A Decoupling Paradigm with Prompt Learning for Remote Sensing Image Change Captioning

Chenyang Liu, Rui Zhao, Jianqi Chen, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi*, *Senior Member, IEEE*

***Abstract*—Remote sensing image change captioning (RSICC) is a novel task that aims to describe the differences between bi-temporal images by natural language. Previous methods ignore a significant specificity of the task: the difficulty of RSICC is different for unchanged and changed image pairs. They process the unchanged and changed image pairs in a coupled way, which usually causes confusion for change captioning. In this paper, we decouple the task into two issues to ease it: whether and what changes have occurred. An image-level classifier performs binary classification to address the first issue. A feature-level encoder contributes to extracting discriminative features to help the caption generation module address the second issue. Besides, for caption generation, we utilize prompt learning to introduce pre-trained large language models (LLMs) into the RSICC task. A multi-prompt learning strategy is proposed to generate a set of unified prompts and a class-specific prompt conditioned on the image-level classifier's results. The strategy can prompt a pre-trained LLM to know whether changes exist and generate captions. Finally, the multiple prompts and the visual features of the feature-level encoder are fed into a frozen LLM for language generation. Compared with previous methods, our method can leverage the powerful abilities of the pre-trained LLM in language to generate plausible captions, which is free of training. Extensive experiments show that our method is effective and achieves state-of-the-art performance. Besides, an additional experiment demonstrates that our decoupling paradigm is more promising than the previous coupled paradigm for the RSICC task. We will make our codebase publicly available to facilitate future research at https://github.com/Chen-Yang-Liu/PromptCC**

***Index Terms*—Change captioning, decoupling paradigm, prompt learning, pre-trained large language model.**

I. Introduction

THE rapid development of remote sensing (RS) technologies empowers today's earth observation. Multi-temporal RS image processing is significant for land change detection and analysis. For land change interpretation, RS Image Change Captioning (RSICC) aims to automatically distinguish and describe the differences between multi-temporal images covering the same area with natural language, which is much more efficient compared with manual analysis. RSICC technology has significant application prospects, such as land planning [1], urbanization analysis [2], and environment monitoring [3].

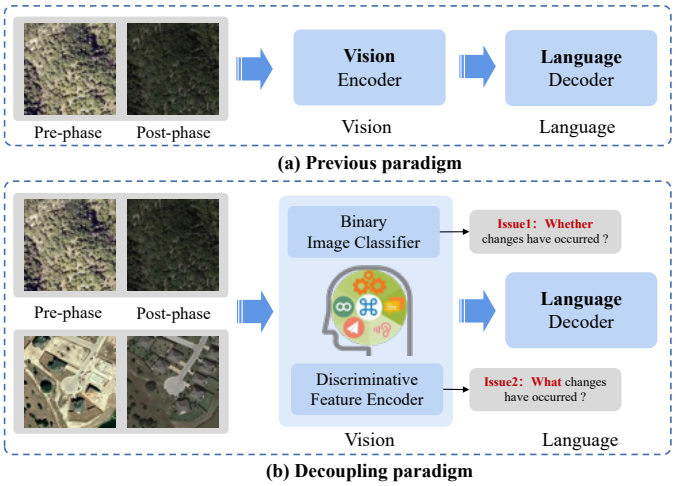


Fig. 1. The comparison of the two paradigms. In the vision stage, the previous paradigm process changed and unchanged image pairs in a coupled way. Unlike that, we decouple the RSICC task into two issues.

The RSICC task is at the intersection of RS image processing and natural language generation. The task requires visually comparing multi-temporal images and understanding semantic changes in the complex scene. Besides, visual understanding needs to be further translated into human language-like sentences. The generated sentences can indicate no change for unchanged image pairs or describe the semantic-level change of interest (e.g., object attributes, positional information, and affiliations between ground objects) for changed image pairs.

Currently, research efforts focusing on the RSICC task remain relatively limited. The existing methods are based on the encoder-decoder framework derived from the image captioning field, as depicted in Fig. 1 (a). For example, [4] and [5] used convolution-based encoders to extract bi-temporal features and recurrent neural network (RNN) based decoders to utilize the visual features to generate sentences word by word. In the case of [6] and [7], an alternate approach is employed, with a vision Transformer-based encoder and a Transformer-based decoder. The subsequent related work section will provide a comprehensive overview of the previous methods. Previous methods have yielded commendable outcomes and contributed to the advancement of RSICC technology.

However, these methods predominantly adhere to the encoder-decoder framework directly borrowed from the image

The work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160401), the National Natural Science Foundation of China under Grant 62125102, the Beijing Natural Science Foundation under Grant JL23005, and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Chenyang Liu, Jianqi Chen, Zipeng Qi, and Zhenwei Shi are with the Image Processing Center, School of Astronautics, with the Beijing Key Laboratory of Digital Media, and with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Rui Zhao is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

captioning field, overlooking the specificity of the RSICC task: the captioning difficulty is different for unchanged and changed image pairs involved in the RSICC task. As illustrated in Fig. 1 (a), the previous methods process the two kinds of image pairs in a coupled way. It is a considerable challenge to design an encoder that can simultaneously highlight changed features of changed image pairs and capture the no-change semantics of unchanged image pairs. In practice, the emphasis of these methods has predominantly been on the feature extraction of changed regions, which might not be conducive to describing unchanged image pairs effectively. In our intuition, the discrimination outcome regarding whether changes exist could serve as a pivotal prompt for human observation of bi-temporal visual differences and the subsequent word-by-word sentence generation. Driven by this notion, unlike the previous paradigm, we propose a decoupling paradigm, as depicted in Fig. 1 (b), which decouples the RSICC task into two issues: whether and what changes have occurred.

For caption generation, conventional methods required retraining a language model from scratch as a decoder, which incurred higher training costs and constrained the language model's learning due to the limited available training data. Inspired by the success of prompt learning and large language models (LLMs) in diverse downstream tasks [8], we introduce them to the RSICC task. Prompt learning aims to contextualize LLMs for specific tasks, leveraging the powerful capabilities of LLMs with minimal learning overhead. In our intuition, we can use a pre-trained LLM as the caption generator and employ prompt learning to prompt the LLM to describe whether and what changes exist in the image pairs.

In this paper, to address the weakness of previous methods, we propose a decoupling paradigm with prompt learning for RSICC (Prompt-CC). We utilize a Vision Transformer for bi-temporal feature extraction. To achieve the task decoupling, we introduce an image-level classifier and a feature-level encoder to further process the extracted features. Specifically, the image-level classifier performs binary classification to determine whether changes exist in the bi-temporal images. The feature-level encoder contributes to extracting discriminative features, aiding in the identification of changes of interest. For caption generation, we leverage prompt learning to utilize a frozen LLM as the caption generator. Precisely, to prompt the LLM to generate captions and understand the decoupling outcome, we propose a multi-prompt learning strategy which generates multiple prompts consisting of a set of unified prompts and a class-specific prompt depending on the classifier's result. Finally, these prompts and the features from the feature-level encoder are sent into a frozen LLM for captioning. Compared with previous methods, our method decouples the RSICC task and exploits the potential of the LLMs for caption generation. Extensive experiments demonstrate the effectiveness and state-of-the-art (SOTA) performance.

Our contributions can be summarized as follows:

- We propose a novel paradigm that decouples the RSICC task into two issues: whether and what changes have occurred. Specifically, we propose a pure Transformer-based model which incorporates an image-level classifier and a feature-level encoder to address those two issues.
- To our knowledge, we are the first to introduce prompt learning and the LLM into the RSICC task. We propose a multi-prompt learning strategy to effectively exploit the potential of the LLM for captioning, and prompt the LLM to know the decoupling result. It also eliminates the necessity of retraining a language decoder from scratch.
- Experiments validate the effectiveness of our decoupling paradigm and the multi-prompt learning strategy and show that our model achieves SOTA performance. Furthermore, an additional experiment demonstrates our decoupling paradigm is more proper than the previous coupled paradigm for the RSICC task.

II. Related work

The RSICC task can be viewed as RS change detection meets RS image captioning. In this section, we will briefly review these three tasks and their relationship in order to promote innovative RSICC methods. Then, we will briefly introduce prompt learning.

A. Remote Sensing Change Detection

RS change detection is a process of comparing bi-temporal images and generating a change map revealing the pixel-level change regions. Current deep learning-based methods mainly include supervised and unsupervised methods.

1) Supervised methods: Supervised methods require a large amount of data with labelled change maps to train deep neural networks. Most existing methods rely on convolution-based or Transformer-based models to learn high-level semantic features from bi-temporal images and then discriminate the features to determine the change regions of interest. Many research works have been devoted to improving the model's feature representation and discrimination ability, such as proposing various bi-temporal fusion strategies [9]–[11], proposing spatial multi-scale feature aggregation strategies [12]–[14], designing various attention modules [2], [15], [16], and introducing the self-attention mechanism or Transformer [17]–[21]. For example, Li *et al.* [9] presented a temporal feature interaction module to facilitate the interaction between multi-level bi-temporal features from the convolutional neural network (CNN) and obtain multi-level differencing features. Besides, a guided refinement module is designed to leverage cross-level complementary information to refine multi-level features. Peng *et al.* [22] proposed a dense attention method in which multiple up-sampling attention units consisting of up-sampling spatial and channel attention use the high-level features with category information to guide the selection of low-level features. Liu *et al.* [21] presented a prior-aware Transformer (PA-Former), which employs a CNN-based feature extractor to extract bi-temporal multi-level features. The low-level features are then fed into a Transformer encoder. The resulting features and multi-level features from the extractor are input into a Transformer decoder followed by a classifier for change discrimination.

2) Unsupervised methods: Although supervised methods can achieve satisfactory results, labelling data requires much human effort. Therefore, a series of unsupervised methods

have been proposed [23]–[26]. For example, Saha *et al.* [23] proposed a deep change vector analysis (DCVA) method, which combines CVA with a pre-trained CNN extracting spatial contextual information to identify changed pixels. Du *et al.* [24] presented a deep slow feature analysis (DSFA) model combining deep networks and slow feature analysis (SFA). The SFA constraint is used to suppress the unchanged components and highlight the changed components of the features extracted by deep networks. CVA and threshold algorithms are then employed to generate the change map. Tang *et al.* [26] proposed a graph convolutional network (GCN) and metric learning-based change detection (GMCD) model. They employ a multi-scale dynamic GCN module to capture contextual information and extract spatial-spectral features. Besides, GMCD employs an effective mechanism to combine spatial-spectral feature analysis and metric learning to generate reliable pseudo labels for unsupervised training.

Both change detection and change captioning aim to recognize changes of interest. However, they are at different levels: the former is at the pixel level, while the latter is at the semantic level.

B. Remote Sensing Image Captioning

RS image captioning is a process of generating sentences describing the visual content of RS images. Current methods can be roughly divided into three categories: retrieval-based, template-based and encoder–decoder-based.

Retrieval-based methods, such as [27], [28], aim to find the image in the database that is most similar to the input image in the latent space and then output the corresponding annotation sentences. Template-based methods, such as [29], generally perform object detection to identify ground elements and fill corresponding object words in the syntax template.

Most existing methods are based on the encoder-decoder framework [30]–[41]. These methods generally perform visual encoding to transform images into high-dimensional feature spaces and then utilize language models to translate visual features into sentences. Most methods concentrate on improving model performance from these two stages. Qu *et al.* [30] and Lu *et al.* [31] tried various CNNs (e.g., AlexNet [42], VGGNet [43], and GoogLeNet [44]) as the encoder to extract features and various RNNs (e.g., naive RNN [45], Long Short-Term Memory (LSTM) [46]) as the decoder to generate sentences. Many existing methods are devoted to improving the visual representation ability of the model. For example, Huang *et al.* [47] proposed a denoising-based multi-scale feature fusion (DMSFF) mechanism to aggregate multiple features at different stages of CNN. DMSFF performs the spatial-wise and channel-wise denoising on multi-scale features before fusing them. Li *et al.* [34] proposed a multi-level attention model. When generating each word, the model performs one attention operation on visual features, one on already generated word embeddings, and one on the resulting features from the first two attentions. Wang *et al.* [38] presented a multi-scale multi-interaction network in which self-attention and cross-attention are used to perform interaction between features of different scales. Besides, more attention-based methods can be found in [35]–[37], [48], [49].

MCCFormer-S: The scene is the same as before.
RSICCFormer: The scene is the same as before.
PSNet: Some houses are built along the road.
Ours: Some houses are built at the top left corner of the scene.

Fig. 2. An example of comparison results between our method and previous methods: MCCFormer-S [52], RSICCFormer [6], and PSNet [7]. More accurate and detailed words are marked in green, while red words are not.

Improving the model at the language decoding stage is also valuable for generating high-quality sentences. Given that Transformer has achieved great success in many sequence generation tasks such as machine translation, the language models adopted by the recent methods have transitioned from RNNs to Transformers, such as [39]–[41], [50], [51]. For example, [40] and [41] use the Transformer's encoder and decoder to process the grid features extracted by the CNN to achieve captioning. Ren *et al.* [51] feed visual grid features and the topic embedding with global information into the Transformer decoder. Chen *et al.* [50] combined a caption-type controller and the Transformer decoder to generate a caption with the specific type. Unlike previous methods that feed the features of the top Transformer encoding layer into the decoder, Liu *et al.* [39] propose an aggregator consisting of LSTM to fuse features from multiple encoding layers.

Although both the above image captioning and the RSICC aim to generate descriptive sentences, their focus is different: the former focus on the content of a single image, while the latter focus on the differences between two images.

C. Remote Sensing Image Change Captioning

Current RSICC methods are mainly based on the encoder-decoder framework, which is similar to most image captioning methods. According to different types of decoders, current methods include RNN-based and Transformer-based. For RNN-based methods, Chouaf *et al.* [4] first explored the RSICC task on a private dataset and proposed a model which uses a pre-trained VGG-16 [43] as the encoder to extract bi-temporal features and an RNN as the decoder to generate sentences. Hoxha *et al.* [5] extended the previous work of [4]. They designed two encoders based on the image-based or feature-based fusion strategy and tried two caption decoders, i.e., RNNs and support vector machines (SVMs). For Transformer-based methods, Liu *et al.* [6] provided a large LEVIR-CC dataset and some benchmark methods. They presented a dual-branch Transformer in which Siamese cross-encoding modules utilize differencing features to focus change regions and multistage bitemporal fusion modules use the features of different encoding stages to capture multiple changes of interest and exclude irrelevant changes. To improve the perception ability for changed objects of different sizes, Liu *et al.* [7] proposed a progressive scale-aware network, which employs multiple difference perception layers and scale-aware reinforcement modules to extract and utilize multi-scale discriminative features for caption decoding sufficiently.

However, the above methods predominantly adhere to the encoder-decoder framework directly borrowed from the image captioning field, overlooking the specificity of the RSICC task. They process unchanged and changed image pairs in a coupled way, so it is difficult to design a visual encoder that balances these two things: highlighting changed features of changed image pairs and capturing the no-change semantics of unchanged image pairs. Besides, they need to retrain a language decoder from scratch as the caption generator. Different from the previous methods, on the one hand, our method decouples the RSICC task by designing an image-level classifier and a feature-level encoder, which facilitates researchers to concentrate on the change captioning of changed and unchanged image pairs distinctly. On the other hand, our method can effectively exploit the powerful abilities of the pre-trained LLM to generate plausible captions without retraining a language decoder from scratch. In Fig. 2, an example of comparison results illustrates the effectiveness of our method.

D. Prompt Learning

Large pre-trained models can learn rich knowledge on large-scale corpora, and their powerful capabilities benefit various downstream tasks [53]. Previous approaches usually fine-tune them to adapt to specific tasks. However, fine-tuning will cost more computing resources as the model's parameters increase. With the emergence of GPT-3 (Generative Pretraining Transformer 3) [54], prompt learning receives growing attention. It aims to design appropriate prompts to remodel downstream tasks to adapt to pre-trained models. It allows researchers to exploit the powerful capabilities of pre-trained models at a low learning cost, while parameters are frozen. Here, we briefly introduce prompt learning in natural language processing (NLP) and computer vision (CV).

1) Prompt Learning in NLP: Generating prompts includes manual and automatic approaches. [54] and [55] designed manual prompts for a specific task. Wallace *et al.* [56] performed a gradient-based search over actual tokens to find discrete prompts iteratively. Li *et al.* [57] used a prefix-tuning method to learn continuous task-specific vectors as the prefix prompt of a language model. Vu *et al.* [58] combined prompt learning and transfer learning. They trained the prompt on some source tasks and then used the result as the initialization of the prompt on the target task. Jiang *et al.* [59] presented a prompt-ensembling method, which combines multiple prompts to obtain the probability for a single token. Tan *et al.* [60] presented a Multi-Stage Prompting (MSP) approach. They divided the language translation process into multiple separate stages. During each stage, MSP independently learns different soft prompts for promoting the language model better adapt to translation tasks. Liu *et al.* [8] conducted a comprehensive survey of the prompt learning methods in NLP.

2) Prompt Learning in CV: Zhou *et al.* [61] proposed Context Optimization (CoOp) to improve the performance of the original Contrastive Language-Image Pretraining (CLIP) [62] while pre-trained parameters are kept fixed. CoOp replaces the manual prompt of the CLIP with a set of learnable vectors. Zhou *et al.* [63] presented CoCoOp as an extension of CoOp, which generates an input-conditional dynamic prompt via a trainable lightweight network. Some recent works have achieved multimodal text generation by prefixing frozen language models with transformed visual features, such as [64]–[67]. For instance, Sollami *et al.* [65] used the product image features extracted by a visual encoder and text embeddings of the product title to form the multimodal conditional prefix of the autoregressive language model for generating product descriptions. Jia *et al.* [68] feed images and some learnable vectors into a pre-trained Vision Transformer backbone for multiple vision tasks. For a specific task, the prompt vectors and the prediction head are trainable while the backbone is frozen. Unlike previous methods adopting uni-modal prompting techniques, Khattak *et al.* [69] employed hierarchical prompts for both image and text branches. The dual-branch prompts are interactive to facilitate the alignment between image and text. Besides, more prompt learning methods for vision tasks can be found in [70]–[75].

III. METHODOLOGY

Fig. 3 shows an overview of our method, which contains two stages: 1) binary image change discrimination and 2) change perception and caption generation. The Vision Transformer of the pre-trained CLIP model extracts visual features from bi-temporal images which will be used in subsequent modules. To achieve the decoupling of the RSICC task, we present an image-level classifier to determine whether changes have occurred and a feature-level encoder to extract discriminative features. For caption generation, we introduce prompt learning to utilize a pre-trained LLM as our caption generator at a low cost. Specifically, we propose a multi-prompt learning strategy to generate multiple learnable prompts consisting of a set of unified prompts and a class-specific prompt depending on the classifier's results. Finally, the generated multiple prompts and the extracted visual features are then concatenated and sent into a frozen pre-trained LLM for caption generation.

A. Preliminary

1) CLIP: CLIP [62] is pre-trained with contrastive learning on large-scale image and text pairs. It consists of a text encoder and an image encoder. The text encoder employs a Transformer to process the text and obtain a textual embedding with global context information. The image encoder employed two different architectures (i.e., ResNet [76] and Vision Transformer [77]) to generate a visual embedding representing the image. Based on a large dataset with 400 million image-text pairs from the internet, CLIP is trained by maximizing the similarity between the visual embedding and corresponding textual embedding. CLIP bridges the gap between images and text and its powerful feature representation capability has benefited many vision-language tasks [78]. Inspired by this, we choose its image encoder to extract rich semantic features from images.

2) Transformer: As a fully-attentional sequence model, Transformer has a global receptive field with the help of its self-attention mechanism and has achieved great success in various RS image processing tasks [79], [80]. In this paper,

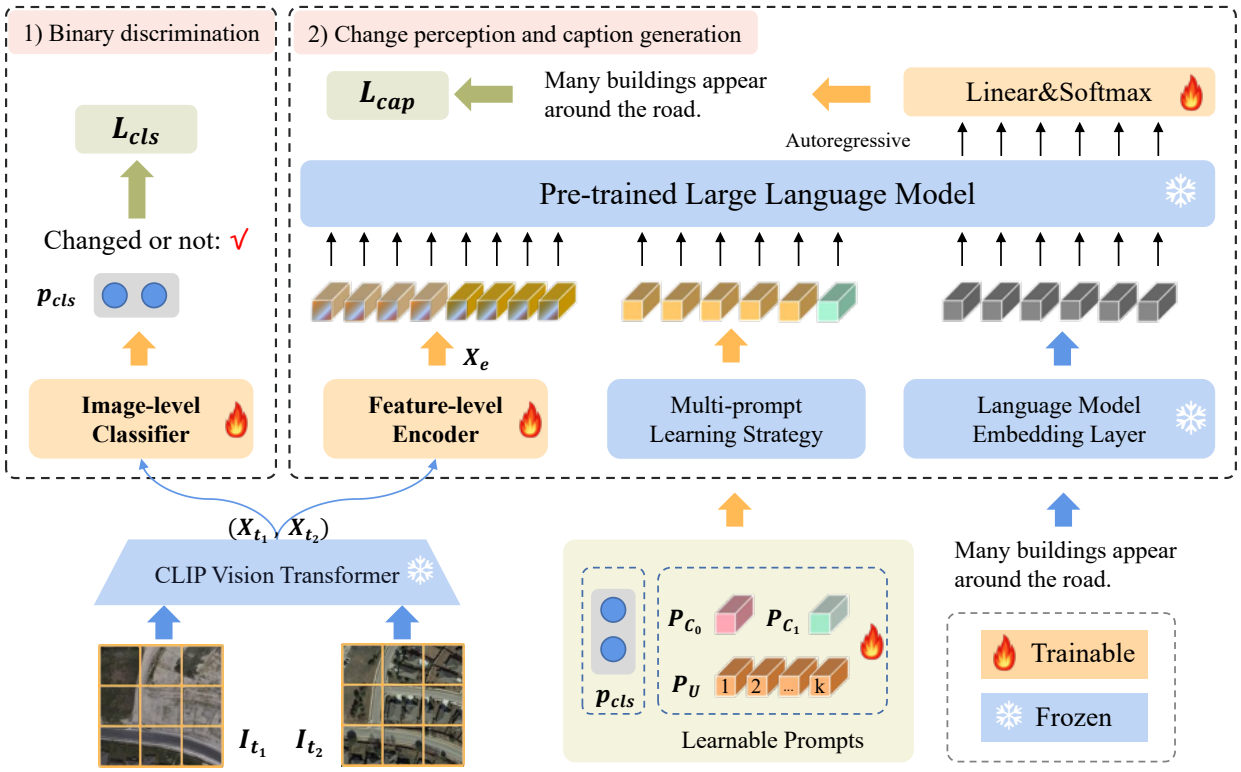


Fig. 3. The overview of our model which consists of two stages: 1) binary image change discrimination and 2) change perception and caption generation. To achieve the decoupling of the RSICC task, an image-level classifier performs binary classification to determine whether changes exist, and a feature-level encoder is used to determine what changes have occurred. Besides, our proposed multi-prompt learning strategy can prompt the LLM to know classification information and exploit the abilities of the LLM to generate plausible captions instead of retraining a caption generator.

our model is purely Transformer based. Both our image-level classifier and feature-level encoder use the Transformer encoder, which is stacked from multiple sublayers. Following most previous Transformer-based methods, for brevity, we will refer to the Transformer encoder as Transformer in the following text. Here, we will briefly explain how it works. For the input $\boldsymbol{X_l} \in \mathbb{R}^{N\times C}$ of the $l-th$ Transformer sublayer, the process is as follows:

$$\boldsymbol{X_s} = LN(\boldsymbol{X_l} + MSA(\boldsymbol{X_l})) \tag{1}$$

$$\boldsymbol{X_{l+1}} = LN(\boldsymbol{X_s} + FFN(\boldsymbol{X_s})) \tag{2}$$

where LN denotes the layer normalization [81], and FFN denotes a feed-forward network consisting of two linear layers with the Rectified Linear Unit (ReLU) activation function. MSA is the multi-head self-attention mechanism and can be formulated as follows:

$$MSA(\boldsymbol{X_l}) = Concat(head_1, ...head_h)\boldsymbol{W^O} \tag{3}$$

$$head_i = \boldsymbol{A_l}(\boldsymbol{X_l}\boldsymbol{W_i^V}) \tag{4}$$

$$\boldsymbol{A_l} = Softmax(\frac{(\boldsymbol{X_l}\boldsymbol{W_i^Q})(\boldsymbol{X_l}\boldsymbol{W_i^K})^T}{\sqrt{d_{model}}}) \tag{5}$$

where $\boldsymbol{W_i^Q} \in \mathbb{R}^{C\times d_{model}}$, $\boldsymbol{W_i^K} \in \mathbb{R}^{C\times d_{model}}$, $\boldsymbol{W_i^V} \in \mathbb{R}^{C\times d_{model}}$, and $\boldsymbol{W^O} \in \mathbb{R}^{(h\times d_{model})\times C}$ are trainable projection matrices to transform the feature dimension. d_{model} is the output dimension of the linear projection $\boldsymbol{W_i^Q}$, $\boldsymbol{W_i^K}$, and $\boldsymbol{W_i^V}$. h is the head number of the multi-head self-attention. $Softmax$ is the activation function.

B. Image-level Classifier

To achieve decoupling of the RSICC task, it is important to distinguish whether changes exist in the bi-temporal images. To this end, we present an image-level classifier to implement a binary classification task (i.e., the input image pair is changed or unchanged). Let $\boldsymbol{X_{t_1}} \in \mathbb{R}^{N\times d}$, $\boldsymbol{X_{t_2}} \in \mathbb{R}^{N\times d}$ be the bi-temporal features extracted by the Vision Transformer, where N is the token number and d is the channel dimension of the visual features. Fig. 4 illustrates the structure of the image-level classifier.

We first perform positional encoding on the $\boldsymbol{X_{t_i}}$ $(i = 1, 2)$ to preserve spatial position information and then concatenate them on the channel dimension of the features. We denote the resulting features as $\boldsymbol{X_f} \in \mathbb{R}^{N\times 2d}$. To obtain a vector for the classification, we add a learnable $\boldsymbol{cls}$ token with a dimension of $2d$. The $\boldsymbol{cls}$ token and the visual features $\boldsymbol{X_f}$ are then sent to a Transformer. The self-attention mechanism will allow the $\boldsymbol{cls}$ to capture the high-level semantics of the image pair. After that, we pass the corresponding output of the $\boldsymbol{cls}$ token through a linear followed by softmax activation to obtain the probability vector $\boldsymbol{p_{cls}} = [\ p_{cls}^{(0)}, p_{cls}^{(1)}]$. The $\boldsymbol{p_{cls}}$ represents the predicted classification probability on whether there is a difference between the bi-temporal images. We can formulate $\boldsymbol{p_{cls}}$ as follows:

$$\boldsymbol{p_{cls}} = Softmax(\boldsymbol{y_{cls}}) = \frac{exp(\boldsymbol{y_{cls}})}{\sum_{i=0}^{1} exp(y_{cls}^{(i)})} \tag{6}$$

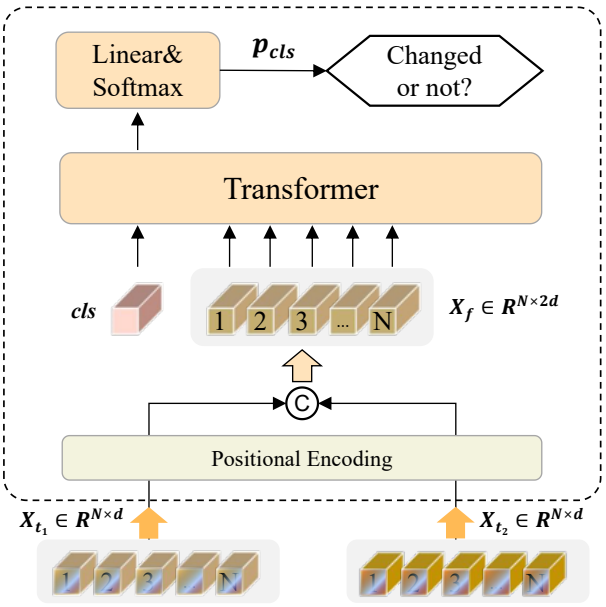


Fig. 4. The overall structure of the image-level classifier. The image-level classifier uses bi-temporal features extracted by the Vision Transformer to distinguish whether changes exist in image pairs.

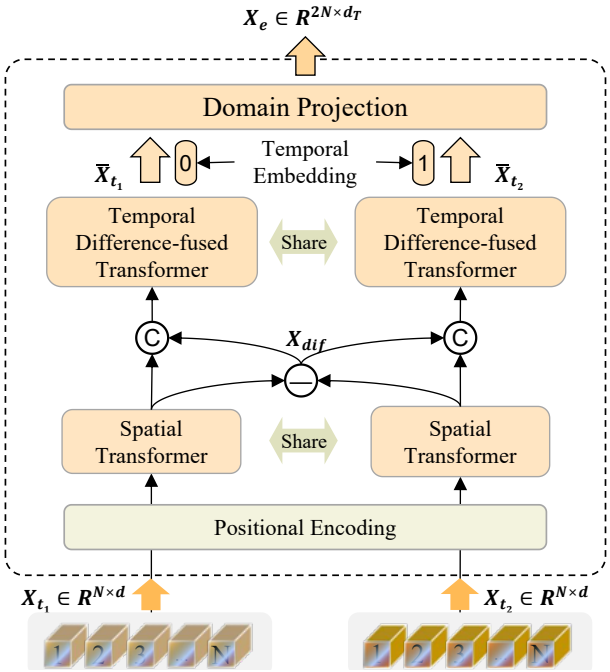


Fig. 5. The overall structure of the feature-level encoder. The differencing features contribute to determining changed regions and extracting discriminative features from bi-temporal images.

where $\boldsymbol{y_{cls}} = [\ y_{cls}^{(0)}, y_{cls}^{(1)}]$ is the output of the linear layer. $Softmax$ is an activation function.

C. Feature-level Encoder

Another issue of decoupling the RSICC task is distinguishing what changes have occurred. To address the issue, we present a feature-level encoder to extract discriminative features. As shown in Fig. 5, our feature-level encoder consists of a spatial Transformer (STr), a temporal difference-fused Transformer (TDTr), and a domain projection module. As with the image-level classifier, the input to the feature-level encoder is visual features $\boldsymbol{X}_{t_i}$ $(i = 1, 2)$ from the Vision Transformer.

After positional encoding, we pass $\boldsymbol{X}_{t_1}$ and $\boldsymbol{X}_{t_2}$ through the STr to extract their respective high-level semantic information. Until this step, there is still no temporal feature interaction. Considering that the bi-temporal images are registered, our intuition is that the differencing information between bi-temporal features will be beneficial to distinguish changed regions and identify changed ground elements. Thus, we concatenate differencing features and the features of each phase and then feed them into the TDTr to focus on changed regions and obtain discriminative semantic representations $\overline{X}_{t_1} \in \mathbb{R}^{2N\times d}$ and $\overline{X}_{t_2} \in \mathbb{R}^{2N\times d}$. To preserve temporal information, we then add temporal embeddings to $\overline{X}_{t_i}$ $(i = 1, 2)$. Since there is a gap between the visual embedding space and the textual embedding space of the GPT-2 model, it is necessary to perform domain transformation. To this end, we employ a linear layer in the domain projection module to transform obtained features into the textual embedding space of GPT-2. The resulting features will be fed into the GPT-2 with frozen pre-trained parameters.

D. Multi-prompt Learning Strategy

The previous prompt learning strategy usually adopts some learnable prompt embeddings as prefixes of LLMs to leverage the powerful abilities of pre-trained LLMs, such as [57], [58], [60]. These prompt embeddings are optimized for specific tasks. After training, the learned prompts will be used for each sample during inference.

Unlike the common strategy, considering the decoupling specificity of the RSICC task, we propose a multi-prompt learning strategy to generate multiple learnable prompts consisting of a set of unified prompts and a class-specific prompt. The unified prompts can be considered as global task-dependent prompts. Similar to the common prompt strategy, the learned unified prompt will be used for each sample in the inference stage. Different from previous methods, our proposed class-specific prompt is conditioned on the image-level classifier's result. Our intuition is that the class-specific prompt can facilitate the LLM to know whether changes exist.

Specifically, the class-specific prompt is generated by combining the two learnable prompt embeddings: $P_{C_0} \in \mathbb{R}^{1\times d_T}$ indicating no change and $P_{C_1} \in \mathbb{R}^{1\times d_T}$ indicating changes, where d_T is the dimension of the textual embedding. Here, we propose to combine them in two ways:

1) Soft class-specific prompt: The soft class-specific prompt $P_{C_s} \in \mathbb{R}^{1\times d_T}$ is generated by weighting P_{C_0} and P_{C_1} with the predicted probabilities of the image-level classifier.

$$P_{C_s} = p_{cls}^{(0)} * P_{C_0} + p_{cls}^{(1)} * P_{C_1} \tag{7}$$

where $p_{cls}^{(0)}$ and $p_{cls}^{(1)}$ are respectively the probability of no change and the probability of change predicted by the image-level classifier in the section III-B.

2) Hard class-specific prompt: The hard class-specific $P_{C_h} \in \mathbb{R}^{1\times d_T}$ are generated by selecting one of P_{C_0} and P_{C_1}

Algorithm 1 The procedure of the caption generation.

Input: $\boldsymbol{X_e}$, MP_i and GT (ground-truth caption).
Output: $Caption$
Define: $\Phi_{De} \leftarrow$ the GPT-2 decoder model.
$\Phi_E \leftarrow$ the text embedding layer of GPT-2.
$\Phi_M \leftarrow$ the embedding-to-word mapping.
// $Mode$ denotes the training stage or testing stage.
if $Mode$ == TRAIN **then**
  $E_{GT} = \Phi_E(GT)$
  $E_{ws} = \Phi_{De}([\boldsymbol{X_e}; MP_i; E_{GT}])$
  $Caption = \Phi_M(E_{ws})$
  return $Caption$
end if
if $Mode$ == TEST **then**
  $E_S = \emptyset$
  $E_w = \emptyset$
  while $E_w \neq \Phi_E(".")$ **do**
    $E_w = \Phi_{De}([\boldsymbol{X_e}; MP_i; E_S])$
    $E_S = [E_S; E_w]$
    $Caption = [Caption; \Phi_M(E_w)]$
  end while
  return $Caption$
end if

based on the classification result of the image-level classifier.

$$cls = Argmax(\boldsymbol{p_{cls}}) \tag{8}$$

$$P_{C_h} = (1 - cls) * P_{C_0} + cls * P_{C_1} \tag{9}$$

where $Argmax$ operation is performed on the channel dimension of $\boldsymbol{p_{cls}}$. cls (i.e. 0 or 1) denotes the predicted change category of the classifier.

Let $P_U \in \mathbb{R}^{K \times d_T}$ denote unified prompts, where K is the number of learnable embeddings. Finally, generated multi-prompt $MP_i (i = s, h)$ that are fed into the LLM can be formulated as follows:

$$MP_i = [P_U; P_{C_i}] \in \mathbb{R}^{(K+1) \times d_T} \tag{10}$$

where [;] denotes the concatenation operation.

E. Pre-trained Large Language Model

The pre-trained LLMs have powerful feature representation abilities due to training on a large-scale corpus. Inspired by the great success of the GPT family (i.e., GPT-2 [82], GPT-3 [54], and ChatGPT [83]) on language generation tasks, we employ available GPT-2 [82] as our caption generator. GPT-2 is an auto-regressive LLM based on Transformer. It is trained on a large-scale Internet text dataset and the training objective is to predict the next word. Our intuition is that we can leverage its powerful text-generation ability to generate captions.

Until now, we have obtained context-rich $\boldsymbol{X_e} \in \mathbb{R}^{2N \times d_T}$ of the feature-level encoder and multi-prompt $MP_i \in \mathbb{R}^{(K+1) \times d_T}$ $(i = s, h)$. We then feed them to the frozen GPT-2 to generate sentences in an auto-regressive manner. The detailed procedure of the caption generation is shown in Algorithm 1.

F. Objective Function

We train our modal in a supervised way. The training objectives include binary classification and caption generation. We adopted the cross-entropy function to compute classification loss and caption loss. Specifically, to constrain the image-level classifier, we define the classification loss as follows:

$$\mathcal{L}_{cls} = -(\widetilde{y}_{cls}^{(0)} \log p_{cls}^{(0)} + \widetilde{y}_{cls}^{(1)} \log p_{cls}^{(1)}) \tag{11}$$

where $\boldsymbol{p}_{cls} = [\ p_{cls}^{(0)}, p_{cls}^{(1)}]$ is the predicted probability vector described in Equation (6). $\widetilde{\boldsymbol{y}}_{cls} = [\ \widetilde{y}_{cls}^{(0)}, \widetilde{y}_{cls}^{(1)}]$ is the one-hot vector representation of the ground-truth.

To constrain the caption-generation-related modules, we define the caption loss as follows:

$$\mathcal{L}_{cap} = -\sum_{t=1}^{L} \sum_{v=1}^{V} \widetilde{y}_t^{(v)} \log p_t^{(v)} \tag{12}$$

where $\boldsymbol{p}_t = [\ p_t^{(1)}, p_t^{(2)}, ..., p_t^{(V)}]$ is the probability vector for predicting the t-th word. $\widetilde{\boldsymbol{y}}_t = [\ \widetilde{y}_t^{(1)}, \widetilde{y}_t^{(2)}, ...\widetilde{y}_t^{(V)}]$ is the t-th word embedding of a ground-truth caption. V is the vocabulary size, and L is the length of the ground-truth caption.

IV. EXPERIMENTS

A. Dataset

1) Dataset description: LEVIR-CC is a large-scale change captioning dataset proposed in the previous work [6]. It contains 10077 image pairs (5038 changed image pairs and 5039 unchanged image pairs) from 20 different regions in Texas, USA. The time span is between 5 and 15 years. The image size is 256×256 pixels with a resolution of 0.5m/pixel. The dataset contains various ground surface changes such as buildings, roads, vegetation, and water. Each image pair of the dataset has five annotated captions, totaling 50385 descriptive captions.

2) Further preparation: Our decoupling paradigm actually requires achieving the binary classification of image pairs and image change captioning. However, there are only descriptive captions in the original annotation files of the dataset, such as "many houses were built beside the road" and "no change has occurred". To compute the classification loss in Equation (11), we collect the corresponding classification labels by analyzing the semantics of all annotated captions. We will made the updated annotations file publicly available at *https://github.com/Chen-Yang-Liu/PromptCC*.

B. Evaluation Metrics

Following previous image change captioning works [4]–[6], we employ four quantitative metrics to measure the quality of generated sentences, including BLEU-n (n=1,2,3,4) [84], ROUGE$_L$ [85], METEOR [86], and CIDEr-D [87]. BLEU-n is a metric based on n-gram precision between the generated sentence and ground-truth sentences. ROUGE$_L$ is more concerned with the recall rate. Different from BLEU and ROUGE$_L$, METEOR introduces a penalty factor to take generated sentence fluency into account. CIDEr-D is a metric specially designed for image captioning. It calculates the term frequency-inverse

TABLE I
COMPARISON RESULTS OF OUR PROPOSED METHOD AND THE PREVIOUS SOTA METHODS ON THE LEVIR-CC DATASET. OUR METHOD ACHIEVES SOTA PERFORMANCE AND OUTPERFORMS PREVIOUS METHODS BY A SIGNIFICANT MARGIN.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE$_L$	CIDEr-D	S_m^*
Capt-Rep-Diff [90]	73.25	65.66	60.02	55.56	33.03	67.22	116.61	68.10
Capt-Att [90]	76.17	68.05	61.74	56.91	34.83	69.38	121.78	70.72
Capt-Dual-Att [90]	78.17	70.73	64.01	58.17	35.23	71.60	127.51	73.13
DUDA [90]	79.64	71.46	64.42	58.42	35.76	71.47	128.24	73.47
MCCFormer-D [52]	76.23	68.14	62.46	58.32	35.18	68.43	121.10	70.76
MCCFormer-S [52]	79.38	73.23	65.38	60.86	36.88	71.06	127.90	74.17
RSICCFormer [6]	81.96	73.71	67.11	61.83	38.16	72.57	132.00	76.14
PSNet [7]	81.97	74.09	67.62	62.22	37.92	73.10	132.87	76.53
Prompt-CC (Ours)	**83.66**	**75.73**	**69.10**	**63.54**	**38.82**	**73.72**	**136.44**	**78.13**

document frequency (TF-IDF) to assign different weights to different n-grams. For all these metrics, higher scores represent better sentence quality. Besides, we follow the previous work [88] to adopt an overall metric S_m^* to integrate the above metrics. It is defined as follows:

$$S_m^* = \frac{1}{4} * (BLEU\text{-}4 + METEOR + ROUGE_L + CIDEr\text{-}D) \quad (13)$$

C. Implementation Details

We use the ViT-B/32 of CLIP to extract grid features instead of a global image embedding. The text embedding layer of GPT-2 is used to convert words into word embeddings with a dimension of 768. Three Transformer layers are used in the image-level classifier. In the feature-level encoder, the STr and TDTr employ two Transformer layers and three Transformer layers, respectively. P_U consists of five learnable embeddings. Positional encoding and temporal encoding are implemented by adding learnable embeddings.

All experiments are implemented in the PyTorch framework. The parameters of pre-trained ViT-B/32 and GPT-2 are kept fixed. Other model parameters are optimized by using the Adam optimizer [89] with an initial learning rate of 0.0001. We set the maximum epoch to 40 and the batch size to 16. In the training stage, we only used one unchanged annotation caption instead of five. The training process will be terminated when the validation BLEU-4 score does not increase for ten consecutive epochs. We employ a two-stage training approach. We first train the image-level classifier with the classification loss $\mathcal{L}_{cls}$. After that, we use the ground-truth classification to generate prompts. The learnable prompt embeddings are initialized randomly and, along with other trainable model parameters, are iteratively updated through backpropagation by minimizing the objective caption loss $\mathcal{L}_{cap}$.

In the test stage, the classifier's prediction is used for the multi-prompt generation. Following previous work [6], we use the beam search strategy with a beam size of 3 when generating captions. Besides, to ensure a fair quantitative comparison, all methods are trained and evaluated based on the publicly available subword-based tokenizer and word mapping of GPT, which are more comprehensive and widely acknowledged. We also recommend that future researchers follow this.

D. Comparison to State-of-the-Art

In this section, we compare our method and several SOTA methods on the LEVIR-CC dataset, including Capt-Rep-Diff [90], Capt-Att [90], Capt-Dual-Att [90], DUDA [90], MCCFormer-S [52], MCCFormer-D [52] and RSICCFormer [6]. These methods use the encoder-decoder framework, in which the encoder visually extracts discriminative features and the decoder utilizes LSTM or Transformer to generate captions. These methods are described in detail as follows:

- Capt-Rep-Diff [90]: A CNN-based encoder extracts visual features from the bi-temporal images. The bi-temporal features and the differencing features are concatenated together and fed into an LSTM decoder for captioning.
- Capt-Att [90]: It employs a spatial attention mechanism. Specifically, differencing features are fed into a CNN-based lightweight network to obtain spatial attention maps, which are attended to bi-temporal features to focus on changed regions. After that, the differencing features of bi-temporal features are used for LSTM decoding.
- Capt-Dual-Att [90]: Unlike Capt-Att, Capt-Dual-Att learns two spatial attention maps on the visual stage, which are applied to bi-temporal features respectively. Specifically, differencing features and features of each phase are concatenated and fed into Siamese lightweight networks to obtain two sets of spatial attention weights.
- DUDA [90]: As an extension of Capt-Dual-Att, DUDA employs a dynamic speaker instead of an LSTM for improved language decoding. This dynamic speaker comprises an LSTM-based dynamic attention module and an LSTM-based caption generator. The dynamic attention module fuses visual and textual features to process bi-temporal and distinctive features, which are then input to the caption generator.
- MCCFormer-S [52]: MCCFormer-S is a change captioning method designed for synthetic images with viewpoint change. It employs a Transformer encoder to handle bi-temporal features extracted by CNNs, facilitating patch correlation across two images through the multi-head attention mechanism. Subsequently, a Transformer decoder translates the resulting features into descriptive sentences.
- MCCFormer-D [52]: Instead of feeding bi-temporal fea-

TABLE II
THE COMPARISON RESULT OF MODEL PARAMETERS, FLOATING-POINT OPERATIONS (FLOPS), TRAINING TIME PER EPOCH, AND TESTING TIME FOR THE LEVIR-CC TESTSET. THESE DATA ARE MEASURED ON A COMPUTING DEVICE EQUIPPED WITH AN INTEL CORE I9-13900K CPU AND AN NVIDIA GTX 4090 GPU.

Method	Parameters			FLOPs	Time	
	Total	Trainable	Trainable/Total		Training	Test
Capt-Rep-Diff [90]	73.21 M	45.67 M	62.38%	19.82 G	3min16s	55s
Capt-Att [90]	73.60 M	46.07 M	62.58%	19.90 G	3min21s	55s
Capt-Dual-Att [90]	75.58 M	48.05 M	63.57%	20.03 G	3min34s	1min20s
DUDA [90]	80.31 M	52.78 M	65.71%	20.28 G	3min43s	1min22s
MCCFormer-D [52]	162.55 M	135.01 M	83.06%	25.09 G	3min30s	1min37s
MCCFormer-S [52]	162.55 M	135.01 M	83.06%	25.09 G	3min36s	1min29s
RSICCFormer [6]	172.80 M	145.26 M	84.06%	27.10 G	4min54s	4min2s
PSNet [7]	319.76 M	231.53 M	72.41%	13.78 G	3min17s	2min5s
Prompt-CC (Ours)	408.58 M (classifier: 56.68 M) (GPT-2: 124 M)	196.28 M (classifier: 56.68 M)	**48.04**%	19.88 G	4min30s	3min14s

tures into a Transformer encoder, MCCFormer-D employs Siamese Transformers with the cross-attention mechanism. Specifically, the visual features of a single phase are used as the query (Q), while the features of another phase are the key (K) and value (V). A Transformer decoder uses obtained features to generate captions.

- RSICCFormer [6]: RSICCFormer is specially designed for registered RS image pairs. Unlike MCCFormer, the cross-attention mechanism leverages the differencing features as the K and V to enhance visual features (Q) from a single-phase image. Multiple Transformer layers with cross-attention are stacked to progressively extract and utilize the differencing information to recognize changes of interest.
- PSNet [7]: Multiple difference perception layers are stacked to use differencing features to sufficiently extract multi-scale features. Scale-aware reinforcement modules leverage the attention mechanism to enhance the interaction between multi-scale features.

1) Quantitative Results: To ensure a fair comparison, we retrained and tested all comparison methods based on the tokenizer and word mapping of GPT-2, rather than directly copying scores from previous papers. Tab. I reports the comparison result of our method and the previous methods on the LEVIR-CC dataset. Unlike the previous methods processing unchanged and changed image pairs in a coupled way, our method exploits the characteristics of the RSICC task to decouple the task. Prompt-CC (ours) uses the hard class-specific prompt. We can observe that it can achieve SOTA performance and outperforms the previous best model PSNet by a significant margin (+1.32% on BLEU-4, +3.57% on CIDEr-D, and +1.60% on S_m^*). It is attributed to the fact that our method decouples the RSICC task and our multi-prompt learning strategy facilitates the model to exploit the powerful capabilities of the pre-trained LLM in language generation.

For a comprehensive understanding of the model complexity, in Tab. II, we report the comparison result of model parameters, floating-point operations (FLOPs), training time per epoch, and testing time for the LEVIR-CC test set. These data are measured on a computing device equipped with an Intel Core i9-13900K CPU and an NVIDIA GTX 4090 GPU. Based on Tab. I and Tab. II, we can observe that our method achieves significant performance improvement compared to the previous method without significantly increasing the complexity of the model. Additionally, it is worth noting that our method has fewer trainable parameters than the previous best model PSNet and has a lower proportion of trainable parameter proportion than previous SOTA methods, such as PSNet and RSICCFormer. For our Prompt-CC model, the additional image-level classifier accounts for 56.68 million parameters and GPT-2 has 124 million parameters. While their utilization results in more parameters, considering the achieved performance improvement and the powerful language generation ability capability of LLMs with prior knowledge, we deem this trade-off acceptable. To mitigate model complexity, on the one hand, future researchers may explore designing a more lightweight and efficient classifier to achieve relatively simple binary image-level change classification. On the other hand, with the recent vigorous emergence of various pre-trained LLMs, future researchers can explore substituting GPT-2 with other pre-trained LLMs to strike a balance between better model performance and lower model complexity. Lastly, it's worth noting that apart from leveraging pre-trained LLM with prompt learning, the decoupling paradigm stands as another significant contribution. An extensive discussion on the advantage of the decoupling paradigm will be presented in Section IV-G.

2) Qualitative Results: Fig. 6 shows qualitative comparison results between our method and previous SOTA methods: MCCFormer-S [52], RSICCFormer [6], and PSNet [7]. Green words are more accurate and detailed, while red words are wrong. It is evident that our method can ignore irrelevant changes and accurately distinguish whether changes of interest have occurred. For unchanged image pairs, such as the second, third, and fifth image pairs, our method can correctly identify cases where no change has occurred. In contrast, the performance of the previous methods was comparatively poorer. For changed image pairs, our method exhibits better perfor-

a) A small track appears.
b) The scene is the same as before.
c) The scene is the same as before.
d) The scene is the same as before.
e) A road has been built on the bareland.

a) The scene is the same as before.
b) Some houses are built at the top of the scene.
c) A house is built at the edge of the scene.
d) The scene is the same as before.
e) The scene is the same as before.

a) The scene is the same as before.
b) Some houses are built along the road.
c) The scene is the same as before.
d) Some houses are built along the road.
e) The scene is the same as before.

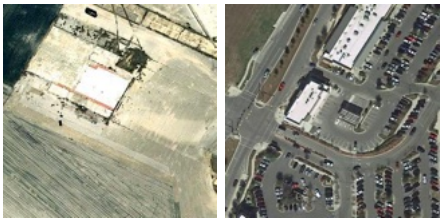
a) Several houses are added to the residential area at the top.
b) The scene is the same as before.
c) The scene is the same as before.
d) Some houses are built along the road.
e) Some houses are built at the top left corner of the scene.

a) The scene is the same as before.
b) Some houses are built along the road.
c) A building appears at the top right corner of the scene.
d) The scene is the same as before.
e) The scene is the same as before.

a) A road with houses on both sides appears in the desert.
b) Some roads and houses are built on bareland.
c) Some roads and houses are built on bareland.
d) Some houses are built along the road.
e) A road with many houses are built on the bareland.

a) Some trees have disappeared and some roads are built.
b) A road is built on the left and a building is built at the bottom.
c) A road appears in the bareland.
d) A road is built and some houses are built along the road.
e) Some roads appear in the bareland.

a) The former forest has been replaced by a residential area with roads and houses.
b) Some houses are built along the road.
c) Two rows of houses are built on both sides of the road.
d) A road with many houses built along appears.
e) Trees are removed and a road with villas built along.



a) Two large buildings with roads and parking lots appear.
b) Some houses are built along the road.
c) Some houses are built along the road.
d) A parking lot is built on the bareland.
e) Some buildings with a parking lot is built on the bareland.

a) A giant building with parking lots appear beside the road.
b) Some houses are built at the bottom corner of the scene.
c) A parking lot appears at the top left corner of the scene.
d) Some houses are built along the road.
e) A parking lot is built on the bareland.

a) A building is built at the top left corner of the scene.
b) The scene is the same as before.
c) Two houses appear at the top left corner of the scene.
d) A house appears at the top left corner of the scene.
e) A building is built at the top corner of the scene.

a) A giant building with parking lots appear beside the road.
b) Some houses are built at the bottom corner of the scene.
c) A large building with a parking lot appears in the bareland.
d) A large building is built on the bareland.
e) A large building with a parking lot is built on the bareland.

Fig. 6. Qualitative comparison of captioning results between our method and previous SOTA methods. (a): one of the five ground-truth captions, (b): MCCFormer-S [52], (c): RSICCFormer [6], (d): PSNet [7], (e): our Prompt-CC model. More accurate and detailed words are marked in green, while red words are not. Compared to the previous method, our method can ignore irrelevant changes and accurately distinguish whether changes of interest have occurred.

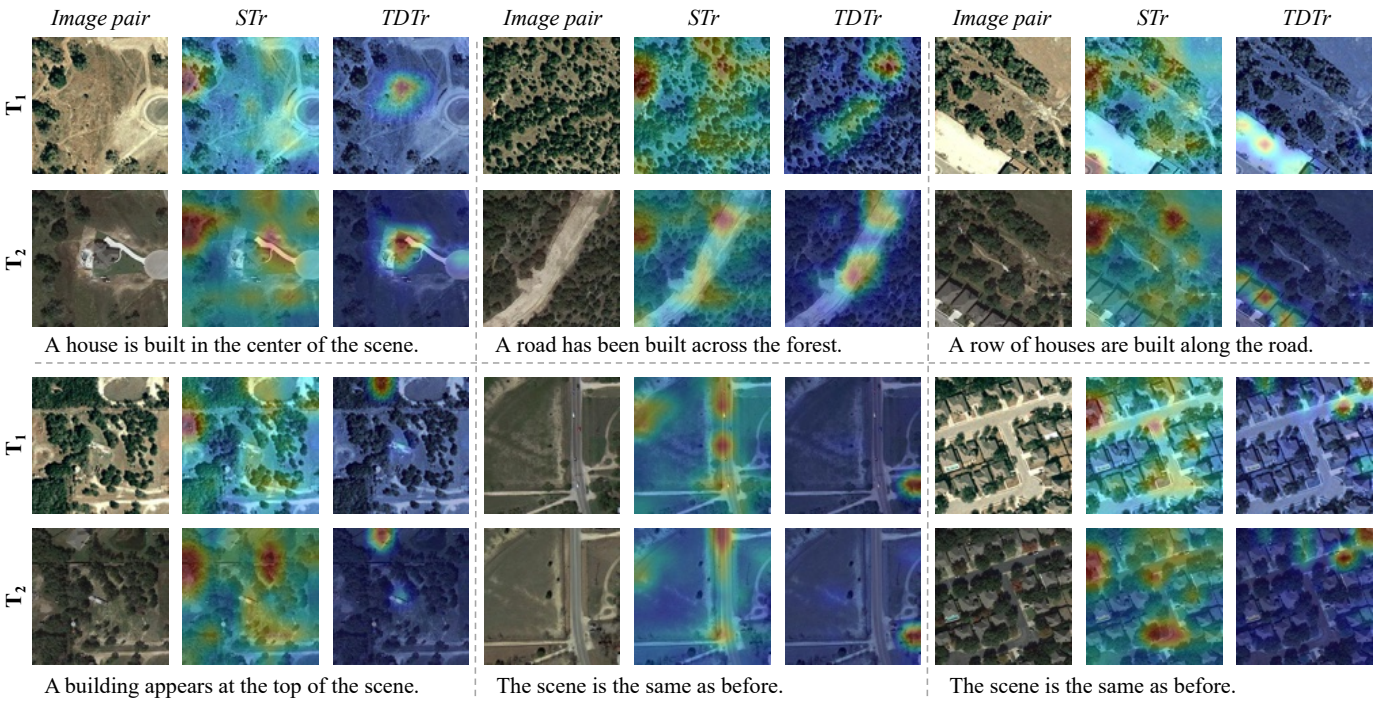


Fig. 7. Visualization of the attention weights of the last layer for the STr and TDTr. The captions are generated by our Prompt-CC model. Compared to the STr, the TDTr can utilize the temporal differencing features to focus on the changed region of interest in two images.

mance in accurately describing changes of interest compared to the previous methods. For example, for the fourth image pair, MCCFormer-S and RSICCFormer fail to identify the presence of any change, and PSNet misidentifies the position of the houses. In contrast, our method accurately recognizes and describes changed houses, change process ("*built*"), and position ("*top left corner*"). Additionally, the sixth and seventh image pairs demonstrate that our method performs better for the accurate use of quantifiers.

E. Ablation Studies

1) Multi-prompt learning strategy: Our multi-prompt learning strategy can exploit the powerful capabilities of the pre-trained LLM and prompt the LLM to know whether changes exist in the images or not. In Tab. III, we conducted ablation studies to validate the effectiveness of our multi-prompt learning strategy. Compared with fine-tuning the GPT-2, prompt learning performs better when leveraging the LLM as our caption generator. The result demonstrates that it is effective to introduce prompt learning into the RSICC task. Furthermore, we can observe that our multi-prompt learning strategy can further improve the model performance and the model performs best when using the hard class-specific prompt. Besides, we report the scores when using a hand-crafted prompt ("*Describe differences between images:*"). We can observe that despite performing less effectively than our multi-prompt learning strategy, the hand-crafted prompt still outperforms fine-tuning GPT-2.

In this paper, to utilize the decoupling results, we propose the multi-prompt learning strategy, which enables the captioning of changed and unchanged image pairs to be achieved by a unified LLM. Apart from that, there is a naive dual-branch approach to utilize the decoupling results: one branch directly generates no-change captions when the classification result is no change, while the other branch uses an LLM to generate captions when the classification result indicates the occurrence of changes. Tab. IV shows the comparison of the two approaches. Two approaches have the same performance in captioning for images predicted to be unchanged while our method outperforms the dual-branch approach for the images predicted to be changed. This demonstrates that joint training on unchanged and changed images is helpful for the feature-level encoder and the learnable prompts to learn semantic context information about visual changes.

2) Feature-level encoder: The feature-level encoder is used to determine what changes have occurred. It utilizes the STr and the TDTr to perform spatial and temporal interactions on the bi-temporal features. In Tab. V, we conducted ablation studies to validate the effectiveness of the structure of the feature-level encoder. Quantitative results show that both spatial feature interaction and temporal feature interaction are effective. When combining the STr and the TDTr, the model can achieve the best performance, which illustrates the structural rationality of our feature-level encoder.

Besides, we conducted qualitative experiments to verify that the TDTr contributes to determining changed regions of changed image pairs. To this end, we visualize the attention weights $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ of the Transformer in Equation (5). Specifically, the attention map can be obtained as follows:

$$Average(\boldsymbol{A}, dim = -1) \rightarrow \boldsymbol{Aver} \in \mathbb{R}^{N\times 1} \tag{14}$$

$$Reshape(\boldsymbol{Aver}) \rightarrow map \in \mathbb{R}^{h\times w} \tag{15}$$

TABLE III

ABLATION STUDIES ON THE MULTI-PROMPT LEARNING STRATEGY. $\boldsymbol{P_U}$ DENOTES THE UNIFIED PROMPTS. P_{C_s} DENOTES THE SOFT CLASS-SPECIFIC PROMPT. P_{C_h} DENOTES THE HARD CLASS-SPECIFIC PROMPT. OUR MULTI-PROMPT LEARNING STRATEGY IS EFFECTIVE AND PERFORMS BEST WHEN USING THE HARD CLASS-SPECIFIC PROMPT.

Fine-tuning	Prompt learning			BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_m^*
	$\boldsymbol{P_U}$	P_{C_s}	P_{C_h}								
✔	–	–	–	79.32	70.53	63.82	58.39	37.19	71.16	128.78	73.88
–	✔	✗	✗	82.78	75.01	68.49	63.07	38.12	72.92	134.08	77.05
–	✔	✔	✗	82.95	74.70	67.93	62.39	38.63	73.37	135.29	77.42
–	✔	✗	✔	**83.66**	**75.73**	**69.10**	**63.54**	**38.82**	**73.72**	**136.44**	**78.13**
–	hand-crafted prompt			82.14	74.39	67.88	62.27	37.83	72.25	132.63	76.25

TABLE IV

COMPARISON OF THE TWO DIFFERENT APPROACHES UTILIZING THE DECOUPLING RESULTS. THE $\boldsymbol{cls}$ DENOTES THE PREDICTED CLASSIFICATION RESULT. FOR THE DUAL-BRANCH APPROACH, ONE BRANCH DIRECTLY GENERATES NO-CHANGE CAPTIONS WHEN THE $\boldsymbol{cls}$ IS 0, WHILE THE OTHER BRANCH USES THE LLM TO GENERATE CAPTIONS WHEN THE $\boldsymbol{cls}$ IS 1.

Evaluation Scope	Captioner	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_m^*
$\boldsymbol{cls}$ = 0 (unchanged)	Dual-branch	90.80	88.58	87.83	87.46	59.24	89.66	181.22	104.39
	Ours	90.80	88.58	87.83	87.46	59.24	89.66	181.22	104.39
$\boldsymbol{cls}$ = 1 (changed)	Dual-branch	75.65	63.16	50.65	40.00	24.84	51.73	55.86	43.11
	Ours	**76.87**	**63.81**	**51.78**	**41.79**	**26.25**	**53.53**	**64.29**	**46.47**
All of Test set	Dual-branch	82.93	75.40	68.70	63.03	38.01	72.93	132.71	76.67
	Ours	**83.66**	**75.73**	**69.10**	**63.54**	**38.82**	**73.72**	**136.44**	**78.13**

TABLE V

ABLATION STUDIES ON THE STRUCTURE OF THE FEATURE-LEVEL ENCODER. QUANTITATIVE RESULTS SHOW THAT THE COMBINATION OF THE STR AND THE TDTR CAN ACHIEVE THE BEST PERFORMANCE.

STr (Spatial)	TDTr (Temporal)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE$_L$	CIDEr-D	S_m^*
✔	✗	83.12	74.93	67.83	61.94	38.28	73.23	133.08	76.63
✗	✔	83.20	75.37	68.67	63.02	38.06	73.21	133.27	76.89
✔	✔	**83.66**	**75.73**	**69.10**	**63.54**	**38.82**	**73.72**	**136.44**	**78.13**

Fig. 7 shows the visualization of the attention weights of the last layer for the STr and TDTr. when changed image pairs are processed. As we can see, the STr focus more on understanding the semantic content in a single image, while the TDTr can focus on the changed regions of interest for changed image pairs. It is attributed to the fact that the TDTr utilize the temporal differencing features. For example, in the second image pair, the STr pays attention to the whole scene of the two images, while the TDTr focuses more on the area of the new road. In the fourth image pair, the STr just pays attention to large areas of trees and fails to attend to the small house, while the TDTr focuses more on the newly built small house at the top of the scene. For the last two unchanged images, although the attention of TDTr attends to slight changes in the trees, it still generates accurate statements revealing that no change occurred. This indicates that our Prompt-CC model is capable of recognizing these changes as irrelevant.

F. Qualitative Analysis on Multi-prompt Learning Strategy

1) Interpretation of prompt: After the training stage, we can obtain the multi-prompt $\boldsymbol{MP_s} \in \mathbb{R}^{(K+1)\times d_T}$ (i.e., using the soft class-specific prompt) and multi-prompt $\boldsymbol{MP_h} \in \mathbb{R}^{(K+1)\times d_T}$ (i.e., using the hard class-specific prompt). During the inference stage, obtained MP_i $(i = s, h)$ will be used to prompt the LLM to know whether changes exist in the input image pairs and generate captions. To further understand the meaning of MP_i $(i = s, h)$, we have tried to interpret them as a sequence of words. Specifically, since the prompt embeddings and textual embeddings share the same latent space of the LLM, we search the vocabulary for the closest word embedding to each embedding in MP_i based on cosine similarity, and use corresponding words as the interpretability words of the prompt embeddings. Tab. VI shows the interpretability of MP_i when the classification result is changed or unchanged. We observe an interesting phenomenon that the learned prompt MP_i is incomprehensible to humans although it is effective for the LLM. This is one reason why many prompt-based methods [8] choose to use learnable prompt embeddings rather than hand-crafted prompts.

2) Captioning comparison results of MP_s and MP_h: Fig. 8 shows some randomly selected captioning results generated by our Prompt-CC when using soft class-specific prompt and hard class-specific prompt. The results demonstrate that our

Fig. 8. Captioning results. (a) is one of the five ground-truth captions. (b) is generated by our Prompt-CC (soft) with the soft class-specific prompt, while (c) is by the Prompt-CC (hard) with the hard class-specific prompt. More accurate and detailed words are marked in green, while red words are not.

TABLE VI
INTERPRETABILITY OF THE PROMPT. $\boldsymbol{p}_{cls}$ IS THE PROBABILITY VECTOR DESCRIBED IN EQUATION (6) AND ILLUSTRATES WHETHER CHANGES EXIST IN IMAGE PAIRS. Ġ DENOTES A SPACE.

$\boldsymbol{p}_{cls}$	Prompt	Interpretability
[1,0]	$\boldsymbol{MP_s}$	[ĠAllah, Ġsuppressing, Elsewhere, ĠCath, ĠMY, ĠPI]
	$\boldsymbol{MP_h}$	[Ky, ACK, emb, gyn, ĠCohn, ĠCategory]
[0,1]	$\boldsymbol{MP_s}$	[ĠAllah, Ġsuppressing, Elsewhere, ĠCath, ĠMY, Ġpokemon]
	$\boldsymbol{MP_h}$	[Ky, ACK, emb, gyn, ĠCohn, ĠQuebec]

Prompt-CC (hard) perform better than Prompt-CC (soft) for most image pairs. Besides, as we can see, our method can ignore irrelevant changes and accurately distinguish whether changes have occurred in a complex scene. For example, the model can ignore illumination changes and small disruptors to determine that there are no changes of interest in the sixth image pairs. Furthermore, for the changed image pairs, our method can recognize changed objects of interest and generate grammatically correct sentences describing the attributes of the changed objects, changing process (i.e., appear, removed, and built), and the relationships between changed objects and surrounding objects. For example, for the second image pair, the model can determine that the woods of the pre-phase image are removed and the roads of the post-phase image are built in the bareland. For the third image pairs, the model knows the positional relationship between the new houses and the unchanged road. For the ninth image pair, the model can realize that the changed region is in the lower-right corner of the scene.

G. Advantage of the decoupling paradigm

Unlike the previous methods that process unchanged and changed image pairs in a coupled way, our new paradigm decouples the RSICC task, which fully exploits the characteristics of the task. The decoupling paradigm allows us to improve the image-level classifier and feature-level encoder

TABLE VII

THE STUDY ON THE ADVANTAGE OF THE DECOUPLING PARADIGM. WE REPORT THE PERFORMANCE IMPROVEMENT WHEN WE IMPROVE THE ACCURACY OF THE IMAGE-LEVEL CLASSIFIER AND THE FEATURE-LEVEL ENCODER SEPARATELY. AS DESCRIBED IN [6], WE DO NOT REPORT CIDER-D SCORE FOR UNCHANGED IMAGE PAIRS OF THE TEST SET. BESIDES, "W." REFERS TO "WITH" AND "W/O." REFERS TO "WITHOUT".

Evaluation Scope	Method	Acc_{cls}	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_m^*
Unchanged image pairs	Classification (*w. Pre*)	97.45%	97.09	96.65	96.43	96.23	76.04	97.83	–	–
	Classification (*w. GT*)	100%	100.00	100.00	100.00	100.00	100.00	100.00	–	–
	Improvement (Δ)	+2.55%	+2.91	+3.35	+3.57	+3.77	+23.96	+2.17	–	–
	Encoder-1 (*w/o. Dif*)	97.45%	97.10	96.67	96.45	96.25	76.22	97.83	–	–
	Encoder-2 (*w. Dif*)	97.45%	97.09	96.65	96.43	96.24	76.05	97.84	–	–
	Improvement (Δ)	–	−0.01	−0.02	−0.02	−0.01	−0.17	+0.01	–	–
Changed image pairs	Classification (*w. Pre*)	85.64%	73.13	59.57	48.12	38.83	24.05	49.58	57.34	42.45
	Classification (*w. GT*)	100%	76.95	63.86	51.71	41.62	26.31	54.05	65.12	46.78
	Improvement (Δ)	+14.36%	+3.82	+4.29	+3.59	+2.79	+2.26	+4.47	+7.78	+4.33
	Encoder-1 (*w/o. Dif*)	85.64%	72.43	58.49	46.22	36.36	23.53	48.61	50.34	39.71
	Encoder-2 (*w. Dif*)	85.64%	73.13	59.57	48.12	38.83	24.06	49.58	57.34	42.45
	Improvement (Δ)	–	+0.70	+1.08	+1.90	+2.47	+0.53	+0.97	+7.00	+2.74
All image pairs	Classification (*w. Pre*)	91.55%	83.66	75.73	69.10	63.54	38.82	73.72	136.44	78.13
	Classification (*w. GT*)	100%	86.49	78.86	71.80	65.74	40.89	77.04	143.81	81.87
	Improvement (Δ)	+8.45%	+2.83	+3.13	+2.70	+2.20	+2.07	+3.32	+7.37	+3.74
	Encoder-1 (*w/o. Dif*)	91.55%	83.12	74.93	67.83	61.94	38.28	73.23	133.08	76.63
	Encoder-2 (*w. Dif*)	91.55%	83.66	75.73	69.10	63.54	38.82	73.72	136.44	78.13
	Improvement (Δ)	–	+0.54	+0.80	+1.27	+1.60	+0.54	+0.49	+3.36	+1.50

independently, which is helpful to concentrate on improving the captioning of changed image pairs and unchanged image pairs. To demonstrate this significant advantage of our decoupling paradigm, we conduct an additional experiment. Tab. VII reports the improvement of model performance when the image-level classification accuracy increases or the feature-level encoder is improved.

1) Image-level classifier: The ground-truth classification labels are used for prompt generation during the training phase, while the predicted results of the classifier are used to generate prompts during the inference phase. To verify the impact of improving the accuracy of the classifier, we have tried to use the ground-truth classification labels to perform the multi-prompt generation during the inference phase, which is equivalent to improving the accuracy of the classifier to 100%. The comparison results are shown in Tab. VII. We can observe that the model performance shows consistent improvement for both changed and unchanged image pairs when the classification accuracy increases. Thus, improving the classifier will directly affect the overall model performance for two kinds of image pairs.

Besides, when the classification accuracy increases, compared with the changed images, unchanged image captioning gains more improvement. As we can see, when accuracy increases by 2.55% for unchanged images, BLEU-4 increases by 3.77, while a 14.36% increase in accuracy of changed images brings only +2.79 on BLEU-4. It is an easy-to-understand phenomenon. For changed image pairs, it is necessary to further recognize the changed objects, which is more complicated and depends on the feature-level encoder.

2) Feature-level encoder: As mentioned above, the feature-level encoder is used to determine what changes have occurred. To explore the detailed impact of improving the feature-level encoder, we have tried applying two different encoders based on the same image-level classifier. The experimental results are reported in Tab. VII. Encoder-1 (*w/o. Dif*) does not utilize the differencing features and the TDTr, while the structure of Encoder-2 (*w. Dif*) is the same as Fig. 5. We observe that improving the feature-level encoder mainly leads to larger performance improvement for changed image pairs but has little effect on the captioning accuracy of unchanged image pairs. It matches our intuition that describing unchanged images does not involve determining the second issue of the decoupling, i.e., what changes have occurred.

H. Parametric Experiments and Analysis

1) Vision Transformer: The Vision Transformer is responsible for processing the input image pair to extract bi-temporal visual features, which will be used in subsequent modules. Its performance will affect the overall captioning performance of the model. In Tab. VIII, we report the performance of the models adopting different Vision Transformers, including two ImageNet-based Vision Transformers and three CLIP-based Vision Transformers. Compared with the Vision Transformer trained on the ImageNet dataset, the Vision Transformer of CLIP performs better. It is attributed to the fact that CLIP is trained on a larger image-text dataset. Besides, the model performs best when using ViT-L/14 with more parameters.

2) Depth of the network: The feature-level encoder contains the STr and the TDTr. The number of Transformer layers is a critical hyperparameter. We have explored diverse configurations of Transformers with varying encoding layer counts in both STr and TDTr. The corresponding captioning

TABLE VIII
THE PERFORMANCE OF THE MODELS USING DIFFERENT VISION TRANSFORMERS, INCLUDING TWO IMAGENET-BASED VISION TRANSFORMERS AND THREE CLIP-BASED VISION TRANSFORMERS. THE EXPERIMENTS ARE BASED ON PROMPT-CC WITH THE HARD CLASS-SPECIFIC PROMPT.

Vision Transformer		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	S_m^*
ImageNet	ViT-B/32	83.05	74.91	67.96	62.17	38.56	73.49	134.95	77.29
	ViT-B/16	83.51	75.41	68.52	62.70	38.79	73.75	135.52	77.69
CLIP	ViT-B/32	83.66	75.73	69.10	63.54	38.82	73.72	136.44	78.13
	ViT-B/16	83.60	75.93	69.15	63.34	38.59	73.62	135.50	77.76
	ViT-L/14	**84.30**	**76.61**	**69.82**	**64.01**	**39.12**	**74.03**	**137.46**	**78.66**

TABLE IX
THE PERFORMANCE OF THE MODELS EMPLOYING TRANSFORMERS WITH DIFFERENT NUMBERS OF ENCODING LAYERS IN THE STR AND TDTR. THE EXPERIMENTS ARE BASED ON PROMPT-CC WITH THE HARD CLASS-SPECIFIC PROMPT. D.STR DENOTES THE DEPTH OF THE STR AND D.TDTR IS THE DEPTH OF THE TDTR.

D.STr	D.TDTr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE$_L$	CIDEr-D	S_m^*
1	2	83.43	75.46	68.60	62.77	38.78	73.72	136.08	77.83
1	3	83.18	75.00	68.20	62.61	38.41	73.19	134.28	77.12
2	2	83.19	75.11	68.26	62.46	38.90	73.71	136.83	77.98
2	3	83.66	75.73	69.10	63.54	38.82	73.72	136.44	78.13
3	2	83.55	75.47	68.75	63.19	39.00	73.92	137.02	78.28
3	3	83.49	75.53	68.71	63.02	38.29	73.13	132.80	76.81

TABLE X
EFFECT OF THE MULTI-PROMPT LEARNING STRATEGY EMPLOYING A SET OF UNIFIED PROMPTS OF DIFFERENT LENGTHS (I.E. THE NUMBER OF LEARNABLE EMBEDDINGS). THE EXPERIMENTS ARE BASED ON PROMPT-CC WITH THE HARD CLASS-SPECIFIC PROMPT.

Length of P_U	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE$_L$	CIDEr-D	S_m^*
1	83.05	75.41	68.72	63.09	38.22	73.17	133.66	77.04
5	83.66	75.73	69.10	63.54	38.82	73.72	136.44	78.13
9	83.22	75.65	68.76	62.90	38.78	73.46	135.86	77.75
15	83.65	75.94	69.03	63.17	38.64	73.69	135.08	77.65

performance is presented in Tab. IX, where D.STr denotes the depth of the STr and D.TDTr is the depth of the TDTr. Finally, we set D.STr and D.TDTr to 2 and 3 in our model.

3) Length of the prompt: The multi-prompt $MP_i(i = s, h)$ consists of a set of unified prompts $P_U \in \mathbb{R}^{K \times d_T}$ and class-specific prompt $P_{C_i} \in \mathbb{R}^{1 \times d_T}(i = s, h)$. As an important hyper-parameter, K denotes the number of learnable embeddings of P_U. We have conducted parametric experiments on the value of K, as shown in Tab. X. We can observe that the model performs best when five learnable embeddings are used in P_U. More prompt embeddings do not lead to further performance gain. Finally, we set K to 5 in our other experiments.

V. DISCUSSION AND FUTURE WORK

This paper proposes a decoupling paradigm that decouples the RSICC task into two issues: whether and what changes have occurred. It enables researchers to concentrate on the change captioning of changed and unchanged image pairs distinctly. Furthermore, we incorporate our proposed multi-prompt learning strategy and LLMs into the RSICC task and validate their effectiveness. Although extensive experiments show the SOTA performance and advantage of our method, several challenges remain to be further addressed:

- We utilize an image-level classifier and a feature-level encoder to address two decoupling issues. To improve the captioning accuracy, future efforts can focus on enhancing the classifier and feature-level encoder by exploring diverse bi-temporal fusion strategies, multi-scale feature aggregation strategies, and attention-based modules.
- The multi-prompt learning strategy utilizes the classification outcomes to prompt an LLM to understand whether changes have occurred and generate captions. Future research can be devoted to devising alternate strategies that efficiently utilize the decoupling result and exploit the potential of pre-trained LLMs for the RSICC task.
- The learnable prompt embeddings are iteratively optimized in a continuous space to align with the GPT-2. If we employ another pre-trained LLM, we should learn different prompts. In other words, the prompts learned for GPT-2 might not be optimal for other LLMs.
- Prompt learning and LLMs have garnered extensive scholarly interest recently. Our work has demonstrated their effectiveness within the RSICC task. Exploring their applicability in other RS tasks is an exciting avenue. With the recent emergence of diverse LLMs, we believe that LLMs will attract broader attention in the RS community

in the forthcoming years. We aspire for our paper to inspire future advancements in RS research.

VI. Conclusion

In this paper, unlike the previous methods that process unchanged and changed image pairs in a coupled way, we propose a decoupling paradigm to decouple the RSICC task into two issues, i.e., whether and what changes have occurred. Furthermore, an image-level classifier and a feature-level encoder are proposed to address these two issues. For caption generation, we introduce prompt learning and LLMs into the RSICC task. A multi-prompt learning strategy is proposed to prompt a pre-trained LLM as the caption generator to understand whether changes exist in the image pairs and generate captions. Compared with previous methods, our method can leverage the powerful language abilities of the pre-trained LLM and it is free of retraining a language model as the caption decoder. Experiments demonstrate that our method can achieve SOTA performance and our decoupling paradigm is more promising than the previous paradigm for the RSICC task.

References

[1] J. Deng, K. Wang, Y. Deng, and G. Qi, "Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data," *International Journal of Remote Sensing*, vol. 29, no. 16, pp. 4823–4838, 2008.

[2] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.

[3] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with landsat," in *LARS symposia*, 1980, p. 385.

[4] S. Chouaf, G. Hoxha, Y. Smara, and F. Melgani, "Captioning changes in bi-temporal remote sensing images," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 2891–2894.

[5] G. Hoxha, S. Chouaf, F. Melgani, and Y. Smara, "Change captioning: A new paradigm for multitemporal remote sensing image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2022.

[6] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[7] C. Liu, J. Yang, Z. Qi, Z. Zou, and Z. Shi, "Progressive scale-aware network for remote sensing image change captioning," *arXiv preprint arXiv:2303.00355*, 2023.

[8] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[9] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[10] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 228–239, 2022.

[11] T. Lei, J. Wang, H. Ning, X. Wang, D. Xue, Q. Wang, and A. K. Nandi, "Difference enhancement and spatial–spectral nonlocal network for change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[12] F. Rahman, B. Vasu, J. Van Cor, J. Kerekes, and A. Savakis, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2018, pp. 958–962.

[13] B. Hou, Q. Liu, H. Wang, and Y. Wang, "From w-net to cdgan: Bitemporal change detection via deep learning techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1790–1802, 2019.

[14] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[15] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2020.

[16] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection," *Remote Sensing*, vol. 12, no. 3, p. 484, 2020.

[17] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[18] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[19] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[20] Q. Li, R. Zhong, X. Du, and Y. Du, "Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[21] M. Liu, Q. Shi, Z. Chai, and J. Li, "Pa-former: Learning prior-aware transformer for remote sensing building change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[22] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7296–7307, 2021.

[23] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in vhr images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3677–3693, 2019.

[24] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9976–9992, 2019.

[25] C. Ren, X. Wang, J. Gao, X. Zhou, and H. Chen, "Unsupervised change detection in satellite images with generative adversarial network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 047–10 061, 2021.

[26] X. Tang, H. Zhang, L. Mou, F. Liu, X. Zhang, X. X. Zhu, and L. Jiao, "An unsupervised remote sensing change detection method based on multiscale graph convolutional network and metric learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[27] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," *Advances in neural information processing systems*, vol. 27, 2014.

[28] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1274–1278, 2019.

[29] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, June 2017.

[30] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International conference on computer, information and telecommunication systems (Cits)*. IEEE, 2016, pp. 1–5.

[31] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2018.

[32] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word-sentence framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[33] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[34] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, "A multi-level attention model for remote sensing image captions," *Remote Sensing*, vol. 12, no. 6, p. 939, 2020.

[35] X. Ma, R. Zhao, and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[36] S. Wu, X. Zhang, X. Wang, C. Li, and L. Jiao, "Scene attention mechanism for remote sensing image caption generation," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[37] Y. Li, X. Zhang, J. Gu, C. Li, X. Wang, X. Tang, and L. Jiao, "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[38] Y. Wang, W. Zhang, Z. Zhang, X. Gao, and X. Sun, "Multiscale multiinteraction network for remote sensing image captioning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2154–2165, 2022.

[39] C. Liu, R. Zhao, and Z. Shi, "Remote sensing image captioning based on multi-layer aggregated transformer," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–1, 2022.

[40] S. Zhuang, P. Wang, G. Wang, D. Wang, J. Chen, and F. Gao, "Improving remote sensing image captioning by combining grid features and transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[41] H. Kandala, S. Saha, B. Banerjee, and X. X. Zhu, "Exploring transformer and multi-label classification for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–1, 2022.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[45] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[47] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 3, pp. 436–440, 2021.

[48] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sensing*, vol. 11, no. 6, p. 612, 2019.

[49] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[50] Z. Chen, J. Wang, A. Ma, and Y. Zhong, "Typeformer: Multiscale transformer with type controller for remote sensing image caption," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[51] Z. Ren, S. Gou, Z. Guo, S. Mao, and R. Li, "A mask-guided transformer network with topic token for remote sensing image captioning," *Remote Sensing*, vol. 14, no. 12, p. 2939, 2022.

[52] Y. Qiu, S. Yamamoto, K. Nakashima, R. Suzuki, K. Iwata, H. Kataoka, and Y. Satoh, "Describing and localizing multiple changes with transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1951–1960.

[53] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[54] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[55] T. Schick and H. Schütze, "Exploiting cloze questions for few shot text classification and natural language inference," *arXiv preprint arXiv:2001.07676*, 2020.

[56] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing nlp," *arXiv preprint arXiv:1908.07125*, 2019.

[57] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[58] T. Vu, B. Lester, N. Constant, R. Al-Rfou, and D. Cer, "Spot: Better frozen model adaptation through soft prompt transfer," *arXiv preprint arXiv:2110.07904*, 2021.

[59] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.

[60] Z. Tan, X. Zhang, S. Wang, and Y. Liu, "Msp: Multi-stage prompting for making pre-trained language models better translators," *arXiv preprint arXiv:2110.06609*, 2021.

[61] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[62] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[63] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 816–16 825.

[64] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021.

[65] M. Sollami and A. Jain, "Multimodal conditionality for natural language generation," *arXiv preprint arXiv:2109.01229*, 2021.

[66] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[67] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, "An empirical study of gpt-3 for few-shot knowledge-based vqa," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3081–3089.

[68] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer, 2022, pp. 709–727.

[69] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," *arXiv preprint arXiv:2210.03117*, 2022.

[70] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Unified vision and language prompt learning," *arXiv preprint arXiv:2210.07225*, 2022.

[71] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T.-S. Chua, and M. Sun, "Cpt: Colorful prompt tuning for pre-trained vision-language models," *arXiv preprint arXiv:2109.11797*, 2021.

[72] H. Yang, J. Lin, A. Yang, P. Wang, C. Zhou, and H. Yang, "Prompt tuning for generative multimodal pretrained models," *arXiv preprint arXiv:2208.02532*, 2022.

[73] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4953–4963.

[74] M. Li, L. Chen, Y. Duan, Z. Hu, J. Feng, J. Zhou, and J. Lu, "Bridge-prompt: Towards ordinal action understanding in instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 880–19 889.

[75] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *arXiv preprint arXiv:2306.16269*, 2023.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[77] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[78] Z. Yu, "A survey on clip-guided vision-language tasks," *Highlights in Science, Engineering and Technology*, vol. 12, pp. 153–159, 2022.

[79] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia *et al.*, "Transformers in remote sensing: A survey," *arXiv preprint arXiv:2209.01206*, 2022.

[80] Z. Qi, H. Chen, C. Liu, Z. Shi, and Z. Zou, "Implicit ray-transformers for multi-view remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

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

[82] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[83] T. OpenAI, "Chatgpt: Optimizing language models for dialogue," *OpenAI*, 2022.

[84] Papineni, Kishore, Roukos, Salim, Ward, Todd, Zhu, and Wei-Jing, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135
[85] Lin and C. Yew, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.
[86] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, ser. StatMT '07. USA: Association for Computational Linguistics, 2007, p. 228–231.
[87] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.
[88] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.
[89] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[90] D. H. Park, T. Darrell, and A. Rohrbach, "Robust change captioning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4623–4632.

Chenyang Liu received his B.S. degree from the Image Processing Center, School of Astronautics, Beihang University in 2021. He is currently working towards the Ph.D. degree in the Image Processing Center, School of Astronautics, Beihang University.

His research interests include machine learning, computer vision and multimodal learning.

Rui Zhao received the B.S. and M.S. degrees from the Image Processing Center, School of Astronautics, Beihang University, Beijing, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

His research interests include computer vision, deep learning, and related problems in remote sensing and video games.

Jianqi Chen received his B.S. degree from the Image Processing Center, School of Astronautics, Beihang University in 2021. He is currently pursuing his M.S. degree in the Image Processing Center, School of Astronautics, Beihang University.

His research interests include deep learning, object detection and artificial intelligence safety.

Zipeng Qi received the B.S. degree from the Hebei University of Technology, Tianjin, China, in 2018. He is currently pursuing the Ph.D. degree with the Image Processing Center, School of Astronautics, Beihang University, Beijing, China.

His research interests include image processing, deep learning, and pattern recognition.

Zhengxia Zou received his BS degree and his Ph.D. degree from Beihang University in 2013 and 2018. He is currently a Professor at the School of Astronautics, Beihang University. During 2018-2021, he was a postdoc research fellow at the University of Michigan, Ann Arbor. His research interests include computer vision and related problems in remote sensing. He has published more than 20 peer-reviewed papers in top-tier journals and conferences, including TPAMI, TIP, TGRS, CVPR, ICCV, AAAI. His research was featured in more than 30 global tech media and was adopted by a number of application platforms with over 50 million users worldwide. His personal website is https://zhengxiazou.github.io/.

Zhenwei Shi (Senior Member, IEEE) is currently a Professor and Dean of the Image Processing Center, School of Astronautics, Beihang University. He has authored or co-authored over 200 scientific articles in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and the IEEE International Conference on Computer Vision (ICCV). His current research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Prof. Shi serves as an Editor for IEEE Transactions on Geoscience and Remote Sensing, Pattern Recognition, ISPRS Journal of Photogrammetry and Remote Sensing, Infrared Physics and Technology, etc. His personal website is http://levir.buaa.edu.cn/.