Revisiting Change Captioning from Self-supervised Global-Part Alignment

Feixiao Lv[1,2], **Rui Wang**[1,2*], **Lihua Jing**[1,2]

[1]Institute of Information Engineering, CAS, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
lvfeixiao@iie.ac.cn, wangrui@iie.ac.cn, jinglihua@iie.ac.cn

Abstract

The goal of image change captioning is to capture the content differences between two images and describe them in natural language. The key is how to learn stable content changes from noise such as viewpoint and image structure. However, current work mostly focuses on identifying changes, and the influence of global noise leads to unstable recognition of global features. In order to tackle this problem, we propose a Self-supervised Global-Part Alignment (SSGPA) network and revisit the image change captioning task by enhancing the construction process of overall image global features, enabling the model to integrate global changes such as viewpoint into local changes, and to detect and describe changes in the image through alignment. Concretely, we first design a Global-Part Transport Alignment mechanism to enhance global features and learn stable content changes through a self-supervised method of optimal transport. Further, we design a Change Fusion Adapter with pre-trained vision-language model to enhance the similar parts features of paired images, thereby enhancing global features, and expanding content changes. Extensive experiments show our method achieves the state-of-the-art results on four datasets.

Introduction

With the popularization of intelligence and the expansion of data in real life, it is crucial to develop automated systems to assist humans in quickly grasping changes in visual content, such as considering changes in monitoring content before and after or changes in image content. Change detection (Radke et al. 2005; Chen et al. 2023), as a method to solve such problems, has received widespread attention from the academic circles. A more advanced and newly emerging task is called change captioning (Jhamtani and Berg-Kirkpatrick 2018; Park, Darrell, and Rohrbach 2019; Kim et al. 2021), which allows for summarizing and describing changes. Compared to traditional change detection, change captioning not only need to locate changes before and after, but also require high naturalness in language description. Therefore, this task poses a more severe challenge for deep understanding of image content, and has profound practical significance for fields such as monitored facilities (Zhu et al. 2024) and pathological change (Li et al. 2023b).

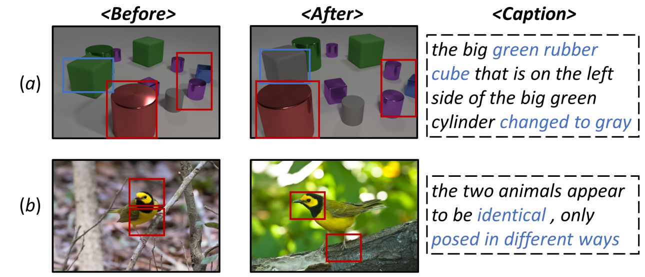
<Before>
<After>
<Caption>
(a)
the big green rubber cube that is on the left side of the big green cylinder changed to gray
(b)
the two animals appear to be identical , only posed in different ways

Figure 1: The examples of image change captioning. (a) An example from CLEVR-Change dataset. (b)An example from Birds-to-Words dataset. The blue box indicates change regions that need attention, and the red box indicates regions that should not be noticed, respectively.

Although significant progress has been made in the fields of change detection and image captioning, change captioning, as an emerging task, faces challenging difficulties. The fundamental challenge is to understand the content of the image, identify the changed content, and describe it in natural language, while filtering out the unchanged content. As shown in Figure 1, the change captioning model needs to have the ability to describe the changes before and after, rather than the entire content of the image. Moreover, there is an advanced difficulty, which is to distinguish the overall pseudo changes in the image, such as changes in image viewpoint and overall chromaticity. As shown in Figure 1, the objects in (a) undergo pseudo changes in size, position, etc. due to changes in viewpoint, and the posture of the bird in (b) change but the class does not.

Despite the pseudo overall changes have received a lot of attention (Kim et al. 2021; Tu et al. 2023a), most existing methods for dealing with such problems are based on simple local object matching. This matching-based method has alleviated the above problems to some extent, but cannot solve the pseudo changes caused by viewpoint changes, such as big when near and small when far. The reason for this is that the matching-based method pays less attention to analyze the global features of the image properly, resulting in unstable recognition of overall features such as viewpoint and image structure, which makes it difficult to decouple the

[*] Corresponding author

local changes caused by the overall changes in the image, thereby reducing the efficiency of describing image changes.

To this end, we propose a new method to revisit the image change captioning task, by enhancing the construction process of overall image global features, enabling the model to integrate global changes such as viewpoint into local changes, and to detect and describe changes in the image through alignment. Our proposed method consists of two main components: Global-Part Transport Alignment and Self-supervised Fusion Change Encoding. Specifically, (1) we propose the Global-Part Transport Alignment strategy, which transfers the global features of the image to the local regions so that the local features can adapt to changes in global, thereby enabling the model to better recognize global changes. Afterwards, we introduce local alignment to align the model with similar regions in paired images; (2) We propose a Self-Supervised Fusion Change Encoding method and design a Fusion Change Adapter module in the vision-language model(VLM) (Li et al. 2023a) to enable the model to fully utilize the knowledge of the VLM while better learning Fusion Change. We also utilize the self-supervised method to distinguish between overall changes that need to be filtered such as viewpoint, and content changes that needs to be understood from the entire dataset. The main contributions of our paper can be summarized as follows:

- We revisit Change Captioning task from the viewpoint of Self-supervised Global-Part Alignment and propose a new framework for describing changes.
- We propose the Global-Part Transport Alignment strategy, which improves the model's perception ability of both the overall and local images through optimal transport and part alignment.
- We propose Self-Supervised Fusion Change Encoding module that introduces Fusion Change Adapter and designs consistency constraints for true and pseudo changes to better learn view-invariant representation.
- We conduct extensive experiments on four public datasets with different change scenarios, and experimental results show that our proposed method surpasses existing image change captioning methods and achieves new state-of-the-arts.

Related Work

Image Change Captioning

More challenging than general image captioning (Farhadi et al. 2010; Kulkarni et al. 2013; Vinyals et al. 2015; Xu et al. 2015), image change captioning is a new task of describing subtle changes between two images from different moments. As one of the earliest studies, DDLA (Jhamtani and Berg-Kirkpatrick 2018) released the Spot-the-diff dataset extracted from the VIRAT dataset. They approximate object-level differences by clustering pixels based on pixel-wise difference of images. Since there usually exist viewpoint changes in a dynamic environment, CLEVR-Change dataset was introduced by Park (Park, Darrell, and Rohrbach 2019) to overcome several limitations of the Spot-the-Diff dataset including lack of viewpoint change and localization ground truth. They proposed the DUDA method and utilized feature-level differences to enhance the robustness of viewpoint change Moreover, more recent work has begun to focus on the challenge of changing viewpoint. VACC (Kim et al. 2021) designs a viewpoint encoding and difference modeling mechanism and tackles the problem using one image each from the before and after scenarios. VARD (Tu et al. 2023a) designs position embedded representation learning module and makes the model adapt to the changes of viewpoints from different angles by minimizing the intrinsic properties and encoding the effective positions for the features in an image pair. VIR-VLFM (Lu et al. 2023) introduces language model and viewpoint registration flow to capture actual changes between multiple images. SCORER (Tu et al. 2023c) proposes using self-supervised methods to distinguish between real changes and pseudo changes, but also a lack of constraints and analysis on the overall features of the image. NCT (Tu et al. 2023b) proposes a Transformer-based architecture to enhance the discernment of the network by generating new attention states for each token. Thus, it performs well in both change captioning and localization.

Although many excellent studies on viewpoint change, they are based on simple matching or subtle changes in viewpoint, and they are inefficient in dealing with pseudo changes caused by viewpoint changes, such as the near large far small and deformation of objects. Unlike these methods, we truly transmit the overall changes of the image to local areas and align different parts of the image through self-supervised way.

Part Alignment

Part Alignment has been applied in many research fields, including Fine-Grained Visual Categorization, Object Re-identification, image/video retrieval, and more. Similar to these studies, the image change captioning task requires Part Alignment to highlight the subtle differences in subtitle parts to distinguish them. many existing methods focus on highlighting discriminative part-level features (Chen et al. 2019; Jia et al. 2022). On the other hand, many earlier methods are based on metric learning (Luo et al. 2019). Some recent methods also leverage the strong feature representation from the spatial attention (Li, Wu, and Zheng 2021) or self-attention mechanism (Zhu et al. 2023). How to design a Part Alignment module to suit corresponding tasks has become a recent trend. Hence, we further designed an Optimal Transport Alignment for fine-level interactions between paired images. This is the key to enabling the network to fuse the overall changes of images into parts and align them.

Method

As shown in Figure 2, our proposed SSGPA considers three parts: (1) The proposed Global-Part Transport Alignment learns the token representation of two images to transfer the overall features of the images to the part features, and aligns the two images based on similarity. (2) The proposed Self-supervised Fusion Change Encoding module performs change encoding on aligned features to improve the model's ability to distinguish change features. (3) A language De-

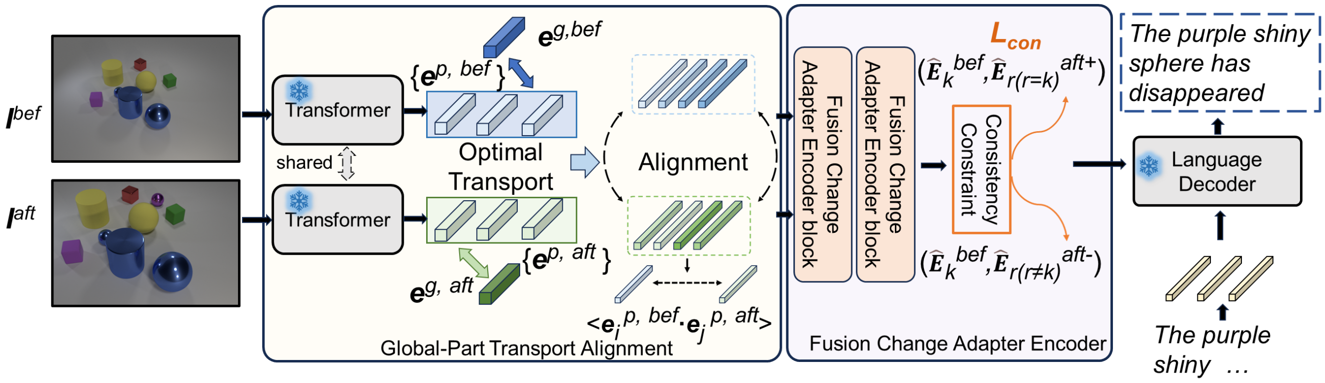
I^{bef}
I^{aft}
Transformer
shared
Transformer
$\{e^{p,\ bef}\}$
$e^{g,bef}$
Optimal Transport
$\{e^{p,\ aft}\}$
$e^{g,\ aft}$
Alignment
$<e_i^{p,\ bef}, e_j^{p,\ aft}>$
Global-Part Transport Alignment
L_{con}
$(\hat{E}_k^{bef}, \hat{E}_{r(r=k)}^{aft+})$
Fusion Change Adapter Encoder block
Fusion Change Adapter Encoder block
Consistency Constraint
$(\hat{E}_k^{bef}, \hat{E}_{r(r\neq k)}^{aft-})$
Fusion Change Adapter Encoder
The purple shiny sphere has disappeared
Language Decoder
The purple shiny …

Figure 2: The overall workflow of our proposed SSGPA model. Our model consists of three parts: (a)Global-Part Transport Alignment, (b)Self-supervised Fusion Change Encoding, (c)Language Decoder, respectively.

coder decodes the final image features into change captions to improve the quality of generated descriptions.

Global-Part Transport Alignment

Input Representation As the general vision and language training model, our SSGPA inputs are the representations of a pair of similar images. Formally, given a pair of images "before" I^{bef} and "after" I^{aft}, we use the pre-trained ResNet101 (He et al. 2016) to extract the grid features of the paired images input and express them as:

$$\begin{aligned}\mathbf{I}^{bef} &= \{[\mathrm{IMG}^{\mathrm{bef}}], i_0^{p,bef}, \ldots, i_{M-1}^{p,bef}\},\\ \mathbf{I}^{aft} &= \{[\mathrm{IMG}^{\mathrm{aft}}], i_0^{p,aft}, \ldots, i_{M-1}^{p,aft}\},\end{aligned} \tag{1}$$

where the two tokens $[\mathrm{IMG}^{\mathrm{bef}}]$ and $[\mathrm{IMG}^{\mathrm{aft}}]$ are markers of paired images, which are used to capture the global features of two images at the same time. In addition, in order to better process the positional information of sequence modality triplet, we also introduce positional encoding (Vaswani et al. 2017) to each input token.

Global-Part Optimal Transport Firstly, we transfer the overall features of the image to the part features. Specifically, given the input visual features I^{bef} and I^{aft}, we utilize a self-attention Transformer block (Vaswani et al. 2017), $\mathcal{T}^o(\cdot)$, to further learn global and local image features. This process can be formulated with:

$$\mathbf{E}^o = \mathcal{T}^o(\mathbf{I}^o), \tag{2}$$

where $o \in \{bef, aft\}$ indicates the paired image.

Then, we introduce Optimal Transport (OT) to build a transport plan according to the probability distribution of two sets and the OT cost. Specifically, in the proposed Global-Part Optimal Transport, the transport plan needs to be built according to the feature embedding from both parts and global features. Besides, the similarity between each part embedding and the global embedding needs to be significantly considered. We present a Global-Part Optimal Transport (GPOT) scheme to address this issue. The Global-Part Optimal Transport plan from e_k^o to $[\mathrm{IMG}^{\mathrm{o}}]$ is formulated as:

$$P = \{p \in \mathbb{R}^{H\times M} \mid p\cdot \mathbf{1}^{\mathrm{T}} = \hat{G}, p^T\mathbf{1}^{\mathrm{T}} = \hat{P}\}, \tag{3}$$

where $\mathbf{1}$ is a vector in which all the element values are one, $\hat{\mathbb{P}}$ and $\hat{\mathbb{G}}$ represent the sets of part features and global features. Then, given an Global-Part cost matrix $C \in \mathbb{R}^{M\times H}$ between $\hat{\mathbb{P}}$ and $\hat{\mathbb{G}}$, the Global-Part Optimal Transport cost $\omega(\cdot,\cdot)$ can be defined as:

$$\omega(\hat{x},\hat{y}) = \min_{p\in P}\langle \mathbf{C}, p\rangle, \tag{4}$$

where$\langle , \rangle$ denotes inner product.

Now we consider the matching distance between each part embedding $e_m^{p,o}$ and the global embedding $e_m^{g,o}$, which two constitute e_m^o. Let $i = 0,\cdots,M-1$ and $j = 0,\cdots,H-1$. Then, the distance between the i-th sample in the part set $\hat{\mathbb{P}}$ and the j-th sample in the global set $\hat{\mathbb{G}}$, denoted as $\mathbf{C}_{i,j}$, is computed as:

$$C_{i,j} = \frac{\left\|\mathbf{e}_i^{p,o} - \mathbf{e}_j^{g,o}\right\|_2^2}{\max_{1\le i\le M, 1\le j\le H}\left\|\mathbf{e}_i^{p,o} - \mathbf{e}_j^{g,o}\right\|_2^2}, \tag{5}$$

where $\||\cdot\||_2^2$ denotes Euclidean distance. Note that, we have $e_m^{g,o}$ as multi-head output of $[\mathrm{IMG}^{\mathrm{o}}]$. On the other hand, the global distribution of part set $\hat{\mathbb{P}}$ and global set $\hat{\mathbb{G}}$, which we denote as $d^p \in \mathbb{R}^M$ and $d^g \in \mathbb{R}^H$, is computed as:

$$\mathbf{d}_m^p = e^{\mathbf{e}_m^p}, \mathbf{d}_m^g = e^{\mathbf{e}_m^g}. \tag{6}$$

Then, the Global-Part Optimal Transport loss $\mathcal{L}_{IoT}$ is defined as the dual, form of the optimal transport problem, given by:

$$\mathcal{L}_{IoT}\left(\mathbf{d}_m^p, \mathbf{d}_m^g\right) = \left\langle \boldsymbol{\lambda}^*, \frac{\mathbf{d}^p}{\left\|\mathbf{d}^p\right\|_1}\right\rangle + \left\langle \boldsymbol{\mu}^*, \frac{\mathbf{d}^g}{\left\|\mathbf{d}^g\right\|_1}\right\rangle \tag{7}$$

where λ^* and μ^* denote the approximated solutions of the OT problem, which are usually approximated by the fast Sinkhorn distances algorithm (Cuturi 2013).

Paired Images Part Alignment Due to viewpoint issues, simply aligning each part in order may raise the feature inconsistency problem. Fortunately, for any specific shape combination, its discriminative appearance usually appears on limited size of local parts. Consequently, we propose an

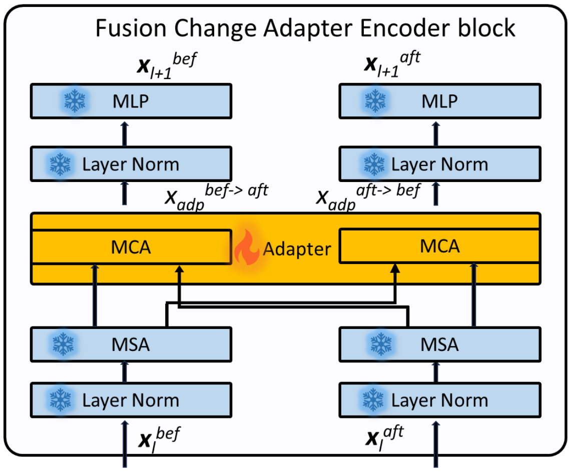

Figure 3: Architecture of the Fusion Change Adapter Encoder block, which is a carefully designed Transformer block with cross-attention adapter.

unsupervised graph matching method to sort the found parts in a unified order based on a basic assumption that the correlation between top N parts is similar across images. we maintain a unified correlation matrix to model latent relations in-between parts. The entry of the correlation matrix is given as follows:

$$\boldsymbol{M}_{ij} = \left\langle \boldsymbol{e}_i^{p,o}, \boldsymbol{e}_j^{p,o} \right\rangle, \tag{8}$$

where $\boldsymbol{M}_{ij}$ denotes the relation score between part $e_i^{p,o}$ and part $e_j^{p,o}$. Given an after image sample, we compute its parts correlation matrix, denoted as $\boldsymbol{M}'$. Then the $\boldsymbol{M}'$ has the largest matching degree with the reference matrix $\boldsymbol{M}$ is considered as the best alignment. The formulation can be simplified as:

$$\hat{\boldsymbol{M}} = \underset{\boldsymbol{M}'}{\operatorname{argmax}} \operatorname{vec}\left(\boldsymbol{M}'\right)^T \operatorname{vec}(\boldsymbol{M}), \tag{9}$$

The permutation with the max matching degree is selected as the correct order and the algorithm returns the resorted parts representations.

Self-supervised Fusion Change Encoding

Fusion Change Adapter Encoder The Fusion Change Adapter Encoder introduces adapter to adapt pre-trained image encoders for encoding image changes. As shown in Figure 3, we insert adapters in each transformer block to extract change features of each image.

In pre-trained image encoders, we freeze the parameters of other parts and design a learnable adapter to calculate changes in paired images. Unlike traditional adapters, our adapter introduces learnable parameters by calculating similarity, which is formulated as follows:

$$\begin{aligned} x_{l+1}^o &= MLP(LN(x_{adp}^{q\to o} + x^o)), \\ x_{adp}^{q\to o} &= MCA\left(MSA(LN(x^q)), MSA(LN(x^o))\right), \end{aligned} \tag{10}$$

where LN is Layer Normalization. q represents another image that is opposite to O, MSA and MCA are multi-head self-attention and multi-head cross-attention, respectively. Only the parameters of the multi-head cross-attention layers in FCA are trained, while the remaining layers are loaded from the pre-trained weights and kept frozen. The parameters of adapter are initialized to 0 to ensure that the output of the pre-trained encoder is not altered in the initial training phase.

Consistency Constraint Given a training batch, we sample $B = \{B_t, B_u\}$ pairs of encoding features. B_u is an unchanged set(including pseudo change), and B_t is a true changed set. We need to increase the distance between unchanged and pseudo change pairs, and decrease the distance between true changes. For the encoding feature in the k-th "before" image $\hat{E}_k^{bef}$, the encoding feature in the r-th "after" image $\hat{E}_r^{aft}(r = k)$ of B_u is its positive, while encoding features in the other "after" images $\hat{E}_r^{aft}(r \neq k)$ will be the negatives. The situation of B_t is reversed. Then, we project these positive/negative pairs into a shared embedding space, normalize them by L2- normalization, and compute their similarity. We introduce the NCE loss (Oord, Li, and Vinyals 2018) to optimize their contrastive alignment:

$$\begin{aligned} \mathcal{L}_{b2at} &= \frac{1}{B_t}\sum_k^{B_t} \log \frac{e^{\left(\operatorname{sim}\left(\hat{E}_k^{bef}, \hat{E}_{r(r=k)}^{aft^-}\right)/\tau\right)}}{\sum_r^B e^{\left(\operatorname{sim}\left(\hat{E}_k^{bef}, \hat{E}_r^{aft}\right)/\tau\right)}}, \\ \mathcal{L}_{b2au} &= -\frac{1}{B_u}\sum_k^{B_u} \log \frac{e^{\left(\operatorname{sim}\left(\hat{E}_k^{bef}, \hat{E}_{r(r=k)}^{aft^+}\right)/\tau\right)}}{\sum_r^B e^{\left(\operatorname{sim}\left(\hat{E}_k^{bef}, \hat{E}_r^{aft}\right)/\tau\right)}}, \\ \mathcal{L}_{a2bt} &= \frac{1}{B_t}\sum_k^{B_t} \log \frac{e^{\left(\operatorname{sim}\left(\hat{E}_k^{aft}, \hat{E}_{r(r=k)}^{bef^-}\right)/\tau\right)}}{\sum_r^B e^{\left(\operatorname{sim}\left(\hat{E}_k^{aft}, \hat{E}_r^{bef}\right)/\tau\right)}}, \\ \mathcal{L}_{a2bu} &= -\frac{1}{B_u}\sum_k^{B_u} \log \frac{e^{\left(\operatorname{sim}\left(\hat{E}_k^{aft}, \hat{E}_{r(r=k)}^{bef^+}\right)/\tau\right)}}{\sum_r^B e^{\left(\operatorname{sim}\left(\hat{E}_k^{aft}, \hat{E}_r^{bef}\right)/\tau\right)}}, \\ \mathcal{L}_{\text{con}} &= \frac{1}{4}\left(\mathcal{L}_{b2at} + \mathcal{L}_{a2bt} + \mathcal{L}_{b2au} + \mathcal{L}_{a2bu}\right), \end{aligned} \tag{11}$$

where "sim" is dot-product function to measure similarity between two features. τ is temperature hyper-parameter.

Language Decoder

As shown in Figure, we use a large pre-trained multi-level Transformer model to decode image features to change captions. The Transformer model consists of multiple layers of self-attention Transformer blocks. The features are combined with the instruction "Describe the differences between the two images." and sent to LLM for predicting change captions. During training, same as the original parameters of Fusion Change Adapter Encoder, the Language decoder parameter is frozen.

Experiments

Datasets and Metrics

Datasets Birds-to-Words dataset (Forbes et al. 2019) consists of 41k sentences that describe fine-grained changes

Model	B4	M	C	R
DUDA 2019	47.3	33.9	112.0	-
VAM+ 2020	51.3	37.8	115.8	70.4
DUDA+Aux 2021	51.2	37.7	115.4	70.5
VARD-T 2023a	55.4	40.1	126.4	73.8
NCT 2023b	55.1	40.2	124.1	73.8
SCORER+CBR 2023c	56.3	41.2	126.8	74.5
DIRL+CCR 2024	54.6	38.1	123.6	71.9
PLC 2022	51.2	36.2	128.9	71.7
CLIP4IDC 2022	56.9	38.4	150.7	76.4
VIR-VLFM 2023	58.2	42.6	153.4	78.9
Ours	**60.9**	**44.2**	**159.1**	**80.2**

Table 1: Comparison with the state of the arts on CLEVR-Change dataset.

Model	B4	M	C	R
Neural Naturalist 2019	22.0	-	25.0	43.0
OneDiff 2024	25.8	15.6	28.0	49.1
Relational Speaker 2019	21.5	22.4	5.8	43.4
DUDA 2019	23.9	21.9	4.6	44.3
L2C 2021	31.3		15.1	45.3
L2C(+CUB) 2021	31.8	-	16.3	45.6
PLC 2022	31.0	23.4	25.3	49.1
Ours	**36.9**	**27.1**	**43.6**	**53.2**

Table 2: Comparison with the state of the arts on Birds-to-Words dataset.

Model	B4	M	C	R
DUDA 2019	8.1	11.8	32.5	29.1
OneDiff 2024	12.8	14.6	56.6	35.8
VAM+ 2020	11.1	12.9	42.5	33.2
DUDA+Aux 2021	8.1	12.5	34.5	29.9
CLIP4IDC 2022	11.6	14.2	47.4	35.0
VIR-VLFM 2023	12.2	15.3	48.9	36.2
DIRL+CCR 2024	10.3	13.8	40.9	32.8
SCORER+CBR 2023c	10.2	12.2	38.9	-
Ours	**13.5**	**16.0**	**63.4**	**42.7**

Table 3: Comparison with the state of the arts on Spot-the-Diff dataset.

between photographs of birds. This yields a total of 3,347 image pairs, annotated with 40,969 sentences. This leads to 12,890/1,556/1,604 captions for train/val/test splits. CLEVR-Change dataset (Park, Darrell, and Rohrbach 2019) is a large-scale synthetic dataset with moderate viewpoint change with 79,606 image pairs and 493,735 captions, including five change types, i.e., "color", "texture", "add", "drop", and "move". Spot-the-Diff dataset (Jhamtani and Berg-Kirkpatrick 2018) includes 13,192 aligned image pairs from surveillance cameras without fine-grained change but only obvious content changes. Image Editing Request dataset (Tan et al. 2019) includes 3,939 aligned image pairs with 5,695 editing instructions produced by image editing.

Metrics We evaluate the performance with four most popular automatic language metrics: CIDEr(C) (Vedantam, Lawrence Zitnick, and Parikh 2015), BLEU-4(B4) (Papineni et al. 2002), METEOR(M) (Banerjee and Lavie 2005) and ROUGE-L(R) (Lin and Hovy 2003).

Implementation Detail

To compare our SSGPA with other methods in a fair way, we perform our main evaluation with ResNet101 (He et al. 2016) to extract paired image features and flatten them to shape of (49, 2048). All hidden size is 512. Both training and inference are implemented with PyTorch (Paszke et al. 2019) on RTX 3090 GPU. We apply EVA-ViT-g/14 (Fang et al. 2023) and Vicuna-7B (Chiang et al. 2023) as image encoder and LLM, respectively. The above models without the proposed GPTA and SSFEC constitute our baseline. The head and layer numbers are set to 8 and 2 for Input Representation step, and to 8 and 4 for Self-supervised Fusion Change Encoding step on the four datasets, respectively. During training, We use Adam optimizer (Kingma and Ba 2014) to minimize the aforementioned losses and all parameters except MCA adapter are frozen.

Comparisons with State-of-the-art Methods

Results on CLEVR-Change Dataset The CLEVR-Change dataset is the most commonly-used dataset for image change captioning task. We compare our SSGPA with other state-of-the-art models, including DUDA (Park, Darrell, and Rohrbach 2019), VAM+ (Shi et al. 2020), DUDA+Aux (Hosseinzadeh and Wang 2021), NCT (Tu et al. 2023b), VARD-T (Tu et al. 2023a), PLC (Yao, Wang, and Jin 2022), SCORER+CBR (Tu et al. 2023c), DIRL+CCR (Tu et al. 2024) and vision-language pre-training methods CLIP4IDC (Guo, Wang, and Laaksonen 2022) and VIR-VLFM (Lu et al. 2023). The results are reported in Table 1.

As it can be seen, compared with VIR-VLFM (Lu et al. 2023), our SSGPA promotes 2.7, 1.6, 5.7, and 1.3 in B4, M, C, and R, respectively. It is worth mentioning that compared to current methods that focus on viewpoint changes, our method places more emphasis on capturing the overall features of the image, freeing it from the limitation of small-scale viewpoint changes. Therefore, our method can outperform state-of-the-art methods on CLEVR-Change dataset.

Model	B4	M	C	R
Relational Speaker 2019	6.7	12.8	26.4	37.5
DUDA 2019	6.5	12.4	22.8	37.3
BiDiff 2022	6.9	14.6	27.7	38.5
SCORER+CBR 2023c	10.0	15.0	33.4	39.6
NCT 2023b	8.1	15.0	34.2	38.8
DIRL+CCR 2024	10.9	15.0	41.0	34.1
VIXEN-C 2024	8.6	15.4	38.1	42.5
VARD-T 2023a	10.0	14.8	35.7	39.0
Ours	**11.2**	**15.8**	**43.4**	**43.1**

Table 4: Comparison with the state of the arts on Image Editing Request dataset.

Method	CLEVR-Change				Birds-to-Words			
	BLEU-4	METEOR	CIDEr	ROUGE-L	BLEU-4	METEOR	CIDEr	ROUGE-L
Baseline	56.7	38.4	149.2	77.2	31.3	24.0	24.2	47.6
+GPOT	58.7	39.2	151.9	78.7	32.3	25.4	32.1	49.7
+PIPA	57.0	38.9	150.6	77.4	31.9	25.8	31.5	48.4
+GPTA(GPOT+PIPA)	59.3	41.5	153.8	79.2	33.6	26.1	39.2	51.5
+FCAE	59.0	41.2	152.7	77.4	31.9	24.7	30.4	49.8
+CC	56.9	39.2	150.7	77.6	31.6	24.5	29.8	48.6
+SSFCE(FCAE+CC)	59.8	42.6	154.4	79.3	33.1	26.0	37.9	51.3
SSGPA	60.9	44.2	159.1	80.2	36.9	27.1	43.6	53.2

Table 5: Ablation studies on CLEVR-Change dataset and Birds-to-Words dataset.

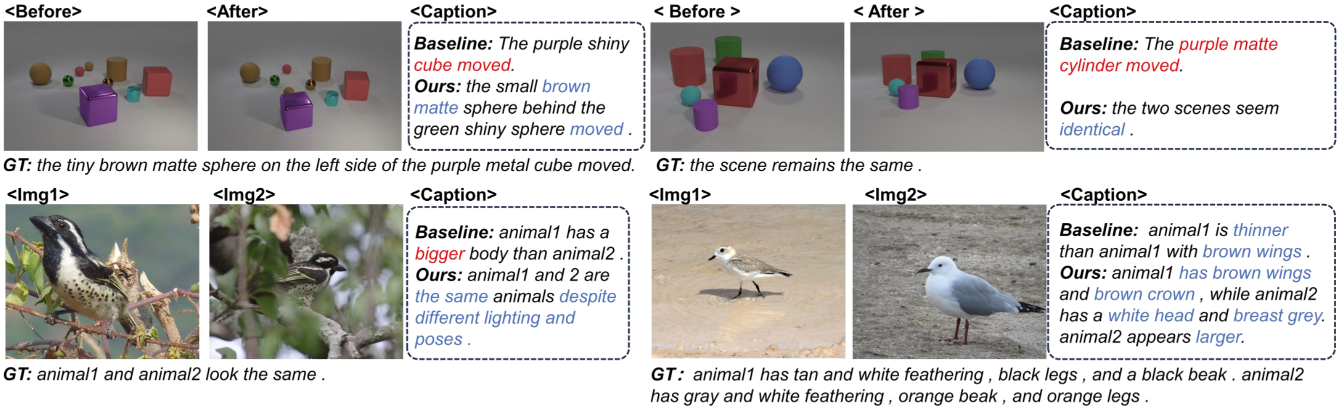
<Before>
<After>
<Caption>
Baseline: The purple shiny cube moved.
Ours: the small brown matte sphere behind the green shiny sphere moved .
GT: the tiny brown matte sphere on the left side of the purple metal cube moved.
< Before >
< After >
<Caption>
Baseline: The purple matte cylinder moved.
Ours: the two scenes seem identical .
GT: the scene remains the same .
<Img1>
<Img2>
<Caption>
Baseline: animal1 has a bigger body than animal2 .
Ours: animal1 and 2 are the same animals despite different lighting and poses .
GT: animal1 and animal2 look the same .
<Img1>
<Img2>
<Caption>
Baseline: animal1 is thinner than animal1 with brown wings .
Ours: animal1 has brown wings and brown crown , while animal2 has a white head and breast grey. animal2 appears larger.
GT : animal1 has tan and white feathering , black legs , and a black beak . animal2 has gray and white feathering , orange beak , and orange legs .

Figure 4: Examples by SSGPA and baseline. The upper part illustrates cases from CLEVR-Change. The lower part illustrates cases from Birds-to-Words. The wrong phrases and fine-grained information are highlighted in red and blue respectively.

Results on Birds-to-Words Dataset Birds-to-Words dataset contains a significant amount of fine-grained changes of birds, where the birds typically have different viewpoint and postures, showing more overall characteristics, thus presenting more challenges than CLEVR-Change dataset. To validate the superiority of our proposed method SSGPA, we compare our method with some other well-known change captioning methods, including Neural Naturalist (Forbes et al. 2019), Relational Speaker (Tan et al. 2019), DUDA (Park, Darrell, and Rohrbach 2019), L2C/L2C(+CUB) (Yan et al. 2021) and PLC (Yao, Wang, and Jin 2022) on Birds-to-Words dataset.

As shown in Table 2, compared with PLC (Yao, Wang, and Jin 2022), our SSGPA promotes 5.9, 3.7, 18.3 and 4.1 in B4, M, C, and R. Our method achieves this from two aspects. Firstly, our Global-Part Transport Alignment transfers the global features to local features through the Optimal Transport method, enhance the relation between subtle parts and fine-grained semantics, which is suitable for fine-grained datasets. Secondly, our Self-supervised Fusion Change Encoding incorporates knowledge of objects to some extent and adapts it to our task through self-supervised methods. Overall, our method achieves new state-of-the-art performance on the Birds-to-Words dataset.

Results on Spot-the-Diff Dataset Spot-the-Diff dataset includes aligned image pairs from surveillance cameras. To further validate the generalization, we compare our method with well-known methods: DUDA (Park, Darrell, and Rohrbach 2019), OneDiff (Hu et al. 2024), VAM+ (Shi et al. 2020), DUDA+Aux (Hosseinzadeh and Wang 2021), CLIP4IDC (Guo, Wang, and Laaksonen 2022), VIR-VLFM (Lu et al. 2023),DIRL+CCR (Tu et al. 2024) and SCORER+CBR (Tu et al. 2023c). As it can be seen, compared with VIR-VLFM (Lu et al. 2023), our SSGPA promotes 1.3, 0.7, 14.5, and 6.5 in B4, M, C, and R.

Results on Image Editing Request Dataset To further validate the generalization of our method, we conduct experiment on challenging Image Editing Request dataset. We compare with following SOTA methods: Relational Speaker (Tan et al. 2019), DUDA (Park, Darrell, and Rohrbach 2019), BiDiff (Sun et al. 2022), SCORER+CBR (Tu et al. 2023c), NCT (Tu et al. 2023b), DIRL+CCR (Tu et al. 2024), VIXEN-C (Black et al. 2024) and VARD-T (Tu et al. 2023a). Compared with VARD-T (Tu et al. 2023a), our SSGPA promotes 1.2, 1.0, 7.7 and 4.1 in B4, M, C, and R, respectively.

Ablation Study

We conduct ablation studies to verify the effectiveness of our proposed SSGPA. The results are shown in Table 5.

Effectiveness of Global-Part Transport Alignment We divide Global-Part Transport Alignment (GPTA) into two parts: Global-Part Optimal Transport (GPOT) and Paired Images Part Alignment (PIPA) to evaluate its performance. As shown in Table 5, when we perform GPOT and compare it with baseline, it can promote the improvements 2.0/1.0 of B4, 0.8/1.4 of M, 2.7/7.9 of C, 1.5/2.1 of R, on CLEVR-Change/Birds-to-Words. It can be seen that the GPOT shows better performance on the fine-grained Birds-to-Words dataset. As mentioned above, the images in this dataset typically have different viewpoints and postures, which shows the ability of our GPOT to handle global-part information and that the GPOT, while also possessing the ability to analyze fine-grained features.

Compared our PIPA with baseline, it can promote the improvements 0.3/0.6 of B4, 0.5/1.8 of M, 1.4/7.3 of C, 0.2/0.8 of R, on CLEVR-Change/Birds-to-Words. The experiment indicates that PIPA shows an effective improvement in both two datasets. Sorting the parts of an image based on their similarity(e.g. head to head) is also more suitable for fine-grained datasets. Overall, compared our GPTA with baseline, it can promote the improvements 2.6/2.3 of B4, 3.1/2.1 of M, 4.6/15.0 of C, 0.2/0.8 of R, on CLEVR-Change/Birds-to-Words. Our proposed Global Part Transport Alignment plays an important role in integrating global features and results in corresponding improvements in the results. More visual details are expanded in Section 4.5.

Effectiveness of Self-supervised Fusion Change Encoding We divide our Self-supervised Fusion Change Encoding(SSFCE) into two parts: Fusion Change Adapter Encoder(FCAE) and Consistency Constraint(CC). FCAE is mainly used to capture changes between paired images, while CC is mainly used to capture different change types. As shown in Table 5, when we perform GPOT and compare it with baseline, it can promote the improvements 2.3/0.6 of B4, 2.8/0.7 of M, 3.5/6.2 of C, 0.2/2.2 of R on CLEVR-Change/Birds-to-Words. The results show significant improvement on both datasets, indicating the effectiveness of our proposed FCAE in capturing changes in paired images.

When we perform CC and compare it with baseline, it can promote the improvements 0.2/0.3 of B4, 0.8/0.5 of M, 1.5/5.6 of C, 0.4/1.0 of R, on CLEVR-Change/Birds-to-Words. The experiment demonstrates the effectiveness of CC for change captioning. Overall, compared SSFCE with baseline, it can promote the improvements 3.1/1.8 of B4, 4.2/2.0 of M, 5.2/13.7 of C, 2.1/3.7 of R, on CLEVR-Change/Birds-to-Words. Due to the use of pre-trained encoders in both our method and baseline, this result demonstrates the effectiveness of our proposed SSFCE in capturing changes. More visual details are expanded in Section 4.5.

Qualitative Results and Visualisation

To intuitively evaluate our method, we first visualise the generation results of our SSGPA and baseline. After that, we visualise the focus areas of the proposed modules.

Figure 4 shows qualitative results by our SSGPA and baseline method. As shown in Figure 4, for CLEVR-Change dataset, the baseline incorrectly identifies the change in viewpoint as a position change of the purple cube of left example, while ignoring the true change in the brown sphere. The same situation occurs in the right example. For the Birds-to-Words dataset, the above issues are more severe due to the fact that the birds have more shooting angles and poses. These examples demonstrate that methods that do not analyze the overall features are difficult to identify pseudo changes caused by changes in viewpoint (such as false movements of objects and changes of size), while our method, as motivated, effectively solves such problems.

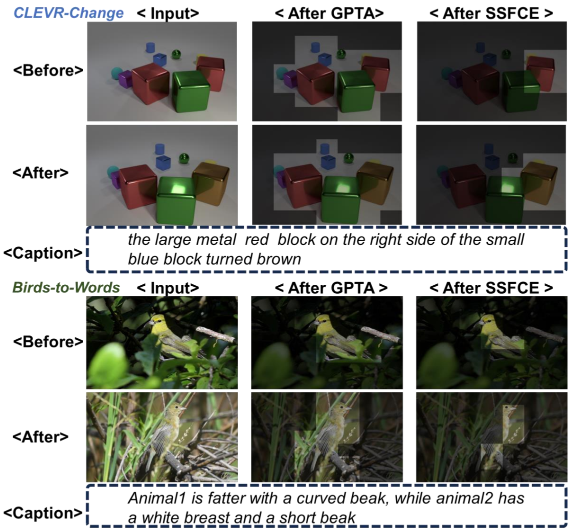
CLEVR-Change
< Input>
< After GPTA>
< After SSFCE >
<Before>
<After>
<Caption>
the large metal red block on the right side of the small blue block turned brown
Birds-to-Words
< Input>
< After GPTA >
< After SSFCE >
<Before>
<After>
<Caption>
Animal1 is fatter with a curved beak, while animal2 has a white breast and a short beak

Figure 5: Visualisation of the process of generating captions from CLEVR-Change dataset and Birds-to-Words dataset.

To better understand the effectiveness of SSGPA, we visually generate the attention weights over tokens of two modules, which are GPTA and SSFCE. As shown in figure 5, our GPTA can focus the model's attention on entire image content, while our SSFCE can effectively distinguish true changes from all content and filter out pseudo changes.

Conclusions

In this paper, we propose a novel change captioning learning paradigm, Self-supervised Global-Part Alignment(SSGPA) to distinguish between pseudo changes and real changes. Our proposed SSGPA consists of two main components, which are Global-Part Transport Alignment (GPTA) for more holistic and consistent paired image features, and Self-supervised Fusion Change Encoding (SSFCE) for localizing the real change regions. For SSGPA, we introduce Global-Part Optimal Transport and Paired Images Part Alignment, firstly transferring the global features to part features and then aligning the features of the two paired images. For GPTA, we introduce trainable adapters to the well-trained vision-language model, allowing the model to learn knowledge of changes and enable it to recognize real or pseudo changes through new self-supervised method. Finally, we conduct extensive experiments on common-used datasets to verify the effectiveness. The results show that our proposed method achieves new state-of-the-art performance.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China Under Grants No. 62176253.

References

Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.

Black, A.; Shi, J.; Fan, Y.; Bui, T.; and Collomosse, J. 2024. VIXEN: Visual Text Comparison Network for Image Difference Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 846–854.

Chen, C.-P.; Hsieh, J.-W.; Chen, P.-Y.; Hsieh, Y.-K.; and Wang, B.-S. 2023. SARAS-net: scale and relation aware siamese network for change detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14187–14195.

Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; and Wang, Z. 2019. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8351–8361.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19358–19369.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, 15–29. Springer.

Forbes, M.; Kaeser-Chen, C.; Sharma, P.; and Belongie, S. 2019. Neural naturalist: generating fine-grained image comparisons. *arXiv preprint arXiv:1909.04101*.

Guo, Z.; Wang, T.-J. J.; and Laaksonen, J. 2022. CLIP4IDC: CLIP for image difference captioning. *arXiv preprint arXiv:2206.00629*.

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

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

Hu, E.; Guo, L.; Yue, T.; Zhao, Z.; Xue, S.; and Liu, J. 2024. OneDiff: A Generalist Model for Image Difference. *arXiv preprint arXiv:2407.05645*.

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

Jia, M.; Cheng, X.; Lu, S.; and Zhang, J. 2022. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Transactions on Multimedia*, 25: 1294–1305.

Kim, H.; Kim, J.; Lee, H.; Park, H.; and Kim, G. 2021. Agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2095–2104.

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

Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12): 2891–2903.

Li, H.; Wu, G.; and Zheng, W.-S. 2021. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6729–6738.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3334–3343.

Lin, C.-Y.; and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, 150–157.

Lu, X.; Yuan, J.; Niu, R.; Hu, Y.; and Wang, F. 2023. Viewpoint Integration and Registration with Vision Language Foundation Model for Image Change Understanding. *arXiv preprint arXiv:2309.08585*.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *ICCV*.

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

Radke, R. J.; Andra, S.; Al-Kofahi, O.; and Roysam, B. 2005. Image change detection algorithms: a systematic survey. *IEEE transactions on image processing*, 14(3): 294–307.

Shi, X.; Yang, X.; Gu, J.; Joty, S.; and Cai, J. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 574–590. Springer.

Sun, Y.; Li, L.; Yao, T.; Lu, T.; Zheng, B.; Yan, C.; Zhang, H.; Bao, Y.; Ding, G.; and Slabaugh, G. 2022. Bidirectional difference locating and semantic consistency reasoning for change captioning. *International Journal of Intelligent Systems*, 37(5): 2969–2987.

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

Tu, Y.; Li, L.; Su, L.; Du, J.; Lu, K.; and Huang, Q. 2023a. Viewpoint-Adaptive Representation Disentanglement Network for Change Captioning. *IEEE Transactions on Image Processing*, 32: 2620–2635.

Tu, Y.; Li, L.; Su, L.; Lu, K.; and Huang, Q. 2023b. Neighborhood Contrastive Transformer for Change Captioning. *ArXiv*, abs/2303.03171.

Tu, Y.; Li, L.; Su, L.; Yan, C.; and Huang, Q. 2024. Distractors-Immune Representation Learning with Cross-modal Contrastive Regularization for Change Captioning. *arXiv preprint arXiv:2407.11683*.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2023c. Self-supervised Cross-view Representation Reconstruction for Change Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2805–2815.

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

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.

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

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

Yan, A.; Wang, X. E.; Fu, T.-J.; and Wang, W. Y. 2021. L2C: Describing visual differences needs semantic understanding of individuals. *arXiv preprint arXiv:2102.01860*.

Yao, L.; Wang, W.; and Jin, Q. 2022. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3108–3116.

Zhu, K.; Guo, H.; Zhang, S.; Wang, Y.; Liu, J.; Wang, J.; and Tang, M. 2023. Aaformer: Auto-aligned transformer for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhu, Y.; Li, L.; Chen, K.; Liu, C.; Zhou, F.; and Shi, Z. 2024. Semantic-CC: Boosting Remote Sensing Image Change Captioning via Foundational Knowledge and Semantic Guidance. *arXiv preprint arXiv:2407.14032*.