Pattern Recognition 171 (2026) 112209

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

DIA: Deriving linguistic information from auxiliary languages for remote sensing image captioning ⋆

Tao Yang, Qing Zhou, Qi Wang *

School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, China

ARTICLE INFO

Keywords:
Remote sensing image captioning
Linguistic-enhancing information
Auxiliary language-enhanced network
Linguistic bridge

ABSTRACT

Remote sensing image captioning (RSIC) is a cross-modal task aimed at describing scene categories, object classes, and their spatial relationships in remote sensing images using natural language. Existing methods typically focus on training models in single language, neglecting the linguistic-enhancing information derived from syntactic structure differences and diverse expressions of the same objects and scenes. This information, present in cross-linguistic annotated data, can significantly enhance language perception and enrich training data. To verify the effectiveness of this information, we propose an auxiliary language-enhanced network called DIA, which leverages linguistic information from auxiliary languages to improve the quality and fluency of target language generation. DIA consists of shared visual feature extractor, target language generator, and auxiliary language generator. The shared visual feature extractor integrates the Linguistic-Irrelevant Feature Enrichment (LiFE) module, while a Linguistic Bridge connects the target and auxiliary language generators. The LiFE module employs linguistic-irrelevant feature extraction and multi-view attention to extract precise visual features, enriching the representations while minimizing language bias. Multi-view attention balances deep semantic expressions and linguistic-irrelevant features. The Linguistic Bridge establishes interactive pathway between the target language generator (ALG) and the auxiliary language generator (TLG), enabling the TLG to learn from the ALG's language modeling capabilities. This interaction allows the TLG to handle complex visual features, improving language generation performance. Extensive experiments demonstrate that our model achieves significant performance improvements on the UCM, Sydney, RSICD, and NWPU datasets. Specifically, on the UCM dataset, BLEU-4 is improved by 5.06 %, and CIDEr is improved by 16.86 %.

1. Introduction

Remote sensing image captioning (RSIC) provides a method for converting geographic information from the image modality into text, reducing the difficulty of non-experts face in understanding remote sensing images. Compared to traditional category-level results, such as remote sensing image object detection [1] and scene classification [2], remote sensing image captioning can generate sentences that describe the scene category, object information, and their spatial relationships in a complete structure [3]. This provides intuitive guidance for a wide range of applications, including disaster monitoring, intelligence generation, smart decision-making, and urban planning [4]. The rapid development of deep learning has significantly advanced remote sensing image captioning technology [5], making intelligent models the mainstream approach for completing this task.

Data is crucial for the development of RSIC intelligent models, and many researchers have contributed to the construction of several popular datasets, such as UCM-captions [6], Sydney-captions [6], RSICD [7], NWPU [8], and RSICN [9]. The existing methods are primarily improvements based on the encoder-decoder architecture, where CNNs are used as the encoder to extract image features, and LSTM networks serve as the decoder to process visual feature and generate sentences. Based on this framework, Huang et al. [10] introduced multi-scale feature fusion to enhance the encoder's ability to extract visual features, improving the feature representation. Some studies involve attention-based applications [11], which uses attention mechanisms to focus on specific visual regions when generating each word. Li et al. [12] introduced a new multi-level attention mechanism to guide the model in generating text hierarchically. Other researches have investigated the use of auxiliary tasks to provide prior textual knowledge to

⋆ This work was supported in part by the National Natural Science Foundation of China under Grant 62471394, and Grant U21B2041.
* Corresponding author.
E-mail addresses: taoytao@mail.nwpu.edu.cn (T. Yang), chautsing@gmail.com (Q. Zhou), crabwq@nwpu.edu.cn (Q. Wang).

https://doi.org/10.1016/j.patcog.2025.112209
Received 25 March 2025; Received in revised form 14 June 2025; Accepted 23 July 2025
Available online 29 July 2025

the model. Ye et al. [13] designed a two-stage joint training framework that jointly optimizes through multi-object detection tasks, providing object information. Zhang et al. [14] introduced a scene classification module to provide scene category information to the model. Additionally, several studies have introduced data augmentation [15] and noise handling techniques [16] to improve the model's generalization ability and robustness. Huang et al. [17] used denoising techniques to reduce interference components in visual features.

Despite the performance improvements achieved through the aforementioned methods that focus on various aspects, these approaches share a common limitation: their predominant reliance on English-annotated data. This leads to the suboptimal utilization of valuable multilingual resources. Furthermore, the abundant information embedded in cross-lingual annotated data remains largely untapped. For the same image annotated in different languages, although the textual expressions may vary, they correspond to the same underlying semantic image features. This correspondence can **enrich the training data of the model**, enhancing the visual feature extractor's ability to capture rich semantic features. Furthermore, different languages have distinct syntactic structures, and these differences **provide diverse sentence constructions**, thereby improving the text generator's contextual perception. These aspects have not been addressed in previous research. As shown in Fig. 1, both the Chinese and English captions contain references to the airport and airplanes, which is marked in blue. When describing the attribute relationship between them, in the Chinese caption, the state of "parked" appears before "airplane", while in the English caption, the order is reversed, which is marked in green.

In light of this, we propose a language-enhanced network called DIA based on auxiliary language annotations to improve RSIC performance. The existing methods follow the approach shown in Fig. 1(a), while Fig. 1(b) illustrates our motivation. Unlike the structure of existing methods that only process single-language annotated data, our network consists of a shared visual feature extractor and Linguistic Bridge-Enhanced language generator. The shared visual feature extractor is used to extract generic visual representations from remote sensing images, while the auxiliary language generator (ALG) is trained on the auxiliary language to help the target language generator (TLG) perceive the underlying relationship between images and text. During inference, the TLG will be masked, ensuring that it does not introduce additional inference overhead.

Concretely, we use pre-trained CNN [18] and CLIP [19] in the shared visual extractor to obtain visual representations of remote sensing images. To alleviate the language bias in CLIP, which arises from its pre-training on English-annotated data, resulting in a preference for English-related patterns and expressions, we introduce a Linguistic-irrelevant Feature Enrichment (LiFE) module in the shared visual feature extractor. This module first extracts inherent visual features (i.e., shape, texture, color, which are the basic visual characteristics of images that can be described independently of any language.) of the remote sensing image through a series of linguistic-irrelevant feature extraction operations. Then, multi-view attention is applied to enhance features related to the text and suppress irrelevant features. Additionally, we introduce a Linguistic Bridge between ALG and TLG, which establishes a communication pathway between the ALG and TLG. This allows ALG to learn the language modeling capabilities that TLG has acquired from the auxiliary language, enabling it to generate more accurate captions. Our network excels in processing cross-linguistic annotated data and extracting pertinent information from such datasets, thereby enhancing the model's linguistic modeling capabilities.

In summary, our contributions are as follows:

(1) To the best of our knowledge, we are the first to construct the DIA method in the field of RSIC, which enhances target language modeling capabilities by utilizing cross-lingual captions.

(2) We propose two modules to enhance the utilization of cross-lingual captions: the LiFE module and the Linguistic Bridge. The LiFE module captures rich representations and reduces linguistic biases in visual features, while the Linguistic Bridge enhances the language modeling capabilities of TLG by learning ALG.

(3) Extensive experiments are conducted on the BRSIC dataset to validate the effectiveness of DIA. Additionally, experiments are carried out on the bilingual NWPU dataset, generated through machine translation, to assess the practicality of DIA.

2. Related works

In this section, we review the related research in the field of image captioning, which can be divided into two parts: natural image captioning (NIC) and remote sensing image captioning (RSIC).

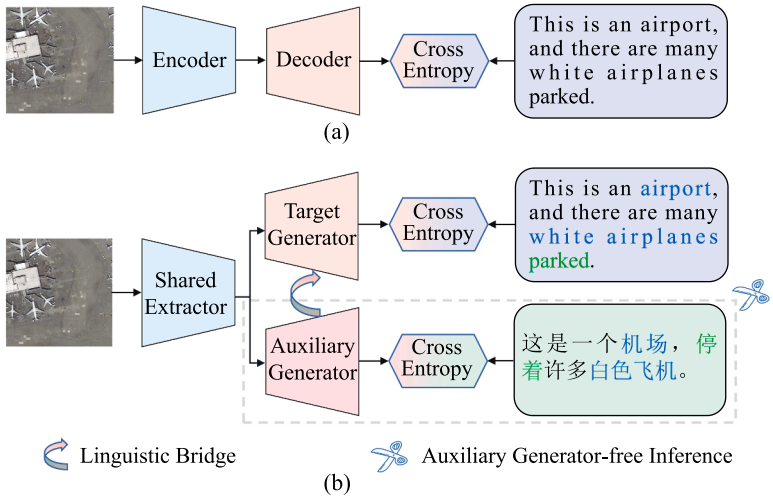

Fig. 1. Comparison between existing methods and our proposed DIA. (a) represents the existing encoder-decoder model trained with English-labeled captions. (b) represents our proposed DIA method with auxiliary language enhancement. Words with the same color represent the same semantics across different languages. During the training stage, the Linguistic Bridge provides a weight interaction pathway between the Target Generator and the Auxiliary Generator, with the direction of the arrow indicating the direction of language modeling capability enhancement. Auxiliary Generator-free Inference indicates that during inference, the Auxiliary Generator is disabled.

2.1. Natural image captioning

In recent years, the encoder-decoder architecture has become the mainstream model structure in the field of NIC. The encoder is used to extract image features, while the decoder converts these visual features into natural language. For example, Vinyals et al. [20] utilized a pre-trained CNN as an encoder to convert images into visual features and then used LSTM to process these visual features to generate captions. Cornia et al. [21] introduced external control signals to manipulate the generation process, effectively improving the quality and diversity of the generated captions. Luo et al. [22] proposed a non-autoregressive image captioning network with semantic conditioning, which offers a new perspective and solution for image captioning tasks using diffusion models. BLIP [23] introduced visual-linguistic alignment mechanisms and performed pretraining on large-scale image-text data, enhancing visual-linguistic alignment and improving caption quality. Huang et al. [24] enhanced the alignment between visual and textual modalities by introducing positive noise.

2.2. Remote sensing image captioning

Remote sensing images differ significantly from natural images, primarily because they often contain complex scenes made up of multiple objects. This characteristic requires model to not only effectively recognize individual objects in the image but also to accurately capture the relationships between these objects. To address this challenge, researchers have proposed various innovative approaches. Wang et al. [25] established a new "World-Sentence" framework, enhancing the interpretability of the captioning process and providing a more intuitive understanding of the image's semantics. Li et al. [26] introduced truncated cross-entropy loss, which successfully alleviated overfitting issues and achieved excellent performance on small datasets. Zhou et al. [27] introduced a contrastive learning-based pretraining image feature extractor to enhance the alignment between visual and textual representations. Yang et al. [28] introduced hierarchical feature aggregation and cross-modal feature alignment, enhancing the alignment between image and text, which significantly improved model performance. Zhou et al. [9] established the first bilingual Chinese-English dataset for RSCI, providing researchers with a multilingual data option.

Although existing RSIC methods continuously improve performance from multiple aspects, they generally lack the ability to enhance the accuracy and quality of generated captions by utilizing cross-lingual data.

3. Methodology

3.1. Overview

The proposed DIA, as shown in Fig. 2, can conceptually be divided into shared visual feature extractor and a Linguistic Bridge-enhanced language generator, corresponding to Fig. 2(a) and (b), respectively. The shared visual feature extractor is used to map remote sensing images into a shared visual feature space, which is jointly occupied by the target language and auxiliary language. It consists of a pre-trained CNN [18], CLIP [19], a LiFE module, and a feature fusion module. The LiFE module is made up of multiple feature extractors and SCP attention blocks. The Linguistic Bridge-enhanced language generator is divided into the TLG and the ALG, with each branch having an identical structure. These structures are formed by stacking a series of Transformer Blocks to create the Memory-Augmented Encoder (MAE) and Meshed Decoder (MD). MAE is used to isolate the unique representation of each language from the shared visual feature space, while MD receives the carefully processed visual features and learns the mapping between remote sensing image visual features and natural language. This process gradually translates high-dimensional image features into grammatically and semantically correct sentences. We build the Linguistic Bridge between the Transformer Blocks of TLG and ALG to enable TLG to learn the feature extraction and language modeling capabilities of ALG.

3.2. Shared visual feature extractor

Given the advantages of pre-trained models on large-scale datasets in extracting image feature representations, we select the pre-trained ResNet [18] model to extract high-level semantic information from remote sensing images, suitable for recognizing object categories and scene backgrounds. Additionally, we use CLIP with a ViT-L/14 [19] backbone to obtain image embeddings rich in semantic information, which are commonly used in image-to-text multimodal tasks. However, we recognize that visual features extracted by models pre-trained on English corpora tend to exhibit bias toward English text, which could negatively affect the language modeling of the ALG, thereby preventing the TLG from gaining more effective assistance from the ALG. To address this, we introduce a LiFE Module to incorporate more linguistic-irrelevant features, thereby enriching the shared visual feature space and providing comprehensive incentives for learning in both the TLG and ALG, as shown in Fig. 2(a).

LiFE module: To better extract the inherent visual features of remote sensing images and address potential biases in pre-trained models, which arise from being trained on English corpora and favoring English-specific syntactic and pragmatic patterns, we propose the LiFE module. This module extracts visual features from multiple aspects to tackle the biases that may introduce challenges in auxiliary language modeling for TLG and ultimately reduce the assistance TLG can provide to ALG. As shown in Fig. 2(c) left, we design five different ways to extract features, aiming to fully explore the multi-level, multi-dimensional, and inherent information in remote sensing images. $\mathrm{BS}_{\mathrm{EX}}$ is designed to preserve the basic structure and richness of visual features in the image, ensuring that the model can capture global features such as color, shape, and texture. $\mathrm{EF}_{\mathrm{EX}}$ enhances the ability to perceive edges and fine details, allowing for better capture of fine-grained features, such as object contours and texture variations. $\mathrm{IC}_{\mathrm{EX}}$ is specifically designed to capture the details at intersections or contact areas in the image, particularly in the presence of complex geometric shapes and object intersections. $\mathrm{HD}_{\mathrm{EX}}$ focuses on variations in the horizontal direction within the image, enhancing the extraction of terrain features with strong horizontal continuity and proving effective in identifying fine-grained terrain variations along the horizontal axis. $\mathrm{VC}_{\mathrm{EX}}$ is highly sensitive to vertical changes, enabling the capture of texture variations in the vertical direction of the terrain.

The above feature extraction transformations are uniquely designed for remote sensing images. By combining these operations, features from different directions and scales are comprehensively enhanced, providing more precise and detailed feature extraction capabilities for representing terrain structures in complex scenes. For the input visual feature x_3, we can obtain the feature x_{MCF}, which contains rich inherent information. The specific computation process is as follows:

$$W'(c_{\mathrm{in}}, c_{\mathrm{out}}, k_1, k_2) = W(c_{\mathrm{in}}, c_{\mathrm{out}}, k_1, k_2) + \Delta W(c_{\mathrm{in}}, c_{\mathrm{out}}, k_1, k_2, A), \tag{1}$$

$$F(x, \tau) = \mathrm{ReLU}\big(W'x + b\big) \oplus x, \quad \tau \in \{(\Delta, A)\} \tag{2}$$

where W is the parameters of Convolution layer. c_{in} and c_{out} are the number of input and output channels, respectively, while k_1 and k_2 are the height and width of the convolution kernel. $\oplus$ represents element-wise vector addition. $\Delta = \{scaling, summing, inverse\}$ represents the transformation of the weights based on specific requirements. $A = \{center, diagonal, horizontal, vertical\}$ is a set of direction that represents the positions of the parameters that need to be transformed. τ is an element in the set composed of Δ and A.

We use Eq. (1) to construct 5 extractors ($\mathrm{BS}_{\mathrm{EX}} = F(\cdot, \emptyset)$, $\mathrm{EF}_{\mathrm{EX}} = F(\cdot, \Delta_1, A_1, A_2)$, $\mathrm{IC}_{\mathrm{EX}} = F(\cdot, \Delta_2, A_1, A_2)$, $\mathrm{HD}_{\mathrm{EX}} = F(\cdot, \Delta_3, A_1, A_3)$, $\mathrm{VD}_{\mathrm{EX}} = F(\cdot, \Delta_4, A_1, A_3)$.) to extract inherent visual features, and the calculation process is shown below:

$$x_{\mathrm{MCF}} = \Gamma\{F(x_3, \tau)\}, \tau \in \{(\Delta, A)\}, \tag{3}$$

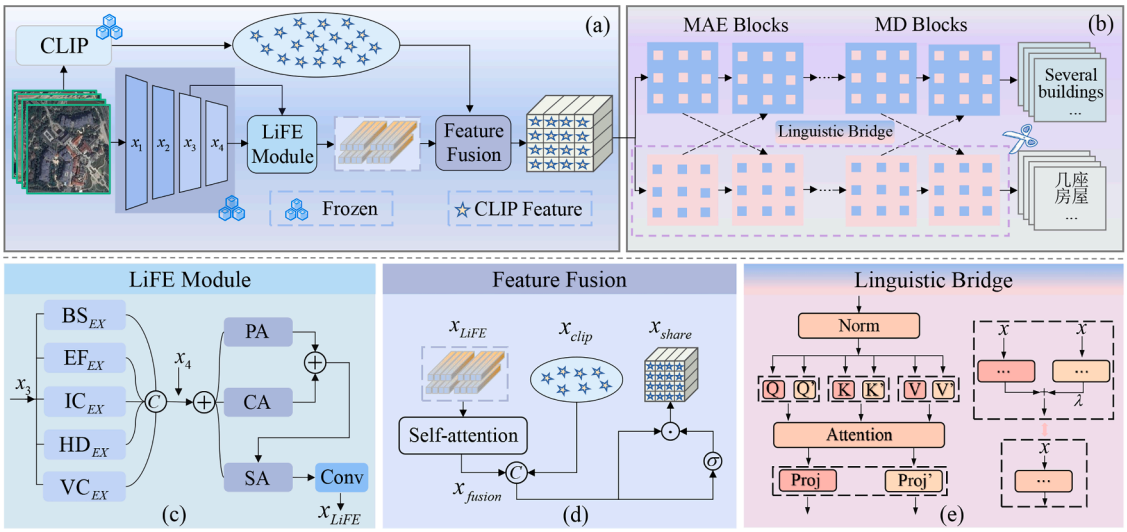

Fig. 2. The overall framework of the proposed DIA method. It mainly includes two steps: (a) Shared Visual Feature Extractor and (b) Linguistic Bridge-Enhanced Text Generator. In addition,(c) shows the specific structure of the LiFE module. (d) illustrates the process of hybrid feature fusion. (e) demonstrates the mechanism by which the Linguistic Bridge functions through reparameterization.

where x_3 represents the third-layer output of ResNet. Γ represents the concatenation operation on all the elements of the set.

To balance the representation of deep semantic features and linguistic-irrelevant features, we introduce Spatial Attention (SA), Channel Attention (CA), and Pixel Attention (PA) mechanisms to fuse language-irrelevant features x_{MCF} with the fourth-layer output x_4 of ResNet, as shown in Fig. 2(c) right. Through fusion, both can complement each other, enhancing the robustness and generalization of the features. The specific computation process is as follows:

$$\tilde{x} = x_4 \oplus x_{\text{MCF}}, \tag{4}$$

$$\beta = \sigma(SA(\tilde{x}, PA(\tilde{x}) \oplus CA(\tilde{x}))), \tag{5}$$

$$x_{\text{LiFE}} = \text{Conv}\big(\tilde{x} \oplus \beta \odot x_4 \oplus (1-\beta) \odot x_{\text{MCF}}\big), \tag{6}$$

where $\odot$ represents element-wise vector multiplication, and σ is the Sigmoid function.

Hybrid feature fusion: We introduce a self-attention mechanism to fuse the CLIP features with the features enriched by the LiFE Module, establishing a unified feature representation, as shown in Fig. 2(d). This effectively reduces language bias between features while enhancing the ability to express complex scene relationships. The specific computation process is as follows:

$$x_{\text{fusion}} = \text{Concat}(\text{SelfAttention}(x_{\text{LiFE}}), x_{\text{clip}}), \tag{7}$$

$$x_{\text{share}} = x_{\text{fusion}} \odot \sigma(x_{\text{fusion}}), \tag{8}$$

where Concat represents the concatenation operation.

3.3. Linguistic bridge-enhanced language generator

Transformer-based language generator: To generate more accurate target language captions for remote sensing images based on shared visual features, the Transformer is used as the language generator, which is divided into two main components: MAE and MD. MAE is used to extract specific language-matched visual representations from the shared visual features, enabling more effective cross-modal information interaction and expression. MD is responsible for converting these extracted specific visual representations into fluent and coherent target text, with the help of multi-layer interactive self-attention mechanisms that enhance the generation quality and semantic consistency. Specifically, MAE consists of Transformer blocks with L_e layers, and MD consists of Transformer blocks with L_d layers. Each block contains multi-head attention and position-wise feed-forward layers, as shown in Fig. 2(b). The multi-head attention in MAE is implemented using global group attention, which effectively captures spatial semantic correlations between landforms. The multi-head attention in MD is implemented using dot-product attention, which can flexibly model both global and local dependencies, enhancing contextual understanding. The computation process can be described as follows:

$$x^i_{\text{att}} = \text{MultiHeadGGA}^i\big(x^{i-1}_{\text{att}}\big),\ i \in [1, \ldots, L_d], \tag{9}$$

$$x^i_{\text{pos}} = \text{PWFF}_{\text{image}}\big(\text{LN}(x^i_{\text{att}})\big), \tag{10}$$

$$y^j_{\text{dec}} = \text{MultiHeadSGA}^j\Big(y^{j-1}_{\text{dec}} v\Big),\ j \in [1, \ldots, L_d], \tag{11}$$

$$y^j_{\text{pos}} = \text{PWFF}_{\text{text}}\Big(\text{LN}(y^j_{\text{att}})\Big), \tag{12}$$

where GGA represents Global Group Attention, SGA represents Pointwise Attention, i denotes the i-th layer of MAE, and j denotes the j-th layer of MD.

Reparameterization-based linguistic bridge: We employ reparameterization approach, which is simple yet effective, to enable the interactive learning between TLG and ALG, with the basic reparameterization process illustrated on the right side of Fig. 2(e). Each transformer block in TLG and ALG is treated as a unit for the reparameterization operation. For the blocks belonging to MAE in TLG and ALG, as well as the blocks belonging to MD, we construct the Linguistic Bridge in an alternating manner, as shown in Fig. 2(b). This approach allows TLG to learn the sequence modeling capabilities of ALG while reducing the risk of overfitting. The computation process can be represented as follows:

$$\text{Block}^k_{\text{TLG}} = \text{Linguistic Bridge}\Big(\text{Block}^k_{\text{TLG}}, \text{Block}^{k-1}_{\text{ALG}}\Big), \tag{13}$$

where k represents the index of the Block, $k \in [1, L_e - 1]$ or $k \in [1, L_d - 1]$, Linguistic Bridge represents the parametric interaction pathway between blocks.

After establishing the interaction pathway between TLG and ALG, we use Cross-Linguistic Re-parameterization to utilize the training weights of ALG between Transformer blocks, parameter interaction takes place between the line layers in each block, as shown in Fig. 2(e) left. Specifically, let w represent the learnable parameters in each layer of Transformer, where l indicates the operation in each block, i.e. $y = l(x; w)$. Let w' represent the corresponding parameters in ALG, which possesses the same structure as w, the Cross-Linguistic reparameterization of the liner layer is as follows:

$$y = l(x; w + \lambda w'), \tag{14}$$

where λ represents the trainable Cross-Linguistic scale.

During training, the parameters in TLG are reparameterized following $Re(w) = w + \lambda w'$. During inference, the TLG is isolated, ensuring that it does not introduce any additional computational overhead.

3.4. Loss function

Following the standard practice in image captioning, we optimize the model parameters through cross entropy loss (XE):

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* | y_{1:t-1}^*)), \tag{15}$$

where $y_{1:t-1}^* = [y_1^*, \ldots, y_{t-1}^*]$ represents the ground truth sequence of words from the start till the $(t-1)$-th time step; T is the maximum time step; θ denotes the trainable parameters of the model.

4. Experiments

In this section, we conduct extensive experiments on four RSIC dataset to demonstrate the effectiveness of our method. First, we introduce the datasets, experimental setup, and evaluation metrics. Then, we compare our method with existing approaches and perform ablation studies. Some visual results are also presented. Finally, we discuss the selection of hyperparameters λ in the Linguistic Bridge.

4.1. Datasets

(1) Sydney-captions: This dataset is derived from the Sydney dataset, with textual annotations provided by Qu et al. [6]. It consists of 613 high-resolution remote sensing images spanning 7 scene categories, sourced from Google Earth. Each image is cropped to a resolution of 500 × 500 pixels, with a resolution of 0.5m. Five distinct captions are provided for each image, giving a total of 3065 sentences.

(2) UCM-captions: Based on the UC Merced Land-Use dataset [6], this dataset features aerial images spanning 21 different scenes, with 100 images per scene. All images have a resolution of 256 × 256 pixels, with a resolution of 0.3048m. Each image is annotated with five descriptions, giving a total of 10,500 sentences.

(3) RSICD: This dataset is proposed by Lu et al. [7], this dataset includes 10,921 remote sensing images covering 30 different scenes, collected from Google Earth, Baidu Map, MapABC, and Tianditu. Each image has dimensions of 224 × 224 pixels. The annotators provided 24,333 unique sentences, and the total number of captions was expanded to 54,605 by duplicating existing sentences randomly, ensuring that each image has five captions.

(4) RSICN: This dataset is proposed by Zhou et al. [9]. Based on Sydney-captions, UCM-captions, and RSICD, they generated five Chinese captions for each remote sensing image, resulting in a total of 68,170 sentences. This dataset is specifically composed of three subsets: Sydney-CN, UCM-CN, and RSICD-CN.

(5) NWPU: The NWPU dataset contains 31,500 images and 157,500 sentences, proposed by Chen et al. [8]. It is currently the largest RSIC dataset, with each image having a resolution of 256x256 pixels and 0.2m resolution, covering 45 complex scenes. To evaluate whether the DIA method depends on high-quality human-annotated auxiliary language, we use machine translation to directly translate the English annotations of the NWPU dataset into Chinese as the auxiliary language.

For both the target language and auxiliary language datasets, we use 80 % of the data for training, 10 % for evaluation, and 10 % for testing. The split follows the publicly available scheme, ensuring a fair comparison of results.

4.2. Implementation details

In our implementation, we use the PyTorch framework and perform training and evaluation on an NVIDIA RTX 3090 GPU. The optimizer used is Adam, with an initial learning rate set to 1e-4. Training runs for up to 30 epochs, with a batch size of 32. We set $L_e = 3$ and $L_d = 3$. At the end of each epoch, we evaluate the model's performance on the validation set and select the model that achieves the highest CIDEr score on the validation data within the 30 epochs as the optimal model.

4.3. Evaluation metrics

To comprehensively evaluate the alignment between the captions generated by our model and the ground-truth, we adopt commonly used evaluation metrics in the image captioning task. These metrics effectively quantify the quality of the generated captions, ensuring the accuracy, fluency, and effectiveness of information conveyed in the model's outputs. BLEU-N (N = 1, 2, 3, 4) [35] is a series of evaluation method based on N-gram overlap, which calculates the score by comparing the word-order similarity between the generated and reference captions. A higher BLEU score indicates that the generated caption is similar to the reference caption in terms of word choice and syntactic structure, reflecting the accuracy and fluency of the content captions. METEOR [36] combines various linguistic features, such as synonym matching, morphological variations, and syntactic structures, to explicitly rank the quality of generated descriptions, allowing for a better evaluation of the linguistic diversity and precision of the generated content. Unlike BLEU and METEOR, ROUGE-L [37] focuses on evaluating recall ability. It measures the structural similarity between the generated and reference captions through the longest common subsequence, enabling the assessment of the completeness and coverage of the generated captions. CIDEr [38] is a consensus-based evaluation metric for image captioning, which measures the consistency between the generated caption and multiple reference captions, reflecting the quality of the generated captions in terms of both content and expression. These evaluation metrics are primarily used to quantify the similarity between the generated and reference captions, serving as standards for assessing the quality of the captions. By combining these evaluation metrics, we can conduct comprehensive evaluation of the model's performance. Higher scores indicate better alignment between the generated captions and the reference captions, reflecting better quality in the generated captions.

4.4. Comparison with existing methods

Effectiveness of DIA: To demonstrate the effectiveness of the proposed DIA method, we compare the performance of existing methods with ours DIA on Sydney, UCM, and RSICD. As shown in Table 1, the proposed DIA method demonstrates leading experimental performance on the Sydney dataset. DIA outperforms all comparison models in BLEU-1 (78.96 %) and BLEU-4 (62.65 %), particularly achieving 1.63 % improvement over the second-best model HCNet (61.02 %) in BLEU-4. In the CIDEr metric, DIA reaches 249.95 %, 2.81 % improvement over HCNet (247.14 %) and 2.08 % increase over SSCPNet (247.87 %), fully demonstrating its ability to better capture key information in images and generate high-quality captions.

As shown in Table 2, the proposed DIA method demonstrates leading experimental performance on the UCM-caption dataset, especially in the BLEU series metrics. DIA outperforms all comparison models with 88.91 % (BLEU-1), 83.79 % (BLEU-2), 79.41 % (BLEU-3), and 75.36 % (BLEU-4), particularly achieving 0.87 % improvement over the second-best model HCNet in BLEU-4, and 2.12 % advantage over SSCPNet (77.29 %) in BLEU-3. This indicates that DIA excels in capturing phrase-level semantic combinations and generating high-quality captions. In the METEOR metric, DIA achieves 48.87 %, surpassing HCNet (48.65 %) by 0.22 % and SSCPNet (48.22 %) by 0.65 %, highlighting improvements in lexical diversity and synonym replacement strategies. The excellent performance in ROUGE-L (83.98 %) further underscores DIA' strong capability in semantic coherence and key information coverage. In the CIDEr metric, DIA reaches 366.23 %, a 14.40 % improvement over HCNet (351.83 %) and 11.41 % increase over SSCPNet (354.82 %), fully

Table 1
Experimental results on the Sydney-captions dataset. Bold indicates the best results. Underlined value denote the second-best results. The * denotes results that we have re-implemented.

Method	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	METEOR	ROUGE-L↑	CIDEr↑
Soft Attention [7]	73.22	66.74	62.23	58.20	39.42	71.27	249.93
Hard Attention [7]	75.91	66.10	58.89	52.58	38.98	71.89	218.19
SAT(LAM) [29]	74.00	65.50	59.00	53.00	36.90	68.10	235.20
Adaptive(LAM) [29]	73.20	63.20	56.30	50.70	36.10	67.80	234.60
Word-sentence [25]	78.91	70.94	63.17	56.25	41.81	69.22	204.11
Structured attention [30]	77.95	70.19	63.92	58.61	39.54	72.99	237.91
RS-CapRet [31]	78.70	70.00	62.80	56.40	38.80	70.70	239.20
MLAT* [32]	73.62	65.75	59.57	54.34	36.51	66.41	217.13
MG* [33]	76.41	68.40	61.20	54.46	38.79	69.73	235.45
HCNet [28]	76.86	71.09	65.73	61.02	39.80	71.72	247.14
SSCPNet [34]	77.70	72.08	65.97	59.98	**40.23**	71.63	247.87
DIA(Ours)	**78.96**	**72.59**	**66.05**	**62.65**	38.66	**73.12**	**249.95**

Table 2
Experiments results on the UCM-captions dataset. Bold indicates the best results. Underlined value denote the second-best results. The * denotes results that we have re-implemented.

Method	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	METEOR↑	ROUGE-L↑	CIDEr↑
Soft Attention [7]	74.54	65.45	58.55	52.50	38.86	72.37	261.24
Hard Attention [7]	81.57	73.12	67.02	61.82	42.63	76.98	271.42
SAT(LAM) [29]	81.95	77.64	74.85	71.61	48.37	79.08	361.71
Adaptive(LAM) [29]	81.70	75.10	69.90	65.40	44.80	78.70	328.00
Word-sentence [25]	79.31	72.37	66.71	62.02	43.95	71.32	278.71
Recurrent-ATT [12]	85.18	79.25	74.32	69.76	45.71	80.72	338.87
Structured attention [30]	85.38	80.35	75.72	71.49	46.32	81.41	334.89
RS-CapRet [31]	84.30	77.90	72.20	67.00	47.20	81.70	354.80
PureT [39]	85.73	80.20	75.62	71.29	46.86	82.01	349.00
MLAT* [32]	85.80	79.77	74.51	69.58	45.56	80.79	319.57
MG* [33]	86.58	81.54	76.94	72.70	47.58	82.27	349.61
HCNet [28]	88.26	83.35	78.85	74.49	48.65	83.91	351.83
SSCPNet [34]	87.22	81.83	77.29	73.02	48.22	83.45	354.82
DIA(Ours)	**88.91**	**83.79**	**79.41**	**75.36**	**48.87**	**83.98**	**366.23**

Table 3
Experiments results on the RSICD dataset. Bold indicates the best results. Underlined value denote the second-best results. The * denotes results that we have re-implemented.

Method	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	METEOR↑	ROUGE-L↑	CIDEr↑
Soft Attention [7]	67.53	53.08	43.33	36.17	32.55	61.09	196.43
Hard Attention [7]	66.69	51.82	41.64	34.07	32.01	60.84	179.25
SAT(LAM) [29]	67.53	55.37	46.86	40.26	32.54	58.23	258.50
Adaptive(LAM) [29]	66.64	54.86	46.76	40.70	32.30	58.43	260.55
Word-Sentence [25]	72.40	58.61	49.33	42.50	31.97	62.60	206.29
Recurrent-ATT [12]	77.29	66.51	57.82	50.62	36.26	66.91	275.49
Structured attention [30]	70.16	56.14	46.48	39.34	32.91	57.06	170.31
RS-CapRet [31]	72.00	59.90	50.60	43.30	37.00	63.30	250.20
MLAT* [32]	75.50	68.27	59.81	52.78	36.66	68.96	224.30
MG* [33]	77.84	67.09	58.30	51.01	37.42	68.10	282.07
DIA(Ours)	**80.56**	**69.57**	**60.65**	**53.23**	**38.40**	**69.57**	**290.69**

demonstrating its ability to better capture key information and generate high-quality captions.

As shown in Table 3, the proposed DIA method demonstrates comprehensive leading experimental performance on the RSICD dataset. Particularly in the BLEU series metrics, DIA outperforms all comparison models with significant advantages: 80.56 % (BLEU-1), 69.57 % (BLEU-2), 60.65 % (BLEU-3), and 53.23 % (BLEU-4). The METEOR score (38.40 %) shows 0.98 % increase over MG, indicating that the model optimization in terms of lexical diversity and synonym replacement strategies has been effective. Notably, DIA achieves 290.69 % in the CIDEr metric, a performance gain of 8.62 % over MG (282.07 %), fully validating the high semantic relevance between the generated captions and human reference texts.

Practicality of DIA: To demonstrate the practicality of the proposed DIA method, we conducted experiments using machine translation to translate the captions in the NWPU dataset into Chinese as an auxiliary language. The experimental results are shown in Table 4. The proposed DIA method demonstrates significant performance advantages on the NWPU dataset, outperforming all comparison models. DIA shows improved performance across the BLEU series of metrics and surpasses the second-place model, SSCPNet, by 0.72 % in the METEOR, validating its significant improvements in vocabulary diversity and synonym replacement strategies. Notably, in the CIDEr metric, DIA achieves 211.51 %, improving by 7.59 % over the second-place model, SSCPNet (203.92 %), highlighting its strength in generating high-quality and diverse descriptive text. The results in Table 4 indicate that the DIA method also achieves performance improvements with the support of machine-translated auxiliary language, further demonstrating the robustness and wide practicality of our approach.

Cross-lingual transferability of DIA: To demonstrate the performance of our DIA after variation in the target language, we conduct extensive experiments with Chinese set as the target language on the RSICN-Sydney, RSICN-UCM, and RSICN-RSICD datase. Due to the lack of research on Chinese remote sensing image captioning, we replicate

Table 4
Experiments results on the NWPU dataset. Bold indicates the best results. Underlined value denote the second-best results. The * denotes results that we have re-implemented.

Method	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	METEOR↑	ROUGE-L↑	CIDEr↑
Multimodal [6]	72.50	60.30	51.80	45.50	33.60	59.10	117.90
Soft Attention [7]	73.10	60.90	52.50	46.20	33.90	59.90	113.60
Hard Attention [7]	73.30	61.00	52.70	46.40	34.00	60.00	110.30
MLCANet [8]	74.40	62.40	54.10	47.80	33.70	60.10	126.40
MLAT* [32]	85.27	76.74	70.07	64.79	43.27	74.96	185.56
PureT [39]	88.80	80.31	73.30	67.50	42.32	75.84	195.12
RS-CapRet [31]	87.10	78.70	71.70	65.60	43.60	77.60	192.90
SSCPNet [34]	88.01	80.81	74.70	69.65	46.46	79.14	203.92
DIA(Our)	**89.64**	**81.36**	**74.87**	**69.83**	**47.18**	**79.67**	**211.51**

Table 5
Experimental results on the RSICN dataset. Bold indicates the best results. Underlined value denote the second-best results. The * denotes results that we have re-implemented.

Method	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	METEOR↑	ROUGE-L↑	CIDEr↑
RSICN-Sydney							
BUTD* [40]	73.01	61.79	51.37	41.01	33.09	60.85	135.81
MLAT* [32]	68.72	55.12	46.55	40.25	30.84	61.05	168.72
MG* [33]	70.59	60.91	52.48	44.88	37.07	60.14	145.72
DIA(Ours)	**73.07**	**61.90**	**53.23**	**46.40**	**37.80**	**61.32**	**173.10**
RSICN-UCM							
BUTD* [40]	70.15	61.71	55.46	50.23	36.68	65.56	190.10
MLAT* [32]	76.71	67.96	60.76	54.87	37.79	**72.02**	233.55
MG* [33]	68.38	61.20	55.05	49.59	39.13	65.20	211.46
DIA(Ours)	**76.84**	**68.58**	**61.49**	**56.58**	**41.96**	68.30	**234.29**
RSICN-RSICD							
BUTD* [40]	60.61	48.42	39.06	31.97	29.74	49.08	149.39
MLAT* [32]	60.84	46.81	37.74	30.88	26.59	50.72	151.61
MG* [33]	61.66	49.59	40.69	33.94	31.76	49.66	157.36
DIA(Ours)	**64.20**	**53.71**	**45.18**	**38.39**	**31.43**	**53.20**	**176.35**

and adjust representative methods, including BUTD, MLAT, and MG. The experimental results are shown in Table 5. The results demonstrate that on the RSICN-Sydney dataset, DIA outperforms all other methods across all metrics, including BLEU-4 (46.40 %), METEOR (37.80 %), ROUGE-L (61.32 %), and CIDEr (173.10 %), with a particularly notable improvement in BLEU-4, surpassing MG (44.88 %) by 1.52 %. This indicates that DIA generates descriptions with better long-sequence matching and semantic coherence. On the RSICN-UCM dataset, DIA significantly outperforms other models in BLEU-3 (61.49 %), BLEU-4 (56.58 %), and METEOR (41.96 %). Notably, the METEOR score is 2.83 % higher than that of MG (39.13 %), which suggests stronger vocabulary selection and diversity generation capabilities, especially in complex scenarios. On the RSICN-RSICD dataset, DIA consistently surpasses the comparison models across all metrics, with significant improvements in BLEU-4 (38.39 %) and CIDEr (176.35 %), which are 4.45 % and 18.99 % higher than MG (33.94 %) and 157.36 %, respectively. This validates the robustness of the model in complex remote sensing scenarios, enabling the generation of more diverse descriptions that closely align with human references.

4.5. Ablation study

Effect of LiFE: The LiFE module consists of multiple convolutional blocks and SCP attention blocks, which collectively enhance the linguistic-irrelevant features of images while maintaining a strong alignment between visual features and textual content. When the LiFE module is introduced in isolation, all evaluation metrics (BLEU-1 through BLEU-4, METEOR, ROUGE-L, and CIDEr) show significant improvements over the baseline, as shown in Table 6. Specifically, the BLEU-4 score increases from 70.30 % to 72.56 % (+2.26 %), and the CIDEr score rises from 349.37 to 361.68 % (+12.31 %). These results indicate that the LiFE module effectively decouples the text dependencies in visual features, thereby enhancing the model's generalization ability and significantly improving the overall quality of the generated captions.

Effect of linguistic bridge: The Linguistic Bridge utilizes reparameterization techniques to establish an interaction pathway between ALG and TLG, with the aim of enabling ALG to learn more comprehensive language relationship modeling capabilities from TLG. After the introduction of the Linguistic Bridge, the model exhibits significant improvements across multiple performance metrics, as shown in Table 6, with notable advancements in BLEU-4 (74.65 %) and CIDEr (362.46 %). Specifically, BLEU-4 and CIDEr improve by 4.35 % and 13.09 %, respectively, compared to the baseline. These results indicate that the Linguistic Bridge allows ALG to capture more complex contextual relationships, thereby enhancing contextual coherence. This improvement leads to better structural and semantic consistency in the generated captions.

Synergistic effect of LiFE and linguistic bridge: When both the LiFE and Linguistic Bridge are introduced simultaneously, all evaluation metrics exhibit significant improvements (BLEU-4: 75.36 %, CIDEr: 366.23 %), with performance gains exceeding those achieved by each module individually, as shown in 6. Notably, the combined CIDEr gain

Table 6
Ablation results on the UCM dataset. The best results are shown in bold. LB represents Linguistic Bridge.

LiFE	LB	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
		85.97	80.12	74.99	70.30	46.33	81.17	349.37
✓		86.63	81.29	76.69	72.56	47.23	82.54	361.68
	✓	88.03	83.03	78.69	74.65	48.01	82.46	362.46
✓	✓	**88.91**	**83.79**	**79.41**	**75.36**	**48.87**	**83.74**	**366.23**

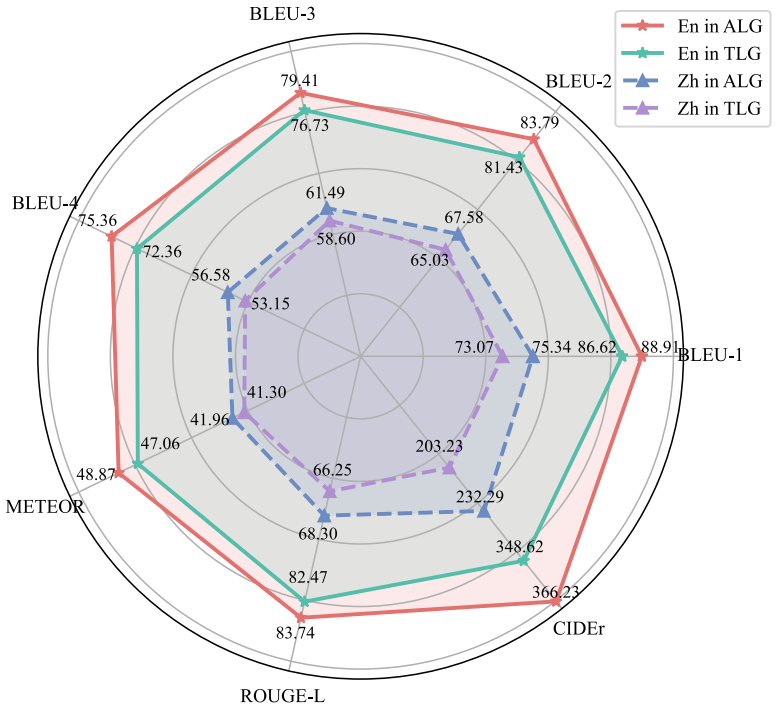

Fig. 3. Experiment results of target language and auxiliary language switching. The red solid line represents the test results of English text generated when English is the target language, the green solid line represents the test results of English text generated when English is the auxiliary language. The blue dashed line represents the test results of Chinese text generated when Chinese is the target language, and the purple dashed line represents the test results of Chinese text generated when Chinese is the auxiliary language. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(+16.86 %) surpasses the individual gains of LiFE (+12.31 %) and LB (+13.09 %). This indicates a complementary relationship between the two modules: the LiFE module focuses on enriching linguistic-irrelevant features and reducing language bias at the feature level, facilitating simultaneous learning by ALG and TLG, while the Linguistic Bridge module leverages TLG to enhance ALG's language modeling capabilities for caption generation. The synergistic effect of both modules comprehensively optimizes the accuracy and fluency of the generated target captions.

Effect of target language transformation: To investigate the impact of switching between the target and auxiliary languages on model performance, we conduct experiments on the UCM dataset by setting both English and Chinese as the target languages, keeping other conditions unchanged. The experimental results are shown in Fig. 3. The findings indicate that the choice of target language significantly affects the model's performance. When English is used as the target language, all metrics outperform the performance when English is set as the auxiliary language, especially CIDEr, which reached 366.23 %, an improvement of 17.61 %. Similarly, when Chinese is used as the target language, all metrics surpass those obtained when Chinese is set as the auxiliary language, with CIDEr reaching 232.29 %, an improvement of 29.06 %.

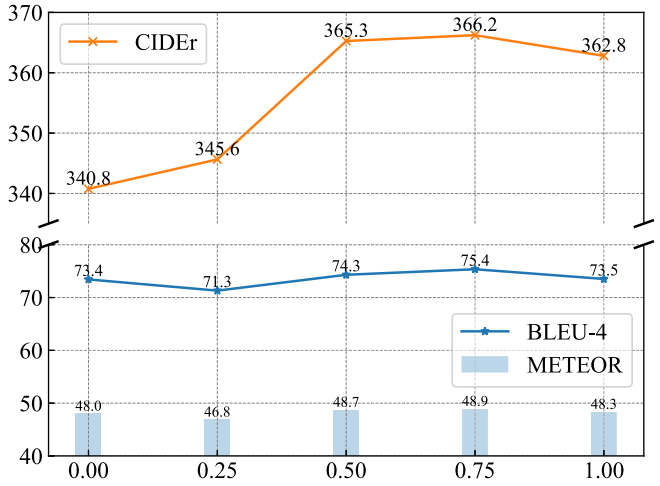

Fig. 4. Experimental results of λ hyperparameter variation. The horizontal axis represents λ, with a range of [0,1] and a step size of 0.25. The vertical axis represents metric scores (e.g., BLEU-4, METEOR, CIDEr).

4.6. Parametric study

The hyperparameter λ determines the extent to which ALG influences TLG by the Linguistic Bridge. To observe this effect, we conducted a hyperparameter selection experiment on the UCM-captions dataset, and the results are shown in Fig. 4. The hyperparameter values are sampled in increments of 0.25 within the range of 0 to 1. The results indicate that, when $\lambda = 0.75$, the model achieves optimal performance with BLEU-4 (75.4 %), METEOR (48.9 %), and CIDEr (366.2 %). Furthermore, the experiment shows that when the hyperparameter value changes, there is a significant fluctuation in the model's performance, particularly in CIDEr, which varies from 340.8 % to 366.2 % (+25.4 %).

4.7. Qualitative visualization analysis

To demonstrate the performance of the proposed DIA method, the qualitative results of the baseline model, the output of DIA, and the manually annotated ground truth (GT) captions are presented in Fig. 5. In the figure, "GT" represents the manually annotated captions of the image, "Base" represents the captions generated by the baseline model, and "Ours" denotes the captions generated by our method. For clar-

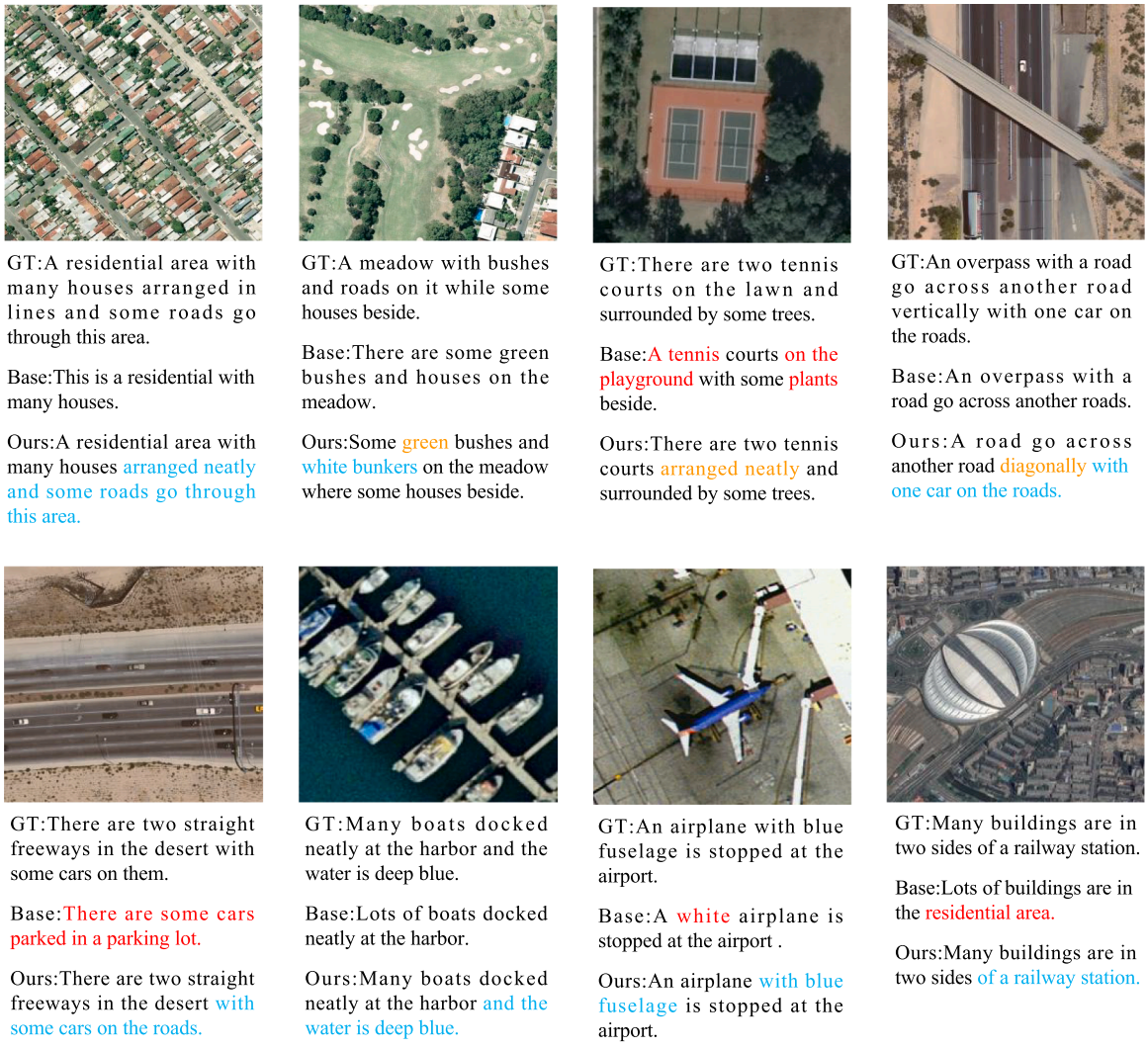

Fig. 5. Example captions generated by the baseline model(Base), our proposed DIA(Our), and the corresponding ground truth annotations(GT). The red words indicate inappropriate descriptions, while the blue words highlight the advantages of our method compared to the baseline model. The orange words emphasize the differences in accurate results between our method and the ground truth. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ity in comparison, we highlight inappropriate captions in red, the comprehensive improvements made by "Ours" relative to the baseline in blue, and the additions or optimizations relative to "GT" in orange. It is clear that the captions generated by our method are more aligned with the image content and significantly outperform those produced by the baseline model. Specifically, for the third image in the first row, our method accurately identifies the number of tennis courts as two and recognizes the trees growing around them. In contrast, the baseline model incorrectly identifies the number of courts and misclassifies the grass as a playground. Similarly, the baseline model misidentifies the freeway as a parking lot and the blue airplane as a white one. In comparison, our method not only correctly identifies the objects and their quantities in the image, but also captures their spatial relationships, such as "houses arranged neatly" and "with one car on the roads". More impressively, our model also provides more comprehensive information relative to "GT". For example, in the second image of the first row, "Ours" accurately describes the color of the "bushes" as "green" and the color of the "bunkers" as "white". In the fourth image of the first row, "Ours" describes the spatial relationship between the two roads as "diagonally", which is a more accurate description compared to the "vertically" relationship in "GT". These results demonstrate that "Ours" not only surpasses the baseline model in accuracy, but also offers more detailed and spatially coherent captions.

5. Conclusion

This paper proposes an enhanced RSIC framework leveraging auxiliary language data to enhance both the precision and language generation capabilities of the model. By incorporating the Linguistic-irrelevant Feature Enhancement (LiFE) module, it effectively mitigates language bias in visual features. Additionally, the Linguistic Bridge facilitates TLG's acquisition of language modeling capabilities from ALG. Experimental results indicate substantial improvements across multiple datasets, particularly in evaluation metrics such as BLEU and CIDEr, highlighting the superior image-text interaction capabilities of this method compared to existing approaches. Furthermore, experiments on machine translation datasets show that the method exhibits strong generalization ability. While the DIA performs well in parallel corpora (i.e., cross-lingual texts are semantically aligned with the same visual features), its effectiveness in non-parallel corpora remains an area that requires further investigation. In parallel corpora, direct alignment information is provided during the training of language models, enabling

the model to establish accurate mapping relationships between auxiliary languages and target languages. However, in non-parallel corpora, the lack of clear semantic alignment makes the challenges more complex. Future research will focus on improving the adaptability of models in scenarios involving non-parallel corpora. Additionally, exploring the use of multimodal information (i.e., visual, auditory, or other types of cross-modal data) for auxiliary training is a potential approach to enhance model performance.

CRediT authorship contribution statement

Tao Yang: Validation, Formal analysis, Conceptualization, Writing – review & editing, Visualization, Methodology, Data curation, Writing – original draft; **Qing Zhou:** Visualization, Project administration, Formal analysis, Methodology, Data curation, Writing – review & editing, Validation, Conceptualization; **Qi Wang:** Methodology, Supervision.

Data availability

Data will be made available on request.

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Q. Zhao, S. Jia, Y. Li, Hyperspectral remote sensing image classification based on tighter random projection with minimal intra-class variance algorithm, Pattern Recognit. 111 (2021) 107635.

[2] R. Damalla, P.A. Bendre, R. Datla, V. Chalavadi, et al., TransRefine: transformer-augmented feature refinement for zero-shot scene classification in remote sensing images, Pattern Recognit. 162 (2025) 111406.

[3] Y. Li, X. Zhang, X. Cheng, X. Tang, L. Jiao, Learning consensus-aware semantic knowledge for remote sensing image captioning, Pattern Recognit. 145 (2024) 109893.

[4] Z. Yang, B. Han, X. Gao, Z.-H. Zhan, Eye-movement-prompted large image captioning model, Pattern Recognit. 159 (2025) 111097.

[5] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, Dense semantic embedding network for image captioning, Pattern Recognit. 90 (2019) 285–296.

[6] B. Qu, X. Li, D. Tao, X. Lu, Deep semantic understanding of high resolution remote sensing image, in: Proc. 2016 International Conference on Computer, Information and Telecommunication Systems, 2016, pp. 1–5.

[7] X. Lu, B. Wang, X. Zheng, X. Li, Exploring models and data for remote sensing image caption generation, IEEE Trans. Geosci. Remote Sens. 56 (2017) 2183–2195.

[8] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, Z. Wang, NWPU-Captions dataset and MLCA-Net for remote sensing image captioning, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–19.

[9] Q. Zhou, T. Yang, J. Gao, W. Ni, J. Wu, Q. Wang, A benchmark for multi-lingual vision-language learning in remote sensing image captioning, arXivpreprintarXiv:2503.04592 (2025).

[10] H. Huang, Z. Shao, Q. Cheng, X. Wu, Multi-scale feature fusion network for remote sensing image captioning, in: Proc. 2023 30th International Conference on Geoinformatics, 2023, pp. 1–4.

[11] Y. Li, S. Fang, L. Jiao, R. Liu, R. Shang, A multi-level attention model for remote sensing image captions, Remote Sens. 12 (2020) 939.

[12] Y. Li, X. Zhang, J. Gu, C. Li, X. Wang, X. Tang, L. Jiao, Recurrent attention and semantic gate for remote sensing image captioning, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–16.

[13] X. Ye, S. Wang, Y. Gu, J. Wang, R. Wang, B. Hou, F. Giunchiglia, L. Jiao, A joint-training two-stage method for remote sensing image captioning, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–16.

[14] Y. Zhang, X. Ding, K. Gong, Y. Ge, Y. Shan, X. Yue, Multimodal pathway: improve transformers with irrelevant data from other modalities, in: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2024, pp. 6108–6117.

[15] H. Zhang, Y. Xu, S. Huang, X. Li, Data augmentation of contrastive learning is estimating positive-incentive noise, arXivpreprintarXiv:2408.09929 (2024).

[16] H. Zhang, S. Huang, X. Li, Variational positive-incentive noise: how noise benefits models, arXivpreprintarXiv:2306.07651 (2023).

[17] Q. Wang, W. Huang, X. Zhang, X. Li, GLCM: global–local captioning model for remote sensing image captioning, IEEE Trans. Cybern. 53 (2022) 6910–6922.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 770–778.

[19] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, Learning transferable visual models from natural language supervision, in: Proc. International Conf. on Machine Learning, 2021, pp. 8748–8763.

[20] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015 (2015) 3156–3164.

[21] M. Cornia, L. Baraldi, R. Cucchiara, Show, control and tell: a framework for generating controllable and grounded captions, in: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019, pp. 8307–8316.

[22] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, T. Mei, Semantic-conditional diffusion networks for image captioning, in: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2023, pp. 23359–23368.

[23] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation, in: Proc. International Conf. on Machine Learning, 2022, pp. 12888–12900.

[24] S. Huang, H. Zhang, X. Li, Enhance vision-language alignment with noise, in: Proc. AAAI Conf. on Artificial Intelligence, 2025, pp. 17449–17457.

[25] Q. Wang, W. Huang, X. Zhang, X. Li, Word–sentence framework for remote sensing image captioning, IEEE Trans. Geosci. Remote Sens. 59 (2020) 10532–10543.

[26] X. Li, X. Zhang, W. Huang, Q. Wang, Truncation cross entropy loss for remote sensing image captioning, IEEE Trans. Geosci. Remote Sens. 59 (2020) 5246–5257.

[27] Q. Zhou, J. Gao, Y. Yuan, Q. Wang, Single-stream extractor network with contrastive pre-training for remote sensing change captioning, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–14.

[28] Z. Yang, Q. Li, Y. Yuan, Q. Wang, HCNet: hierarchical feature aggregation and cross-modal feature alignment for remote sensing image captioning, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–11.

[29] Z. Zhang, W. Diao, W. Zhang, M. Yan, X. Gao, X. Sun, LAM: remote sensing image captioning with label-attention mechanism, Remote Sens. 11 (2019) 2349.

[30] R. Zhao, Z. Shi, Z. Zou, High-resolution remote sensing image captioning based on structured attention, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–14.

[31] J.D. Silva, J. Magalhães, D. Tuia, B. Martins, Large language models for captioning and retrieving remote sensing images, arXivpreprintarXiv:2402.06475 (2024).

[32] C. Liu, R. Zhao, Z. Shi, Remote-sensing image captioning based on multilayer aggregated transformer, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.

[33] L. Meng, J. Wang, R. Meng, Y. Yang, L. Xiao, A multiscale grouping transformer with CLIP latents for remote sensing image captioning, IEEE Trans. Geosci. Remote Sens. 11 (2024) 1–12.

[34] Q. Wang, Z. Yang, W. Ni, J. Wu, Q. Li, Semantic-spatial collaborative perception network for remote sensing image captioning, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–12.

[35] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proc. 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[36] S. Banerjee, A. Lavie, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, in: Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.

[37] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Proc. of Text Summarization Branches Out, 2004, pp. 74–81.

[38] R. Vedantam, L.C. Zitnick, D. Parikh, CIDEr: consensus-based image description evaluation, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.

[39] Y. Wang, J. Xu, Y. Sun, End-to-end transformer based model for image captioning, in: Proc. AAAI Conf. on Artificial Intelligence, 2022, pp. 2585–2594.

[40] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.