilde{X}_{aft}$. Then, the common features are extracted from $\tilde{X}_{aft}$ under the guidance of the learned similarity score matrix B:

$$
\tilde{X}_u = B \cdot \tilde{X}_{aft}.
\tag{6}
$$

Next, we remove the common features $\tilde{X}_u$ from $\tilde{X}_{bef}$ to distill the change features:

$$
\tilde{X}^{bef}_c = \tilde{X}_{bef} - \tilde{X}_u.
\tag{7}
$$

By that analogy, we distill the change features $\tilde{X}^{aft}_c$ in $\tilde{X}_{aft}$ with reference to $\tilde{X}_{bef}$. Finally, we construct the contrastive representation between two images by fusing the above change features, which is implemented by a fully-connected layer with the ReLU activation function:

$$
\tilde{X}_c = \text{ReLU} \left(\left[\tilde{X}^{bef}_c; \tilde{X}^{aft}_c \right] W_h + b_h \right).
\tag{8}
$$

C. Contrastive Change Localizer

After learning the contrastive representation $\tilde{X}_c$, we introduce a contrastive change localizer based on spatial attention mechanism, which is used to pinpoint change on the two images. Concretely, it first generates two attention maps by using $\tilde{X}_c$ to query each image representation, respectively:

$$
\begin{aligned}
\gamma_{bef} &= \sigma \left(MLP \left(\left[\tilde{X}_c; \tilde{X}_{bef} \right] \right) \right), \\
\gamma_{aft} &= \sigma \left(MLP \left(\left[\tilde{X}_c; \tilde{X}_{aft} \right] \right) \right),
\end{aligned}
\tag{9}
$$

where MLP is a two-layer multi-layer perceptron with the ReLU activation function in between. [;] and σ denote concatenation operation and sigmoid activation function. Further, the specific feature of change is localized via implementing a weighted-sum pooling on each image representation over the spatial dimensions, respectively:

$$l_{bef} = \sum_{H,W} \gamma_{\text{bef}} \odot \tilde{X}_{bef}, l_{bef} \in \mathbb{R}^D, \\ l_{aft} = \sum_{H,W} \gamma_{\text{aft}} \odot \tilde{X}_{aft}, l_{\text{aft}} \in \mathbb{R}^D. \tag{10}$$

D. Syntax-aware Language Decoder

With the pooling change features l_{bef}, l_{aft}, and their difference feature l_{diff}, we first concatenate them as $V \in \mathbb{R}^{3 \times D}$. Then, the decoder of transformer learns the cross-modal alignment between the word embedding features $E[W] = \{E[w_1], ..., E[w_m]\}$ and visual features V. Finally, the decoder exploits attended features of change to generate sentences, during which we introduce the syntax knowledge of dependencies between words to calibrate the decoder. This aims to solve the problem of syntax ambiguity in change descriptions.

1) Background Knowledge: We first briefly review the framework of standard transformer decoder. The key module is the scaled dot-product attention. Given a query matrix $Q \in \mathbb{R}^{T_q \times d_k}$, key matrix $K \in \mathbb{R}^{T_v \times d_k}$ and value matrix $V \in \mathbb{R}^{T_v \times d_v}$, the attention result is computed as:

$$\text{Attention}\,(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}, \dim = 1\right) V. \tag{11}$$

The multi-head attention is based on the scaled dot-product attention. It consists of h different "heads". For each head, the attention result is computed by:

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \tag{12}$$

Afterward, the multi-head attention operation is to concatenate all the heads, which is defined as:

$$\text{MultiHead}\,(Q, K, V) = \text{Concat}_{i=1...h}\,(\,\text{head}_i)\, W^O. \tag{13}$$

Further, the output of each attention layer x is fed into a feed-forward network (FFN) based on a non-linear transformation:

$$\text{FFN}(\boldsymbol{x}) = \text{GELU}\,(\boldsymbol{x}W_{f1} + \boldsymbol{b}_{f1})\, W_{f2} + \boldsymbol{b}_{f2}. \tag{14}$$

Then, we give a introduction about the dependencies between words. In natural language processing, dependency parsing refers to the process of examining the dependencies between the linguistic units (*e.g.*, words) of a sentence, in order to determine its grammatical structure. That is, syntax dependency is the notion that words are connected to each other by directed links. The verb is taken to be the structural center of clause structure and tagged as "root". All other syntactic words are either directly or indirectly connected to the "root" in terms of the directed links. In Fig. 2, we mainly illustrate the main words of this change caption. A dependence tag indicates the relationship between two words. For example, the word "moved" changes the meaning of the noun "cylinder". Therefore, we can find that a dependency from "moved" to "cylinder", where "moved" is the pinnacle and "cylinder" is the kid or dependent. The tag of this dependency is "nsubj", which stands for nominal subject of this sentence. The verb "moved" is the root in this dependency structure. In addition, we notice that there is no directed link between the other "cylinder" and the "moved". Based on these directed links, the model can better understand complex structure in a sentence and thus identify which object changed.

2) Decoding Stage: The decoder contains a stack of N identical layers. At the l-th decoder layer, the masked self-attention layer, which prevents the model from seeing future words, first takes the word embedding features $E[W] = \{E[w_1], ..., E[w_m]\}$ as the inputs and models their relationships. The operation is defined as:

$$\hat{E}[W] = \text{LN}\,(E[W] + \text{MultiHead}\,(E[W], E[W], E[W])), \tag{15}$$

where LN is short for layer normalization [34]. Then, the decoder utilizes the attended features $\hat{E}[W]$ to query the most related features from V based on the cross-attention layer:

$$\hat{H} = \text{LN}\,(E[\hat{W}] + \text{MultiHead}\,(E[\hat{W}], V, V)). \tag{16}$$

Afterward, the $\hat{H}$ is passed to a feed-forward layer:

$$\tilde{H} = \text{LN}(\hat{H} + \text{FFN}(\hat{H})). \tag{17}$$

Finally, the probability distributions of target words and dependencies are calculated via two separate single hidden layers:

$$W = \text{Softmax}\left(\tilde{H}W_c + b_c\right), \\ D = \text{Softmax}\left(\tilde{H}W_d + b_d\right), \tag{18}$$

where $W_c \in \mathbb{R}^{D \times U}$, $W_d \in \mathbb{R}^{D \times n}$, $b_c \in \mathbb{R}^U$, and $b_d \in \mathbb{R}^n$ are the parameters to be learned. U is the dimension of vocabulary size; n is the number of dependency relations.

E. Joint Training

We jointly train the caption generator and dependency predictor in an end-to-end manner by maximizing the likelihood of the observed word sequences and dependency relations. Given the target ground-truth caption words $(w_1^c, \ldots, w_m^c)$ and dependency relations $(w_1^d, \ldots, w_m^d)$, we minimize the negative log-likelihood loss of caption generator and dependency predictor, respectively:

$$L_{cap}(\theta_c) = -\sum_{t=1}^{m} \log p\,(w_t^c \mid w_{<t}^c; \theta_c), \\ L_{dep}(\theta_d) = -\sum_{t=1}^{m} \log p\,(w_t^d \mid w_{<t}^d; \theta_d), \tag{19}$$

where θ_c and θ_p are the parameters of the caption generator and dependency predictor, respectively. m is the length of the caption and dependencies. The final loss function is optimized as follows:

$$L(\theta) = L_{cap} + \lambda L_{dep}, \tag{20}$$

where λ is a trade-off parameter to balance the contributions from the caption generator and dependency predictor.

IV. EXPERIMENTS

A. Datasets

Image Editing Request dataset [7] is comprised of 3,939 real image pairs with 5,695 editing instructions. Each image pair in the training set has one instruction, and each image pair in the validation and test sets has three instructions. The changed objects in this dataset are usually inconspicuous and vague. We use the official split with 3,061 image pairs for training, 383 for validation, and 495 for testing.

CLEVR-Change is a large-scale synthetic dataset [9] with moderate viewpoint change. It has 79,606 image pairs and 493,735 captions, including five change types, *i.e.*, "Color", "Texture", "Add", "Drop", and "Move". It has two change settings: both scene and pseudo change and only pseudo change. We use the official split with 67,660 for training, 3,976 for validation and 7,970 for testing.

CLEVR-DC is a large-scale synthetic dataset [11] to simulate extreme viewpoint shifts. It consists of 48,000 pairs with the same change types as CLEVR-Change. We use the official split with 85% for training, 5% for validation, and 10% for test, respectively.

B. Evaluation Metrics

Following the state-of-the-art methods [8], [10], [35], five metrics are used to evaluate the generated sentences, *i.e.*, BLEU-4 (B) [36], METEOR (M) [37], ROUGE-L (R) [38], CIDEr (C) [39], and SPICE (S) [40]. BLEU-4 is exploited for corpus level comparisons of 4-gram matches and has been widely used in machine translation task. METEOR is designed to measure the relationship between candidate and reference sentences based on exact token matching. ROUGE-L computes the word correlations that co-exist in two sentences in the same order, based on the Longest Common Sub-sequence (LCS). CIDEr is recently proposed and especially designed for the captioning task to capture human judgment of consensus. SPICE is also a new metric and designed for captioning task, which compares semantic propositional content between candidate and reference sentences. We compute results based on the Microsoft COCO evaluation server [41].

C. Implementation Details

For a fair comparison, we follow the state-of-the-art methods to use a ResNet-101 model [42] pre-trained on the Imagenet dataset [43] for extracting grid features of an image pair, with the dimension of $1024 \times 14 \times 14$. We first project these features into a lower dimension of 512. The hidden size in the overall model and word embedding size in the decoder are set to 512 and 300. To obtain the ground-truth dependencies, we exploit a pre-trained Biaffine Parse [44] to extract the explicit dependency relations of each sentence in the training sets. We set the layer number of neighborhood feature aggregating as 1; the number of dependency tags as 49; the neighborhood range r as 3; the layer number of decoder as 2; the number of attention head as 8.

During training, on CLEVR-Change and CLEVR-DC, we set the batch size and learning rate as 128 and 2×10^{-4}. On

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON CLEVR-CHANGE ON TOTAL PERFORMANCE. "*" REPRESENTS THIS MODEL IS TRAINED WITH THREE PRE-TRAINING TASKS.

Method	Total				
	B	M	R	C	S
DUDA (ICCV'19)	47.3	33.9	-	112.3	24.5
M-VAM (ECCV'20)	50.3	37.0	69.7	114.9	30.5
DUDA+TIRG (CVPR'21)	51.2	37.7	70.5	115.4	31.1
IFDC (TMM'21)	49.2	32.5	69.1	118.7	-
R^{3}Net+SSP (EMNLP'21)	54.7	39.8	73.1	123.0	32.6
VACC (ICCV'21)	52.4	37.5	-	114.2	31.0
SRDRL+AVS (ACL'21)	54.9	40.2	73.3	122.2	32.9
SGCC (ACM MM'21)	51.1	40.6	73.9	121.8	32.2
MCCFormers-D (ICCV'21)	52.4	38.3	-	121.6	26.8
PCL w/o PT (AAAI'22)	32.7	27.7	57.2	89.8	-
PCL w/ PT (AAAI'22) *	51.2	36.2	71.7	128.9	-
NCT (Ours)	55.1	40.2	73.8	124.1	32.9

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON CLEVR-CHANGE ON SCENE CHANGE.

Method	Scene Change				
	B	M	R	C	S
DUDA (ICCV'19)	42.9	29.7	-	94.6	19.9
DUDA+TIRG (CVPR'21)	49.9	34.3	65.4	101.3	27.9
IFDC (TMM'21)	47.2	29.3	63.7	105.4	-
R^{3}Net+SSP (EMNLP'21)	52.7	36.2	69.8	116.6	30.3
SRDRL+AVS (ACL'21)	52.7	36.4	69.7	114.2	30.8
NCT (Ours)	53.1	36.5	70.7	118.4	30.9

Image Editing Request, the batch size and learning rate are set to 32 and 2×10^{-4}. We use Adam optimizer [45] to minimize the negative log-likelihood loss of Eq. (20). In the inference phase, the greedy decoding strategy is used to generate target captions. Both training and inference are implemented with PyTorch [46] on an RTX 3090 GPU.

D. Performance Comparison

1) Results on the CLEVR-Change Dataset.: We compare the proposed method with the state-of-the-art methods in: 1) total performance evaluating the overall performance under both scene and pseudo changes; 2) scene change; 3) different change types. The ten comparison methods are DUDA [9], M-VAM [8], IFDC [22], DUDA+TIRG [10], R^{3}Net+SSP [35], VACC [11], SRDRL+AVS [20], MCCFormers-D [23], SGCC [24], and PCL w/ and w/o PT (pre-training) [12]. Herein, PCL designs three pre-training tasks to enhance the fine-grained alignment between image differences and captions. The authors of PCL pre-train the model with 8K warm-up steps and 250K iterations in total. In contrast to them, the other compared methods are trained in an end-to-end manner. Therefore, we compare PCL with and without pre-training for a fair comparison. The results are shown in Table I - V.

In Table I, we can observe that 1) NCT achieves superior results on most metrics; 2) note that MCCFormers-D is also based on transformer and identifies change based on feature similarity. There are two major differences between it and ours. First, it implements individual feature matching between two sets of features. Instead, our NCT aims to compare two images at neighborhood level to capture contrastive properties between them, which helps perceive fine-grained change while being

TABLE III
A DETAILED BREAKDOWN OF EVALUATION ON CIDEr WITH DIFFERENT CHANGE TYPES: "(C) COLOR", "(T) TEXTUR", "(A) ADD", "(D) DROP", AND "(M) MOVE".

| Method | CIDEr | | | | |
	C	T	A	D	M
DUDA (ICCV'19)	120.4	86.7	108.3	103.4	56.4
M-VAM (ECCV'20)	122.1	98.7	126.3	115.8	82.0
DUDA+TIRG (CVPR'21)	120.8	89.9	119.8	123.4	62.1
IFDC (TMM'21)	133.2	99.1	128.2	118.5	82.1
R^{3}Net+SSP (EMNLP'21)	139.2	123.5	122.7	121.9	**88.1**
SRDRL+AVS (ACL'21)	136.1	122.7	121.0	126.0	78.9
SGCC (ACM MM'21)	128.0	122.9	117.1	116.9	77.1
PCT w/ PT (AAAI'22)	131.2	101.1	**133.3**	116.5	81.7
NCT (Ours)	**140.2**	**128.8**	<u>128.4</u>	**129.0**	<u>86.0</u>

TABLE IV
A DETAILED BREAKDOWN OF EVALUATION ON SPICE.

| Method | SPICE | | | | |
	C	T	A	D	M
DUDA (ICCV'19)	21.2	18.3	22.4	22.2	15.4
M-VAM (ECCV'20)	28.0	26.7	30.8	32.3	22.5
DUDA+TIRG (CVPR'21)	29.7	27.4	31.4	30.8	23.5
R^{3}Net+SSP (EMNLP'21)	31.6	30.8	<u>32.3</u>	31.7	25.4
SRDRL+AVS (ACL'21)	**32.4**	30.9	**33.0**	32.4	25.4
SGCC (ACM MM'21)	30.0	31.1	30.8	30.1	25.3
NCT (Ours)	**32.4**	**31.8**	<u>32.3</u>	**32.6**	**25.5**

TABLE V
A DETAILED BREAKDOWN OF EVALUATION ON METEOR.

| Method | METEOR | | | | |
	C	T	A	D	M
DUDA (ICCV'19)	32.8	27.3	33.4	31.4	23.5
M-VAM+RAF (ECCV'20)	35.8	32.3	37.8	36.2	27.9
DUDA+TIRG (CVPR'21)	36.1	30.4	37.8	36.7	27.0
IFDC (TMM'21)	33.1	27.9	36.2	31.4	31.2
R^{3}Net+SSP (EMNLP'21)	38.9	35.5	38.0	37.5	30.9
SRDRL+AVS (ACL'21)	39.0	35.6	38.9	**38.0**	30.1
SGCC (ACM MM'21)	37.8	36.1	38.9	36.7	**32.8**
NCT (Ours)	**39.1**	**36.3**	**39.0**	37.2	30.5

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON CLEVR-DC.

Method	B	M	R	C	S
DUDA	40.3	27.1	-	56.7	16.1
DUDA + CC	41.7	27.5	-	62.0	16.4
M-VAM	40.9	27.1	-	60.1	15.8
M-VAM+CC	41.0	27.2	-	62.0	16.4
VA	44.5	29.2	-	70.0	17.1
VACC	45.0	29.3	-	71.7	**17.6**
NCT	**47.5**	**32.5**	**65.1**	**76.9**	15.6

2) Results on the CLEVR-DC Dataset: The experiment is also carried out on a newly released synthetic dataset (ICCV'21) with extreme viewpoint changes. We compare with six state-of-the-art methods: DUDA/DUDA+CC [9], M-VAM/M-VAM+CC [8], and VA/VACC [11].

The comparison results are shown in Table VI. We find that NCT outperforms the state-of-the-art methods on most metrics by a large margin. This validates that our method has a good robustness in any viewpoint change. This mainly benefits from the fact of capturing contrastive information between a pair of images at neighborhood level, because under viewpoint changes, the object relations are stable within/between local neighborhoods of no change.

TABLE VII
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON IMAGE EDITING REQUEST DATASET.

Method	B	M	R	C	S
multi-head att	6.1	11.8	35.1	22.8	-
static rel-att	5.8	12.6	35.5	20.7	-
dynamic rel-att	6.7	12.8	37.5	26.4	-
NCT	**8.1**	**15.0**	**38.8**	**34.2**	**12.7**

3) Results on the Image Editing Request Dataset: We conduct the experiment on another challenging dataset, Image Editing Request. The changed objects in this dataset are usually vague and inconspicuous. We compare NCT with three state-of-the-art methods reported by Tan *et al.* [7]: multi-head att, static rel-att, and dynamic rel-att.

Table VII shows that NCT outperforms the state-of-the-art methods by a large margin. This indicates that the proposed method can accurately describe which part of the "source" image has been edited by capturing neighborhood contrastive features and achieving syntax disambiguity based on explicit dependencies between words.

In short, experiments on the above three datasets show that our method has a good generalization of change localization and description on different change scenarios.

immune to viewpoint change. Second, different from it based on the standard transformer for caption generation, we exploit syntax dependencies to calibrate decoder, which helps better understand complex syntax structure of change descriptions. 3) Compared with SGCC, the proposed NCT is a little lower on the metrics of METEOR and ROUGE-L. Our conjecture is that SGCC exploits more visual modalities than ours, such as semantic attributes extracted by Yolov4 [47] and image depth maps that are computed by Monodepth2 [48]. In contrast, our NCT surpasses SGCC on the other metrics by a large margin. 4) Compared with PCL, NCT surpasses it without pre-training by a large margin. For PCL with pre-training, NCT also outperforms it on the three metrics. For CIDEr, NCT is a little lower. Our conjecture is that PCL leverages three pre-training tasks (with 8K warm-up steps and 250K iterations in total) to augment the model.

In Table II, it is noted that NCT outperforms the state-of-the-art methods on every metrics, especially improving CIDEr score by a large margin. In Table III - V, we compare NCT with state-of-the-art methods under the specific change types using the metrics of CIDEr, SPICE and METEOR. Especially, CIDEr and SPICE are especially designed for evaluating captioning performance. The results show that our NCT achieves the superior results over the state-of-the-art methods in almost every category. This shows that our method has a good generalization ability under different change types.

In a word, compared to the state-of-the-art methods in different situations, the proposed NCT achieves the encouraging performance. This superiority results from that 1) the neighborhood feature aggregating and common feature distilling help learn reliable contrastive features and resist irrelevant viewpoint changes; 2) the syntax dependencies can solve the problem of structure ambiguity in change descriptions.

TABLE VIII
ABLATION STUDIES BASED ON TOTAL PERFORMANCE ON
CLEVR-CHANGE

Method	B	M	R	C	S
Diff-sub	53.3	38.8	72.1	119.7	31.8
NFA	54.3	39.7	73.1	121.9	32.0
CFD	54.1	39.6	73.1	122.8	32.2
ST	53.7	39.4	72.7	120.7	31.9
NCT w/o S	54.6	40.0	73.6	123.4	32.5
NCT	**55.1**	**40.2**	**73.8**	**124.1**	**32.9**

E. Ablation Studies

We carry out ablation studies to validate the effectiveness of each proposed module and the full model. (1) Diff-sub is a transformer-based baseline model which computes difference features by direct subtraction. Specifically, it first directly subtracts two image features to obtain the difference representation. Then, it uses the spatial attention mechanism to select the most relevant features on each image based on the subtracted representation. Finally, the specifically changed features are fed into a standard transformer decoder for caption generation. (2) NFA performs neighborhood feature aggregating before subtraction. (3) CFD distills common features to construct contrastive features, instead of using direct subtraction. (4) ST refers to syntax-aware transformer decoder that uses syntax dependency relation to augment the baseline model. (5) NCT w/o S is the neighborhood contrastive transformer without syntax dependency. (6) NCT is the proposed full model: neighborhood contrastive transformer with syntax dependency.

The ablation studies are based on the total performance on CLEVR-Change, and the results are shown in Table VIII. We observe that 1) compared to the baseline model, each module and the full model achieve consistent improvements; 2) the performance of NFA and CFD are close, and better performance is achieved by their combination; 3) only augmenting the baseline model with syntax dependency, the improvement is slight, and the best performance is achieved through combining it with NCT. The above observations indicate that 1) the effectiveness of each proposed module and the full model; 2) each module not only plays its unique role, but also supplements the other; 3) only if the model learns effective contrastive features by neighborhood feature aggregating and common feature distilling, the syntax-aware decoder would use these features to yield correct sentences.

F. Evaluating the Model's Robustness under Various Degrees of Viewpoint Changes

To evaluate the robustness of NCT under different degrees of viewpoint changes, following the pioneer work [9], we compute the IoU of the bounding boxes of the objects (except the changed object) across the two images, where the lower IoU refers to higher difficulty. Herein, we employ SPICE to evaluate the sentences generated by Diff-sub (baseline model) and NCT with respect to different IoU of the object's bounding boxes. The results are shown in the left sub-figure of Fig. 3. It is noted that NCT consistently outperforms the baseline by a large margin. This indicates that the proposed method
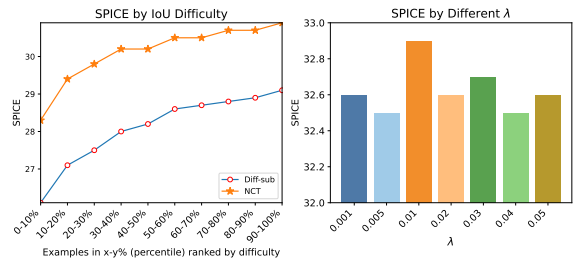

Fig. 3. Left sub-figure is the visualization of captioning performance (SPICE) that is breakdown by viewpoint change (measured by IoU); right sub-figure is the effects of the trade-off parameter λ on CLEVR-Change.
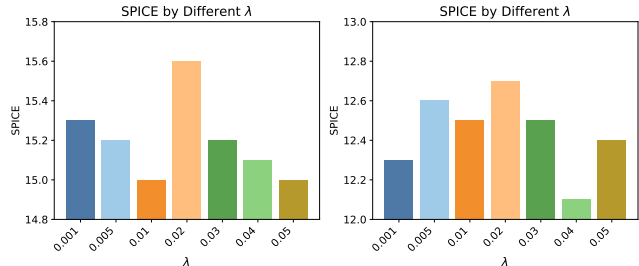

Fig. 4. The effects of the trade-off parameter λ on CLEVR-DC (left sub-figure) and Image Editing Request (right sub-figure).

can identify reliable change and handle the varying degrees of viewpoint changes.

G. Study on the Trade-off Parameter λ

In this section, we discuss the effect of the trade-off parameter λ in Eq. (20) on CLEVR-Change, CLEVR-DC and Image Editing Request. This parameter is to balance the contributions from the caption generator and dependency predictor. On CLEVR-Change, with different values, the obtained SPICE scores are shown in the right sub-figure of Fig. 3. We find that as the values of λ increasing or decreasing, the performance of NCT changes. This is mainly because the whole model will focus much on one part but ignore the supervision signal from the other. Based on this, we empirically set λ to 0.01. In addition, for other two datasets, CLEVR-DC and Image Editing Request, the results are shown in Fig. 4. With different values, the obtained SPICE scores on CLEVR-DC in the left sub-figure, and on Image Editing Request in the right sub-figure. It is noted that similar to the experimental results on CLEVR-Change, as the values of λ increasing or decreasing, the performance of NCT changes. On the both datasets, the better value is 0.02. The above analysis shows that the value of this trade-off parameter λ is close on different datasets, which validates that the proposed method has a good robustness on different change scenarios.

H. Study on the Parameter of Neighborhood Range r

In this section, we will analyze the effect of neighborhood range r. Herein, we set it as 3 and 5, respectively. Note that as this value is larger than 5, the computation cost increases sharply and is more than one RTX 3090 GPU. With different
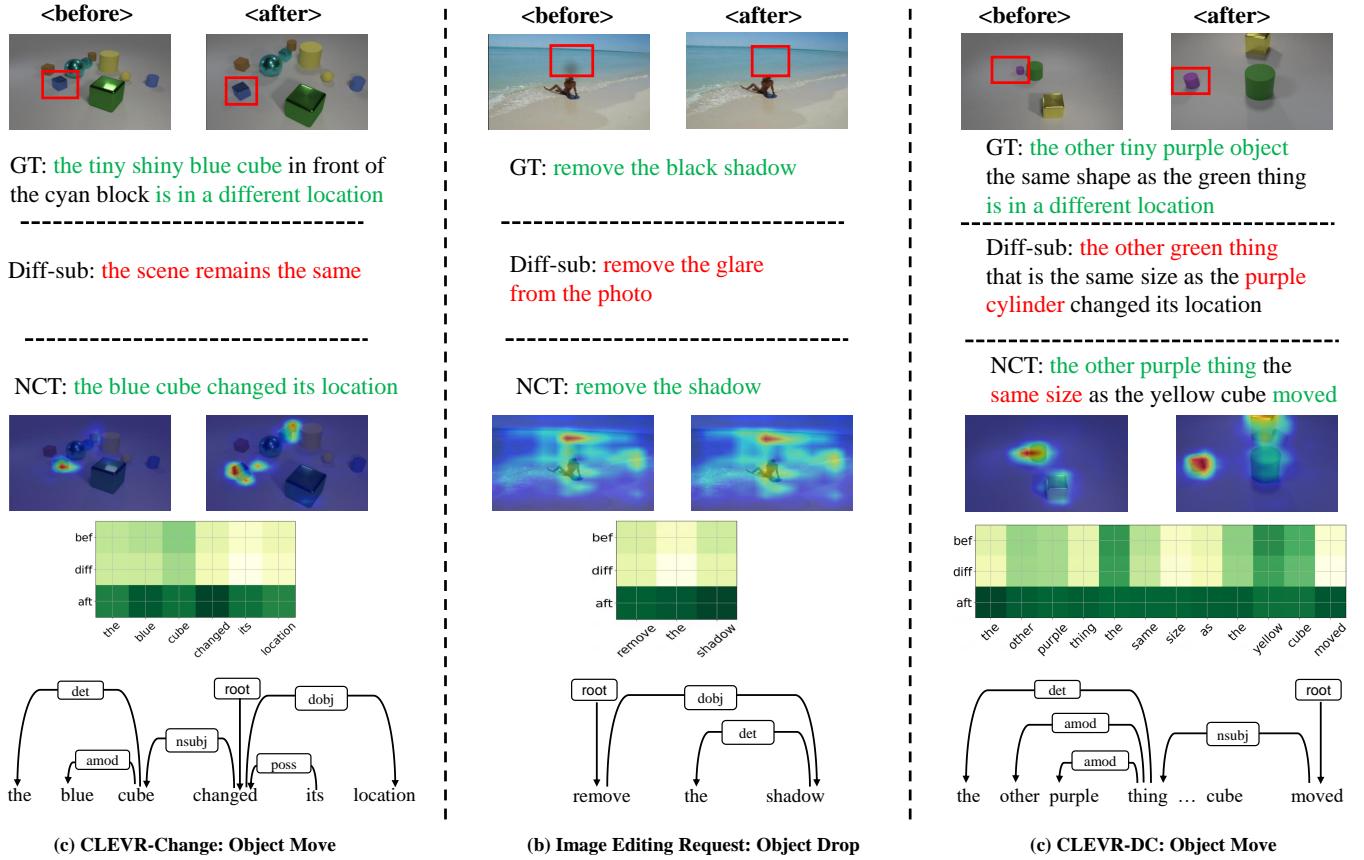
<before> **<after>** **<before>** **<after>** **<before>** **<after>**

GT: the tiny shiny blue cube in front of the cyan block is in a different location

Diff-sub: the scene remains the same

NCT: the blue cube changed its location

GT: remove the black shadow

Diff-sub: remove the glare from the photo

NCT: remove the shadow

GT: the other tiny purple object the same shape as the green thing is in a different location

Diff-sub: the other green thing that is the same size as the purple cylinder changed its location

NCT: the other purple thing the same size as the yellow cube moved

(c) CLEVR-Change: Object Move **(b) Image Editing Request: Object Drop** **(c) CLEVR-DC: Object Move**

Fig. 5. Qualitative examples on CLEVR-Change, Image Editing Request, and CLEVR-DC. For each example, we report the captions generated by Diff-sub and NCT along with the ground-truth (GT) captions. Correct and incorrect parts of the captions are in green and red, respectively. We visualize the results of change localization on the "before" and "after" images, and illustrate the heat map visualization to show the semantic alignment between changed features and corresponding words. We visualize the predicted dependencies of each example. The ground-truth changes are shown in red boxes.

values, the captioning performance and parameter number are shown in Table IX.

We find that there is no obvious performance increase as we enlarge neighborhood range, and the results on most metrics even decrease. Our conjecture is that the model only needs the closet referents to guide where the changed object is, so the neighborhood range of 3×3 is suitable. Based on this, we empirically set r to 3 on the three datasets.

TABLE IX
STUDY THE EFFECTS OF THE PARAMETER OF NEIGHBORHOOD RANGE r ON THE THREE DATASETS, WHERE CC, CD, AND IER ARE SHORT FOR CLEVR-CHANGE, CLEVR-DC, AND IMAGE EDITING REQUEST.

r	Set	Params	B	M	C	S
3 × 3	CC	26.65M	**55.1**	**40.2**	124.1	**32.9**
5 × 5	CC	26.74M	54.9	39.9	**124.8**	32.7
3 × 3	CD	26.70M	**47.5**	**32.5**	**76.9**	**15.6**
5 × 5	CD	26.79M	44.4	31.3	71.1	14.5
3 × 3	IER	34.07 M	8.1	**15.0**	34.2	**12.7**
5 × 5	IER	34.17M	**9.5**	14.7	**36.9**	11.9

I. Qualitative Analysis

To evaluate the overall performance of NCT about change localization and caption generation, we conduct qualitative analysis on CLEVR-Change, Image Editing Request, and CLEVR-DC, as shown in Fig. 5. For each image pair, we report the captions yielded by the baseline model of Diff-sub and our NCT along with the ground truth (green words). To evaluate the accuracy of changed objects, we also visualize the changed results based on the attention weights of change detection. For the first example, the object movement is slight, which makes Diff-sub misjudge that there is nothing changed. For the second example, the removed object is too faint to notice. In this case, Diff-sub directly subtracts two images to compute change features, which wrongly judges the "shadow" as "glare" and fails to generate the accurate sentence. For the third example, extreme viewpoint change results in pseudo movements of all objects, which makes Diff-sub misidentify really changed object. Besides, another possible reason of this failure is that the syntax structure in the ground-truth sentence is complex. That is, the referent "green thing" is closer to changed type than "purple thing", which might make Diff-sub misjudge the changed object as the "green thing". In contrast to Diff-sub, the proposed NCT can accurately localize and describe changed objects. This mainly benefits from that 1) the neighboring feature aggregating helps the model identify the real change while being immune to viewpoint change; 2) the common feature distilling can effectively summarize common properties of the image pair and extract differentiating features from each image, so as to construct constrictive
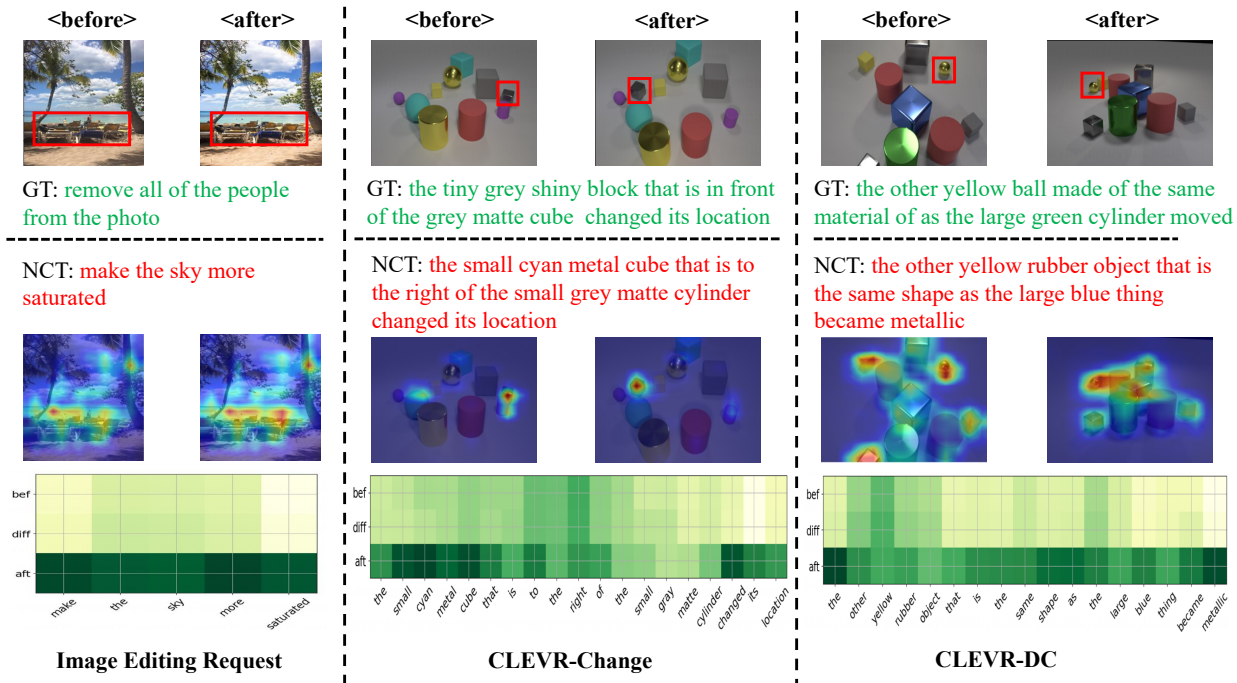
Fig. 6. Failure examples obtained by NCT on the test split of Image Editing Request, CLEVR-Change, and CLEVR-DC.

features between them; 3) introducing dependency relations between words helps solve syntax ambiguity in sentences and understand their complex syntax structure. For instance, in the third example, our NCT can predict the directed link between really changed object "purple thing" and changed type "moved", so as to identify the really changed object.

In addition, we find that in the third example, NCT predicts that the tiny purple object is the same size as the yellow cube. The possible reasons for this misunderstanding are that 1) NCT compares their size mainly based on the "after" image. 2) The change types in this dataset do not include size change. In this case, the model has a limited ability to accurately compare the size between two objects. Therefore, further exploration to identify the changes of objects' size is warranted in future research. More qualitative examples are shown in the supplementary material.

J. Discussion

Fig. 6 illustrates the failure cases obtained by NCT on the three datasets with different change scenarios. For all the change scenes, we can observe that the proposed NCT successfully localizes the changed objects. However, it fails to describe them in accurate sentences. For the failure cases, our conjecture is that the visual signal of change appears in a inconspicuous region with weak feature in each example. This makes it overwhelmed by most unchanged objects. As such, the decoder cannot receive sufficient visual information for caption generation. In our opinion, there are two possible solutions for this challenge. One solution is to exploit other visual modalities to augment grid features, such as semantic segmentation features which can capture more fine-grained visual information [49], so as to enhance the feature represen-

tation for these objects with weak change signals. The other solution is to take advantage of the paradigm of pre-training to fine-tuning, such as fine-tuning the visual features on change captioning datasets. Specifically, we can exploit a pre-trained feature extractor that coordinates with our framework (*e.g.*, Vision Transformer [50]). Then, we do not freeze its parameters and jointly train it with the proposed NCT. In this way, the training loss can be propagated back to the feature extractor, so as to enhance the representation ability of image features. We will try these strategies in the subsequent work. In addition, we notice that the visualized attention weights of change localization are with noises on CLEVR-DC. we will try to address the problem of extreme viewpoint changes from the perspective of leveraging 3D knowledge in the future.

V. CONCLUSION

In this paper, we propose a Neighborhood Contrastive Transformer (NCT) to pinpoint and describe the change under different change scenes. In NCT, the neighborhood feature aggregating module can help overcome the influence of viewpoint change, and quickly find the inconspicuous change under the guidance of surrounding conspicuous referents. The common feature distilling module can capture common properties from each image and learn contrastive representation between the image pair. Furthermore, we introduce the explicit dependencies between words to calibrate the decoder of transformer, which helps understand complex syntax structure in change descriptions during training. Extensive experiments demonstrate that NCT outperforms the state-of-the-art methods by a large margin on the three public datasets with different change scenarios, which also shows that it has a good generalization ability to deal with various change settings.
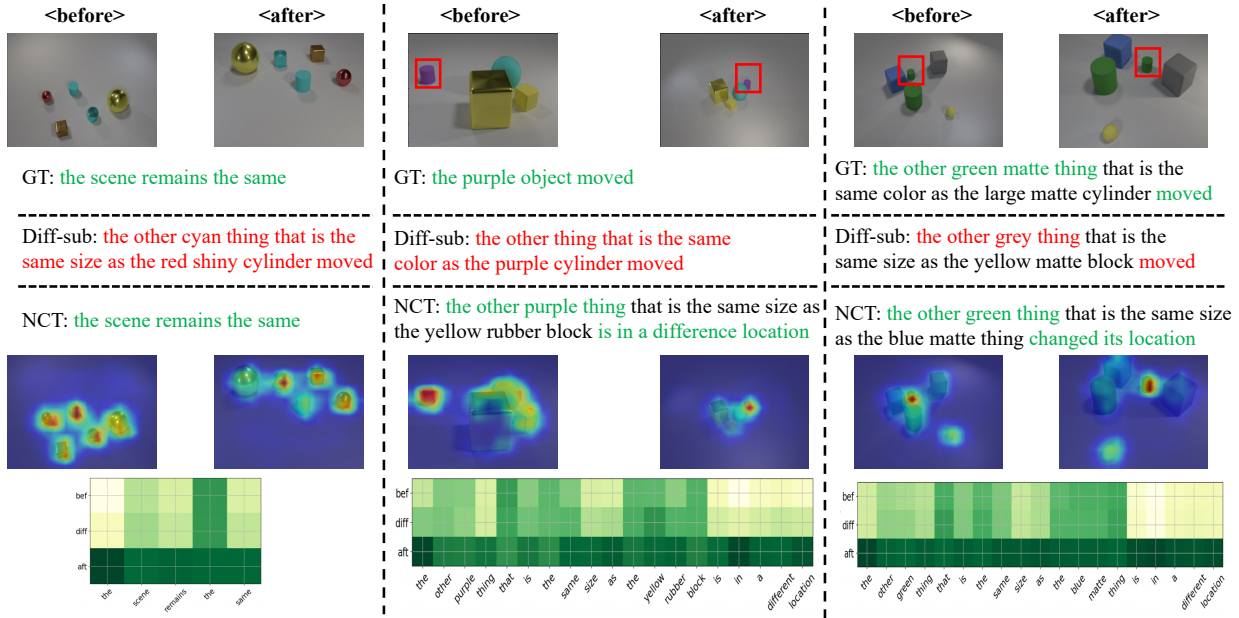
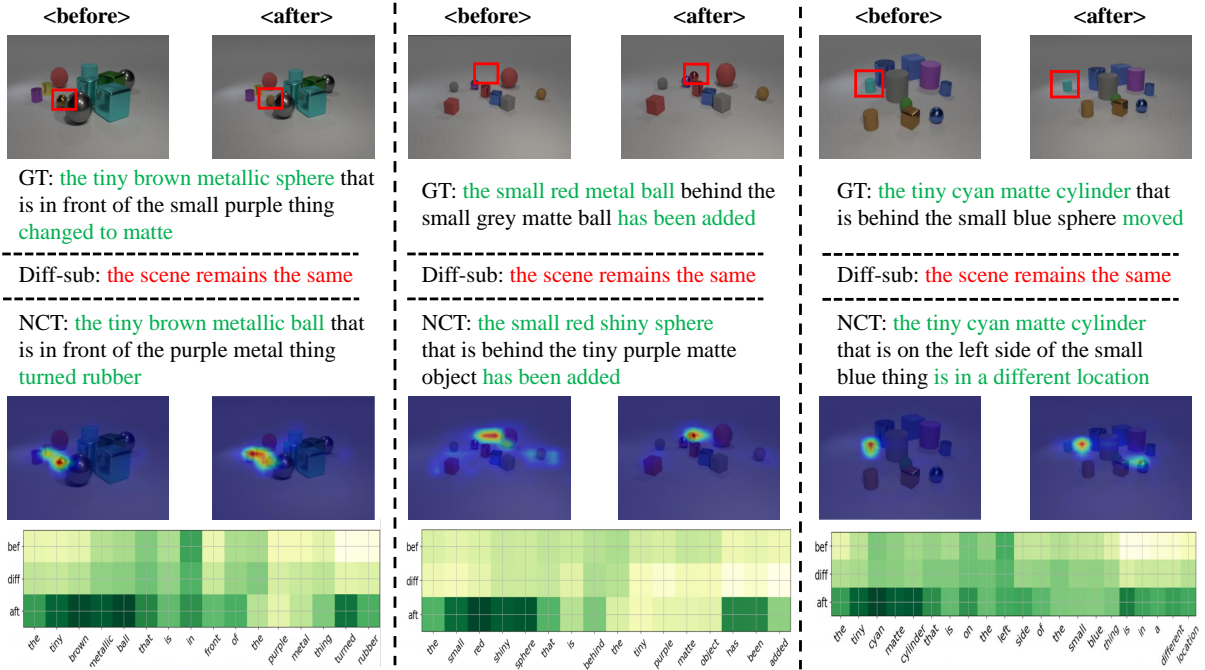Fig. 7. Qualitative examples on the test split of CLEVR-DC.

Fig. 8. Qualitative examples on the test split of CLEVR-Change.

APPENDIX
IMPLEMENTATION DETAILS AND MORE QUALITATIVE EXAMPLES ON THE THREE DATASETS

In the appendix, we first provide more implementation details of our method. On the three datasets, we train the model to convergence with 10K iterations in total. Both training and inference are implemented with PyTorch on an RTX 3090 GPU. In the training stage, the used resources on the three datasets are shown in Table X. We can find that our method does not need much resources and training time, so it can be easy reproduced by other researchers.

TABLE X
THE USAGE OF TRAINING TIME AND GPU MEMORY ON THE THREE DATASETS

	Training Time	GPU Memory
CLEVR-Change	4 hours	13G
CLEVR-DC	2 hours	8G
Image Editing Request	30 minutes	4G

Then, we illustrate more qualitative examples about change localization and caption generation on the three datasets, which
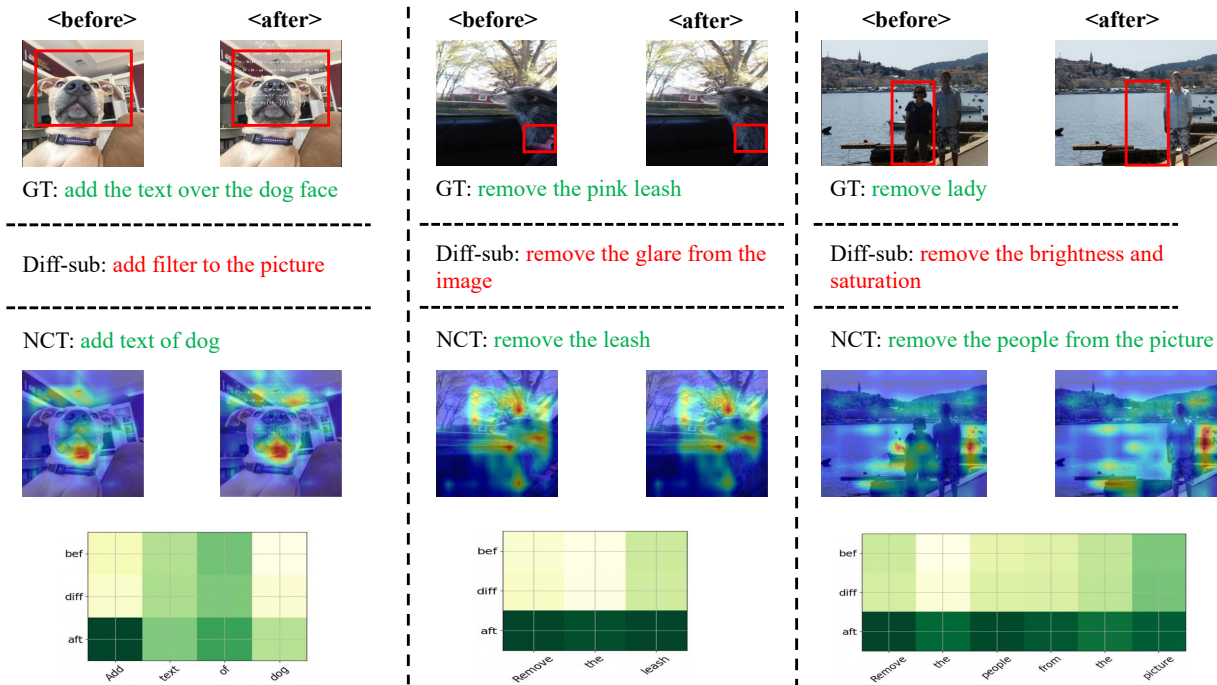
Fig. 9. Qualitative examples on the test split of Image Editing Request.

are shown in Fig. 7 - 9. For the CLEVR-DC dataset that is to stimulate extreme viewpoint changes, there exist obvious pseudo movements for all the objects in a scene, as shown in Fig. 7. This misleads the baseline model of Diff-sub into yielding wrong results. The qualitative examples on CLEVR-Change are shown in Fig. 8. Since the changed objects are partially occluded or inconspicuous, the baseline model cannot locate these changes and misjudges nothing has changed. Instead, the proposed NCT accurately distinguishes these fine-grained changes from pseudo changes and generates related sentences. It is noted that in Fig. 7, the heat maps in the left-hand side example highlight all the five objects. Our conjecture is that the heat map is generated based on the attention weights of contrastive change localizer (Sec. III-C). When nothing has changed, the learned contrastive representation of the image pair would not contain the information of changed object. And the features of background are much weaker than the object features. In this case, the localizer would attend to object features and assign similar attention weight for each object feature, so the visualized heat maps highlight all the five objects. On Image Editing request from Fig. 9 we can observe that in each example, the change information is so vague that it is hard to find, but our model still locates the changed object, so as to generate the desirable caption compared to the baseline model. The superior results of our method mainly benefits from that 1) the neighborhood feature aggregating helps the model handle irrelevant viewpoint change and locate fine-grained change; 2) the common feature distilling can capture joint information of the image pair and extract differentiating properties from each image, which constructs constrictive features between them; 3) introducing explicit dependency relations between words helps disambiguate complex syntax structure in change sentences during training.
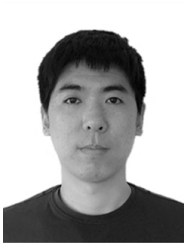
REFERENCES

[1] J. Wang, B.-K. Bao, and C. Xu, "Dualvgr: A dual-visual graph reasoning unit for video question answering," *IEEE Transactions on Multimedia*, vol. 24, pp. 3369–3380, 2022.

[2] S. Liu, A. Li, J. Wang, and Y. Wang, "Bidirectional maximum entropy training with word co-occurrence for video captioning," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[3] X. Liu, L. Li, S. Wang, Z.-J. Zha, Z. Li, Q. Tian, and Q. Huang, "Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3003–3018, 2023.

[4] J. Wu, T. Chen, H. Wu, Z. Yang, G. Luo, and L. Lin, "Fine-grained image captioning with global-local discriminative objective," *IEEE Transactions on Multimedia*, vol. 23, pp. 2413–2427, 2021.

[5] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "Exploring pairwise relationships adaptively from linguistic context in image captioning," *IEEE Transactions on Multimedia*, vol. 24, pp. 3101–3113, 2022.

[6] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4024–4034.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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, pp. 3108–3116.

[13] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[14] 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.

[15] 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*, vol. 31, pp. 2726–2738, 2022.

[16] 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*, vol. 31, pp. 3565–3577, 2022.

[17] L. Yu, J. Zhang, and Q. Wu, "Dual attention on pyramid feature maps for image captioning," *IEEE Transactions on Multimedia*, vol. 24, pp. 1775–1786, 2022.

[18] J. Ji, X. Huang, X. Sun, Y. Zhou, G. Luo, L. Cao, J. Liu, L. Shao, and R. Ji, "Multi-branch distance-sensitive self-attention network for image captioning," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[19] Y. Tu, C. Zhou, J. Guo, H. Li, S. Gao, and Z. Yu, "Relation-aware attention for video captioning via graph learning," *Pattern Recognition*, vol. 136, p. 109204, 2023.

[20] 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.

[21] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image retrieval-an empirical odyssey," in *CVPR*, 2019, pp. 6439–6448.

[22] Q. Huang, Y. Liang, J. Wei, Y. Cai, H. Liang, H.-f. Leung, and Q. Li, "Image difference captioning with instance-level fine-grained feature representation," *IEEE Transactions on Multimedia*, vol. 24, pp. 2004–2017, 2021.

[23] 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.

[24] 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.

[25] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun, "Contrastive attention for automatic chest x-ray report generation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 269–280.

[26] Z. Li, Q. Tran, L. Mai, Z. Lin, and A. L. Yuille, "Context-aware group captioning via self-attention and contrastive features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3440–3450.

[27] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, "Joint syntax representation learning and visual cue translation for video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, October 2019.

[28] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, "Controllable video captioning with pos sequence guidance based on gated fusion network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2641–2650.

[29] J. Zhang, K. Mei, Y. Zheng, and J. Fan, "Integrating part of speech guidance for image captioning," *IEEE Transactions on Multimedia*, vol. 23, pp. 92–104, 2021.

[30] J. Deng, L. Li, B. Zhang, S. Wang, Z. Zha, and Q. Huang, "Syntax-guided hierarchical attention network for video captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 880–892, 2022.

[31] Q. Zheng, C. Wang, and D. Tao, "Syntax-aware action targeting for video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 096–13 105.

[32] W. Zhao, X. Wu, and J. Luo, "Multi-modal dependency tree for video captioning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6634–6645, 2021.

[33] D. Zhao, Z. Song, Z. Ji, G. Zhao, W. Ge, and Y. Yu, "Multi-scale matching networks for semantic correspondence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3354–3364.

[34] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[35] Y. Tu, L. Li, C. Yan, S. Gao, and Z. Yu, "R^3Net: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.

[36] 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.

[37] 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.

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

[39] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[40] 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.

[41] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[42] 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.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[44] T. Dozat and C. D. Manning, "Deep biaffine attention for neural dependency parsing," *arXiv preprint arXiv:1611.01734*, 2016.

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

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[47] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[48] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[49] M. Wu, X. Zhang, X. Sun, Y. Zhou, C. Chen, J. Gu, X. Sun, and R. Ji, "Difnet: Boosting visual information flow for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 020–18 029.

[50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR*, 2021.

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 Jiao-tong 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.

Li Su received the Ph.D. degree in computer science from the Graduate University of Chinese Academy of Sciences, Beijing, in 2009. She is currently a Professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences. Her research interests include media computing and content analysis.

Ke Lu (Senior Member, IEEE) was born in Ningxia on March 13, 1971. He received the master's and Ph.D. degrees from the Department of Mathematics and Department of Computer Science, Northwest University, Xi'an, Shanxi, China, in July 1998 and July 2003, respectively. He worked as a Post-Doctoral Fellow with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, from July 2003 to April 2005. Currently, he is a Distinguished Professor with the University of the Chinese Academy of Sciences, Beijing. He is also a Double-hired Professor with the Pengcheng Laboratory, Shenzhen, Guangdong, China. His current research areas focus on computer vision, 3-D image reconstruction, and computer graphics.

Qingming Huang received the B.S. degree in computer science and the Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Chair Professor and the Deputy Dean with the School of Computer Science and Technology, University of Chinese Academy of Sciences. He has coauthored over 400 academic papers in international journals, such as IEEE TPAMI, TIP, TKDE, TMM and TCSVT, and top level international conferences, including NeurIPS, ACM Multimedia, ICCV, CVPR, ECCV, VLDB, AAAI and IJCAI. He is a Fellow of IEEE. His current research interests include multimedia computing, image/video processing, pattern recognition, and computer vision.