SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

An Enhanced Deep Learning Framework For Land Target Detection And Captioning In Remote Sensing Imagery

Ashok Kumar M[1] , Dr. K. Mahesh[2] , Dr.Eswaran Perumal[3]

[1]Research Scholar, Department of Computer Applications, Alagappa University, Karaikudi, m.ashokkumar81190@gmail.com
[2]Professor, Department of Computer Applications, Alagappa University, Karaikudi. mahesh.alagappa@gmail.com
[3]Associate Professor, Department of Computer Applications, Alagappa University, Karaikudi. eswaranperumal@gmail.com
Corresponding author mail id : m.ashokkumar81190@gmail.com

Abstract
Remote sensing image analysis supports ecological observing, town development, and tragedy organization. Though existing methods separately handle target detection and caption generation, limiting contextual understanding and producing less informative descriptions of complex aerial scenes. To address these challenges, this research proposes a Scene-Aware Swin Transformer Captioning Network (SASTCN) for integrated land target detection and remote sensing image captioning. The proposed model utilizes a ST encoder to acquire hierarchical visual features for the effective representation of land targets. Moreover, the scene-aware embedding module acquires high-level contextual information in relation to various land use categories. Additionally, a semantic attribute learning module detects discriminative attributes of the detected land targets. This way, these complementary visual, scene, and semantic attributes are merged into a multimodal feature fusion approach to form a comprehensive semantic representation. Finally, the captions are generated using a transformer-based decoder, which includes positional encoding and beam search decoding to ensure contextual and semantic coherency in the produced sentences. Moreover, comprehensive text preprocessing, vocabulary formation, and sequence encoding are included to improve the language modeling and caption generation process. The proposed model is programmed in Python programming language and tested on the benchmark dataset called Remote Sensing Image Caption Dataset (RSICD). Experiments prove that the proposed SASTCN recognizes various geographical targets and produces correct captions and reaches a Semantic Propositional Image Caption Evaluation (SPICE) score of 0.6031 and increases the contextual and semantic coherency of the produced sentences.

Keywords: Image Captioning, Multimodal learning, Geospatial analysis, Land-use classification, Scene interpretation, Visual semantics.

1. Introduction

The system of remote sensing (RS) is an interdisciplinary system based on the use of sensors to detect and analyse objects at a distance without physical contact. It provides valuable information sources to military intelligence gathering, monitoring and data gathering [1]. Multi-platform aerial RS has become a key technology for the capture of geospatial data, where higher-level systems, ranging from outposts to Unmanned Aerial Vehicles (UAVs), play a crucial role in tasks as diverse as wide area tracking and detection, traffic tracking, shipping and aircraft security, and rescue operations, among others [2]. For RS image analysis, segmentation of land cover is crucial. It involves large amounts of data and calls for sophisticated pixel-level semantic segmentation techniques. It enhances the efficiency of using resources and reduces disaster risk by providing information for ecological condition detection and urban planning [3]. The applications of RS are important in meteorology, agriculture, urban planning, environmental resource analysis and military fields. It is indispensable to solve the ecological problems in the region and country, foster economic development and improve public service [4]. Target identification is the function of identifying particular items (e.g., people, cars, planes and ships) in digital images and videos found in RS images. This assessment is used for financial and social purposes, and is a vital piece of information for

making decisions [5]. In optical RS imaging, target pixels are scattered in many directions as a result of the diversity of targets' orientation over diverse overhead views. As such, they are susceptible to their environment [6].Remote Sensing Image Captioning (RSIC) is a critical problem at the intersection of Artificial Intelligence (AI) and Earth observation. Its goal is to translate intricate geometric and radiometric data from satellite or aerial images into descriptions that are understandable by humans [7]. It is important to detect the number and shape of buildings in RS images for land planning, mapping and post-disaster reconstruction. The number of surviving buildings and the extent of damage are crucial indications, especially when it comes to earthquake mitigation and disaster response [8]. Since land-use changes entail interrelated items, effectively capturing semantic links between detected land targets and scene features is a major issue in RS image captioning. To generate coherent, semantically relevant captions, captioning algorithms must grasp the context and object-level information [9].

Research aim and organization

The aim of this research is to enhance the performance of remote sensing image analysis through the integration of effective land target identification along with automatic generation of descriptions to facilitate the task of semantically understanding the scene. A network called the Scene-Aware Swin Transformer Captioning Network (SASTCN) that can learn the visual, scene and semantic information together to generate context-aware descriptions is proposed.

The outline of this research is as follows: Section 1 is about the history of research, motivation and objectives of the study. Section 2 is about the literature review. Section 3 describes the proposed method and architecture. Section 4 is about the results of the experiments. Section 5 talks about the limitations of the study.

2. Related works

Existing remote sensing methods use Deep Learning (DL) for target detection, segmentation, and captioning; however, problems remain in contextual comprehension, small target detection, and semantic caption synthesis.

For satellite image processing, Convolutional Neural Network (CNN)-based object detection methods were used [10]. Even though Satellite Imagery Multiscale Rapid Detection with Windowed Networks (SIMRDWN) achieved a 97% accuracy rate, small object detection, differences in resolutions and inefficiency of SIMRDWN in real-time operation are some of the weaknesses. For mapping urban surface water, a Sparse Superpixel-Based Water Extraction (SSWE) technique combines multitarget sparse detection, feature improvement, and scale-adaptive simple non-iterative clustering (SA-SNIC) segmentation [11]. It obtained 0.942 kappa. Spectral constraints, shadows, and intricate urban settings are the limitations of this research.

The Deep Convolutional Generative Adversarial Network-Single-Shot Detector (DCGAN + SSD) architecture with Particle Swarm Optimization (PSO)-based hyperparameter tuning improves underwater target recognition for automated underwater vehicles by enhancing deteriorated image quality [12]. Better detection accuracy was shown in the experimental findings. Image deterioration and difficult underwater penetration are some of the constraints. Feature Enhancement Feature Pyramid Network (FE-FPN) optimizes multiscale feature representation for RS target detection [13]. It outperformed Cascade Region-based Convolutional Neural Network (R-CNN) with FPN by 2.0% in average precision. For very dense small targets and significant object-scale fluctuations, performance remain difficult.

For flat surface semantic classification from ultra-resolution RS photos, DeepLabVersion3Plus (semantic segmentation network) +-MobileNetV2-Convolutional Block Attention Module (DeepLabv3+-M-CBAM) combines MobileNetV2 with the CBAM[14]. With quicker segmentation, it obtained an F1-score of 88.42%. Additional validation under various environmental circumstances is necessary to ensure transferability across various satellite datasets. For the purpose of detecting RS targets, You Only Look Once Version 7 (YOLOv7) integrates Bottleneck Transformers and the CBAM [15]. Accuracy increased by 4.2% and 6.3%, respectively. For incredibly dense objects, detection remain difficult. Deformable Convolutional Hybrid-You Only Look Once Version 7 (DCH-YOLOv7) progresses small-target localization and multiscale feature extraction in difficult backdrops, making optical RS target recognition more effective [16]. It outperformed YOLOv7 by 3.1%, achieving 90.6% mAP@0.5. Validation of generalization to various RS datasets yet to be completed.

The Hierarchical Filtering Feature Pyramid Network (HF-FPN) was created to detect small targets in RS images by dynamically filtering and fusing multi-scale features. It improved mAP@0.5 by 3.5–8.1% [17]. Cross-scenario generalization and computational overhead continue to be difficult. Boots trapping Language-Image Pre-training. (BLIP) is used to learn image–text correlations to produce captions for RS images [18]. It showed meaningful caption embeddings, but variations in the intricacy of the land cover

have an impact on caption quality. To learn scene attributes for RS image captioning in unseen scenes, Scene Attribute Learning Captioning (SAL Cap) uses a global object scene attribute extractor [19]. It improved the generalizability, flexibility, and correctness of captions. Effective attribute inference under sparsely annotated training information is critical to effectiveness. To enhance the creation of captions for RS images, Cross-Modal Retrieval and Semantic Refinement (CRSR) [20] associates cross-modal retrieval and semantic refinement modules. The model produced better captioning results, but the quality of the retrieved semantic information can affect robustness.

2.1 Research gap

Existing RS approaches are excellent at detecting and segmenting targets, but they have difficulties in capturing complicated spatial relationships, multiscale objects, and a wide range of environmental variables [10, 13]. Recent captioning techniques increase image-text alignment, but they lack scene-level knowledge and semantic descriptions [18, 19]. Cross-modal retrieval strategies improve caption generation, but their effectiveness is dependent on the quality of the information obtained [20]. Therefore, an integrated model is necessary for proper land interpretation and context-aware caption production. The proposed SASTCN addresses these limitations by combining Swin Transformer-based hierarchical feature extraction, scene-aware embedding, semantic attribute learning, and multimodal fusion to accurately understand land targets and generate contextually rich captions for complex RS images.

3. Methodology

The proposed methodology comprises the SASTCN model, which combines ST-based visual feature extraction, scene-aware embedding, semantic attribute learning, multimodal fusion, and transformer decoding. Preprocessing, vocabulary building, and encoding are applied to RSICD images and captions. The model uses beam search, optimization techniques, and attention processes to generate context-aware captions and depict land targets; the overall design is represented in Figure 1.

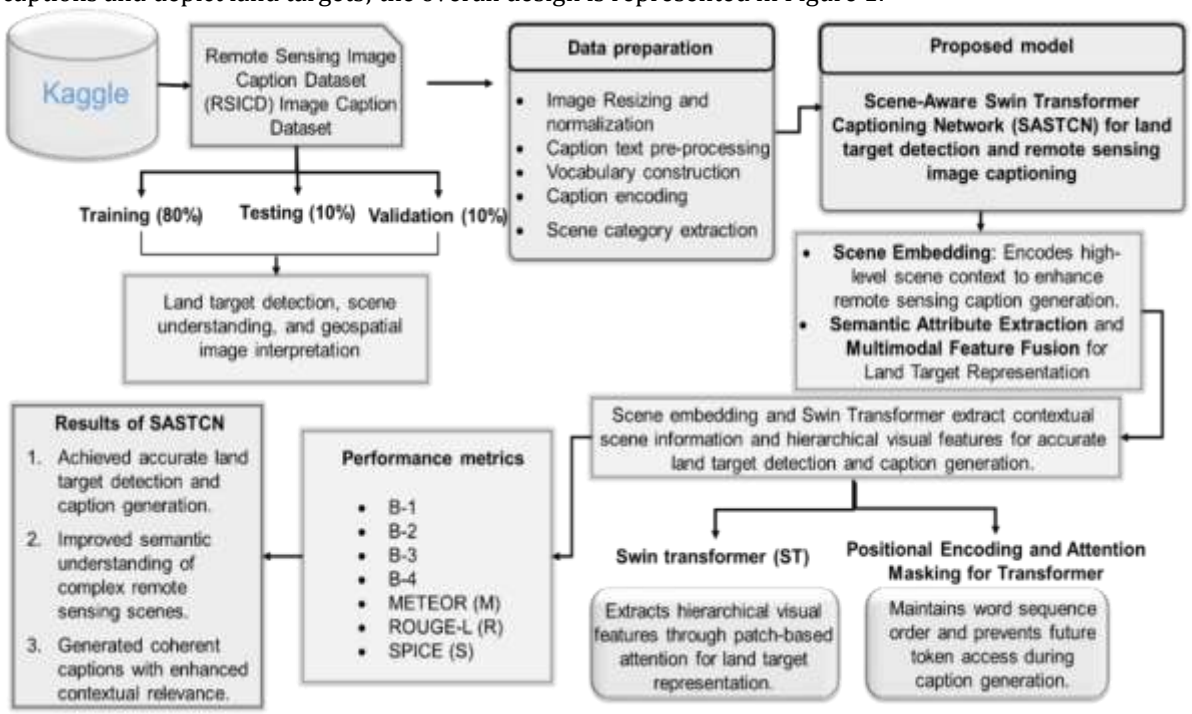

Figure 1: Proposed SASTCN Pipeline for Multimodal Land Target Detection

3.1 Remote Sensing Image Caption Dataset (RSICD)

The RSICD dataset utilized in this research was sourced from Kaggle (https://www.kaggle.com/datasets/thedevastator/rsicd-image-caption-dataset). The dataset comprises RS images combined with textual captions for building and assessing image captioning models. The dataset is split into three subsets: training (train.csv-80%), validation (valid.csv-10%), and testing (test.csv-10%). The testing subset contains 1,093 RS images distributed among 96 distinct land-use categories, such as airports, bridges, rivers, industrial locations, housing developments, agricultural fields, and infrastructure buildings. It facilitates learning visual-semantic links for accurate land targeting and context-aware caption creation. The dataset properties, structure, and use in scene-aware RS image caption production are summarized in Table 1.

Table 1: Description of the RSICD Dataset Used for SASTCN Training and Evaluation

Parameter	Description
Dataset Name	RS Image Caption Dataset (RSICD)
Data Type	RS images with textual captions
Data Format	CSV files with image filenames and captions
Dataset Splits	Training (train.csv), Validation (valid.csv), Testing (test.csv)
Input Data	Satellite/aerial RS images
Text Information	Multiple captions associated with each image
Primary Task	RS image caption generation
Applications	Land target detection, scene understanding, and geospatial image interpretation
Usage in SASTCN	Learning visual, scene-aware, and semantic relationships for caption generation

3.2 Data preparation

The RSICD data preparation entails image transformation, caption processing, vocabulary construction, encoding, and scene extraction to provide effective visual-semantic learning for SASTCN caption production.

Image Pre-processing: The RSICD images are normalized and transformed to standardize pixel intensity distributions and improve the reliability of visual feature extraction with the ST encoder. To preserve spatial consistency and increase visual diversity, image transformation uses resizing and augmentation techniques, such as flipping and rotation. Through better visual feature extraction for land target representation, these modifications improve the robustness of the model. By scaling the intensity values according to the image's mean and standard deviation, the pixel values are normalized, as shown in Equation (1).

$$J_o = \frac{J-\mu}{\sigma} \quad (1)$$

Where J_o is the normalized land image pixel value,J is the originalRSICD pixel value, μ defines the mean pixel intensity, and σ is the standard deviation of pixel intensity. RSICD images are resized to a predetermined input resolution range of 224×224 to 256×256 pixels for feature extraction. This method ensures spatial consistency, minimizes computing complexity, and provides consistent image dimensions for effective hierarchical feature learning with the Swin Transformer encoder.

Caption pre-processing: It prepares RSICD textual descriptions for transformer-based caption creation. To better semantic representation, tokenization and keyword removal are implemented. Tokenization separates captions into individual word tokens, removing unnecessary keywords, symbols, and rare terminology while retaining vital descriptive information about land targets and scene surroundings. To create fixed-length textual inputs for transformer-based caption creation, special tokens($< START >, < END >$) and numerical encoding are used. The caption sequence that has been processed is shown in Equation (2).

$$D_j = [< START >, y_1, y_2, \dots, y_U, < END >] \quad (2)$$

Where D_j is the processed caption sequence for the j^{th} image, $< START >$ denotes the caption beginning token,$y_1, y_2, \dots, y_U$ represents the encoded caption word tokens, U is the total number of caption tokens, and $< END >$is the caption ending token.

Vocabulary construction: The vocabulary construction finds the most pertinent terms needed to describe detected land targets and the context of the surrounding landscape in RSICD data. To improve the creation of useful captions, the created vocabulary combines semantic properties, scene-aware information, and visual aspects. For a specific RS image, the hierarchical visual features denoted by w_j are extracted by the ST encoder. The visual representation is used to determine the importance of scene-specific and land target-related seed words, are shown in Equation (3).

$$T_j^{(I)} = \sigma\left(N_1(w_j)\right) \quad (3)$$

Where $T_j^{(I)}$ represents the initial relevance score of the jth vocabulary word, w_j denotes the hierarchical visual feature extracted by the ST encoder, $N_1(.)$ is the feature transformation network, and $\sigma(.)$is the sigmoid activation function. To combine individual word embeddings with the original image-level semantic representation, the semantic relevance of each vocabulary word are improved. Regarding the jth word, w_j, the refined word relevance score is computed as shown in Equation (4).

$$T_j^{(W)} = \sigma\left(N_2\left(\left|T_j^{(I)}, w_j\right|\right)\right) \quad (4)$$

Where $T_j^{(W)}$ denotes the refined vocabulary relevance score, $T_j^{(I)}$indicates the initial word relevance, w_j denotes the word-level semantic embedding, $N_2(.)$ is the semantic refinement network, σ denotes the Sigmoid activation function, and $\left|T_j^{(I)}, w_j\right|$represents the fused visual-semantic land detection features.

Caption encoding: Caption encoding transforms pre-processed captions into numerical sequences for the SASTCN Transformer decoder. Each token is provided a vocabulary index, allowing semantic learning across images and descriptions, while fixed-length padding provides consistent input representation.

Scene category extraction for High-Level Scene Understanding: Scene category extraction uses semantic categories like metropolitan areas, woodlands, agricultural fields, rivers, and airports to extract high-level land-use information from RS images. Scene-aware caption generation is enabled by DL-based scene classification techniques that extract hierarchical visual elements to capture spatial patterns and environmental relationships. A scene-to-index mapping is used to extract scene labels from RSICD captions and translate them into numerical representations. Every image-caption pair is arranged in a unique dataset structure with its matching encoded caption and scene name. While DataLoader effectively creates shuffled mini-batches for training and validating the suggested SASTCN architecture, the Dataset class handles image retrieval, caption encoding, and scene assignment.

3.3 Scene-Aware Swin Transformer Captioning Network (SASTCN) for land target detection and RS image captioning

The proposed SASTCN integrates a ST encoder, scene-aware embedding, attribute extraction, multimodal feature fusion, and a Transformer decoder into a single method. Accurate land target detection and semantically relevant caption production are made possible by positional encoding, masked self-attention, beam search decoding, cross-entropy loss, and Adam optimization. A scene embedding module is incorporated into the proposed SASTCN to encode high-level scene category information related to each RS image. A trainable embedding layer maps scene labels taken from the RSICD dataset into dense embedding vectors after they have been transformed into numerical indices. These scene embeddings supplement the visual characteristics that the ST encoder extracts and offer contextual information about various land-use categories.

3.3.1 Swin Transformer (ST) Encoder for Land Target Detection

For efficient land target detection and semantic representation, the ST-based encoding module in SASTCN extracts hierarchical visual features from RS images. The module uses shifted-window self-attention and patch-based representation to collect global scene contexts and local target information. Land image patches are created and converted into feature embeddings for an input RS image, are shown in Equation (5).

$$Y_q = Flatten\left(J_q\right)X_f \quad (5)$$

Where Y_q is the land target patch feature embedding,J_q is the qth image patch extracted from the RS image, $Flatten(\cdot)$ denotes the operation converting image patches into one-dimensional feature vectors,X_f is the learnable feature projection matrix, and q signifies the index of the image patch. The spatial location of identified land objectives is maintained by integrating positional data, is shown in Equation (6).

$$X_0 = Y_q + Q_f \quad (6)$$

Where X_0 is the position-aware land target feature representation,Y_q denotes the patch feature embedding,$and Q_f$ represents the learnable positional feature embedding.To find discriminative target features, the ST uses window-based self-attention represented in Equation (7).

$$\hat{X}^m = V - MSA\left(MO(X^{m-1})\right) + X^{m-1} \quad (7)$$

Where $\hat{X}^m$ represents the output feature representation after window-based self-attention,$V - MSA$represents the window-based multi-head self-attention operation,$andMO(.)$ defines the normalization operation applied before attention computation, X^{m-1} denotes the input feature representation from the previous ST block, and m represents the ST block index. Interactions between nearby regions are improved by shifted-window attention defined in Equation (8).

$$\hat{X}^{m+1} = SW - MSA\left(MO(X^m)\right) + X^m \quad (8)$$

Where $\hat{X}^{m+1}$ represents the output feature after shifted-window attention, $SW - MSA$ denotes the shifted-window multi-head self-attention operation, $MO(.)$ defines the normalized feature, X^m is the feature representation from the current transformer block, and $m + 1$ represents the nextST block index. Figure 2 shows how to extract hierarchical features for caption synthesis utilizing shifted-window attention, fusion modules, and transformer decoding.

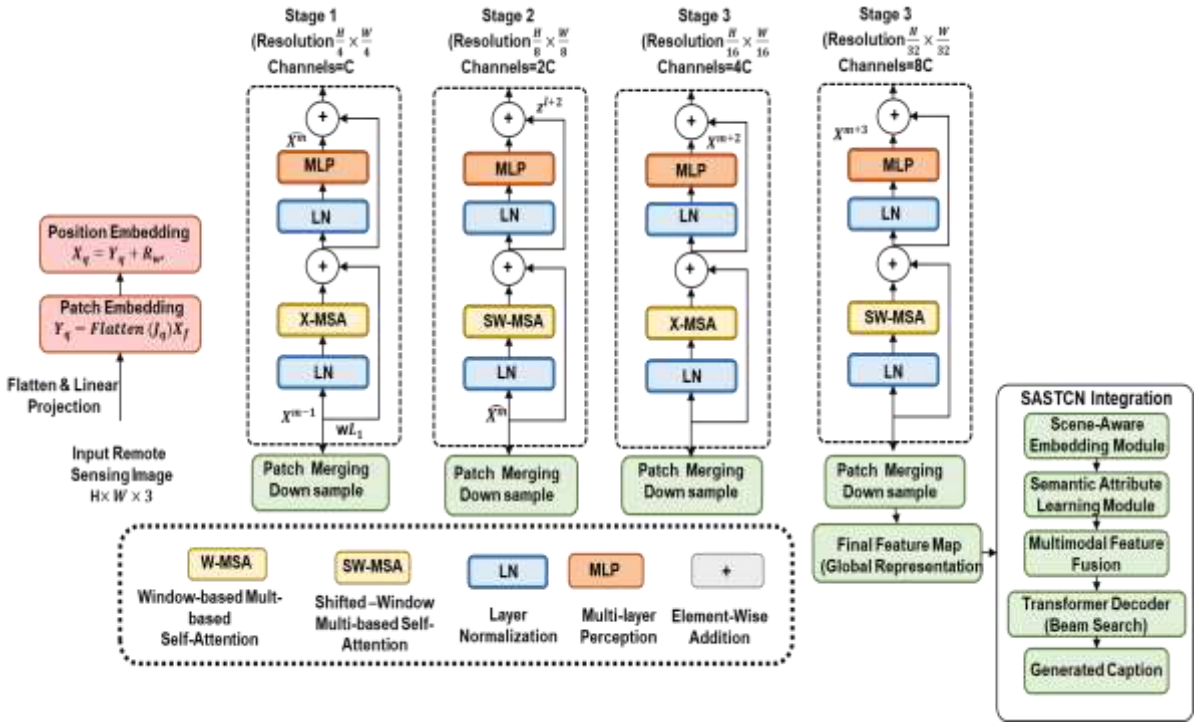

Figure 2: ST Encoder Architecture of the Proposed SASTCN model

3.3.2 Semantic Attribute Extraction and Multimodal Feature Fusion for Land Target Representation

The attribute extraction tool extracts discriminative semantic properties of land targets from ST image information. Important visual characteristics are captured through a 50-dimensional attribute representation produced by a fully connected layer with sigmoid activation. The multimodal feature fusion module uses feature concatenation to combine the extracted image features, scene embeddings, and attribute representations. The combined representation is subsequently converted into a single 256-dimensional semantic feature for caption creation using a fully connected layer.

3.3.3 Positional Encoding and Attention Masking for Transformer-Based Caption Generation

The proposed SASTCN uses a 2-layer Transformer decoder with 8-head multi-head attention and positional encoding to produce captions for RS images. Positional encoding is integrated into word embeddings to identify the location of each token in the resulting caption sequence because self-attention processes all tokens concurrently. Sinusoidal functions are used to create the positional encoding matrix for a caption sequence of length, as expressed in Equations (9) and (10).

$$PE(pos, 2j) = sin\left(\frac{pos}{10000^{2j/d}}\right) \quad (9)$$

$$PE(pos, 2j+1) = cos\left(\frac{pos}{10000^{2j/d}}\right) \quad (10)$$

Where $PE$is the positional encoding value.posdefines the position of the caption token in the sequence, $2j$ signifies the even dimension index, d denotes the dimension size of the word embedding, 10000 is the scaling factor used for sinusoidal position encoding, and $sin(\cdot), cos(\cdot)$ represent the sinusoidal functions used to generate position information. The Prior to entering the transformer decoder, the caption word embeddings are supplemented with the generated positional information, are shown in Equation (11).

$$Y_u = F_u + PF_u \quad (11)$$

Where Y_u represents the position-aware caption representation for the uth token,F_u is the word embedding feature of the uth caption token,PF_u represents the positional feature added to the word embedding, and u demonstrates the caption token position index. During training, the attention mask is created to ensure autoregressive narrative production by preventing the transformer decoder from obtaining future caption tokens. The formulation of the attention mask is calculated by shown in Equation (12).

$$N(j,k) = \begin{cases} 0, & j \geq k \\ -\infty, & j < k \end{cases} \quad (12)$$

Where $N(j,k)$ is an attention mask value between the j and k caption tokens, j is the current position of the caption token, k is the position of the future caption token, 0 makes the previous and current tokens

visible, and ∞ prevents the subsequent tokens from being noticed. The masked self-attention is then computed as shown in Equation (13).

$$Attention(R, L, W) = softmax\left(\frac{QK^U}{\sqrt{e_k}} + N\right)^W \quad (13)$$

Where R is the query matrix, L is the key matrix, W is the value matrix, QK^U is the resemblance mark among enquiry and crucial vectors,e_k is the dimension of key vectors, N is the attention mask matrix, and $Softmax(\cdot)$ normalizes attention scores. Figure 3 depicts transformer decoding for RS captions using self-attention, cross-attention, and positional encoding.

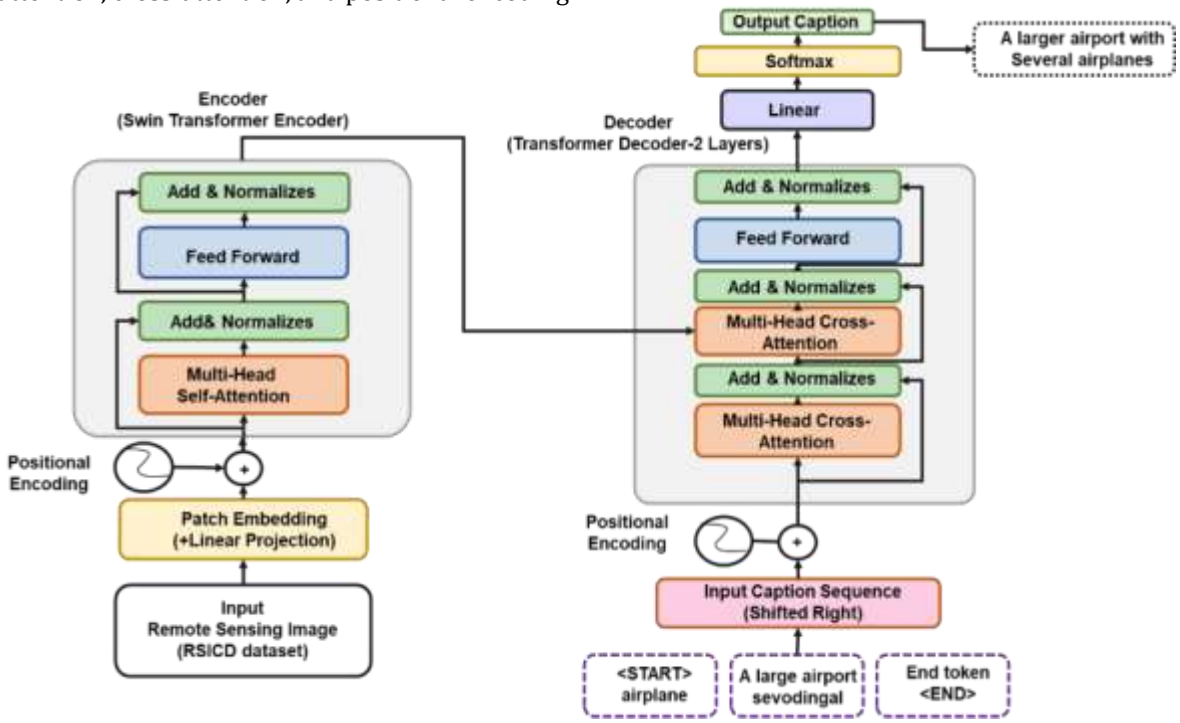

Figure 3: Transformer Decoder Architecture for Caption Generation in the Proposed SASTCN model

3.3.4 Beam Search, Loss Function, and Optimization Strategy

Beam Search with Length Penalty is used during inference to produce the final captions by keeping track of several candidate sequences and choosing the caption with the highest score. To attack a compromise between statistical competence and group excellence, a beam size of three is employed. Padding tokens (<pad>) are ignored while using Cross-Entropy Loss as the caption generation target. For efficient training convergence, the Adam optimizer modifies model parameters at a learning rate of 1×10^{-4}. To precisely identify land targets and produce contextually relevant captions for RS images, the proposed SASTCN combines a ST encoder, scene-aware embedding, semantic attribute extraction, multimodal feature fusion, and Transformer-based caption production, and the steps were shown in Algorithm 1.

Algorithm 1: SASTCN model

Input: RSICD dataset D = {(I,C)}
Output: Generated caption Cg
Begin
1. Load RS images I and captions C from RSICD.
2. Image Preprocessing:
   Resize I → Ir
   Normalize pixels → In = (I-μ)/σ
3. Caption Processing:
   Tokenize captions C → Ct
   Remove irrelevant keywords → Cr
   Encode caption tokens → Ce
4. Swin Transformer Feature Extraction:
   Divide In into patches Jq
   Generate patch embedding Yq = Flatten(Jq)Xf
   Add positional embedding X0 = Yq + Qf
   Apply W-MSA and SW-MSA

Extract hierarchical visual feature Fv
5. Scene Feature Extraction:
Convert scene labels → S
Generate scene embedding Fs
6. Attribute Learning:
Extract land target attributes → Fa
7. Feature Fusion:
Combine Fv, Fs, and Fa
Generate fused semantic feature Fm
8. Caption Generation:
Add positional encoding to Ce → Cp
Apply masked self-attention and cross-attention
Generate caption probability Pg
9. Optimize using Cross-Entropy Loss and Adam optimizer.
10. Generate final caption Cg using Beam Search.
End

4. Result

The proposed SASTCN model was implemented in Python with PyTorch, with GPU acceleration via CUDA when available. The tests used DL libraries such as TorchVision and TIMM, as well as evaluation tools for BLEU, METEOR, and CIDEr metrics, to train and evaluate RS image-caption generating performance.

4.1 Convergence Behavior and Contextual Caption Understanding Evaluation

The training loss convergence plot of the proposed SASTCN model (shown in Figure 4(a)) demonstrates the model's performance as the loss value consistently decreases starting from 4.38 to 0.64 over 40 epochs, indicating a steady optimization and stability of performance. For the scene-wise caption generation, Figure 4(b) illustrates the performance, where images with higher scores (such as mountain, airport, and beach images) demonstrate well contextual awareness over diverse RS contexts.

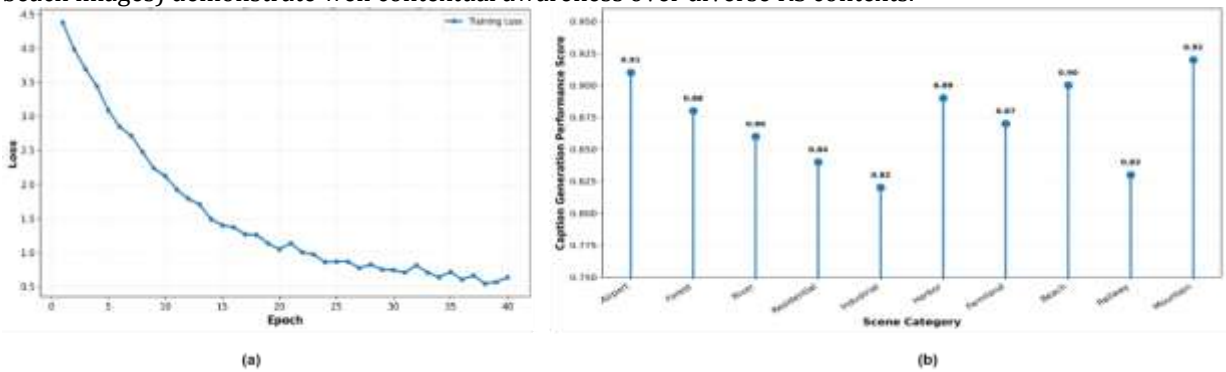

Figure 4: Visualization of (a) Training Loss Convergence; and (b) Scene-Wise Caption Generation Performance

4.2 Multi-Metric Evaluation of RS Image Captioning

The SASTCN model is evaluated in terms of conventional metrics for image captioning from RS images, which include Bilingual Evaluation Understudy (BLEU-n, where n=1, 2, 3, 4 or B-1, B-2, B-3, and B-4), where BLEU measures the n-gram accuracy between generated and reference captions; Metric for Evaluation of Translation with Explicit Ordering (METEOR (M)) measures semantic similarity via word alignment and synonyms; Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence (ROUGE-L (R))evaluates the sequence similarity in terms of longest common subsequence; Semantic Propositional Image Caption Evaluation (SPICE (S))measures semantic consistency through scene level concepts; Semantic Propositional Image Caption Evaluation + Consensus-based Image Description Evaluation (SPIDE+ CIDEr (Sm)) or SPIDEr (Sm) where the metric takes into account SPICE and CIDEr scores to evaluate both semantic relevance and human-like caption generation quality.

4.3 Quantitative Analysis of Land Target Semantic Representation and Caption Generation

The proposed Scene-Aware Swin Transformer Captioning Network (SASTCN) is compared to the current Cross-modal Retrieval and Semantic Refinement (CRSR) technique [20] for RS image caption invention.The valuation focuses on the ability of both techniques to comprehend complicated landscapes and create relevant descriptions for RS images. The proposed SASTCN model achieves BLEU-4 (0.7611), METEOR (0.4728), and SPICE (0.6031), which indicates that it can produce accurate and semantically relevant

captions, outperforming the current ones. Table 2 and Figure 5 illustrate these results as improvements in the recognition of RS land targets, the interpretation of the context of the scene, and the quality of the description produced.

Table 2: Semantic Caption Generation Capability of the Proposed SASTCN model for RS Land Targets

Metrics	CRSR [20]	SASTCN [Proposed]
B-1	0.7994	0.8916
B-2	0.7440	0.8356
B-3	0.6987	0.7972
B-4	0.6602	0.7611
METEOR (M)	0.4150	0.4728
ROUGE-L (R)	0.7488	0.7845
SPICE (S)	0.4845	0.6031
SPIDEr (Sm)	1.0397	1.1876

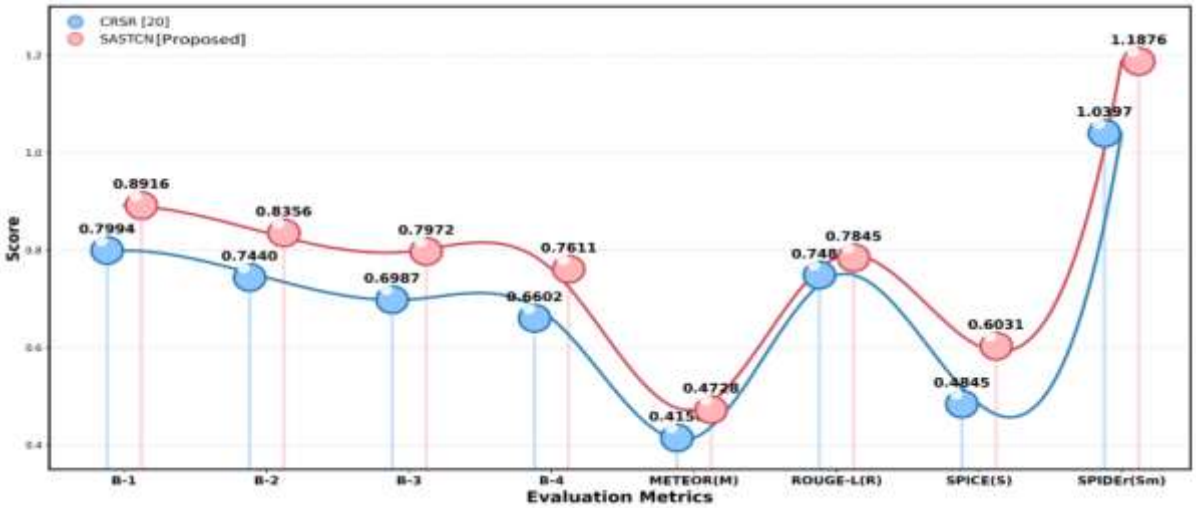

Figure 5: Overall performance comparison of the various methos

4.4 Discussion

However, the spatial differences and lack of contextual knowledge make RS image analysis challenging to achieve multiple targets on the terrain and generate appropriate captions. The proposed SASTCN tackles these challenges by leveraging ST-based hierarchical feature representation, scene-aware contextual understanding, and semantic attribute learning. This model is able to detect complex spatial dependence, land target properties and relationships between environment and RS images and to interpret the land and generate descriptive, semantically rich and contextually relevant captions for a broad set of geographical scenes. This existing CRSR [20] is not effective to obtain complicated spatial connections and precise land target characteristics in RS images. It is not very contextually aware at the scene level; descriptions of scenes in captions are limited and geographical scenes are described with a limited semantic description of multiple geographical landscapes. In this regard, the proposed SASTCN aims to overcome these limitations by adopting a hierarchical feature extraction approach based on ST, embedding using a scene-aware approach, and learning semantic attributes. The model is able to well represent complex geographical dependency, land target characteristics and relationships among contexts in RS images, which are used to identify land targets accurately and generate semantically rich, accurate and contextually meaningful captions for a vast variety of RS scenes.

5. Conclusion

By identifying land targets and generating semantic descriptions, RS image analysis can be used to better comprehend geographical contexts. The proposed model SASTCN will be able to enhance the scene interpretation of complicated aerial images, which combines visual feature learning, context visualization, and automatic caption generation. After normalization, the RSICD data undergo the following pre-processing steps: Caption normalization, Vocabulary construction, Sequence encoding, Pixel normalization,

Pixel scaling and Augmentation of data to increase the consistency of the data. For land target detection and multimodal feature fusion in caption generation, it is important to achieve success in these three aspects: ST-based hierarchical visual representation learning, scene category embedding, and semantic attribute extraction. To accurately locate the land target and to generate context-rich captions for RS images, the proposed SASTCN model adopts ST encoding, scene-aware embedding, semantic attribute learning, multimodal feature fusion, and Transformer decoding. The suggested SASTCN outperforms BLEU-4 (0.7611), and SPICE (0.6031) for semantic RS interpretation. The shortcomings of SASTCN are the use of an annotated dataset and complexity of computation. Future research will concentrate on lightweight architectures, and integration with larger multimodal RS datasets to improve generalizability.

Reference

1. Sun, X. J., & Lin, J. C. W. (2022). A target recognition algorithm of multi-source remote sensing image based on visual Internet of Things. Mobile Networks and Applications, 27(2), 784-793. https://doi.org/10.1007/s11036-021-01907-1
2. Zhou, P., Guo, X., Sun, X., Sun, B., Su, S., Jiang, W., ...& Huang, S. (2026). RAPT-Net: Reliability-Aware Precision-Preserving Tolerance-Enhanced Network for Tiny Target Detection in Wide-Area Coverage Aerial Remote Sensing. Remote Sensing, 18(3), 449. https://doi.org/10.3390/rs18030449
3. Wang, Z., Xia, M., Weng, L., Hu, K., & Lin, H. (2023). Dual encoder–decoder network for land cover segmentation of remote sensing image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 2372-2385. https://doi.org/10.1109/JSTARS.2023.3347595
4. Hu, W., Jiang, X., Tian, J., Ye, S., & Liu, S. (2025). Land target detection algorithm in remote sensing images based on deep learning. Land, 14(5), 1047. https://doi.org/10.3390/land14051047
5. Zakria, Z., Deng, J., Kumar, R., Khokhar, M. S., Cai, J., & Kumar, J. (2022). Multiscale and direction target detecting in remote sensing images via modified YOLO-v4. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1039-1048. https://doi.org/10.1109/JSTARS.2022.3140776
6. Zhou, L., Zheng, C., Yan, H., Zuo, X., Liu, Y., Qiao, B., & Yang, Y. (2022). RepDarkNet: A multi-branched detector for small-target detection in remote sensing images. ISPRS International Journal of Geo-Information, 11(3), 158. https://doi.org/10.3390/ijgi11030158
7. Zuo, Q. (2026). Cross-Region Few-Shot Remote Sensing Image Captioning via Adaptive Vision-Language Feature Fusion. Journal of Computer, Signal, and System Research, 3(1), 51-64. https://doi.org/10.71222/kt0ew220
8. Han, Q., Yin, Q., Zheng, X., & Chen, Z. (2022). Remote sensing image building detection method based on Mask R-CNN. Complex & Intelligent Systems, 8(3), 1847-1855. https://doi.org/10.1007/s40747-021-00322-z
9. León, J. L., Nogueira, V., Salgueiro, P., &Quaresma, P. (2026). Describing Land Cover Changes via Multi-Temporal Remote Sensing Image Captioning Using LLM, ViT, and LoRA. Remote Sensing, 18(1), 166. https://doi.org/10.3390/rs18010166
10. Tahir, A., Munawar, H. S., Akram, J., Adil, M., Ali, S., Kouzani, A. Z., & Mahmud, M. P. (2022). Automatic target detection from satellite imagery using machine learning. Sensors, 22(3), 1147. https://doi.org/10.3390/s22031147
11. Liu, Q., Tian, Y., Zhang, L., & Chen, B. (2022). Urban surface water mapping from VHR images based on superpixel segmentation and target detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 5339-5356. https://doi.org/10.1109/JSTARS.2022.3181720
12. Dinakaran, R., Zhang, L., Li, C. T., Bouridane, A., & Jiang, R. (2022). Robust and fair undersea target detection with automated underwater vehicles for biodiversity data collection. Remote Sensing, 14(15), 3680. https://doi.org/10.3390/rs14153680
13. Wei, R., Feng, Z., Wu, Z., Yu, C., Song, B., & Cao, C. (2023). Optical remote sensing image target detection based on improved feature pyramid. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 7507-7517. https://doi.org/10.1109/JSTARS.2023.3303692
14. He, C., Liu, Y., Wang, D., Liu, S., Yu, L., & Ren, Y. (2023). Automatic extraction of bare soil land from high-resolution remote sensing images based on semantic segmentation with deep learning. Remote Sensing, 15(6), 1646. https://doi.org/10.3390/rs15061646
15. Xu, S., Chen, Z., Zhang, H., Xue, L., & Su, H. (2024). Improved remote sensing image target detection based on YOLOv7. Optoelectronics letters, 20(4), 234-242. https://doi.org/10.1007/s11801-024-3063-z
16. Cui, C., Wang, R., Wang, Y., Zhou, F., Bian, X., & Chen, J. (2024). Research on optical remote sensing image target detection technique based on DCH-YOLOv7 algorithm. IEEE Access, 12, 34741-34751. https://doi.org/10.1109/ACCESS.2024.3368877

17. Lu, Y., Zhang, B., Zhang, C., He, Y., & Wang, Y. (2025). HFEF2-YOLO: Hierarchical Dynamic Attention for High-Precision Multi-Scale Small Target Detection in Complex Remote Sensing. Remote Sensing, 17(10), 1789. https://doi.org/10.3390/rs17101789
18. Huang, X., Lu, K., Wang, S., Lu, J., Li, X., & Zhang, R. (2024). Understanding remote sensing imagery like reading a text document: What can remote sensing image captioning offer?. International Journal of Applied Earth Observation and Geoinformation, 131, 103939. https://doi.org/10.1016/j.jag.2024.103939
19. Guo, Z., Liu, H., Ren, Z., Jiao, L., Gou, S., & Li, R. (2025). Attribute-Based Learning for Remote Sensing Image Captioning in Unseen Scenes. Remote Sensing, 17(7), 1237. https://doi.org/10.3390/rs17071237
20. Li, Z., Zhao, W., Du, X., Zhou, G., & Zhang, S. (2024). Cross-modal retrieval and semantic refinement for remote sensing image captioning. Remote Sensing, 16(1), 196. https://doi.org/10.3390/rs16010196