scientific reports

Check for updates

OPEN

Feature refinement and rethinking attention for remote sensing image captioning

Yunpeng Li[1,2], Chengjin Tao[1,2,5], Meng Liu[1,2,5], Xiangrong Zhang[3], Guanchun Wang[3], Tianyang Zhang[3], Dong Zhao[1,2]✉ & Dabao Wang[4]

Effectively recognizing different regions of interest with attention mechanisms plays an important role in remote sensing image captioning task. However, these attention-driven models implicitly hypothesize that the focused region information is correct, which is too restrictive. Furthermore, the visual feature extractors will fail when facing weak correlation between objects. To address these issues, we propose a feature refinement and rethinking attention framework. Specifically, we firstly construct a feature refinement module by interacting grid-level features using refinement gate. It is noticeable that the irrelevant visual features from remote sensing images are weakened. Moreover, different from one attentive vector for inferring one word, the rethinking attention with rethinking LSTM layer is developed to spontaneously focus on different regions, when rethinking confidence is desirable. Thus, there are more than one region for predicting one word. Besides, the confidence rectification strategy is adopted to model rethinking attention for learn strongly discriminative contextual representation. We validate the designed framework on four datasets (i.e., NWPU-Captions, RSICD, UCM-Captions and Sydney-Captions). Extensive experiments show that our approach have superior performance and achieved significant improvements on the NWPU-Captions dataset.

Keywords Remote sensing image captioning, Visual perception, Feature refinement, Rethinking attention mechanism, Vision-language

Remote sensing image captioning (RSIC) task is a crucial task in remote sensing image (RSI) field, which contributes to a variety of applications, such as conservancy construction[1], urban planning[2], disaster assessment[3] and battlefield environment monitoring[4]. Specifically, RSIC task aims to understand the content[5] of a RSI and generate a comprehensive and appropriate natural language like ground-truth sentences, which is also a challenging task. There two examples from RSIC datasets in Fig. 1. In recent years, the prevailing methodologies for generating textual descriptions have centered around attention-driven design frameworks. These frameworks typically involve a feature extractor that encodes the input remote sensing imagery and a language decoder, which employs an attention mechanism[6–8] to generate a coherent sequence of words. However, these methods often come with low performance, due to worthless visual features, simple feature fusion and semantic comprehension. Hence, designing a RSIC network with powerful discrimination to solve these problems is crucial.

The most existing methods have been proposed with powerful convolutional neural network (CNN) with long-short term memory (LSTM) or Transformer, namely CNN-LSTM or CNN-Transformer framework. For example, Shi et al.[9] generated captions with a CNN-LSTM framework, which not only "describe" visual contents[10] but also "read" the label information with fully convolutional network. However, the encoder only transformed the input RSI into global feature, which was then decoded for generating all words in predicted sentence. Generally, RSIs usually involve objects with different scales[11,12], resulting in varied scene range and target size. Therefore, there are complex objects in a RSI[13–15], the global feature is not sufficient to represent contents in the RSI. The appeared attention mechanism can face this issue and benefit RSIC task, which selectively process visual features and provide more valuable information at a decoding time.

[1]The Jiangsu Province Engineering Research Center of Integrated Circuit Reliability Technology and Testing System, Wuxi University, Wuxi 214105, China. [2]The Jiangsu Province Engineering Research Center of Photonic Devices and System Integration for Communication Sensing Convergence, Wuxi University, Wuxi 214105, China. [3]Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China. [4]Remote Sensing Satellite Department, China Academy of Space Technology, Beijing 100094, China. [5]These authors contributed equally: Chengjin Tao and Meng Liu. ✉email: dzhao@cwxu.edu.cn

(a)

Ground-Truth
(1): Three small planes parked in a line on the airport and a big plane behind them.
(2): There are four aircraft on the open ground, the largest of which is three times large as the smallest one.
(3): There are many planes of different sizes in a clearing.
(4): Four planes are parked on the runway.
(5): Four planes of different sizes were on the marked ground.

(b)

Ground-Truth
(1): The island has dense vegetation and a beach and the waters around the island is light blue.
(2): The shape of the island is nearly circular.
(3): The green island, which is shaped like a crescent moon, is surrounded by pale blue water.
(4): Many trees are on the island.
(5): The green island full of trees is on a blue sea.

Figure 1. The two examples are from the NWPU-Captions[16] dataset, each image has human-marked five sentences.

The novel RSIC model with spatial attention (i.e., hard attention and soft attention) was exploited by Lu et al.[17], in which regions of RSI were given different weights. The decoder could bridge the correspondence between each attend vector and one predicted word. Scene attention[18] extracted more scene information before guiding designed attention, which aimed to discover a sequence of key regions and describes them using coherent words. Yuan et al.[19] focused more on the location of objects and the underutilized high-level information. Their work proved that the image features extracted by each layer in CNN contained different information. The shallower layers is for low-level visual information such as color, edges and corners. While deeper layers can extract high-level semantic information such as the category. Therefore, Zhang et al.[20] represented the high-level features as semantic attributes[21,22] of RSIs and proposed attribute attention. These semantic attributes[23,24] not only represent important visual information, but also important component of one sentence. The success of the label attention mechanism[25] also demonstrated the importance of semantic attributes. In addition, the generated sentences also include high-level understanding of RSIs, the proposed summarization driven RSIC model[26], multi-level attention mechanism[27], word-sentence framework[28] and recurrent attention with semantic gate framework[29] explored how to generate some corresponding nouns, prepositions and relational words based on some generated words. To achieve more effective feature representation, Shen et al.[30] combined pre-trained CNN for feature extraction, the adopted Transformer decoder made an impact on semantic inferring. Besides, some approaches[31–33] focused on object-level region features to augment RSIC model perfomance. For example, Luo et al.[31] utilized segmentation network for structured features as object-level region information and combined with spatial features. Furthermore, current RSIC benchmark datasets are deficient, Cheng et al.[16] tackled this limitation by presenting a more challenging NWPU-Captions dataset. Following that, some model[34,35] was proposed to capture and utilize multi-scale features, which conducted on the NWPU-Captions dataset and achieved superior performance. In a global-local captioning model, Wang et al.[36] obtained global-local visual feature representation for RSIC task. Recently, Du et al.[37] designed CNN-Transformer framework to recognise semantic content for the NWPU-Captions dataset.

Despite existing RSIC methods have achieved successful performance, there are still some limitations about the learning of effective visual features. (1) The feature representations contain worthless information. If terrain information contained in the scene (such as “road”, “car”, “building”, etc.) appears in an inconspicuous corner, these weak features would be hard to be represented by current feature extractor, which may still contain some redundant information in obtained visual features. We argue that it is helpful to extract potential relationships between objects and filter out invalid visual information. (2) The visual information in decoder will accumulate over time, but the early decoding requires more visual information about RSIs, one attention step provides less information at beginning time. Obviously, current attention module[31,37] is a one-to-one mapping from attentive map to predicted word, which cannot be absolute. (3) There may be no content in visual features that meets the decoder, but the attention module still returns some vectors, which is the weighted average of candidate features. The forced decoding process is completely irrelevant independent of expected reasoning at current time. That is, the attention results are not expected, the decoder may be misled with imprecise guidance.

In this paper, a RSIC model based on feature refinement and rethinking attention has been developed. The flowchart is shown in Fig. 2. Our feature refinement is based on CNN visual features for refined visual features. Generally, the feature extractors[25,30] directly provided the captured visual representations for sentence prediction. This risks unnecessary or even misleading visual information. However, in our feature refinement module, the refinement gate directs the focus towards visually relevant information to model the relationships between pixel-level features. And rethinking attention is used to achieve adaptive alignment from single or multiple regions to one word prediction. In other words, the decoder achieves the cross-modal interaction by analyzing more than one contextual vector at each predicting word position. When predicting the next word, the rethinking attention will face new rethinking attention steps. The termination of the rethinking attention mechanism depends on the learnable confidence and set max steps. Note that the confidence rectification strategy serves as an effective supervision to guarantee the optimal rethinking steps.

The major contributions can be summarized as follows:

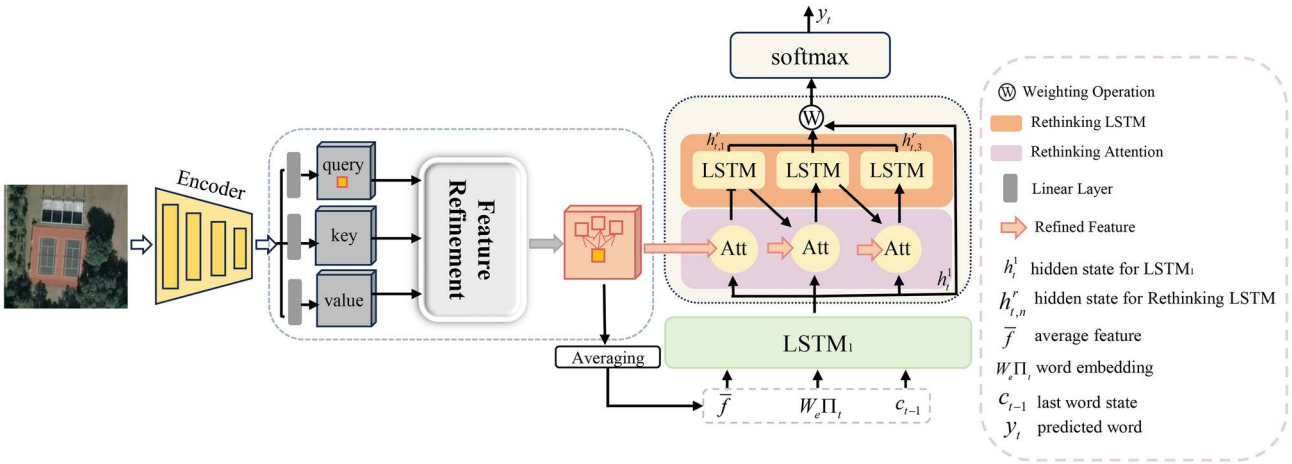

Figure 2. The overview of our proposed RSIC model. It consists of the visual feature refinement module and rethinking attention. The input RSI is from the UCM-Captions[38] dataset.

(1) A novel feature refinement and rethinking attention RSIC model is proposed, which effectively addresses the limitations of unnecessary visual information and one-step attention-based approaches. Our model can preserve meaningful visual details at each decoding stage, and obtain state-of-the-art performance on four public datasets (NWPU-Captions, RSICD, UCM-Captions and Sydney-Captions).
(2) A feature refinement module is introduced into our encoder, which can create relation awareness with refinement gate based on CNN visual features. It has more information exchanges between objects to improve discrimination ability and removes redundant information.
(3) Unlike existing attention-based RSIC models, our rethinking attention derives one or multiple attention steps for inferring each word, the attention steps are controlled by the learnable confidence at different decoding stages. Specifically, the multi-head attention or spatial attention can be is adopted in the rethinking attention for enhancing cross-modal inference.
(4) For constraining the rethinking attention, the adopted confidence rectification strategy optimizes the rethinking attention in the training stage, which is an alignment supervision strategy to enforce interaction between vision and semantics.

Related work

RSIC task is a multi-modal task involving RSI encoding and natural language generation. The attention-driven RSIC models can make comprehensive cross-modal interpretation between vision and semantic information. Prevailing methodologies can be categorized by decoder structure (i.e., LSTM, Transformer). In this section, we will concisely review these approaches.

LSTM-based methods

In LSTM-based RSIC methods, the encoder (i.e., VGG and ResNet) is responsible for extracting features from RSIs, which can be enhanced for more comprehensive representations. The LSTM-based decoder converts visual features into natural language. The multi-modal method was proposed by Qu et al.[38], a global feature obtained by CNN encoder could be decoded by multi-modal layer for predicting word step by step. Category-aware features from final layer of feature extractor were considered by Zhang et al.[39], one LSTM preserved the high-level features to improve the sentence quality. The category-aware feature contains more semantic clues than a global feature. However, inferring the global feature and category-aware feature for a sentence is not reasonable.

Later, the welcoming attention-based RSIC architectures have witnessed the performance enhancement. Lu et al.[17] boosted the performance by using spatial attention. The attention-based methods can realize one attentive visual region for a word, which facilitates powerful cross-modal interaction. Wu et al.[18] proposed a scene attention mechanism, which utilized the fusion feature of scene features and state vector from LSTM to guide attention module for predicting the next word. To improve the semantic-awareness of a RSI, the attribute attention[20] and label-attention mechanism[25] exploited high-level semantic information to locate different concerned regions. In details, a pre-trained classification network could obtain label information for visual features. Motivated by semantic-level features, Li et al.[27] proposed multi-level attention that mainly exploited relationships between generated words and visual features to sequentially describe objects in RSI. Recurrent attention mechanism[29] was proposed for exploiting both semantic context around each candidate region and visual context at the certain linguistic state, which introduced previous attentive results to select the discriminative contextual features for each word prediction. Zhang et al.[40] adopted a global visual feature-guided attention in encoder for enhancing feature associated with the object itself, a linguistic state-guided attention specifically provided textual features in each decoding time step. There would be some inaccurate words in generated sentence. Thus, the summarization-driven method[26] summarized five ground-truth captions as semantic features, and adaptive weighting integrated standard CNN-LSTM semantic inference for improving generalization capability. Li et

al.[21] designed a trainable semantic concept module, the matched semantic concepts were beneficial for learning consensus-aware semantic knowledge in cross-modal features[22]. Besides, Zhao et al.[31] used segmentation branch to provide object-level structure region and the novel structure attention guided to predict the objects. The other region features based on region proposal networks[33] were explored like Faster-RCNN encoder. It provides new framework for the RSIC task. Instead of only employing visual information, Lu et al.[41] proposed a sound active attention framework which effectively utilized sound and visual cues for more accurate sentences. On the other hand, most existing methods optimize for converting into longer sentences with cross entropy (CE). The truncating cross entropy[42] loss with CNN-LSTM structure had achieved great success by solving the interaction between training stage and annotations. Similarly, Chavhan et al.[43] proposed a dual supervised training policy to improve performance for RSIC task. These improved losses ameliorate the robustness for RSIC task. Recently, the biggest RSIC dataset is published for further research, Cheng et al.[16] also adopted the multi-level and contextual attention network (MLCA-Net) for multi-scale feature extraction and latent context features. Following that, Huang et al.[34] proposed a multi-scale contextual information aggregation network (MC-Net) could dynamically focus on multi-scale information for the NWPU-Captions dataset, visual-text alignment LSTM was helpful for exploring deeper semantic information of multi-scale features. Yang et al.[35] considered the effectiveness of hierarchical features and explored efficient utilization of visual texture and semantic features.

Essentially, attention mechanisms in previous approaches are responsible for cross-modal interaction between visual feature and linguistic state. However, one attentive region generally matches one word based on fragmented visual information. While the insufficient, unnecessary or misleading visual information may be provided for word prediction. In fact, one weighted contextual vector is also hard to achieve the cross-modal interaction.

Transformer-based methods

With the application of Transformer in various fields[13,14], researchers have carried out Transformer-based architectures for RSIC task, which have achieved excellent performance than CNN-LSTM framework. These methods generally adopt an improved CNN encoder with a Transformer-based decoder ("CNN-Transformer") . As we know, Transformer has the significant advantage of using self-attention to learn long-range dependencies at the level of intra-vision or intra-text. The global-local captioning model (GLCM)[36] introduced how to make full use of the advantages of both global and local features in CNN based encoder, their Transformer decoding network contained self-attention and co-attention for complex cross-modal reasoning. Gajbhiye et al.[44] proposed a memory-guided Transformer, which jointly modeled multi-attentive features and inferred ordered words with Transformer decoder from a structural and global perspective. The model designed by Zia et al.[45] was similar with method[44], however, they considered the multi-scale features in feature extractor. Specially, "Transformer-Transformer" framework was proposed by Wang et al.[28], Transformer-encoder encoded semantic words learned through a word generator and Transformer-decoder generated description sentences, which only transferred in semantic domains. Employing Transformer encoder for a RSI is a basic necessity in RSIC task. Thus, the mask-guided Transformer network[46] took the raw RSI as input for generating the patch-level visual representations, including semantic information from topic token and patch-level visual features, which showed great potential in Transformer-based decoder. Li et al.[32] explored patch-level salient features and object-level labels from Transformer-based visual features, two parallel cross-modal attention in Transformer decoder could measure the alignment between the representations of captured patch-level region and label knowledge. Wu et al.[47] proposed swin-Transformer encoder for multi-scale visual feature with shifted window partitioning scheme, the Transformer decoder was for visual-linguistic reasoning. Based on the NWPU-Captions dataset, the plane to hierarchy (P-to-H)[37] model adopted selective search to obtain visual and semantic maps, then a deformable transformer learned multi-scale feature and performed intra-class interactive learning. Inspired by diffusion based decoder in natural image captioning, Cheng et al.[48] proposed an innovative diffusion model, in which the diffusion RSIC model with non-autoregressive decoder were fed with the discernible visual context features extracted by a refined multi-scale feature extraction.

It is important for RSIC task that how to enhance the capability of complex cross-modal reasoning through multi-level interaction. These methods attempt to form more powerful feature representation to guide the visual-textual attention process (i.e., spatial attention and self-attention).

Proposed approach

Our designed model consists of two core modules: the feature refinement module and the rethinking attention module. In the feature refinement module, CNN features are refined with multi-head attention (MHA) for interaction and deeper refinement. The rethinking attention establishes correlations among different regions, using one or more weighted region features for inferring a word at *t* time. In addition, the confidence rectification strategy contributes to control and optimize appropriate attention steps.

Firstly, the RSI is converted into a feature map F through a CNN. Each position in F represents a region and contains objects, our decoder should understand what objects are included in the RSI and the semantic relationship among objects. The calculation formula is written as follows:

$$F = CNN\,(I) \tag{1}$$

where I represents the input RSI, $CNN\,(\cdot)$ represents VGG16[49] network, and the obtained feature $F \in R^{k \times dim}$ is defined as follows:

$$F = \{f_1, f_2, ..., f_k\} \tag{2}$$

where k represents grid number in feature F, dim is feature dimension. The mean calculation $\overline{f} \in R^{dim}$ of feature F is calculated as follows:

$$\overline{f} = \frac{1}{k}\sum_{i}^{k} f_i \tag{3}$$

Visual feature refinement

Effective visual features play a critical role in RSIC task. Thus, a available feature extractor for visual feature enhancement is a potential research. Due to the rich complexity for a RSI, the interaction between visual features can provide inter-feature relation and incorporate each region's strengths. Motivated by this, the feature refinement module is designed to refine CNN visual features (as shown in Fig. 3). It adopts MHA to achieve interaction between grid-level features while suppressing redundant visual information. Firstly, F is subjected via linear layer to three independent mappings, namely Q (query), K (key), and V (value):

$$\begin{aligned} Q &= F(W_Q) \\ K &= F(W_K) \\ V &= F(W_V) \end{aligned} \tag{4}$$

where W_Q, W_K, and W_V are learnable parameters. The interaction process of the feature refinement module is to perform similarity calculation between Q and K, the similarity score $s_i \in R^{k\times dim/H}$ in $head_i$ is defined as follows:

$$s_i = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i \tag{5}$$

where V_i, Q_i and K_i represent value, query and key in $head_i$, $\sqrt{d}$ is the scaling factor, T is a symbol of matrix transposition, H is the head number in MHA. The interactive visual features $F' \in R^{k\times dim}$ are defined as follows:

$$F' = f_{MHA}(Q, K, V) \tag{6}$$

where $f_{MHA}(Q, K, V) = W_s Concat(head_1, ..., head_H)$. Note that W_s is learnable parameter, and H is set to 8 in the experiment.

The MHA layer realizes pairwise interaction of grid features, and spatial feature can be strengthened. Although there is no strong correlation between the two grid features, correlation weights will still be constructed. Thus, the weighted features still include irrelevant or even misleading information. The designed refinement gate re-measures the correlation between the output attention results and the query Q. Specifically, two parallel linear transformations with unshared parameters generate vector $x \in R^{k\times dim}$ and the refinement gate $g \in R^{k\times dim}$, respectively. The calculation process is as follows:

$$x = W_q^x Q + W_f^x F' + b^x \tag{7}$$

$$g = \sigma\left(W_q^g Q + W_f^g F' + b^g\right) \tag{8}$$

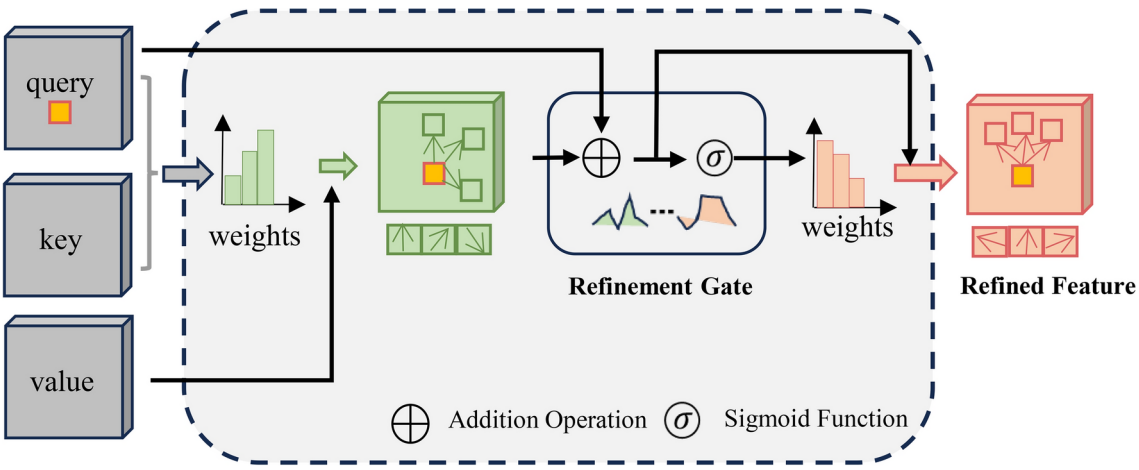

Figure 3. The visual feature refinement is leveraged to efficiently and effectively achieve refined features with refinement gate.

where $W_q^x, W_f^x, b^x, W_q^g, W_f^g$ and b^g are learnable parameters. $\sigma(\cdot)$ is sigmoid function for refinement coefficient (from 0 to 1). Then, applying g to vector x with element-wise multiplication obtains the final refinement features $Z \in R^{k \times dim}$:

$$Z = g \odot x \tag{9}$$

where $\odot$ represents element-wise multiplication, which can suppress information for small value in g and enhance the features with strong signals. And $Z = \{z_1, z_2, ..., z_k\}$ contains k vector features. The mean calculation $\overline{z} \in R^{dim}$ of feature Z is as follows:

$$\overline{z} = \frac{1}{k} \sum_{i}^{k} z_i \tag{10}$$

Rethinking visual information and propagating confidence orientation

Numerous studies have shown that applying attention mechanism can realize visual features analysis, enhancing reasoning capabilities through the interaction between visual and semantic features and extracting more accurate vision-to-language information. However, these attention mechanisms[25,34,36] only carry out the interaction or fusion for multi-modal features with one attentive map at current time, which limits the capability of description generation. In fact, visual information are variously required in different reasoning stages. For example, being able to infer some word at sentence head, the sufficient visual features should be concerned for perception of salient content in RSI. The similar phenomenon is appeared for learning relation semantics among objects. Thus, multiple attentive vectors should be generated for one word. Over time, predicted object requires precise visual information for one word (i.e., one attention vector). In some case, model should weaken attention on visual information for some non-visual words. Therefore, in this work, we seek to rethinking visual information with rethinking attention mechanism, which has greater than or equal to one attention step for predicting a word.

Standard attention

Taking the baseline model as an example[50], a brief principle is given in Fig. 4. The decoder combines two stacked LSTMs with standard attention mechanism. Note that the following symbols for 1 and 2 are used to distinguish each LSTM layer at current time. The first LSTM layer is named $LSTM_1$, the processing procedure is defined as follows:

$$\left(h_t^1, m_t^1\right) = LSTM_1\left(\left[\mathrm{W}_e \Pi_t, \bar{f}, c_{t-1}\right], \left(h_{t-1}^1, m_{t-1}^1\right)\right) \tag{11}$$

where $h_t^1 \in R^{dim}$ and $m_t^1 \in R^{dim}$ represent the hidden state and memory vector of $LSTM_1$ at current time, respectively. h_{t-1}^1 and m_{t-1}^1 represent the hidden state and memory vector of $LSTM_1$ at previous time, respectively. $\mathrm{W}_e \Pi_t$ represents embedding one-hot encoded word, $\overline{f}$ represents mean feature for CNN feature F, and $c_{t-1} \in R^{dim}$ is predicted word vector at $t-1$ time. Specifically, decoder provides a query (such as h_t^1) to calculate attention weights on feature F for attentive results. The calculation formula is defined as follows:

$$\hat{a}_t = f_{att}\left(h_t^1, F\right) \tag{12}$$

where $f_{att}(\cdot)$ represents standard attention function. Then, $\hat{a}_t \in R^{dim}$ and h_t^1 are input into the second LSTM layer, named $LSTM_2$. The calculation process is defined as follows:

$$\left(h_t^2, m_t^2\right) = LSTM_2\left(\left[\hat{a}_t, h_t^1\right], \left(h_{t-1}^2, m_{t-1}^2\right)\right) \tag{13}$$

where $h_t^2 \in R^{dim}$ and $m_t^2 \in R^{dim}$ represent the current hidden state and memory state of $LSTM_2$, respectively. and h_{t-1}^2 and m_{t-1}^2 represent the last hidden state and memory state of $LSTM_2$, respectively. The output h_t^2 of $LSTM_2$ will be used as the word state vector $c_t = h_t^2$ for word prediction. Finally, $c_t \in R^{dim}$ is passed through softmax layer to obtain probability distribution $\phi_t \in R^{W dim}$:

$$\phi_t = \mathrm{softmax}\left(W_\phi c_t + b_\phi\right) \tag{14}$$

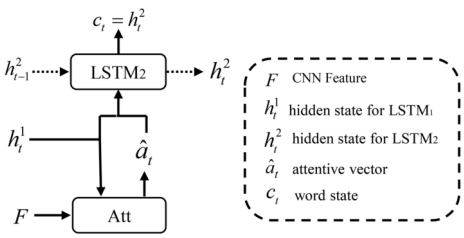

Figure 4. The standard attention structure.

where W_ϕ and b_ϕ are learnable parameters, *Wdim* is equal to the number of word in created vocabulary.

Rethinking attention

Differently, our decoder relies on a standard LSTM layer, the designed rethinking LSTM layer (named $LSTM_r$) and rethinking attention. The semantic vectors, such as $h_t^1 \in R^{dim}$ from $LSTM_1$ and $h_{t,n-1}^r \in R^{dim}$ from rethinking LSTM, is "memorized", updated, and transmitted as query for guiding rethinking attention. Our rethinking attention is shown in Fig. 5. It is possible to focus on different visual regions at *t* time, in which different attention steps vary for predicting a word. Specifically, at *t* time, $LSTM_1$ is defined as follows:

$$\left(h_t^1, m_t^1\right) = LSTM_1\left(\left[\mathrm{W}_e\Pi_t, \bar{z}, c_{t-1}\right], \left(h_{t-1}^1, m_{t-1}^1\right)\right) \tag{15}$$

where $\bar{z}$ represents mean feature for refinement feature *Z*. The *n*-th hidden state $h_{t,n}^r \in R^{dim}$ and memory vector $m_{t,n}^r \in R^{dim}$ of $LSTM_r$ are as follows:

$$h_{t,n}^r, m_{t,n}^r = LSTM_r\left(\left[\hat{a}_{t,n}, q_{t,n}\right], \left(h_{t,n-1}^r, m_{t,n-1}^r\right)\right) \tag{16}$$

where $h_{t,n-1}^r \in R^{dim}$ and $m_{t,n-1}^r \in R^{dim}$ represent $n-1$ step hidden state and memory vector in rethinking LSTM at *t* time, respectively. let $h_{t,0}^r = h_{t-1}^r$ and $m_{t,0}^r = m_{t-1}^r$. In order to construct the *n*-th query $q_{t,n}$ for the rethinking attention, we set same dimension with different attention steps at *t* time. The *n* step query $q_{t,n} \in R^{dim}$ is performed by the following equations:

$$q_{t,n} = \left[h_t^1, h_{t,n-1}^r\right] W_q + b_q \qquad n > 0 \tag{17}$$

where b_q is the learnable deviation. Then, $q_{t,n}$ drives *n* attention step, the calculation formula $\hat{a}_{t,n} \in R^{dim}$ is as follows:

$$\hat{a}_{t,n} = f_{att}\left(q_{t,n}, Z\right) \tag{18}$$

where $f_{att}(\cdot)$ can be standard attention or MHA.

Confidence rectification strategy

At *t* time, our rethinking attention can model attention steps. A learned confidence condition plays a controlling role, which can adaptively determine whether to stop the rethinking attention at current time. Once the confidence condition fails, the attention steps will be prevented. The calculation of confidence rectification is updated by:

$$p_{t,n} = \begin{cases} \sigma\left(\max\left(0, W_p h_t^1\right) + b_p\right) & n = 0 \\ \sigma\left(\max\left(0, W_p h_{t,n}^r\right) + b_p\right) & n > 0 \end{cases} \tag{19}$$

where W_p and b_p are learnable parameters, and $\sigma(\cdot)$ is the activation function sigmoid. The number of attention step for each inference stage is determined by two important factors:

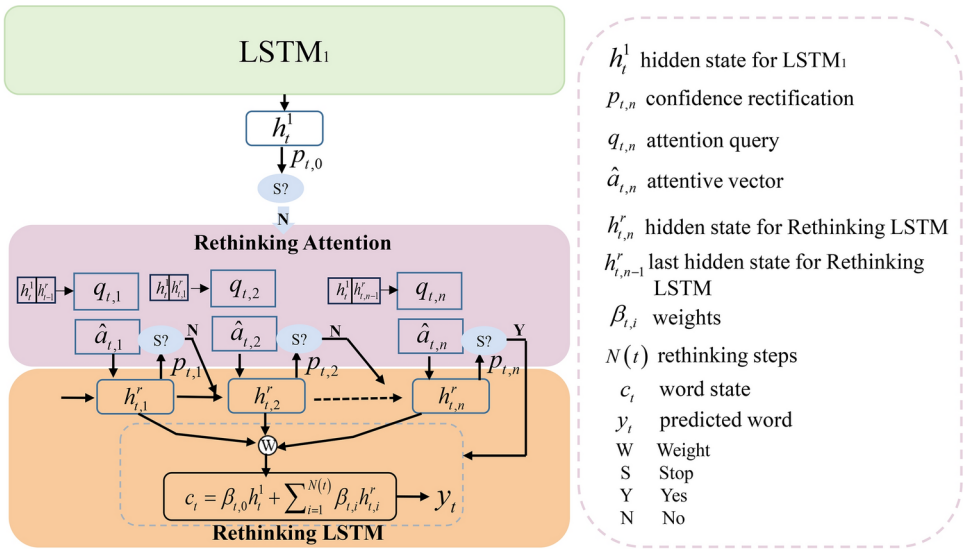

Figure 5. During the decoding stage, an improved decoder consists of the rethinking attention and rethinking LSTM.

$$N(t) = \min\left\{\mathrm{M}, \min\left\{n : \prod_{i=0}^{n}(1-p_{t,i}) < \varepsilon\right\}\right\} \tag{20}$$

where M is set to 4. ε is a threshold value, which is set to 1×10^{-4} in our experiments. Yet, rethinking LSTM is performed with $N(t)$ hidden states, these are underlying semantic information as well as h_t^1. With confidence rectification $p_{t,n}$, a confidence weight is thought of as calculating the relevance of semantic information associated with the word states at each time step, the $\beta_{t,n}$ is defined as:

$$\beta_{t,n} = \begin{cases} p_{t,0} & n = 0 \\ p_{t,n}\Pi_{i=0}^{n-1}(1-p_{t,i}) & n > 0 \end{cases} \tag{21}$$

where the maximum value of n is $N(t)$. For ensuring the sum of the weights $\beta_{t,n}$ to 1 at t time, $\beta_{t,n}$ is normalized:

$$\beta_{t,n} = \frac{\beta_{t,n}}{\sum_{i=0}^{N(t)}\beta_{t,i}} \tag{22}$$

Therefore, the final hidden state $h_t^r \in R^{dim}$ and memory vector $m_t^r \in R^{dim}$ of rethinking LSTM are computed as:

$$\begin{cases} h_t^r = \beta_{t,0}h_t^1 + \sum_{n=1}^{N(t)}\beta_{t,n}h_{t,n}^r \\ m_t^r = \beta_{t,0}m_t^1 + \sum_{n=1}^{N(t)}\beta_{t,n}m_{t,n}^r \end{cases} \tag{23}$$

where let $c_t = h_t^r$.

Training strategy

For designed rethinking attention, a loss function with confidence rectification strategy is adopted in our model. The formula is defined as follows:

$$L_t = \lambda_{\mathrm{att}}\left(N(t) + \sum_{i=0}^{N(t)}(i+1)(1-p_{t,i})\right) \tag{24}$$

where λ_{att} is a hyperparameter, which is set to 1×10^{-4}. And $(i+1)(1-p_{t,n})$ can promote larger $p_{t,n}$, learning appropriate attention steps.

The overall loss function is defined as follows:

$$loss = \frac{1}{T}\sum_{t=1}^{T}\left[-\log\left(\phi_t^{\theta}\left(y_t \,|y_{1:t-1}, Z\right)\right) + \mu L_t\right] \tag{25}$$

where T represents the length of predicted sentence, θ represents all parameters of designed model, y_t represents generated words, $y_{1:t-1}$ represents predicted words, the hyperparameter μ is set to 0.2.

Experiments and analysis

In this section, we firstly introduce four public datasets, evaluation metrics, experimental settings and compared models. Following that, we provide performance of our model and state-of-the-art approaches with analyses. Additionally, we also perform a series of visualization analyses, ablation experiments and parameter analyses.

Dataset and setting

Datasets

We conduct the experiments and evaluate our model on the widely used four datasets, which the open source benchmark datasets are available for RSIC task. Each image in four datasets is annotated manually with five different sentences.

(1) NWPU-Captions[16]: The NWPU-Captions dataset is the largest RSIC dataset published by Cheng et al.[16], which is measuring 500×500 pixels. This dataset includes a total of 45 scenes, 31,500 images.
(2) RSICD[17]: Each image in the RSICD is with the size of 224×224 pixels. The dataset contains 30 scenes, 10921 images.
(3) UCM-Captions[38]: The UCM-Captions dataset contains 2100 images of 21 types of scenes, each of which is 256×256 pixels in size.
(4) Sydney-Captions[38]: The Sydney-Captions dataset contains 613 images with 7 categories. The size of each image is 500×500 pixels.

Evaluation metrics

In order to make a fair comparison with other methods, we verify on seven automatic evaluation metrics, i.e., BLEU-n[51], METEOR[52], ROUGE_L[53] and CIDEr[54]. We pay more attention on CIDEr during experiments, since CIDEr considers the word frequency as the weight and measures the weighted cosine similarity of words in different n-grams, which can better reflect the capability on generating sentences. BLEU-n is a n-gram precision

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
SAT	0.7340	0.6120	0.5280	0.4690	0.3370	0.6010	1.1090
SM-Att	0.7390	0.6160	0.5340	0.4690	0.3380	0.5950	1.1370
Struc-Att	0.7442	0.6091	0.5193	0.4557	0.3087	0.6064	1.2270
MLCA-Net	0.7450	0.6240	0.5410	0.4780	0.3370	0.6010	1.2640
MC-Net	0.7410	0.6260	0.5440	0.4780	0.3470	0.6110	1.1590
VRTMM	0.8116	0.7033	0.6213	0.5570	0.3660	0.6845	1.5885
GLCM	0.5536	0.4228	0.3353	0.2720	0.2789	0.5042	1.2774
P-to-H	0.7571	0.6291	0.5457	0.4828	0.3187	0.5858	1.2071
DiffNet	0.7972	0.6635	0.5604	0.4793	0.3059	0.6182	1.2324
Ours	**0.8490**	**0.7620**	**0.6957**	**0.6441**	**0.4198**	**0.7468**	**1.8167**

Table 1. Comparison scores of our method and other state-of-the-art methods on the NWPU-Captions dataset. Significant values are given in bold.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
SAT	0.6707	0.5438	0.4550	0.3870	0.3203	0.5724	2.4686
SM-Att	0.6699	0.5523	0.4703	0.4068	0.3255	0.5802	2.5738
Struc-Att	0.7016	0.5614	0.4648	0.3934	0.3291	0.5706	1.7031
MLCA-Net	0.7570	0.6340	0.5390	0.4610	0.3510	0.6460	2.3560
MC-Net	0.7280	0.6060	0.5110	0.4330	0.3600	0.6410	2.4540
VRTMM	0.7813	0.6721	0.5645	0.5123	0.3737	0.6713	2.7150
GLCM	0.7767	0.6492	0.5642	0.4937	0.3627	0.6769	2.5491
P-to-H	0.7581	0.6416	0.5585	0.4923	0.3550	0.6523	2.5814
DiffNet	**0.7858**	**0.6807**	**0.5928**	**0.5202**	**0.3947**	**0.7018**	**2.9074**
Ours	0.7836	0.6679	0.5774	0.5042	0.3672	0.6730	2.8436

Table 2. Comparison scores of our method and other state-of-the-art methods on the RSICD. Significant values are given in bold.

score widely adopted for vision-to-language tasks, where n is from 1 to 4. ROUGE_L is similar with the concept of BLEU-n, which calculates recall rate of the longest common subsequence L between candidate and reference sentences. METEOR is capable of generating an alignment on all references based on WordNet synonyms and stemmed tokens for judging the word correlation.

Training details and experimental setup

In our experiments, all RSIs are resized to 224×224 before entering the pre-trained VGG16[30], the extracted CNN features are with size of $7 \times 7 \times 512$. For four public datasets, the proportions of the training set, verification set and test set are 80%, 10% and 10% for training, validation and testing. And our experiments are completed on a single NVIDIA GeForce GTX 1080Ti.

Specifically, the embedded dimension of all LSTMs is 512. Word embedding is also represented as 512. Each attention step obtains 512-dimensional vector. The largest attention step is set to 4, and the optimization parameter of rethinking attention is set as 1×10^{-4}. The initial learning rate of the encoder and decoder is set to 1×10^{-5} and 4×10^{-4}, respectively. We use the Adam optimizer[55] for training. The batch size is set to 32, the maximum epoch is set to 35 and the beam is set to 3.

Compared models

We present a quantitative comparison with representative attention-based RSIC approaches: soft attention with CNN-LSTM framework (SAT)[17], an improved attribute attention (FC-Att/SM-Att)[20], a CNN-Transformer framework (VRTMM)[30], a novel structured attention with CNN-LSTM framework (Struc-Att)[31], a multi-level and contextual attention network (MLCA-Net)[16], a novel multi-scale contextual information aggregation network (MC-Net)[34], an attention-based global-local RSIC model (GLCM)[36], a novel Deformable Transformer RSIC model with deformable scaled dot-product attention (P-to-H)[37], a diffusion network (DiffNet) for RSIC task[48].

Evaluation results and analysis

Tables 1, 2, 3 and 4 shows the results of all algorithms on the dataset of NWPU-Captions, RSICD, UCM-Captions and Sydney-Captions. The best results are marked in bold. Firstly, we can see that our model scores higher than SAT, SM-Att, Struc-Att, MLCA-Net, and MC-Net on four datasets. Different with standard attention, SM-Att introduced high-level attribute features into calculating weights in attention layers, which measured semantic relation between high-level semantics and image features. Thus, higher scores can be achieved by SM-Att on the

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
SAT	0.7995	0.7365	0.6792	0.6244	0.4171	0.7441	3.1044
SM-Att	0.8115	0.7418	0.6814	0.6296	0.4354	0.7793	3.3860
Struc-Att	0.8538	0.8035	0.7572	0.7149	0.4632	0.8141	3.3489
MLCA-Net	0.8260	0.7700	0.7170	0.6680	0.4350	0.7720	3.2400
MC-Net	0.8450	0.7840	0.7320	0.6790	0.4490	0.7860	3.3550
VRTMM	0.8394	0.7785	0.7283	0.6828	0.4527	0.8026	3.4948
GLCM	0.8182	0.7540	0.6986	0.6468	0.4691	0.7524	3.0279
P-to-H	0.8230	0.7700	0.7228	0.6792	0.4439	0.7839	3.4629
DiffNet	**0.8712**	**0.8177**	**0.7695**	**0.7243**	**0.4777**	**0.8210**	**3.6631**
Ours	0.8534	0.7968	0.7487	0.7058	0.4665	0.8118	3.4900

Table 3. Comparison scores of our method and other state-of-the-art methods on the UCM-Captions dataset. Significant values are given in bold.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
SAT	0.7391	0.6402	0.5623	0.5248	0.3493	0.6721	2.2015
SM-Att	0.7430	0.6535	0.5859	0.5181	0.3641	0.6772	2.3402
Struc-Att	0.7795	0.7019	0.6392	0.5861	0.3954	0.7299	2.3791
MLCA-Net	0.8310	0.7420	0.6590	0.5800	0.3900	0.7110	2.3240
MC-Net	0.8340	0.7500	0.6780	0.6070	0.4060	0.7390	2.5640
VRTMM	0.7443	0.6723	0.6172	0.5699	0.3748	0.6698	2.5285
GLCM	0.8041	0.7305	0.6745	0.6259	0.4421	0.6965	2.4337
P-to-H	**0.8373**	**0.7771**	**0.7198**	**0.6659**	**0.4548**	**0.7860**	**3.0369**
DiffNet	0.8011	0.7283	0.6598	0.5981	0.4216	0.7490	2.7442
Ours	0.7815	0.6994	0.6257	0.5569	0.4000	0.7167	2.4808

Table 4. Comparison scores of our method and other state-of-the-art methods on the Sydney-Captions dataset. Significant values are given in bold.

dataset of RSICD and UCM-Captions than SAT, Struc-Att, MLCA-Net and MC-Net. More importantly, MLCA-Net and MC-Net implemented multi-scale structure for feature representation. Further, multi-scale attention is designed for the connection among visual features at different levels. MLCA-Net performs better on the NWPU-Captions dataset. In Tables 2 and 4, the scores of MC-Net sightly exceed MLCA-Net. Although these designed attention can infer words with accurate objects and relationships in the RSI, there are still redundant or wrong information. VRTMM, GLCM and P-to-H tried to filter out redundant information with Transformer structure. VRTMM adopted pre-trained CNN with Transformer decoder, which fused semantic information and visual features, thereby enhancing the feature representation. As seen on the dataset of NWPU-Captions, RSICD and UCM-Captions, VRTMM surpasses GLCM and P-to-H methods on all metrics. This improvements indicate that the redundant information is needed to be filtered out. Therefore, our model performs refinement gate in the encoder for refined visual features and rethinking attention for more sensible cross-modal interaction. It can be seen that our method achieves the best performance in most of the metrics. For example, on the NWPU-Captions dataset, our method is with ROUGE_L and CIDEr scores of 0.7468 and 1.8167, while VRTMM is with ROUGE_L and CIDEr scores of 0.6845 and 1.5885. On ROUGE_L and CIDEr, these scores of our method achieve an improvement of 13.4% and 58.43% compared with DiffNet. However, the DiffNet is the previous highest-performing and most competitive algorithm on the dataset of RSICD, UCM-Captions and Sydney-Captions. It is because that DiffNet adopts a refined multi-scale feature extraction for discernible visual features. Additionly, the diffusion model-based non-autoregressive decoder can boost the precision of sentence-level semantic analysis. Thus, the scores of the DiffNet exceed our model on the dataset of RSICD, UCM-Captions and Sydney-Captions.

As shown in Fig. 6, it can be seen two visualization examples for the caption generation process with the rethinking attention. Note that maximum attention step is set to 4. For each word generation, attention steps are greater than one in Fig. 6a,b. We observe that four attention steps are appeared for generating the first word, like "some" and "many" in Fig. 6. At the beginning, more visual features are attended by the rethinking attention. Therefore, the more observing leads to better comprehension and high quality sentence generation. Besides, the 1-th attention step is coarse, later attention steps will gradually enhance the interested areas. This similar phenomenon for "many" word is also shown in Fig. 6b. There are three attention steps for the second word. It is clear that the number of attention steps is dynamically changing at different decoding steps. More steps are also taken at inferring semantic relationship among objects (i.e., "around"). For some objects, two attention steps discriminately allocate weights on the image, then the rethinking attention are terminated. It indicates that adaptive alignment is effective and realized by the rethinking attention.

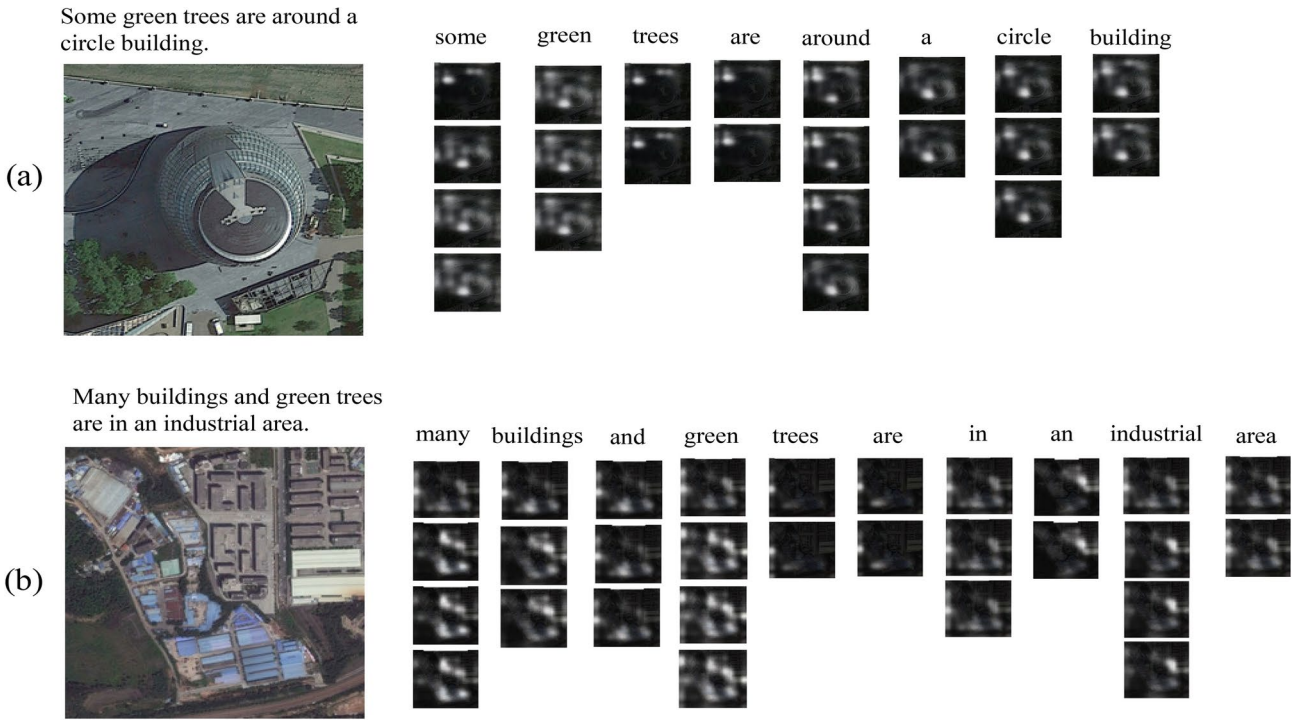

Figure 6. Visualization examples[17] for the caption generation process of our rethinking attention. We show the attention steps taken at each decoding step with the weighted regions.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
A1	0.8247	0.7289	0.6557	0.5997	0.3878	0.7046	1.6499
A2	0.8315	0.7368	0.6654	0.6100	0.3949	0.7125	1.6828
A3	0.8397	0.7475	0.6768	0.6211	0.4049	0.7274	1.7555
A4	**0.8490**	**0.7620**	**0.6957**	**0.6441**	**0.4198**	**0.7468**	**1.8167**

Table 5. Ablation performance of our designed model on the NWPU-Captions dataset. Significant values are given in bold.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
A1	0.7679	0.6579	0.5699	0.4962	0.3534	0.6590	2.6022
A2	0.7711	0.6645	0.5777	0.5048	0.3574	0.6674	2.7288
A3	0.7712	0.6636	0.5762	0.5020	0.3577	0.6664	2.6860
A4	**0.7836**	**0.6679**	**0.5774**	**0.5042**	**0.3672**	**0.6730**	**2.8436**

Table 6. Ablation performance of our designed model on the RSICD. Significant values are given in bold.

Ablation experiments

We study different ablations to analyze the effectiveness of our feature refinement and rethinking attention framework, ablation experiments are conducted on two public datasets. The experimental results are given in Tables 5-6. VGG16 combined with top-down attention is baseline model, named A1. A2 represents that baseline employs feature refinement based on CNN features. Adding rethinking attention into A2 model denotes our full model. Note that MHA (A2) or standard attention (A3) are adopted in the rethinking attention. There are clear performance improvements between A1 and A4 on two datasets. The highest performing compared with baseline on the NWPU-Captions dataset is increased by 4.22% on ROUGE_L and 16.68% on CIDEr.

For A1 and A2 models, they are trained with CE loss, while A3 and A4 models are optimized with formula 25. Interestingly, owing to feature extractor in baseline can not strengthen visual features, A2 model with refinement gate verifies that the interaction in visual domain helps to improve the quality of the visual representation. The results in Table 5 indicate that A2 improves performance (CIDEr) from 1.6499 to 1.6828 on the NWPU-Captions dataset, while it surpasses baseline on the RSICD in Table 6. Importantly, from the experimental results

Methods	Latency(s)	Fps(f/s)
Baseline	3.60	9.89
Ours	5.41	5.58

Table 7. Comparison between our method and baseline model on latency and fps. All results are reported on the NWPU-Captions dataset.

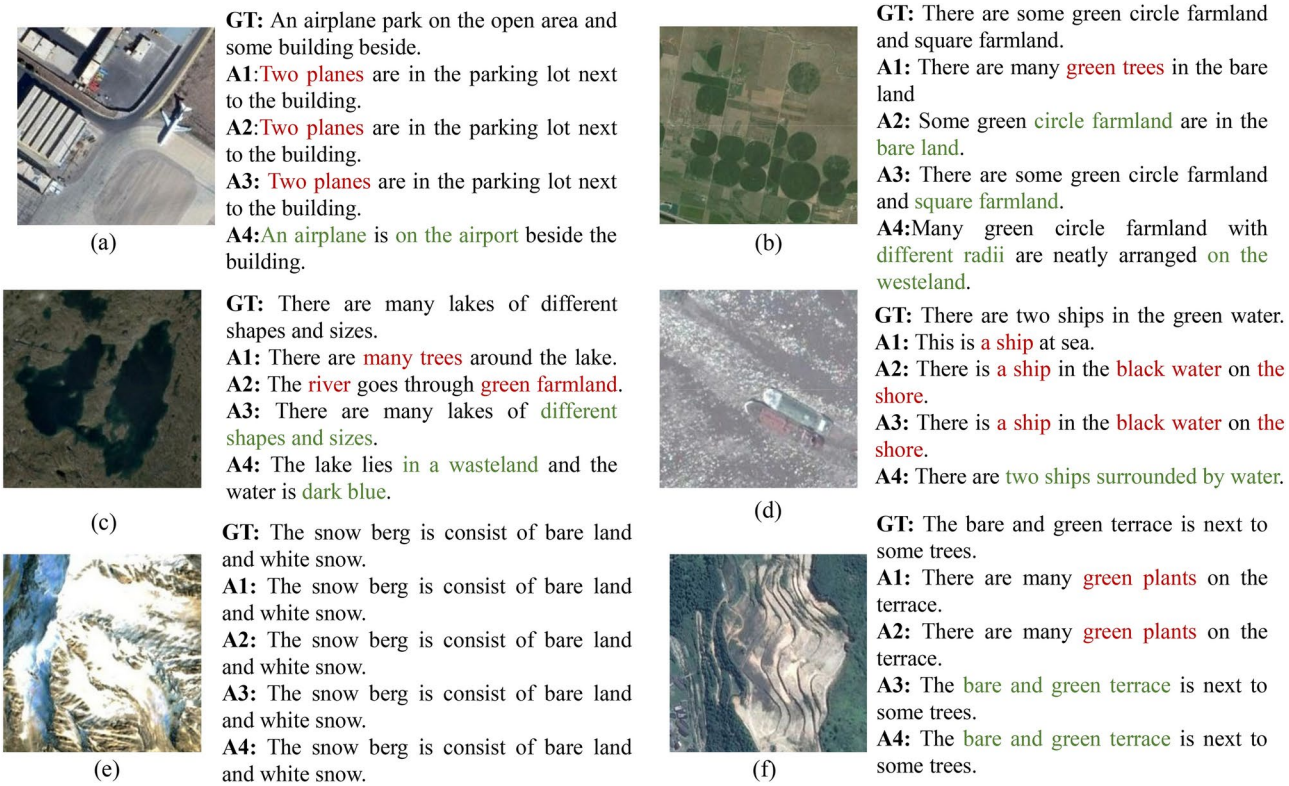

Figure 7. The output sentences are generated by one selected ground-truth (GT) sentence, baseline model and ablation models on the NWPU-Captions dataset[16]. The red words indicate mismatching in the generated images, and the green ones are precise words with our model.

shown in Tables 5 and 6, it can be seen that the rethinking attention improves performance compared with the standard attention adopted in A1 and A2. The excellent scores of A3 and A4 are applicable to all metrics, especially on CIDEr. It is because that the rethinking attention can adaptively focus on multiple image area at t time and effectively align these semantics for a word. However, the effect of MHA and standard attention adopted by A3 and A4 is obvious. As we know, the MHA focuses on learning feature diversity. In other task, it may be meaningful in network. However, for the rethinking attention, it may affect the number of steps, which are not easy to stop attention steps. According the experimental results shown in Tables 5 and 6, A4 is better than A3 in all metrics but inferior to A3 on BLEU4 and METEOR metrics. Standard attention in A4 is attributed to simple attention control, benefiting for modeling and controlling rethinking attention. Therefore, A4 can get a higher score. However, the comparison results on latency and fps in Table 7, our designed model is poor than baseline. This is because attention module are adopted in feature refinement and rethinking attention, this increases complexity of information processing and transmission in our network.

In addition, Fig. 7 shows ground-truth sentence and the other sentences generated by baseline and ablation models. For Fig. 7a, the "an planes" is mistaken with "two planes" generated by A1, A2 and A3. And Fig. 7c is similar to Fig. 7a, in which "two ships" are corrected. This indicates that full model can correctly do counting at some scene. In addition, the "shore" in the generated sentence is not related to the RSI. And the relationship between predicted "ship" and "water" by A2 and A3 model is messy. For Fig. 7b, the word "farmland" is neglected in description generated by A1, which may be due to high frequency "trees" in vocabulary and the stereotype of visual features without feature refinement. A3 and A4 are good at semantic inspiration with rethinking attention, which can infer the correlation between the most relevant areas with words. The A4 model is an extension of A3 model, which further boosts the generated sentence-level coherence (i.e. "Many green circle farmland with different radii are neatly arranged on the westeland"). As shown in Fig. 7c, the descriptions generated by A3 and A4 are richer and more accurate. Compared with A1, they also contain a clear and coherent syntax structure. A2

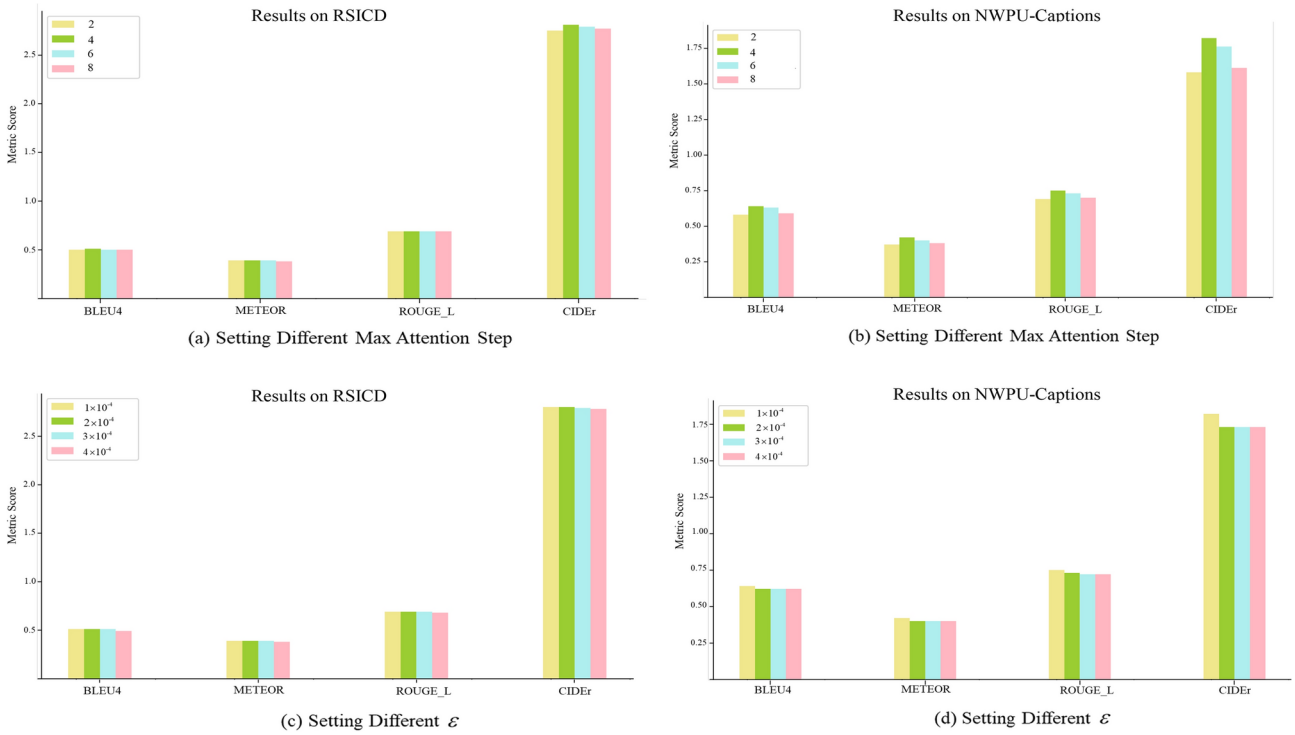

Figure 8. The visualizations of our model performance (i.e., BLEU4, METEOR, ROUGE_L, CIDEr) that are affected by parameters with different parameters ε and max attention step on the dataset of RSICD and NWPU-Captions.

infers wrong semantic information like A1 in Fig. 7b, "river" and "green farmland" are not suitable for the scene content. On the contrary, A3 and A4 describe the "lakes of different shapes and sizes" and "water is dark blue". In Fig. 7f, the wrong descriptions "green plants" appear in A1 and A2 respectively. The descriptions generated by A3 and A4 have a high consistency with annotated reference sentence. It is inferred that the rethinking attention in A3 and A4 can better control over the generated description. For simple scene in Fig. 7e, all models are enable effective captioning. Above all, this shows that the proposed model not only focuses on comprehensive target information, but also improves the overall quality of generated description.

Parameter analysis

To evaluate the impact of the maximum attention steps and the learning rate ε in the rethinking attention, we set different maximum attention steps and ε, in which the consistency of other parameters are ensured. Figure 8 shows the experimental results for two concerned parameters on the dataset of RSICD and NWPU-Captions.

Figure 8a,b shows the influence of different maximum attention steps on the dataset of RSICD and NWPU-Captions. The experimental results show that our model performs best when the maximum steps is as 4. It is clear that green column is highest on two datasets. When the value is 2, the experimental results are the lowest on CIDEr, especially on the NWPU-Captions dataset. This phenomenon shows that the rethinking attention module is sensitive to attention steps. With the increase of attention step, the performance of our model will be further improved. Our model can focus on multiple image areas for predicting a word state. If the parameter is greater than 4, the performance deterioration will appear. Excessive attention steps not only increase calculation cost, but also accept redundant visual information in current decoding process. It can conclude that more attention steps (less than 4) in the rethinking attention help to obtain a better performance. To accurately convert visual features into text description and reduce calculation costs, setting maximum steps as 4 is the best choice, which is also applied to the dataset of Sydney-Captions and UCM-Captions.

Figure 8c,d shows the influence of parameter ε on the dataset of RSICD and NWPU-Captions, assigning 1×10^{-4}, 2×10^{-4}, 3×10^{-4} and 4×10^{-4} for ε. The experimental results are obtained with maximum attention step set as 4. Obviously, it's better to set an appropriate ε for excellent attention attention than undesigned ε parameter. The experimental results shown in Fig. 8c,d, it is suitable to choose a small rate for two datasets. This is because the smaller ε can better control distribution of the rethinking attention weight and affect attention steps. However, the performance increase stops when ε is greater than 1×10^{-4}. While the performances degradation with 2×10^{-4}, 3×10^{-4} and 4×10^{-4} for ε are not serious. For the RSICD, four CIDEr socres almost equal, $\varepsilon = 1 \times 10^{-4}$ performs well on the NWPU-Captions dataset.

Conclusions

In this paper, we engineer a novel RSIC model to caption the content under different RSI scenes. The proposed feature refinement module can help handle the correlation between objects, using the correlation of grid-level feature weakens valueless features for the rethinking attention. We think that one-step is not reliable for effective attentive maps. Specifically, the rethinking attention can adaptively map different regions for semantic inference at t time, which considers issues such as insufficient visual information, misguided features and more meaningful guidance. Moreover, the confidence rectification strategy is encouraged for optimizing the rethinking attention, which is essentially to match appropriate visual information and semantic reasoning. We verify the advantages of our proposed method by comprehensive experiments on four benchmark datasets, which outperforms the baseline model by 4.22% on ROUGE_L and 16.68% on CIDEr, on the NWPU-Captions dataset. In addition, visualized weighted regions are strongly activated, which demonstrate that features learned from the rethinking attention are effective.

Data availability

The NWPU-Captions, RSICD, UCM-Captions and Sydney-Captions datasets can be obtained from (https://github.com/HaiyanHuang98/NWPU-Captions, https://pan.baidu.com/s/1bp71tE3#list/path=%2F, https://pan.baidu.com/s/1mjPToHq#list/path=%2F, https://pan.baidu.com/s/1hujEmcG#list/path=%2F).

Received: 5 October 2024; Accepted: 4 March 2025
Published online: 13 March 2025

References

1. Yin, J. et al. Integrating remote sensing and geospatial big data for urban land use mapping: A review. *Int. J. Appl. Earth Obs. Geoinf.* **103**, 102514 (2021).
2. Lu, D., Moran, E. & Hetrick, S. Detection of impervious surface change with multitemporal landsat images in an urban-rural frontier. *ISPRS J. Photogramm. Remote. Sens.* **66**, 298–306 (2011).
3. Kucharczyk, M. & Hugenholtz, C. H. Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. *Remote Sens. Environ.* **264**, 112577 (2021).
4. Zhang, L. & Zhang, L. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.* **10**, 270–294 (2022).
5. Song, H., Yuan, Y., Ouyang, Z., Yang, Y. & Xiang, H. Quantitative regularization in robust vision transformer for remote sensing image classification. *Photogram. Rec.* **39**, 340–372 (2024).
6. Zhang, X., Wang, Q., Chen, S. & Li, X. Multi-scale cropping mechanism for remote sensing image captioning. In *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 10039–10042 (IEEE, 2019).
7. Wang, B., Zheng, X., Qu, B. & Lu, X. Retrieval topic recurrent memory network for remote sensing image captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **13**, 256–270 (2020).
8. Vaswani, A. et al. Attention is all you need. *Proc. Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
9. Shi, Z. & Zou, Z. Can a machine generate humanlike language descriptions for a remote sensing image?. *IEEE Trans. Geosci. Remote Sens.* **55**, 3623–3634 (2017).
10. Song, H. Mbc-net: long-range enhanced feature fusion for classifying remote sensing images. *Int. J. Intell. Comput. Cybern.* (2023).
11. Ma, X., Zhao, R. & Shi, Z. Multiscale methods for optical remote-sensing image captioning. *IEEE Geosci. Remote Sens. Lett.* **18**, 2001–2005 (2020).
12. Wang, Y., Zhang, W., Zhang, Z., Gao, X. & Sun, X. Multiscale multiinteraction network for remote sensing image captioning. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* **15**, 2154–2165 (2022).
13. Liu, C. et al. Rscama: Remote sensing image change captioning with state space model. *IEEE Geosci. Remote Sens. Lett.* (2024).
14. Li, Y., Zhang, X., Cheng, X., Chen, P. & Jiao, L. Inter-temporal interaction and symmetric difference learning for remote sensing image change captioning. *IEEE Trans. Geosci. Remote Sens.* (2024).
15. Song, H., Wei, C. & Yong, Z. Efficient knowledge distillation for remote sensing image classification: a cnn-based approach. *Int. J. Web Inf. Syst.* (2023).
16. Cheng, Q. et al. Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–19 (2022).
17. Lu, X., Wang, B., Zheng, X. & Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **56**, 2183–2195 (2018).
18. Wu, S., Zhang, X., Wang, X., Li, C. & Jiao, L. Scene attention mechanism for remote sensing image caption generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7 (IEEE, 2020).
19. Yuan, Z., Li, X. & Wang, Q. Exploring multi-level attention and semantic relationship for remote sensing image captioning. *IEEE Access.* **8**, 2608–2620 (2019).
20. Zhang, X., Wang, X., Tang, X., Zhou, H. & Li, C. Description generation for remote sensing images using attribute attention mechanism. *Remote Sens.* **11**, 612 (2019).
21. Li, Y., Zhang, X., Cheng, X., Tang, X. & Jiao, L. Learning consensus-aware semantic knowledge for remote sensing image captioning. *Pattern Recogn.* **145**, 109893 (2024).
22. Wang, Q., Yang, Z., Ni, W., Wu, J. & Li, Q. Semantic-spatial collaborative perception network for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* (2024).
23. Zhang, Y., Shi, X., Mi, S. & Yang, X. Image captioning with transformer and knowledge graph. *Pattern Recogn. Lett.* **143**, 43–49 (2021).
24. Tian, C., Tian, M., Jiang, M., Liu, H. & Deng, D. How much do cross-modal related semantics benefit image captioning by weighting attributes and re-ranking sentences?. *Pattern Recogn. Lett.* **125**, 639–645 (2019).
25. Zhang, Z. et al. Lam: Remote sensing image captioning with label-attention mechanism. *Remote Sens.* **11**, 1–15 (2019).
26. Sumbul, G., Nayak, S. & Demir, B. Sd-rsic: Summarization-driven deep remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **59**, 6922–6934 (2020).
27. Li, Y., Fang, S., Jiao, L., Liu, R. & Shang, R. A multi-level attention model for remote sensing image captions. *Remote Sens.* **12**, 939 (2020).
28. Wang, Q., Huang, W., Zhang, X. & Li, X. Word-sentence framework for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **59**, 10532–10543 (2020).
29. Li, Y. et al. Recurrent attention and semantic gate for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–16 (2021).

30. Shen, X., Liu, B., Zhou, Y., Zhao, J. & Liu, M. Remote sensing image captioning via variational autoencoder and reinforcement learning. *Knowl. Based Syst.* **203**, 105920 (2020).
31. Zhao, R., Shi, Z. & Zou, Z. High-resolution remote sensing image captioning based on structured attention. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2021).
32. Li, Y. et al. A patch-level region-aware module with a multi-label framework for remote sensing image captioning. *Remote Sensing* **16**, 3987 (2024).
33. Zhao, K. & Xiong, W. Exploring region features in remote sensing image captioning. *Int. J. Appl. Earth Obs. Geoinf.* **127**, 103672 (2024).
34. Huang, H. et al. Mc-net: multi-scale contextual information aggregation network for image captioning on remote sensing images. *Int. J. Digit. Earth* **16**, 4848–4866 (2023).
35. Yang, Z., Li, Q., Yuan, Y. & Wang, Q. Hcnet: Hierarchical feature aggregation and cross-modal feature alignment for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* (2024).
36. Wang, Q., Huang, W., Zhang, X. & Li, X. Glcm: Global-local captioning model for remote sensing image captioning. *IEEE Trans. Cybern.* **53**, 6910–6922 (2022).
37. Du, R. et al. From plane to hierarchy: Deformable transformer for remote sensing image captioning. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* (2023).
38. Qu, B., Li, X., Tao, D. & Lu, X. *Deep semantic understanding of high resolution remote sensing image* (In Proc. Int. Conf. Comput. Inf. Telecommun, Syst, 2016).
39. Zhang, X. et al. Natural language description of remote sensing images based on deep learning. *In Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS).* 4798–4801 (2017).
40. Zhang, Z. et al. Global visual feature and linguistic state guided attention for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–16 (2021).
41. Lu, X., Wang, B. & Zheng, X. Sound active attention framework for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **58**, 1985–2000 (2019).
42. Li, X., Zhang, X., Huang, W. & Wang, Q. Truncation cross entropy loss for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **59**, 5246–5257 (2020).
43. Chavhan, R., Banerjee, B., Zhu, X. X. & Chaudhuri, S. A novel actor dual-critic model for remote sensing image captioning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 4918–4925 (IEEE, 2021).
44. Gajbhiye, G. O. & Nandedkar, A. V. Generating the captions for remote sensing images: A spatial-channel attention based memory-guided transformer approach. *Eng. Appl. Artif. Intell.* **114**, 105076 (2022).
45. Zia, U., Riaz, M. M. & Ghafoor, A. Transforming remote sensing images to textual descriptions. *Int. J. Appl. Earth Obs. Geoinf.* **108**, 102741 (2022).
46. Ren, Z., Gou, S., Guo, Z., Mao, S. & Li, R. A mask-guided transformer network with topic token for remote sensing image captioning. *Remote Sens.* **14**, 2939 (2022).
47. Wu, Y. et al. Trtr-cmr: Cross-modal reasoning dual transformer for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* (2024).
48. Cheng, Q., Xu, Y. & Huang, Z. Vcc-diffnet: Visual conditional control diffusion network for remote sensing image captioning. *Remote Sens.* **16**, 2961 (2024).
49. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556 (2014).
50. Anderson, P. et al. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition. (CVPR)*, 6077–6086 (2018).
51. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318 (2002).
52. Banerjee, S. & Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72 (2005).
53. Lin, C. *Rouge: A package for automatic evaluation of summaries* (In Proc. Assoc. Comput, Linguist, 2004).
54. Vedantam, R., Zitnick, C. & Parikh, D. Cider: Consensus-based image description evaluation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4566–4575 (2015).
55. Diederik, P. K. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 7–9 (2014).

Acknowledgements

The authors would like to express their gratitude to the editors and the anonymous reviewers for their insightful comments.

Author contributions

Conceptualization, Y.L.; funding acquisition, Y.L. and D.Z.; methodology, Y.L., C.T. and M.L.; software, Y.L., C.T. and M.L.; supervision, D.Z., X.Z., G.W. and T.Z; writing-original draft, Y.L., C.T. and M.L.; writing-review and editing, X.Z., Y.L., G.W. and T.Z. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by the Wuxi Innovation and Entrepreneurship Fund "Taihu Light" Science and Technology (Fundamental Research) Project under Grant K20241045 and K20221046, the Wuxi University Research Start-up Fund for Introduced Talents under Grant 2024r011, the Postdoctoral Fellowship Program of CPSF under Grant GZC20241321, the Natural Science Foundation of Jiangsu Province under Grant BK20210064, the 111 Project under Grant B17035.

Declarations

Competing interests

The authors declare no competing interest.

Additional information

Correspondence and requests for materials should be addressed to D.Z.