Pattern Recognition Letters 186 (2024) 164–169

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Discrete diffusion models with Refined Language-Image Pre-trained representations for remote sensing image captioning ☆

Guannan Leng [a], Yu-Jie Xiong [a,*], Chunping Qiu [b,*], Congzhou Guo [b]

[a] School of Electric and Electronic Engineering, Shanghai University of Engineering Science, ShangHai, 201600, China
[b] Institute of Geospatial Information, Information Engineering University, Zhengzhou, 450001, China

ARTICLE INFO

Editor: Gangyi Jiang

Keywords:
Discrete diffusion model
Contrastive Language-Image Pre-training
Transformer
Remote sensing image captioning

ABSTRACT

RS image captioning (RSIC) utilizes natural language to provide a description of image content, assisting in the comprehension of object properties and relationships. Nonetheless, RS images are characterized by variations in object scales, distributions, and quantities, which make it challenging to obtain global semantic information and object connections. To enhance the accuracy of captions produced from RS images, this paper proposes a novel method referred to as Discrete Diffusion Models with Refined Language-Image Pre-trained representations (DDM-RLIP), leveraging an advanced discrete diffusion model (DDM) for nosing and denoising text tokens. DDM-RLIP is based on an advanced DDM-based method designed for natural pictures. The primary approach for refining image representations involves fine-tuning a CLIP image encoder on RS images, followed by adapting the transformer with an additional attention module to focus on crucial image regions and relevant words. Furthermore, experiments were conducted on three datasets, Sydney-Captions, UCM-Captions, and NWPU-Captions, and the results demonstrated the superior performance of the proposed method compared to conventional autoregressive models. On the NWPU-Captions dataset, the CIDEr score improved from 116.4 to 197.7, further validating the efficacy and potential of DDM-RLIP. The implementation codes for our approach DDM-RLIP are available at https://github.com/Leng-bingo/DDM-RLIP.

1. Introduction

Remote sensing image captioning (RSIC) plays an important role in geographic information retrieval, disaster monitoring, and image understanding. Yet, it is also a challenging task that requires describing rich image content using natural language, compared with image object detection, classification, and segmentation tasks.

Motivated by the captioning algorithms in the Computer Vision (CV) field, many researchers have developed deep learning methods for RSIC by utilizing classical encoder–decoder networks coupled with attention modules. Qu et al. [1] were the first to propose a CNN-RNN architecture for RSIC and constructed the UCM-Captions and Sydney-Captions datasets. To further enrich the dataset, Cheng et al. [2] annotated the NWPU-RESISC45 dataset [3], which is currently the largest RSIC dataset in terms of the number of images, captions, and categories. These datasets form the foundation for subsequent remote sensing image captioning tasks.

Zhao et al. [4] proposed the structured attention approach, which uses selective search to find segmentation proposals and multiplies them with CNN features to obtain structured features that well reflect local region characteristics. However, when there are a large number of objects in the image, the selective search cannot work well. Wang et al. [5] proposed the multiscale multi interaction network to account for the differences between natural images and RS images. Zhang et al. [6] proposed a novel attention mechanism, namely the label-attention mechanism, which mainly uses the label information of RS images to guide the generation of image descriptions. Zhang et al. [7] proposed a multi-source interactive stair attention mechanism, which utilizes previous semantic vectors as queries and obtains the next word vector using attention on region features. Shen et al. [8] propose to finetune the CNN jointly with the variational autoencoder, and use a transformer to generate the text description with both spatial and semantic features. By employing CNN to extract image features, it becomes possible to delve deeper into the internal information of the images. In our paper, we use Vision Transformer (ViT) for feature extraction, which allows the extraction of higher-quality features, thereby providing a solid foundation for image captioning.

☆ This work is supported by the National Natural Science Foundation of China under Grant Nos. 62006150 and 42201513, and the Science and Technology Commission of Shanghai Municipality, China under Grant No. 21DZ2203100.

* Corresponding author.
E-mail addresses: 805477481@qq.com (G. Leng), xiong@sues.edu.cn (Y. Xiong), chunping.qiu@aliyun.com (C. Qiu), czguo0618@sina.cn (C. Guo).

https://doi.org/10.1016/j.patrec.2024.09.019
Received 22 September 2023; Received in revised form 9 February 2024; Accepted 26 September 2024
Available online 5 October 2024

Despite the promising performance of the above autoregressive methods, there are still some improvements due to the differences between RS images and natural pictures. Text generation methods mostly adopt the autoregressive way (AR) that generates the output tokens one by one. Such a way is able to capture the sequential dependency relations among tokens, but would be time-consuming when generating long texts. Thus, non-autoregressive (NAR) generation methods, which generate all tokens in parallel and greatly reduce the inference latency, have been proposed [9]. Moreover, text generation requires relatively less computational and resource-intensive compared to image generation. This opens up more possibilities for parallel computing. However, NAR models generally underperform AR ones on text generation accuracy, since the token dependency relations cannot be well captured by the parallel generation. To narrow the performance gap, previous works have proposed various improvement techniques for NAR methods, e.g., knowledge distillation [10] and large-scale pre-training [11]. Our work is inspired by the advanced diffusion models which have achieved astonishing performance in many vision tasks [12]. Compared to the currently dominant autoregressive methods, the discrete diffusion model (DDM) is more flexible, as it can predict multiple tokens at once. We are aiming to explore this new branch of discrete DDM-based captioning methods for RS images.

It is, however, not possible to directly apply the powerful diffusion-based methods for our tasks. This is because RS images, taken from a high-altitude perspective, have significant differences from natural images, such as larger scale and imbalanced foreground–background ratios. Therefore, our goal is to improve the model architecture and investigate how to enable DDM to generate accurate descriptions of RS images.

Inspired by a DDM-based image caption method for natural pictures [13], we tried to integrate DDM with RS-specific representations by fine-tuning language-image pre-trained models on RS datasets, followed by an attention module to encourage the features to focus on the descriptions-related regions. This way, the DDM is coupled with specifically refined RS language-image pre-trained representations, i.e., DDM-RLIP, which can be enhanced to generate captions for RS images. We instanced CLIP to extract pre-trained representations, and our main contributions are as follows.

- We are the first to apply DDM on the challenging task of RSIC, and our results demonstrate its comparative performance to the traditional autoregressive-based methods.
- To enhance the feature's representation ability and align it more closely with the text descriptions, we fine-tuned pre-trained language-image models on RS datasets and employed an attention module to refine the features.
- Our proposed model achieved superior performance compared to state-of-the-art (SOTA) models on three challenging RSIC datasets. Additionally, we conducted an investigation on the effects of our adaptations.

2. Related work

Image captioning serves as a crucial link between images and text, making it a prominent research area in the field of artificial intelligence. In recent years, some research [14–16] efforts begin to focus on remote sensing image captioning tasks, aiming to introduce natural language generation into the field of remote sensing image processing. Remote sensing image captioning is characterized by unique aspects, including geographical information, spatial relationships, and domain-specific terminology within the images. It has provided significant assistance to the remote sensing image domain, as image captions enhance people's understanding of remote sensing images and improve the quality of descriptions. Furthermore, inspired by the advancements in powerful diffusion models, there have been new developments in the field of computer vision as well.

2.1. Remote sensing image captioning

The size of objects in remote sensing image varies, which can lead to omissions during feature extraction. To address this problem, the denoising-based multi-scale feature fusion network [14] first filters noise in the image with two fully connected layers, following which the encoder output is obtained by fusing the outputs of CNN features at three scales. The Multi-Level Attention Model [15] uses three attention blocks to represent attention to different areas of the image, attention to different words, and attention to vision and semantics.

Considering the large-scale variation and richness of objects in remote sensing images, researchers have attempted to extract global semantic information to facilitate word generation. The mean pooling operation is a common way to capture such information [1,16]. VRTMM [8] (Variational autoencoder and reinforcement-learning-based two-stage multitask learning model) uses the output of VGG16 [17] (visual geometry group 16) with a soft-max layer to represent the semantic features. RASG [18] (recurrent attention and semantic gate) utilizes a recurrent attention mechanism and semantic gate to generate better image features corresponding to the current word.

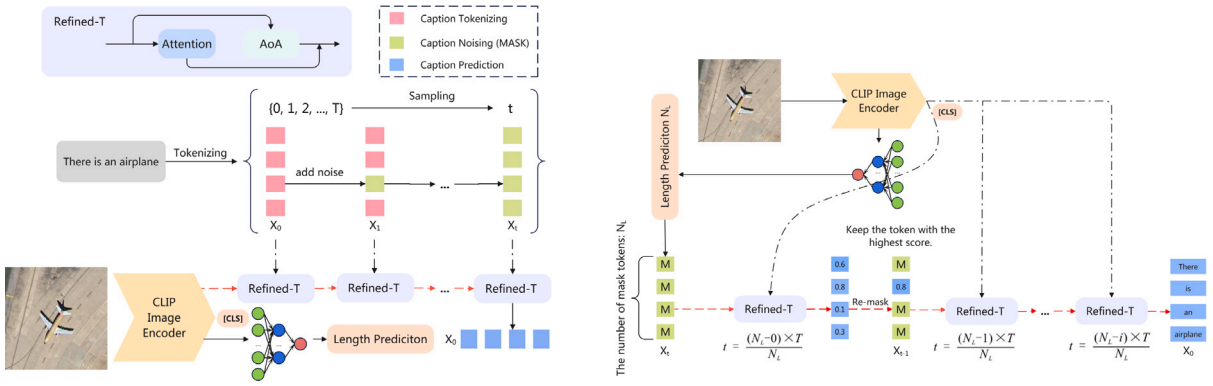
Considering the relationship between foreground and background of remote sensing images, Zheng et al. [19] introduced a foreground-aware relationship framework that explicitly uses foreground modeling to perform remote sensing target segmentation. The foreground-scene relationship module learns the symbiotic relationship between the foreground and the scene, which associates context to enhance foreground features, thereby reducing false alarms. Zhang et al. [20] handled the classification of hyperspectral remote sensing images with a dense network. They collected multiscale features from different layers of all the network, and these features with scale information were used for classification guidance.

Moreover, an attention module was leveraged to preserve the aspect ratio of the object. Yu and Koltun [21] introduced a novel convolutional module that can mix multiscale contextual information without loss of resolution, and this module could be inserted into the existing structure at any resolution. Qiu et al. [22] proposed the gated multiscale module to integrate features of different levels. Cheng et al. [23] proposed a new discriminative loss function to address the problems of intraclass diversity and interclass similarity in remote sensing images.

2.2. Diffusion-based image captioning

The original work introducing the Denoising Diffusion Probabilistic Models (DDPM) was conducted by Ho et al. [12]. DDPM comprises two primary phases: firstly, the forward process, also termed the diffusion process, gradually alters the original image, transitioning it into a fully noisy image. Secondly, there is the inverse process, commonly known as the denoising process, which systematically reverts the noisy image to its initial state. Regardless of the chosen direction (whether forward or backward), this process is conceptualized as a parameterized Markov chain. To accelerate the generation process, Song et al. [24] introduced the Denoising Diffusion Implicit Models (DDIM) . Notably, DDIM shares an identical training objective with DDPM but does not enforce the Markov chain constraint on the diffusion process, enabling the utilization of smaller sampling steps during the generation phase.

Incorporating the diffusion model into the realm of image captioning, Xu [25] presented an innovative approach. This method seamlessly integrates the strengths of the diffusion model with CLIP, circumventing the necessity for a distinct alignment process between images and text. The CLIP model's output serves as the initial state, embarking on a stepwise diffusion journey across multiple iterations. During each iteration, the diffusion model introduces a degree of randomness, generating text and subsequently refining the state based on this generated text and the current state. Austin et al. [26] proposed a structured denoising diffusion model based on a discrete-space formulation. Li et al. [27] introduced a text generation model grounded in the principles of the



(a) The process of training. (b) The process of inference.

Fig. 1. Overview of the proposed DDM-RLIP, consisting of a DDM sub-network, an image encoder such as CLIP-based ones, and a transformer-based sub-network. (a) During training, the caption is tokenized and gradually converted to $[mask]$ by adding noise that depends on the sampled step t. Then, these noisy tokens are fed into a transformer model for clean text token prediction, together with image features. The predicted tokens are used for loss calculating together with GT. The $[CLS]$ token of the CLIP model is used to predict the length N_L of the caption. (b) During inference, all $[mask]$ tokens X_T is the input and the caption length N_L is first predicted. For each step t, we have three inputs for the transformer's Adaptive LayerNorm layer: t depending on N_L and the total noise length T, image features, and previous text tokens X_{t-1}. We retain the token with the highest score each time and gradually infer the initial caption X_0.

diffusion model, aiming for enhanced controllability. Their approach addresses the challenges associated with minimizing extraneous content and improving fluency. It initiates with a random noise vector as input and orchestrates a diffusion process to systematically craft text imbued with precise and targeted semantics.

3. Method

The overview of DDM-RLIP is illustrated in Fig. 1, consisting of training and inference process. The training process gradually transforms each text token into a mask token by applying noise with a certain probability. Then, the noisy tokens, combined with image features from pre-trained CLIP encoders, are used to predict clean tokens with a refined transformer. The $[CLS]$ token (from CLIP) is used to predict the corresponding caption text length N_L using MLP. Predicted and ground truth (GT) of caption tokens are used for loss calculation. After training, the inference process removes noise from a sequence of all $[mask]$ tokens based on the confidence level in a step-by-step manner, using the text length as the diffusion step T. For each step t, we have three inputs for the transformer's Adaptive LayerNorm layer: t depending on N_L and the total noise length T, image features, and previous text tokens X_{t-1}. To enhance the correlation between image and text features for RS image description, the utilized transformer is enhanced with an additional attention module, and the pre-trained CLIP image encoders are fine-tuned before freezing the parameters in the training process.

3.1. Noising and denoising process

In the nosing process, DDM randomly and stochastically transforms the text tokens into all $[mask]$ tokens within T steps at the level of text.

We represent each token in the caption as a discrete state x, and denote the noisy version token at tth diffusion step as x_t, and there are a total of T_t steps. The noising process at each step depends on whether x_{t-1} is a $[mask]$ token. When it is not, token x_{t-1} could be replaced by a special $[mask]$ token with a probability of ϵ_t, or remain unchanged with a probability of η_t, or could be replaced by another token from the vocabulary except $[mask]$ with a probability of $\theta_t = 1 - \epsilon_t - \eta_t$, namely:

$$p\left(x_t \mid x_{t-1}\right) = \begin{cases} \eta_t, x_t = x_{t-1} \\ \epsilon_t, x_t = [mask] \\ 1 - \epsilon_t - \eta_t, \text{ otherwise} \end{cases} \tag{1}$$

When x_{t-1} is already a $[mask]$ token, the transition probability from Step $t-1$ to t is:

$$p\left(x_t \mid x_{t-1}\right) = \begin{cases} 1, x_t = [mask] \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

The aforementioned transition method will eventually convert the caption tokens into a sequence of special $[mask]$.

The denoising process gradually removes noise to restore the initial caption token. We start from an all $[mask]$ sequence and use a transformer network for inverse projection, i.e., $p(x_{t-1} \mid x_t, y)$, where y is the image features. In addition, we use a sine function to encode the time step t as the position encoding:

$$p = t * \text{step}_{\text{scale}}/T, \tag{3}$$

$$\text{PE}_i = \begin{cases} \sin\left(p/10000^{2i/d_{\text{model}}}\right), i < d_{\text{model}}/2 \\ \cos\left(p/10000^{2i/d_{\text{model}}}\right), i \geq d_{\text{model}}/2 \end{cases} \tag{4}$$

where $\text{step}_{\text{scale}}$ is the wavelength, and d_{model} is the hidden dimension.

3.2. Training and inference

The DDM uses masking at the sentence level in the noising process and directly predict the initial text token instead of noise distribution [13]. We transform the process from Step t to $t-1$ into directly computing the initial text, setting Step to 0, and training with image features. The process from Step $t-1$ to t can be described as follows:

$$q(x_t \mid x_{t-1}) = \text{Cat}(x_t; p = x_{t-1}\boldsymbol{Z}_t) \tag{5}$$

where $\text{Cat}(x; p)$ is a categorical distribution over the one-hot row vector x. And the transition vector $\boldsymbol{Z}$ is: $[\boldsymbol{Z}_t]_{ij} = q(x_t = j \mid x_{t-1} = i)$. The process from Step 0 to Step t is:

$$q\left(x_t \mid x_0\right) = \text{Cat}\left(x_t; p = x_0\overline{\boldsymbol{Z}}_t\right), \quad \overline{\boldsymbol{Z}}_t = \boldsymbol{Z}_1 \ldots \boldsymbol{Z}_t \tag{6}$$

In summary, we can train the network by adding noise to the initial text token in the nosing processing, and combining the all $[mask]$ tokens with image features to predict clean tokens. The predicted tokens and GT are used for loss calculation.

During the inference, we first use the $[CLS]$ token from the CLIP encoder as a predicted feature to predict the text length T corresponding to the image. Then, starting from the $[mask]$ token x_T with length T, we directly predict x_0 using the trained denoising network and the image representations y: $p_\theta(x_0 \mid x_t, y)$. Subsequently, we obtain x_{t-1} by adding noise on the predicted raw text $\hat{x_0}$ through a Markov chain. After T steps, we can gradually recover the original text x_0.

3.3. Refined Language–Image Pre-trained (RLIP) representations

To obtain accurate descriptions for RS images, we empirically found that improving the representations is important. Thus, we proposed to refine the pre-trained representations in two different ways. One is re-sorting to a refined transformer with two distinct attention modules to encourage the representations to be more relevant to the text contents. This module can be trained together with the DDM sub-network. The other refinement is related to the pre-trained image encoders, which are not easy to train from scratch. The performance of CLIP-like pre-trained encoders is very powerful and can be used in a zero-shot manner. However, significant differences between RS images and natural images motivated us to fine-tune CLIP-encoders with RS data.

Adapted Transformer with an attention-on-attention module. The traditional attention module in transformer $f_{att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ operates on Queries, Keys, and Values. It first computes similarity scores between $\boldsymbol{Q}$ and $\boldsymbol{K}$, and then performs a weighted average vector calculation with the scores and $\boldsymbol{V}$, which can be expressed as:

$$\hat{\boldsymbol{v}}_i = f_{att}(\boldsymbol{q}_i, \boldsymbol{k}_j, \boldsymbol{v}_j) \tag{7}$$

where $\boldsymbol{q}_i \in \boldsymbol{Q}$ is the ith query, $\boldsymbol{k}_j \in \boldsymbol{K}$ and $\boldsymbol{v}_j \in \boldsymbol{V}$ are the jth key/value pair; f_{sim} is a function that computes the similarity score of each $\boldsymbol{k}_j$ and $\boldsymbol{q}_i$; and $\hat{\boldsymbol{v}}_i$ is the attended vector for the query $\boldsymbol{q}_i$.

Our utilized transformer consists of adapted Attention modules that measure the correlation between the attention result in $\hat{\boldsymbol{v}}_i$, calculated using Eq. (7), and the query $\boldsymbol{q}$, as shown in Fig. 2. This module generates a "information vector" and an "attention vector" through two independent linear transformations, both of which are based on the attention results and the current query:

$$\boldsymbol{i} = W_q^i \boldsymbol{q} + W_v^i \hat{\boldsymbol{v}} + b^i \tag{8}$$

$$\boldsymbol{g} = \sigma\left(W_q^g \boldsymbol{q} + W_v^g \hat{\boldsymbol{v}} + b^g\right) \tag{9}$$

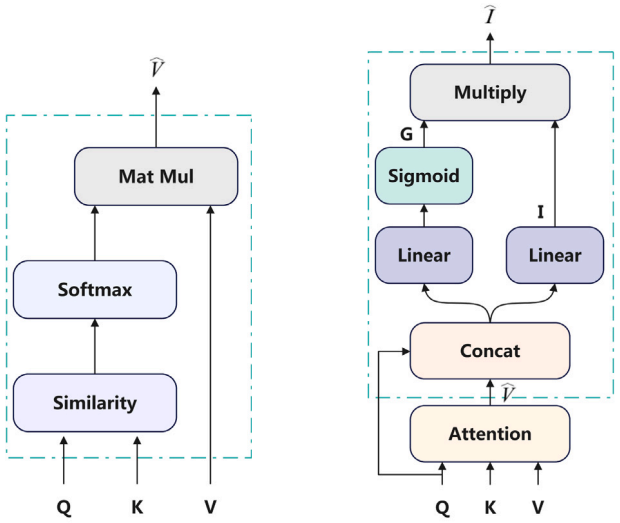
where $W_q^i, W_v^i, W_q^g, W_v^g \in R^{D\times D}, b^i, b^g \in R^D$, and D is the dimension of $\boldsymbol{q}$ and $\boldsymbol{v}$; σ denotes the sigmoid activation function. Then, there is a additional attention matching the "information vector" and the "attention vector", obtaining the participating information $\hat{\boldsymbol{i}}$, with element-wise multiplication $\odot$:

$$\hat{\boldsymbol{i}} = \boldsymbol{g} \odot \boldsymbol{i} \tag{10}$$

The output of the second attention mechanism is combined with the output of the primary attention mechanism to process the predicted captions and the image embeddings.

Fine-tuning CLIP encoders with RS datasets. Contrastive learning is utilized as the original training process of CLIP. Image–text pairs are fed into the text encoder and image encoder respectively, the embeddings of which are used to calculate similarities and cross-entropy losses. We fine-tuned CLIP encoders on RS image scene classification datasets and constructed text by using the categories of RS images, such as "An aerial photograph of {category}". We fine-tuned for 10 epochs using a Adam optimizer with a learning rate of 1×10^{-6}.

To mitigate the overfitting resulting from the limited size of the available datasets, we incorporated data augmentation techniques for both images and text. Our approach included image augmentation methods such as random cropping, random resizing, and horizontal and vertical flipping, as well as text augmentation techniques such as back-translation. The back-translation method involved translating the existing captions into Chinese, French, Italian, and Portuguese and then converting them back into English.



(a) Attention (b) Attention on Attention

Fig. 2. (a) The attention module generates some weighted average $\hat{\boldsymbol{V}}$ based on the similarity scores between $\boldsymbol{Q}$ and $\boldsymbol{K}$; (b) The Attention-on-Attention first generates a "information vector" $\boldsymbol{I}$ and an "attention vector" $\boldsymbol{G}$, followed by an element-wise multiplication.

Table 1
Comparative results on the UCM-Captions dataset.

Model	Backbone	B@1	B@4	M	R	C
SAT [28]	VGG+RNN	79.9	62.4	41.7	74.4	310.4
FC-ATT [29]	VGG+LSTM	81.3	63.5	41.7	75.0	299.9
SM-ATT [29]	VGG+LSTM	81.5	64.5	42.4	76.3	318.6
LAM [6]	VGG+LSTM	81.9	71.6	48.3	79.0	361.7
MLA [15]	ResNet+LSTM	84.0	69.1	53.3	81.9	311.9
Struc-ATT [4]	ResNet+LSTM	85.3	71.4	46.3	81.4	334.8
SCST [7]	VGG+LSTM	87.2	70.3	46.5	82.5	371.2
RASG [18]	VGG+LSTM	85.1	69.7	45.7	80.7	333.8
VRTMM [8]	VGG+Transformer	83.9	68.2	45.2	80.2	349.4
JTTS [30]	ResNet+LSTM	86.9	73.7	49.0	83.6	371.0
MMN [5]	ResNet+LSTM	83.0	65.1	45.3	78.5	338.1
DDM-RLIP	Vit-b/16+DDM	**89.7**	**77.2**	**48.9**	**85.1**	**372.6**

B@1, B@4, M, R and C denote BLEU-1, BLEU-4, METEOR, ROUGE_L and Cider respectively.

Table 2
Comparative results on the Sydney-Captions dataset.

Model	Backbone	B@1	B@4	M	R	C
SAT [28]	VGG+RNN	79.0	54.7	39.2	72.0	220.1
FC-ATT [29]	VGG+LSTM	80.7	55.4	40.9	71.1	220.3
SM-ATT [29]	VGG+LSTM	81.4	58.0	41.1	71.9	230.2
LAM [6]	VGG+LSTM	74.0	53.0	36.8	68.1	235.1
MLA [15]	ResNet+LSTM	81.5	**61.3**	**45.6**	70.6	199.2
Struc-ATT [4]	ResNet+LSTM	77.9	58.6	39.5	72.9	237.9
SCST [7]	VGG+LSTM	76.4	57.2	39.4	71.7	281.2
RASG [18]	VGG+LSTM	80.0	59.0	39.0	72.1	263.1
VRTMM [8]	VGG+Transformer	74.4	56.9	37.4	66.9	252.8
JTTS [30]	ResNet+LSTM	84.9	64.9	44.5	76.6	280.1
MMN [5]	ResNet+LSTM	**84.2**	60.1	42.1	60.1	**285.1**
DDM-RLIP	Vit-b/16+DDM	79.7	59.5	41.6	**74.4**	274.5

4. Datasets and experimental setup

This study evaluates DDM-RLIP's performance using three popular datasets: Sydney-Captions (613 images and 3065 sentences) [1], UCM-Captions (2100 images and 10,500 sentences) [1], and NWPU-Captions (31,500 images and 157,500 sentences) [2]. The datasets were randomly split into 8:1:1 ratios for training, validation, and testing, and the evaluation metrics included BLEU, METEOR, ROUGE_L, and

Table 3
Ablation results on all three datasets.

	Sydney-Captions					UCM-Captions					NWPU-Captions				
	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr
DDM+CLIP	79.6	**60.8**	40.7	73.4	270.8	87.8	74.3	47.1	84.0	364.8	87.9	62.8	39.9	74.2	189.2
DDM+CLIP+Refined-T	78.7	60.0	39.8	71.9	271.5	89.0	75.4	48.3	84.5	381.8	88.8	68.9	42.1	76.1	195.2
DDM+Refined-C	78.0	60.8	39.9	71.9	270.3	87.9	74.5	47.3	83.9	**388.4**	89.2	68.6	41.5	75.4	195.0
DDM-RLIP	**79.7**	59.6	**41.6**	**74.4**	**274.5**	**89.7**	**77.2**	**48.9**	**85.1**	372.6	**89.6**	**69.2**	**42.5**	**76.8**	**197.7**

CIDEr [5,30]. The model was trained on 4 A100 GPUs, with a batch size of 128 per GPU, using ViT-B/16 as an exemplary implementation, with a maximum sentence length of 20. The images were resized to 256 × 256 pixels, and the training lasted for 200 epochs with a warm-up period of 5 epochs, using the AdamW optimizer with a weight decay of 0.01 and a learning rate ranging from 1×10^{-4} to 0 with cosine annealing.

4.1. Experiments results

On the NWPU-Captions dataset, we achieve BLEU-4 scores of 69.2 and CIDEr scores of 197.7, respectively, compared to 47.8 and 116.4 in the latest literature [2]. On the UCM-Captions and Sydney-Captions datasets, there are more related studies for comparison, as listed in Tables 1 and 2, respectively. A CNN-based image feature extractor can capture deep-level image information, while LSTM can capture and store long-term temporal information, making it more suitable for text generation. Transformer introduces attention mechanisms to identify the most relevant text to the image. The structure of the diffusion model is flexible, the sampling process is deterministic, the training process is stable, and it is easy to train, resulting in more accurate image captionings.

Compared to SOTA models, our proposed model demonstrated performance improvements in terms of BLEU-4 and CIDEr scores by 6.9 and 1.4 on the UCM-Captions dataset, and by 21.4 and 81.3 on the NWPU-Captions dataset, respectively. Additionally, our model achieved a 1.5 ROUGE_L score improvement on the Sydney-Captions dataset, with other results comparable to the baselines. Notably, JTTS [30] utilized additional labels for guided training, and DDM-RLIP achieved competitive results, with higher scores on the UCM-Captions dataset and comparable results on the Sydney-Captions dataset. Fine-tuning is based on contrastive learning of the multimodal model CLIP. CLIP's training data consists of text-image pairs: an image and its corresponding text captioning. The model can learn the matching relationship between text and images, obtaining features of images that are more relevant to the text. The diffusion model exhibits strong noise resistance and smoothness, with a rapid descent rate of the loss function, making it easier to converge and obtain more accurate image captionings.

4.2. Discussion

To validate the effect of RLIP representations, i.e., the adapted transformer with an additional attention module and the fine-tuned CLIP image encoder in our proposed captioning method, we carried out some ablation experiments on the three datasets, as listed in Table 3.

Our ablation analysis demonstrated that fine-tuning the CLIP image encoder, in combination with the adapted transformer, notably enhances the performance of RSIC. Specifically, on the Sydney-Captions, UCM-Captions, and NWPU-Captions datasets, the CIDEr scores improved by 3.7, 7.8, and 8.5, respectively. One possible explanation is the use of an attention module in the transformer, which aligns the textual content with the related image regions. This enables the adapted transformer to suppress irrelevant or erroneous information and retain only relevant information from the extracted image representations, thereby empowering the model to generate more accurate descriptions. Another reason can be attributed to the fine-tuned CLIP image encoders, which are better equipped than the original models trained on natural images to output RS image representations of varying scales. Overall, RLIP representations facilitate DDM in gradually recovering the original text tokens.

GT: A baseball diamond composed of sand and weeds.

Ours: It is a very old baseball diamond compose of sand and weeds.

GT: It is a dense residential area with lots of houses arranged in lines.

Ours: It is a dense residential area with lots of houses arranged neatly.

GT: Lot of green trees in the dense forest.

Ours: This is a dense forest with lots of dark green trees.

Fig. 3. Examples of the UCM RSIC dataset.

Our study was the first to apply DDM-based models to RSIC and achieved promising results, despite significant room for improvement. Notably, the Sydney-Captions dataset did not experience significant performance gains with our proposed method, which could be attributed to the dataset's limited size. This drawback is linked to the DDM-RLIP's dependence on the training data volume. Compared to the methods in Tables 1 and 2, DDM-RLIP exhibits superior performance, with a substantial improvement in evaluation metric results. The combination of the diffusion model and self-attention mechanism enables more accurate image captioning for remote sensing. Moving forward, we aim to improve DDM-RLIP by optimizing the diffusion process, developing high-performance feature relation extractors, and addressing the data dependence issue.

In addition, our research highlights the necessity to create a comprehensive and diverse RSIC dataset containing specific content descriptions of the images. Fig. 3 showcases selected GT and predicted captions from the UCM-Captions dataset, indicating that the current GT descriptions are insufficiently informative. For instance, the cars located at the corner are overlooked in the first example. This does not only constrain the supervised learning process but also complicates the evaluation of the performance of different methods. Thus, developing a more extensive and representative dataset is a crucial aspect for future research.

5. Conclusion

Precise generation of descriptions for RS images is an imperative and challenging research concern. This letter presents the novel application of DDM to RSIC, DDM-RLIP. This method is on the basis of a DDM-based captioning method for natural pictures. DDM-RLIP

utilizes an adapted transformer with an additional Attention module to attend to the important words in the generated caption, after fine-tuning the CLIP image encoder on RS datasets to better consider RS image characteristics. Our method demonstrates superior effectiveness in processing intricate targets within multi-scale RS images, yielding results that surpass those established by SOTA models on three common RS captioning datasets.

CRediT authorship contribution statement

Guannan Leng: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft. **Yu-Jie Xiong:** Funding acquisition, Writing – review & editing. **Chunping Qiu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing. **Congzhou Guo:** Conceptualization, Data curation, Validation, Writing – review & editing.

Declaration of competing 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.

Data availability

Data will be made available on request.

References

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

[2] 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.

[3] G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: Benchmark and state of the art, Proc. IEEE 105 (10) (2017) 1865–1883.

[4] 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.

[5] Y. Wang, W. Zhang, Z. Zhang, X. Gao, X. Sun, Multiscale multiinteraction network for remote sensing image captioning, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15 (2022) 2154–2165.

[6] Z. Zhang, W. Diao, W. Zhang, M. Yan, X. Gao, X. Sun, LAM: Remote sensing image captioning with label-attention mechanism, Remote Sens. 11 (20) (2019) 2349.

[7] X. Zhang, Y. Li, X. Wang, F. Liu, Z. Wu, X. Cheng, L. Jiao, Multi-source interactive stair attention for remote sensing image captioning, Remote Sens. 15 (3) (2023) 579.

[8] X. Shen, B. Liu, Y. Zhou, J. Zhao, M. Liu, Remote sensing image captioning via variational autoencoder and reinforcement learning, Knowl.-Based Syst. 203 (2020) 105920.

[9] J. Gu, J. Bradbury, C. Xiong, V.O. Li, R. Socher, Non-autoregressive neural machine translation, in: International Conference on Learning Representations, 2018, URL: https://openreview.net/forum?id=B1l8BtlCb.

[10] C. Zhou, G. Neubig, J. Gu, Understanding knowledge distillation in non-autoregressive machine translation, 2019, arXiv preprint arXiv:1911.02727.

[11] W. Qi, Y. Gong, J. Jiao, Y. Yan, W. Chen, D. Liu, K. Tang, H. Li, J. Chen, R. Zhang, et al., Bang: Bridging autoregressive and non-autoregressive generation with large scale pretraining, in: International Conference on Machine Learning, PMLR, 2021, pp. 8630–8639.

[12] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[13] Z. Zhu, Y. Wei, J. Wang, Z. Gan, Z. Zhang, L. Wang, G. Hua, L. Wang, Z. Liu, H. Hu, Exploring discrete diffusion models for image captioning, 2022, arXiv preprint arXiv:2211.11694.

[14] W. Huang, Q. Wang, X. Li, Denoising-based multiscale feature fusion for remote sensing image captioning, IEEE Geosci. Remote Sens. Lett. 18 (3) (2020) 436–440.

[15] Y. Li, S. Fang, L. Jiao, R. Liu, R. Shang, A multi-level attention model for remote sensing image captions, Remote Sens. 12 (6) (2020) 939.

[16] X. Lu, B. Wang, X. Zheng, X. Li, Exploring models and data for remote sensing image caption generation, IEEE Trans. Geosci. Remote Sens. 56 (4) (2017) 2183–2195.

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

[18] 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.

[19] Z. Zheng, Y. Zhong, J. Wang, A. Ma, Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4096–4105.

[20] C. Zhang, G. Li, S. Du, Multi-scale dense networks for hyperspectral remote sensing image classification, IEEE Trans. Geosci. Remote Sens. 57 (11) (2019) 9201–9222.

[21] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, 2015, arXiv preprint arXiv:1511.07122.

[22] H. Qiu, H. Li, Q. Wu, F. Meng, K.N. Ngan, H. Shi, A2RMNet: Adaptively aspect ratio multi-scale network for object detection in remote sensing images, Remote Sens. 11 (13) (2019) 1594.

[23] G. Cheng, C. Yang, X. Yao, L. Guo, J. Han, When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs, IEEE Trans. Geosci. Remote Sens. 56 (5) (2018) 2811–2821.

[24] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, 2020, arXiv preprint arXiv:2010.02502.

[25] S. Xu, CLIP-diffusion-LM: Apply diffusion model on image captioning, 2022, arXiv preprint arXiv:2210.04559.

[26] J. Austin, D.D. Johnson, J. Ho, D. Tarlow, R. van den Berg, Structured denoising diffusion models in discrete state-spaces, Adv. Neural Inf. Process. Syst. 34 (2021) 17981–17993.

[27] X.L. Li, J. Thickstun, I. Gulrajani, P. Liang, T.B. Hashimoto, Diffusion-lm improves controllable text generation, 2022, arXiv preprint arXiv:2205.14217.

[28] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Int. Conf. Mach. Learn., PMLR, 2015, pp. 2048–2057.

[29] X. Zhang, X. Wang, X. Tang, H. Zhou, C. Li, Description generation for remote sensing images using attribute attention mechanism, Remote Sens. 11 (6) (2019) 612.

[30] 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.