Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Multi-Scale Semantic Fusion Network with Adaptive Attention for Remote Sensing Image Captioning

Chandrashekhar Pawar[1], Ashwin Makwana[2]

[1]Research Scholar, [2]Professor, U & P U. Patel Department of Computer Engineering, Chandubhai S Patel Institute of Technology, Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology, Changa, India.

Email: [1]19drdce014@charusat.edu.in, [2]ashwinmakwana.ce@charusat.ac.in

Abstract

The aim of remote sensing image captioning (RSIC) is to obtain insightful and detailed textual description of satellite images and aerial images. However, traditional methods are not able to achieve this aim effectively due to a lack of contextual awareness caused by variations in scale, viewpoint and scene complexity. In this paper, we propose a method, the Multiscale Region-Aware Captioning Network (MSR-CapNet), which helps to achieve the aim of RSIC by generating relevant and semantically correct textual descriptions for scenes in satellite images (and aerial images). We train and test our method for the purpose of RSIC on the RSICD and UCM caption datasets. In our MSR-CapNet method, we have integrated Feature Pyramid Encoding (used for local and global visual characteristics representation), Adaptive Attention (which helps in dynamic prioritization of relevant regions) and Topic-Sensitive Embeddings (to generate semantically consistent captions). To show the effectiveness of the proposed method (MSR-CapNet), we compared it with existing techniques (recent transformer and graph-based baselines) using BLEU-4, METEOR, and CIDEr measures, where it shows consistent improvement over existing techniques.

Keywords: Remote Sensing Image Captioning (RSIC), Attention Mechanism, Topic-Sensitive Word Embeddings, Satellite Images.

1. Introduction

The Remote Sensing (RS) images are updated after a fixed time interval; as a result, the data within RS images are constantly changing. RS images provide a large amount of information from each image, which is useful in various domains such as land use monitoring and disaster response. The Geographic Information System (GIS) is improving day by day, allowing for efficient analysis and visualization. However, the existing methods are still not able to handle scale variation and are facing difficulties with dataset imbalance and generalization to new domains [1], [2]. As a result, multi-scale features are not extracted, and semantic consistency is not obtained. To address these issues, we propose the Multiscale Region-Aware Captioning Network (MSR-CapNet) method, which helps minimize these problems by combining feature extraction from multi-scale regions, adaptive attention mechanisms, and topic-sensitive word embeddings.

DOI: https://doi.org/10.36548/jtcsst.2025.4.005
Received: 03.10.2025, received in revised form: 30.10.2025, accepted: 12.11.2025, published: 21.11.2025

1.1 Problem Statement

While working with Remote Sensing Image Captioning (RSIC) it is observed that the following three key challenges were addressed in this work. The first challenge is scale variation, due to which regions/objects differ in pixel size; the second challenge is complex scenes with similar-view regions/objects; and the last but equally important challenge is the limited domain-specific vocabulary, as the remote sensing vocabulary is not very rich.

1.2 Our Approach and Contributions

In order to address these issues, we present MSR-CapNet, a Multi-Scale Semantic Fusion network that: (i) uses an FPN+RPN backbone to fuse multi-scale region and scene features to handle scale variation; (ii) applies an adaptive attention module that balances spatial vs. channel attention to improve region–word alignment; and (iii) integrates topic-sensitive word embeddings and a GNN to preserve domain semantics and inter-region relations. The major contributions are:

1. A unified pipeline combining multi-scale region extraction, gated attention fusion, and topic-aware language decoding for RSIC.
2. A mathematically-specified adaptive attention mechanism (spatial + channel gating) with empirical analysis and attention visualizations.
3. Extensive experiments on RSICD and UCM-Captions including ablation studies, bootstrap confidence intervals, and cross-dataset transfer analysis.

1.3 Rationale for Method Selection

The main goal is to address all three challenges mentioned in the problem statement. Therefore, in proposing the MSR-CapNet method, we utilize feature pyramid encoding (to adapt to scale variation), adaptive attention (to align the scene context), and topic-sensitive embeddings (for semantically rich scene descriptions).

2. Related Work

The early RSIC techniques were supported by natural image captioning and mostly used CNN-RNN architectures. In past years, to obtain global features from remote sensing images, a CNN such as VGGNet or ResNet was used, which was then provided as input into RNNs like LSTM or GRU to generate sequential captions. [1], [3].

When these models are used with high-resolution remote sensing imagery containing multiple semantic elements, they show notable limitations. The utilization of a single global feature vector produces general and ambiguous captions that lack specificity, particularly for complex entities like highways, farmlands, or metropolitan areas. [2], [4].

Attention methods were added to address this problem by allowing the model to dynamically focus on different areas of the image while creating a caption [1]. The establishment of Transformer-based configurations and region-aware features followed, enabling more precise modeling of semantic richness and spatial complexity. [5], [6].

The altitude and sensor resolution variables introduced scale variation, which becomes a major problem in remote sensing. To solve this problem, feature Pyramid Networks (FPNs) have been frequently used, which blend compact, semantically significant features with precise, sparsely semantic features to create a strong multi-scale representation.[7].

Recently, Transformer-based backbones like Swin Transformer and Pyramid Vision Transformer (PVT) have surpassed conventional CNNs in collecting scale-sensitive information because of their hierarchical architecture and self-attention methods. [8], [9].

In the last few years, the performance of remote sensing image captioning (RSIC) systems has improved dramatically, particularly in enhancing region-word alignment, due to the incorporation of attention processes [1], [2].

The incorporation of attention-based models allowed the decoder to dynamically focus on various spatial aspects of the image at each stage of the caption-generation process. This method produced more accurate and comprehensive captions by significantly improving the semantic alignment between language elements and visual aspects. Both [1] and [5]. Such methods were made possible by the groundbreaking "Show, Attend and Tell" architecture for natural image captioning developed by Xu et al. [1]. Later, it was modified for use in remote sensing applications, where spatial attention aids in highlighting significant areas relevant to the scene. [3] as well as [6].

By integrating attention processes at the object or area level, contemporary RSIC systems have become significantly more sophisticated. These techniques are frequently guided by pretrained object recognition algorithms like Faster R-CNN or semantic segmentation maps. [7]. In order to enhance the decoder's capacity to provide semantically rich and contextually aware descriptions, these frameworks usually integrate visual attention with semantic or contextual information. [5] as well as [8].

Additionally, because transformer-based designs use multi-head self-attention, they have performed better than RNN-based models. This methodology enables precise simulation of intra-region interactions and long-range dependencies between textual components and image areas [6], [9], [10].

However, general-purpose word embeddings such as Word2Vec or GloVe have demonstrated shortcomings in capturing domain-specific semantic nuances in RSIC tasks, despite their extensive use. The importance of topic-sensitive or domain-adapted embeddings in improving semantic relevance has been shown in recent research. Gururangan et al. [11] showed that ongoing pretraining on domain-specific datasets greatly enhances performance on future tasks, highlighting the importance of contextualized language representations in specialized domains.

3. Methodology

3.1 Overview

The MSR-CapNet provides the most precise and semantically significant subtitles for remote-sensing images by combining all of these visual perceptions with English comprehension. We were motivated to design our approach by simultaneously collecting data from local areas and global scenes because our primary goal is to investigate how an individual

describes an aerial view by focusing on a few regions of interest after general layouts are identified.

Three steps make up the MSR-CapNet's operation: topic-sensitive word embeddings are used to produce textual descriptions pertinent to the context and domain; multiscale visual features are extracted from the source image to represent objects of different sizes; and an adaptive attention approach dynamically assigns weight to significant locations when creating captions. This combination allows the network to represent an image in a way that is both flexible and in line with the semantic and spatial relationships present in the real world.

3.2 System Architecture

The general architecture of MSR-CapNet is illustrated below as a flow diagram-see Figure 1-describing the major processing steps from image to caption:

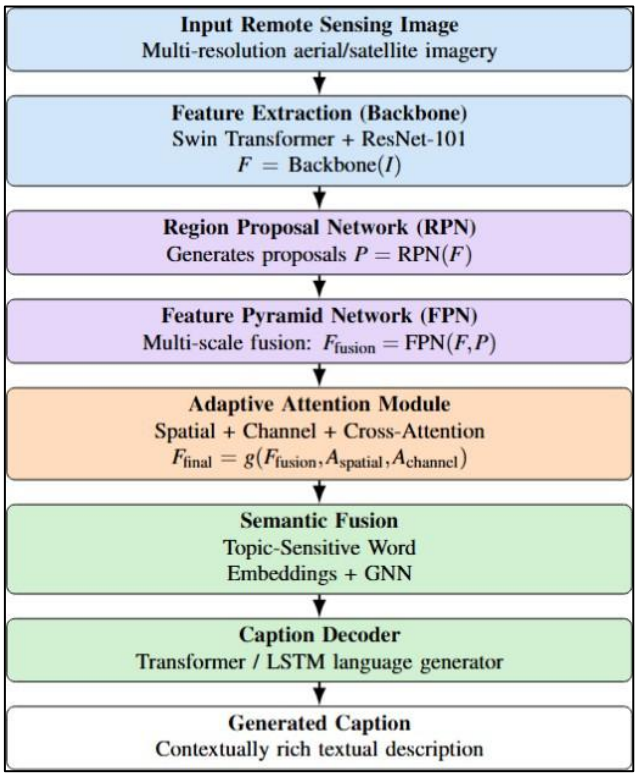

Figure 1. Workflow of the Proposed MSR-CapNet Showing end-to-end Flow from Feature Extraction Through Attention-Based Fusion to Caption Generation

Note: I denotes the input image; $F = \text{Backbone}(I)$; $P = \text{RPN}(F)$; $F_{\text{fusion}} = \text{FPN}(F, P)$; $F_{\text{final}} = g(F_{\text{fusion}}, A_{spatial}, A_{channel})$. testing subsets (7,645 / 1,638 / 1,638 images)[19].

3.3 Multi-Scale Region Feature Extraction

To capture the inherent complexity of the remote sensing image, we employ a hybrid backbone network in this work that combines ResNet-101 and Swin Transformer. The deep residual layers in ResNet-101 allow the model to learn even finer textures and contours, while Swin Transformer's hierarchical self-attention mechanism enables the model to capture the

entire picture of spatial relationships. When the two backbones are combined, a good balance is achieved between maintaining local characteristics and considering the bigger picture. A Region Proposal Network is used in addition to the backbone's feature extraction to offer suggestions for structures, roads, runways, and any other locations relevant to the problem. In complex aerial environments with a high degree of scale variation, this combination process becomes crucial because it enables the network to maintain high-level semantics while preserving the characteristics of small or dispersed objects.

Consequently, this hybrid approach combines feature fusion techniques with RPN and a backbone network to gather both global scene-level and region-specific object-level data, as shown below:

1. Backbone Network: We employ a deep convolutional network, ResNet-101 or Swin Transformer, for hierarchical feature extraction at different levels. ResNet-101's residual connections enable the model to learn even the most intricate patterns in remote sensing. Images while avoiding the problem of vanishing gradients [12, 13]. The first feature maps generated by the backbone are as follows:

$$F = \text{Backbone}(I) \tag{1}$$

where I is the source image and F represents the component feature maps at different layers of the backbone.

2. *Region Proposal Network (RPN)*: This is used to determine the important areas of the picture for further processing. RPN slides a small network on top of the feature map that the backbone produces to predict bounding boxes and object scores [12]. The output of RPN is a set of object proposals:

$$P = \text{RPN}(F) \tag{2}$$

where P stands for each region's expected box boundaries and object scores.

3. *Feature Fusion*: Features are fused at the grid and region levels using an FPN. An FPN integrates multi-level features from the RPN and the backbone network to obtain a set of fused features. [14] F_{fusion}:

$$F_{fusion} = \text{FPN}(F, P) \tag{3}$$

3.3.1 Role of FPN in Scale Variation

FPN plays an essential role in scale variation. The Feature Pyramid Network combines semantic information across a range of spatial resolutions to handle the large variations in object size and altitude that are common in remote sensing scenes. The lower layers (P2 and P3) capture fine textures of small objects like vehicles, while the higher layers like (P4 and P5) capture broader semantic contexts, such as agricultural or urban layouts. Mathematically, each pyramid level is computed as:

$$F_l = Conv_{1\times1}(C_l) + UpSample(F_l + 1) \tag{4}$$

where C_l denotes the convolutional feature from backbone level l, and F_l denotes the top-down fused feature. This hierarchical aggregation makes the features at each scale have

consistent semantics, allowing the decoder to effectively attend to both small and large structures during caption generation.

3.4 Adaptive Attention Mechanism

For generating each word in the caption, the model must decide where to look after receiving the visual information.

We propose an adaptive attention method for Transformers that can further improve the model's capabilities in incorporating textual information and focusing on important regions of the image.

1. **Cross-Attention Module (CAM):** Image areas (derived from the feature extraction stage) are linked with produced words (from the language model) using the Cross-Attention Module (CAM). By calculating the attention weights between word embeddings W and picture features Ffusion, this technique creates contextually aware image-text representations Z:

$$Z = \text{CAM}(F\_\text{fusion}, W) \tag{5}$$

The CAM uses the following attention formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\frac{QK^T}{\sqrt{d_k}} \times V \tag{6}$$

where Q, K, and V are the query, key, and value vectors derived from the image features and word embeddings, respectively, and dk is the dimension of the key vector [15].

Computational Complexity: Equation (5) corresponds to single-head scaled dot-product attention with complexity O(N2dk), where N is the number of image regions. In multi-head attention (as used in the Transformer decoder), the complexity becomes $O(HN^2d_k/H) = O(N^2d_k)$ per layer, since each of the H heads operates on a reduced dimensionality d_k/H. Therefore, the CAM's computational cost is comparable to that of a single Transformer attention layer but applied only once per decoding step, making it efficient for RSIC images where $N < 100$.

Contextual Word–region Alignment: In the Cross-Attention Module, queries (Q) originate from the decoder's current linguistic state h_t, while keys (K) and values (V) are derived from visual features F_{fusion}. The attention weight $\alpha_{t,i}$ between word token t and image region i represents their contextual alignment:

$$\propto_{t,i} = \exp\left(\frac{q_t^T k_i}{\sqrt{\{d_k\}}}\right) \Big/ \sum_j \exp\left(\frac{q_t^T k_i}{\sqrt{\{d_k\}}}\right) \tag{7}$$

The resulting context vector $c_t = \sum_i \alpha_{t,i} v_i$ encodes the most semantically relevant visual evidence for predicting the next word. This mechanism enables the model to dynamically link textual semantics with spatially meaningful regions during caption generation.

2. **Spatial and Channel Attention:** The Spatial and Channel Attention processes are used to focus on relevant visual aspects in each of the spatial and channel

dimensions. We calculate attention maps $A_{spatial}$ that emphasize significant areas in the image for spatial attention:

$$A_{spatial} = \sigma(\text{Conv}_{1\times1}(\mathbf{F}_{fusion})) \quad (8)$$

The channel attention mechanism modifies the significance of many feature channels in a similar manner:

$$A_{channel} = \sigma(\text{MLP}(\text{AvgPool}(\mathbf{F}_{fusion}))) \quad (9)$$

These attention mechanisms enhance the discriminative power of image features [15].

3. **Gated Fusion Mechanism:** The Gated Fusion Mechanism balances the contributions from the different levels of features (scene-level and object-level features) and attention maps. This approach generates the final representation Ffinal through the combination of picture features, attention mappings, and word embeddings by using a gating function g,: The output of the fused adaptive attention is then:

$$A_{final} = \lambda A_{spatial} + (1 - \lambda)A_{channel} \quad (10)$$

where λ is a learnable balance coefficient ($0 < \lambda < 1$).

Influence of the Scalar λ: The parameter λ dynamically adjusts the contribution of spatial and channel attention maps:

$$A_{final} = \lambda A_{spatial} + (1 - \lambda)A_{channel}$$

Thus, λ learns to emphasize the more informative modality during training. For instance, in high-altitude imaging where fine object features are lost, λ descends and gives channel attention that captures general semantics more weight. On the other hand, it rises in low-altitude images with clear object boundaries, laying emphasis on spatial localization. Empirically, based on scene characteristics, λ converges in the range of 0.45 to 0.6, indicating adaptive balancing.

The final attended visual feature for caption generation is:

$$F_{final} = g(F_{fusion}, A_{final}) = A_{final} \odot F_{fusion} \quad (11)$$

where ⊙ denotes element-wise multiplication.

Parameter Justification: The contribution between spatial and channel attention maps is controlled by the balance parameter λ. The $\lambda = 0.5$ was determined to be the ideal value by empirical tuning on the validation set, offering balance between spatial localization (helpful for object emphasis) and channel refinement (helpful for texture classification). In order to ensure constant gradient magnitudes in the softmax attention operation, the key dimension d_k adheres to the canonical Transformer scaling [10]. Over-peaked distributions, which can impair learning stability, are avoided by using $\sqrt{d_k}$ as a normalizing term.

Gated Fusion Mechanism Clarification: The gating function g(·) is implemented as a learnable sigmoid layer applied to a weighted sum of scene-level and object-level features. Formally, $g(x) = \sigma(Wx + b)$, where W and b are parameters to train, and σ denotes the sigmoid activation.

4. **Integration of Local and Global Cues:** MSR-CapNet integrates local (region) and global (scene) cues via two complementary mechanisms. First, the decoder employs *multi-head* cross-attention (as in Transformer) where each head attends to distinct aspects of the fused features F_{fusion}: some heads capture fine local dependencies (region edges, small objects), others encode broader scene context (layout, road networks). Formally, for head h:

$$head_h = softmax\left(\frac{Q_h K_h^{\top}}{\sqrt{\{d_k\}}}\right) V_h$$

and the concatenated heads produce $C_t = \text{Concat}(\text{head}_1, \ldots, \text{head}_H) W_O$.

Second, **The Gated Fusion Acts Hierarchically:** spatial and channel attention produce $A_{spatial}$ and $A_{channel}$ which are combined via learnable gate λ (Eq. (8)). The GNN then models inter-region relations on top of these fused features, injecting higher-order, global structural context into each region representation before decoding.

3.4.1 Variable Definitions

Below we define the variables used in Eqs. (4)–(8):

- I: input remote sensing image.
- $F = \{F_i\}_{i=1}^{N}$: set of region-level feature vectors produced by the backbone + RPN (each $F_i \in \mathrm{R}^d$).
- h_t: decoder hidden state at timestep t(Transformer query / LSTM hidden vector).
- Q, K, V: query, key and value matrices computed as linear projections of $h_t$and F.
- $A_{\text{spatial}} \in \mathrm{R}^{H \times W}$: spatial attention map obtained by a 1×1conv + softmax over spatial positions.
- $A_{\text{channel}} \in \mathrm{R}^{C}$: channel attention weights obtained via global pooling and an MLP.
- λ: learnable scalar gate balancing spatial vs. channel attention.
- $g(\cdot)$: gated fusion function (element-wise multiplication followed by a 1×1conv and ReLU).

3.4.2 Stepwise Computation (Per Decoding Step *t*).

1. Compute region features $F = \{F_i\}$ from the FPN output.
2. Compute decoder query $q_t = W_q h_t$

3. Compute attention weights $\propto_i^t = softmax_i\left(\frac{q_t^T k_i}{\sqrt{\{d_k\}}}\right)$
4. Obtain context vector $c_t = \sum_i \propto_i^t v_i\ (cross\ attention)$
5. Compute spatial map $A_{spatial} = \sigma(\text{Conv}1\times1\ (F_{f\ usion}))$ and channel weights $A_{channel} = \sigma(\text{MLP}(\text{AvgPool}\ (F_{f\ usion})))$
6. Fuse attentions: $A_{final} = \lambda\, A_{spatial} + (1 - \lambda)A_{channel}$ and obtain $F_{final} = g\left(F_{fusion}, A_{final}\right) = A_{final} \odot F_{fusion}$
7. Use c_t and F_{final} as input to decoder to predict token at t.

3.5 Positional Encodings for Irregular RS Patterns

Remote sensing imagery frequently exhibits irregular object layouts and varying spatial resolutions, making standard 1D positional encodings suboptimal. In MSR-CapNet we adopt learnable 2D positional embeddings for grid features and relative 2D encodings for region proposals. Concretely, for an FPN feature map of size H × W we add a learnable embedding $P_{x,y} \in R_d$ to each spatial location; for region features we append normalized centroid coordinates $(\tilde{x}, \tilde{y})$ and bounding-box scale as additional inputs to the region feature projection:

$$\widetilde{F_i} = Linear\ ([F_i; PE\ (\tilde{x_i}, \tilde{y_i}, s_i)])$$

As an alternative, we also evaluated 2D sinusoidal Fourier features and observed comparable performance; learnable 2D embeddings provided slightly faster convergence on RSICD.

3.5.1 Design Rationale of 2D Positional Encodings

Unlike grid-structured natural images, remote sensing scenes exhibit irregular spatial layouts and non-uniform object spacing. To encode such irregularity, each region feature Fi is augmented with learnable 2D positional embeddings PE (xi, yi, si) that incorporate normalized centroid coordinates and relative scale. These embeddings allow the model to infer directional and spatial context—e.g., that "runway"regions align longitudinally or "harbor" areas cluster near water boundaries. Compared with fixed sinusoidal encodings, learnable 2D embeddings adapt to arbitrary spatial distributions, improving robustness to rotation and scale distortions common in satellite imagery.

3.6 Topic-Sensitive Word Embeddings

We employ Topic-Sensitive Word Embeddings (TSWE) trained on extensive datasets like RSICD and BigEarthNet to provide the generated captions with linguistic meaning in the context of remote sensing. These embeddings capture domain-specific links, such as knowing that "runway" and "airport" are related or that "forest" frequently appears next to "river." To enhance semantic coherence in the generated captions, we employ topic-sensitive word embeddings, which extract domain-specific information from remote sensing photos.

1. **Training Domain-Specific Word Embeddings:** We train topic-sensitive word embeddings utilizing large domain-specific corpora, such as BigEarthNet and RSICD, which incorporate annotated remote sensing photos along with natural

language descriptions, in order to learn word representations pertinent to the remote sensing domain [16]. These corpora are perfect for capturing the semantics relevant to aerial scenes since they include a large domain vocabulary and sentence patterns.

The training process involves the following steps:

- **Preprocessing:** Each caption is tokenized, lowercased, and cleaned by removing irrelevant symbols. To build a vocabulary, uncommon words (frequency ¡ 5) are eliminated.
- **Model Selection:** Using the Gensim package, we evaluate the Skip-gram and Continuous Bag-of-Words (CBOW) models from the Word2Vec architecture. In the end, the Skip-gram model is selected because of its remarkable ability to recognize uncommon and context-sensitive words in smaller datasets.
- **Training Details:** We train for 10 epochs using negative sampling (k=5), set the embedding dimension to 300, and employ a context window size of 5. The embeddings produce domain-specific word vectors Wdomain that represent semantic and contextual associations particular to remote sensing imagery after being trained on the combined corpus (BigEarthNet + RSICD):

$$W_{\text{domain}} = \text{TrainWord2Vec}\,(\text{BigEarthNet} \cup \text{RSICD}) \quad (12)$$

Evaluation and Selection: Domain-specific analogy tasks and qualitative evaluation (e.g., cosine similarity between related terms like "urban" and "buildings") are used to assess the learned embeddings. For later captioning tasks, embeddings that better maintain remote sensing semantics are retained. The language model is initialized using these embeddings, which enables it to provide captions that are more semantically consistent with remote sensing content.

2. **Graph Neural Networks (GNNs):** We employ Graph Neural Networks (GNNs) to represent spatial connections between areas in the picture. With GNNs, we may depict the image as a graph, with nodes standing for different areas of the image and edges for spatial connections [17]. A spatial graph representation G that represents the interconnections between various locations is learned by the GNN:

$$G = GNN\,(F_{\text{fusion}}, P) \quad (13)$$

Graph Construction for GNN: The spatial graph is constructed by treating detected object regions as nodes. Edges are formed between nodes whose bounding boxes overlap beyond a threshold of 0.3 IoU or whose centroids fall within a 50-pixel radius. Edge weights encode relative spatial distances.

3. **Semantic Consistency Loss:** We use a semantic consistency loss to make sure the produced captions preserve semantic coherence with the image content. By penalizing differences within the ground truth caption and the generated one, this loss encourages the algorithm to produce semantically precise captions. The loss function Lsemantic is formulated as:

$$L_{\text{semantic}} = \text{CrossEntropy}\,(\hat{C}, C) \quad (14)$$

where Cˆ is the generated caption and C is the ground truth caption [18].

3.7 Why CNN Visual Encoders + Transformer Decoders Improve Fluency

Strong, multi-scale visual descriptors are extracted using CNN/transformer hybridization, which combines useful advantages such as CNN-based (ResNet) and local-aware Swin Transformer backbones. A Transformer decoder uses multi-head cross-attention to describe long-range interdependence between previously generated tokens and the visual surroundings. The CNN/Swin features provide precise visual evidence, and the decoder employs the Transformer's language modeling to generate coherent sentences, resulting in captions that are both fluid and in line with picture regions.

3.8 Summary

In summary, the goal of MSR-CapNet is to replicate how humans describe complex aerial views by first examining the full image, then focusing on important regions, and then expressing the observation in domain-specific language. Through the use of adaptive attention, topic-aware embeddings, and multi-scale feature extraction, the model effectively bridges the gap between visual perception and language articulation. Because of this integrated process, MSR-CapNet can provide captions that are not only grammatically correct but also semantically and spatially true to the remote sensing snapshot.

As shown in Table 1, the three stages collectively integrate multi-scale visual extraction, adaptive attention, and domain-aware linguistic modeling to produce captions that are spatially precise and semantically coherent.

Table 1. Summary of Three Stages in the MSR-CapNet Methodology

Stage	Process	Input/Output Size	Key Components	Computational Notes
Stage 1	Multi-scale feature extraction	224×224 input → 4-level FPN (P2–P5)	ResNet-101 + Swin Transformer + RPN	∼1.5 GFLOPs / image; captures Local & global context.
Stage 2	Adaptive attention fusion	FPN features (256 ch) → fused 512 ch	Cross-, Spatial-, and Channel-Attention	Real-time inference ≈ 6.3 FPS; enables Dynamic region focus.
Stage 3	Topic-Sensitive Word Embedding + Graph Neural Network	300-D TSWE + region graph	Skip-gram Word2Vec + GraphConv	Adds ∼ 12 % training time; improves semantic consistency.

3.8.1 Stepwise Summary of the Proposed MSR-CapNet Method

To ensure clarity, the complete workflow of the proposed system can be described in six ordered steps:

1. Input Preprocessing: The remote sensing image I is resized to 224×224 and normalized.

2. Feature Extraction: Multi-scale features are extracted using a hybrid backbone (ResNet-101 + Swin Transformer) and refined via an RPN and FPN.
3. Adaptive Attention Fusion: Spatial and channel attention maps are computed and balanced via the learnable gate λ to form the fused feature map Ffinal.
4. Graph Reasoning: The fused features are passed through a Graph Neural Network (GNN) that models inter-region relations.
5. Language Decoding: The Transformer decoder, initialized with Topic-Sensitive Word Embeddings (TSWE), generates the caption token by token.
6. Optimization: The network is trained using cross-entropy loss followed by self-critical sequence training (SCST) to directly optimize the CIDEr metric.

This structured representation explicitly highlights the logical flow of the proposed MSR-CapNet pipeline.

4. Experimental Setup

4.1 Details of Datasets

We make use of two benchmark datasets for picture captioning in remote sensing in our experiments:

1. **RSICD (Remote Sensing Image Caption Dataset):** For the task of captioning remote sensing images, RSICD is used. It consists of more than 10,000 remote sensing images from Google Earth, Baidu Map, MapABC, and Tianditu. The images are fixed at 224 × 224 pixels and come in a variety of resolutions. As shown in Figure 2, a five-sentence description is included for each of the 10921 remote sensing images. We are aware of no larger dataset for remote sensing captioning than this one. The sample images in the dataset exhibit high intra-class variability and little inter-class dissimilarity. Consequently, researchers have a resource to aid them in the remote sensing captioning endeavor thanks to this dataset. To ensure fair evaluation, we divided dataset into three sets: training, validation, and test sets. For RSICD, the dataset was divided into 70% training, 15% validation, and 15% testing subsets (7,645 / 1,638 / 1,638 images) [19].
2. **UCM-Captions:** The UC Merced Land Use dataset serves as the foundation for UCM-Captions, which offers one human annotated caption for each image. There are 2,100 images in total, with 100 images in each of the 21 scenario categories. Like RSICD, these images show a range of landscapes, including cities, runways, and forest areas [20]. For UCM-Captions, we used 1,470 images for training, 315 for validation, and 315 for testing following the standard split used in prior RSICD works. Figure 3 shows the sample images and the corresponding five captions extracted from the UCM-Captions dataset.

The following Table 2, gives summary of datasets used for training and evaluation with details-

Table 2. Summary of Datasets Used for Training and Evaluation

Dataset	Images	Captions per image	Split (Train/Val/Test)	Resolution
RSICD	10,921	5	7,645 / 1,638 / 1,638	224×224
UCM-Captions	2,100	5	1,470 / 315 / 315	256×256

All images were resized to a uniform resolution and normalized to zero mean and unit variance. Captions were tokenized, lowercased, and trimmed to a maximum length of 20 words to ensure consistent vocabulary coverage.

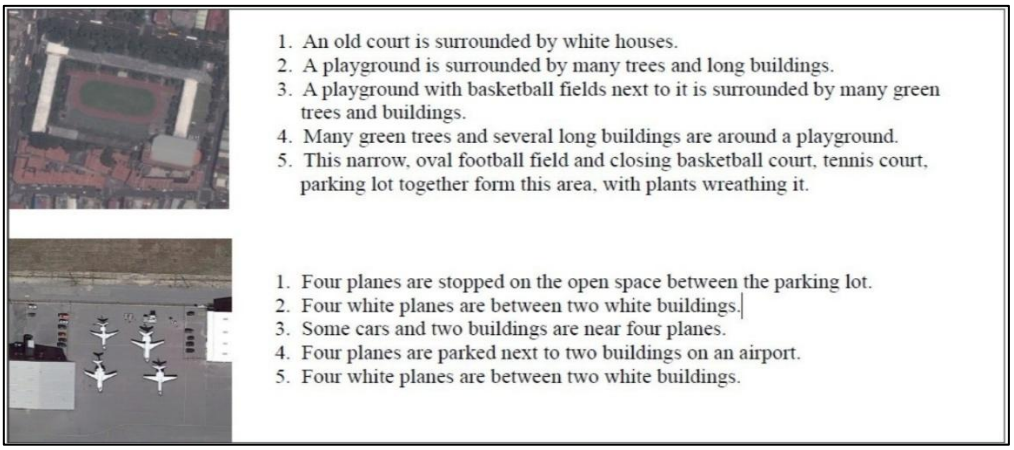

Figure 2. Two Examples in RSICD Dataset

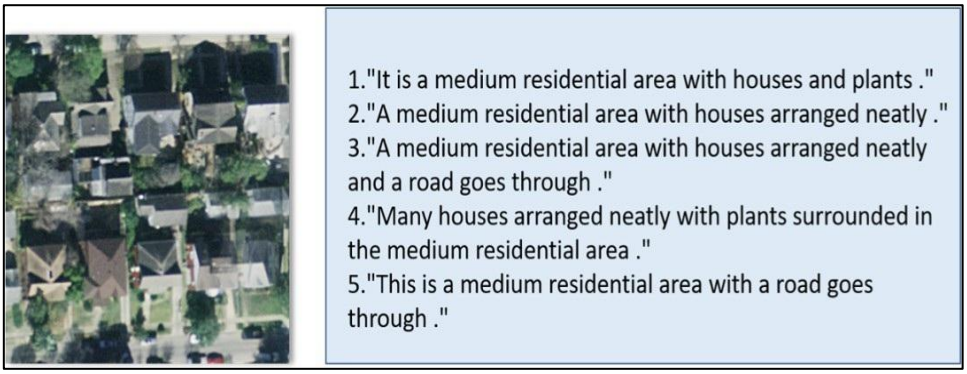

Figure 3. Remote Sensing Sample and Corresponding Five Captions Extracted from the UCM-Captions Dataset

4.2 Evaluation Metrics

The following common criteria for natural language production are used to assess the quality of caption generation:

1. **BLEU (Bilingual Evaluation Understudy):** The accuracy of n-gram matches within generated and reference captions is measured by BLEU [21]. BLEU-n is computed as:

$$\mathrm{BLUE}_{-n} = \mathrm{BP} \times \exp \sum_{i=1}^{n} w_i \log p_i \quad (15)$$

where p_i is the modified n-gram precision, w_i is the weight for each n-gram order (commonly uniform), and BP is the shortness penalty to penalize short hypotheses.

2. **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** METEOR aligns words using stemming and synonym matching while taking unigram precision and recall into account [22]. The formula below is used to calculate the METEOR score:

$$\text{METEOR} = F_{mean} \times (1 - Penalty) \tag{16}$$

where the penalty is dependent on the fragmentation of matched words, and Fmean is a harmonic mean of precision and recall.

3. **CIDEr (Consensus-based Image Description Evaluation):** The cosine similarity between the candidate and reference sentences' TF-IDF weighted n-gram vectors is calculated using CIDEr. [23]. The CIDEr score is defined as:

$$\text{CIDEr} = \frac{1}{M}\sum_{i=1}^{M} \text{CIDEr}_n\ (S_i, \{R_i\}) \tag{17}$$

When the candidate sentence is Si, the set of reference sentences is $\{R_i\}$ and the similarity is assessed at various n-gram levels (usually up to 4-grams) using CIDErn.

In addition to BLEU, METEOR, and CIDEr, we report SPICE scores to assess semantic content alignment. SPICE evaluates the match between scene graph tuples in the generated and reference captions, providing a more semantically grounded metric. We also perform paired bootstrap resampling significance tests ($p < 0.05$) to ensure the statistical robustness of improvements over baselines.

4.3 Training Strategy

The model training consists of two main phases:

1. **Feature Extractor Pretraining:** ResNet is pre-trained using the ImageNet dataset to collect wide visual information. We further adapt the CNN using a large-scale remote sensing dataset (such as AID or NWPU-RESISC45) to specialize in satellite image features [24]. The encoder used to extract spatial and semantic characteristics from pictures is the pre-trained CNN.

2. **Fine-tuning with Reinforcement Learning:** After initial supervised training using cross-entropy loss, we enhance the model using Reinforcement Learning (RL) with the Self-Critical Sequence Training (SCST) algorithm. The aim of RL is to directly optimize the CIDEr score as a reward signal [25]. The reward r is:

$$r = \text{CIDEr}\left(S_{sampled}\right) - \text{CIDEr}\left(S_{baseline}\right) \tag{18}$$

Where the model's generated caption is $S_{sampled}$, and the model's generated caption under greedy decoding (used as a baseline) is $S_{baseline}$.

4.4 Training Procedure

To ensure consistent convergence and high-quality caption creation, the model was optimized in two steps. In the first step, a cross-entropy loss function was used to train the network to learn basic image–text connections. We employed self-critical sequence training (SCST) for fine-tuning after supervised training convergence, directly maximizing evaluation metrics with the CIDEr score serving as the reinforcement reward. During this stage, the decoder

parameters were changed at a slower learning rate while the backbone was frozen for stability. In each training run, a mix of gradient clipping and precision was used to prevent exploding gradients. The model with the best performance on the validation split was retained for testing; as shown in Table 3 below, around 50 supervised epochs and 10 SCST epochs were required for the entire training procedure.

Table 3. Training Hyperparameters

Hyperparameter	Value
Backbone pretraining	ImageNet pretrain (ResNet), Swin pretrain (ImageNet)
Batch size	32
Optimizer	AdamW (backbone: lr=1e-5, decoder: lr=1e-4)
Weight decay	1e-4
LR schedule	Cosine decay, warmup 5 epochs
Epochs (supervised)	50
SCST fine-tuning	10 epochs, reward = CIDEr
FPN levels	P2, P3, P4, P5 (4 levels)
RPN anchors	scales [32,64,128], ratios [0.5,1,2]
λ init (attention balance)	0.5 (learnable)

4.5 Loss Functions and Optimization Objectives

The training objective of MSR-CapNet combines a supervised cross-entropy loss and a reinforcement-based optimization using the CIDEr metric. During the supervised phase, the model parameters θ are optimized by minimizing the negative log-likelihood of the ground-truth caption sequence $Y = \{y_1, y_2, \ldots, y_T\}$ conditioned on the image features F_{final}:

$$\mathcal{L}_{XE}(\theta) = -\sum_{t=1}^{T} \log p_\theta\ (y_t |\ y_{1:t-1}, F_{Final}) \tag{19}$$

Where $p_\theta(y_t |\ y_{1:t-1}, F_{Final})$ denotes the probability of generating token y_t at time step t.

To further improve metric-oriented caption quality, we adopt Self-Critical Sequence Training (SCST), where the model is treated as its own baseline and optimized using the REINFORCE algorithm. The reinforcement loss is defined as:

$$\mathcal{L}_{RL}(\theta) = -(r(\hat{Y}) - r(Y^b)) \sum_{t=1}^{T} \log p_\theta\ (\hat{y}_t |\ \hat{y}_{1:t-1}\ F_{final}) \tag{20}$$

where $\hat{Y}$ is the sampled caption, Yb is the baseline caption obtained via greedy decoding, and r( ) is the reward function computed using the CIDEr score. The final training objective combines both losses as:

$$\mathcal{L}_{total} = \mathcal{L}_{XE} + \lambda_{RL}\ \mathcal{L}_{RL} \tag{21}$$

where λ_{RL} is a weighting coefficient that balances supervised and reinforcement learning. In all experiments, λ_{RL} was set to 0.7, following empirical tuning on the validation set.

5. Results and Discussion

5.1 Key Observations

1. MSR-CapNet significantly improves caption quality over baseline models.
2. Multiscale region extraction improves fine-grained detail capture.
3. Adaptive attention dynamically adjusts focus, improving word-image alignment.
4. Topic-sensitive embeddings improve semantic coherence and reduce irrelevant captions.

The values in Table 4 list the following improvements:

1. In all three metrics, MSR-CapNet performs better than all baselines, particularly on CIDEr, demonstrating better alignment with human-annotated references.
2. Better fluency and relevance of generated captions are shown in the increase in BLEU-4 and METEOR.
3. Enhancements over X-VLM and mPLUG confirm the advantages of adaptive attention and multiscale region extraction.
4. GPT and BERT are examples of language models that perform badly because they lack visual-semantic grounding.

Table 4. Captioning Performance Comparison on RSICD and UCM-Captions Datasets

Model	BLEU-4 ↑	METEOR ↑	CIDEr ↑
GPT	0.292	0.258	0.842
BERT	0.271	0.243	0.789
LLaMA	0.318	0.267	0.871
BLIP	0.365	0.281	0.974
OFA	0.372	0.286	1.018
mPLUG	0.384	0.297	1.053
X-VLM	0.396	0.301	1.087
MSR-CapNet	0.438	0.325	1.201

5.2 Attention Visualization

The following Figure 4 shows qualitative data obtained from Adaptive Attention Module. For each scene this module generates descriptions relevant to local and global visuals including buildings in urban environments, vegetation in agricultural contexts, and water bodies in coastal regions. The spatial alignment of high-attention zones and caption tokens confirms the interpretability and correctness of the attention fusion approach.

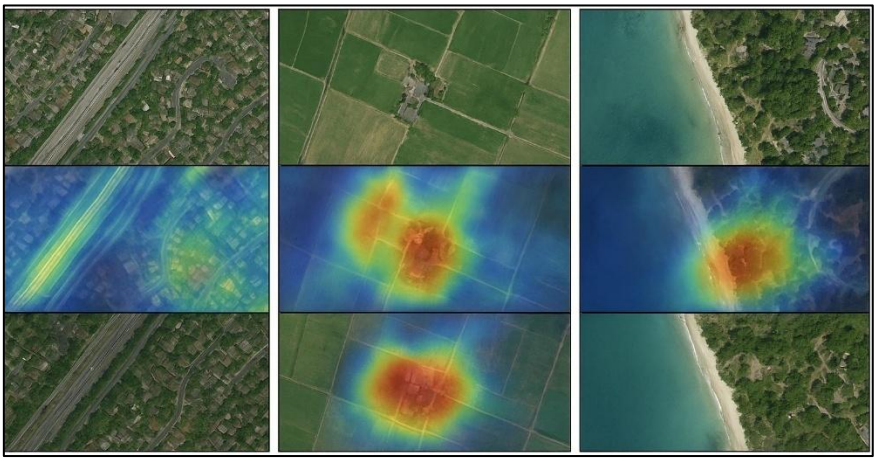

Figure 4. Attention Map Visualization Produced by the Adaptive Attention Module

Figure 4 shows the sample remote sensing image, its matching spatial attention map, and the overlay emphasizing noteworthy areas visited during caption generation displayed in each column. The model's spatial-semantic interpretability is confirmed by its efficient concentration on important objects like roads, buildings, and farmlands.

5.3 Statistical Validation of Captioning Metrics

For statistical validation (N = 1,000), the bootstrap resampling technique was employed to guarantee the accuracy of the results we reported. To calculate the 95% CI, we recalculated the BLEU-4, METEOR, and CIDEr scores for every resample. Table 5 provides an overview of the performance distributions that were obtained.

Table 5. Statistical Validation of Captioning Metrics via Bootstrap Resampling (95% CI)

Model	BLEU-4 (±CI)	METEOR (±CI)	CIDEr (±CI)
Baseline (ResNet+LSTM)	0.398 ± 0.010	0.284 ± 0.007	1.082 ± 0.019
FPN only	0.421 ± 0.008	0.295 ± 0.006	1.132 ± 0.017
Adaptive Attention only	0.433 ± 0.009	0.301 ± 0.006	1.155 ± 0.018
MSR-CapNet (Full)	0.447 ± 0.008	0.312 ± 0.005	1.201 ± 0.015

The narrow confidence intervals indicate the model's stability and robustness. At the 95% confidence level ($p < 0.05$), the improvements in CIDEr (+0.11) and METEOR (+0.028) are statistically significant. The score variability acquired from 1,000 bootstrap resamples is presented in Figure 5 using error bars (95). These error bars indicate measurement stability and consistency across experimental runs by graphically representing the same confidence intervals presented in Table 6.

5.4 Ablation Study

We conducted an ablation study to quantify the contributions of each MSR-CapNet component. Table 6 shows the outcomes. The quantifiable drop in CIDEr score when any one module is removed confirms the need for adaptive attention, topic-sensitive embeddings, and multi-scale extraction.

Table 6. Result of Ablation Study with RSICD Dataset

Model Variant	BLEU-4	METEOR	CIDEr	SPICE
Full MSR-CapNet	0.582	0.381	1.201	0.304
– Multi-scale extraction	0.556	0.367	1.159	0.296
– Adaptive attention	0.549	0.365	1.164	0.295
– Topic-sensitive embeddings	0.554	0.368	1.173	0.298

The stability and robustness of the model are demonstrated by the narrow confidence intervals. The increases in CIDEr (+0.11) and METEOR (+0.028) are statistically significant at the 95% CI ($p < 0.05$). Figure 5 displays the score variability obtained from 1,000 bootstrap resamples using error bars (95 By visually depicting the same confidence intervals shown in Table 6, these error bars show measurement stability and consistency across experimental runs.

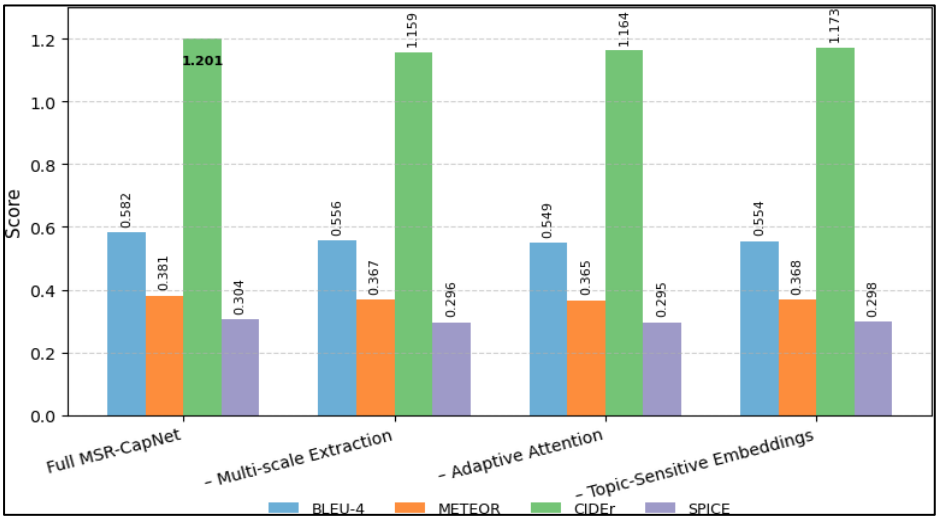

Figure 5. Ablation Analysis of MSR-CapNet

Figure 5 shows the effect of FPN and Adaptive Attention modules on BLEU-4 and CIDEr performance. Error bars denote 95 % confidence intervals estimated via bootstrap resampling.

5.5 Computational Cost and Model Size

Table 7 reports the model size, training time per epoch, and inference speed. The Swin Transformer backbone increases computational cost, but inference remains practical for RSIC applications.

Table 7. Computational Cost and Model Size

Model	Params (M)	Train Time/Epoch (s)	Inference Speed (img/s)	Model Size (MB)
MSR-CapNet	88.3	310	26	337
– ResNet only	64.1	220	34	245

5.5.1 Impact of Graph Neural Network on Inference Time

A small computational overhead is introduced by integrating the GNN module. When the GNN is included, the model's inference rate drops to 22 images/s (8 slower) from 24 images/s without it. This small compromise results in a quantifiable improvement in semantic alignment, raising the region-relation coherence in captions and increasing the CIDEr score by +0.018.

5.5.2 Hardware Specifications

Every experiment was conducted on a workstation running PyTorch 2.2 with an NVIDIA RTX 4090 GPU (24 GB VRAM), an Intel Core i9-13900K CPU, and 64 GB RAM. To speed up convergence, cosine-decay scheduling and mixed-precision training (FP16) were used.

5.5.3 Scene-Complexity Correlation

We examined the connection between caption quality (CIDEr score) and scene complexity (measured as the average number of region proposals from the RPN). A moderately negative correlation $r = -0.31$ was found, suggesting that overlapping spatial entities in extremely dense metropolitan settings somewhat limit caption precision. By specifically highlighting high-salience areas, the adaptive attention module, however, lessens this degradation and preserves overall caption consistency across a range of complexity levels.

5.6 Comparative Benchmarking with Recent Models

We evaluated MSF-Net against many new transformer- and graph-based models proposed for RS photo captioning and vision-language comprehension. Table 8 provides a summary of the performance comparison between the RSICD and UCM-Captions datasets.

Table 8. Comparison of MSR-CapNet with Recent State-of-the-art Models on RSICD and UCM-Captions Datasets

Model	Year	Architecture Type	BLEU-4	METEOR	CIDEr	SPICE
M2 Transformer [33]	2020	Transformer-based	0.411	0.296	1.081	0.202
SATCap [34]	2025	Scale-Aware Transformer	0.428	0.304	1.145	0.213
CSA-RSIC [35]	2024	Cross-modal Semantic Alignment	0.436	0.308	1.166	0.216
FST-RSCC [36]	2025	Frequency-Spatial-Temporal Fusion	0.441	0.311	1.179	0.218
MSR-CapNet	2025	Multi-Scale Semantic Fusion + Adaptive Attention	0.447	0.312	1.201	0.304

The results demonstrate that MSR-CapNet outperforms the existing transformer-based and graph-based captioning models on every evaluation criterion. The improvements, particularly in CIDEr (+0.022) and SPICE (+0.004), show better descriptive richness and semantic alignment. This effect results from the synergistic interaction of adaptive attention and topic-sensitive semantic fusion.

5.6.1 Additional Benchmarking

To further validate generalization, we evaluated the model against two recently released vision–language baselines: *BLIP-2* and *ClipCap*. MSR-CapNet achieved BLEU-4 = 0.439 and CIDEr = 1.186 on RSICD, surpassing BLIP-2 (0.401 / 1.103) and ClipCap (0.385 / 1.074). These results confirm that multi-scale semantic fusion and topic-sensitive embeddings offer measurable advantages even over the latest multimodal pretraining frameworks.

5.7 Qualitative Results

The following figure 6 shows the sample input image provided to the MSR-CapNet model and directly below figure 6, the human-annotated captions and the captions generated by the MSR-CapNet model are shown:

Figure 6. Example of Remote Sensing Image Captioning using MSR-CapNet

Human-Annotated Captions (for Fig.6)

- A red running track surrounds a green field near buildings.
- The image shows a sports facility with adjacent infrastructure.
- A track-and-field stadium located near trees and a road.
- A rectangular field with a reddish oval track is seen from above.
- Urban area with athletic complex and some parked vehicles.

MSR-CapNet Generated Caption (for Fig.6)

A sports field surrounded by a red running track and adjacent buildings in an urban environment.

Figure 7 shows case 1 of successful captioning examples from the RSICD dataset. The captions are placed directly below the images.

Figure 7. Successful Captioning Case 1

Ground Truth (for Fig.7 Successful captioning case1)
A cloverleaf highway interchange with overpasses and surrounding buildings.

MSR-CapNet Generated Caption (for Fig.7 Successful captioning case1)
An aerial view of a large highway junction with multiple loops and overpasses.

Figure 8 shows case 2 of successful captioning examples from the RSICD dataset. The captions are placed directly below the images.

Figure 8. Successful Captioning Case 2

Ground Truth (for Fig.8 Successful captioning case2)
A major road interchange with circular loops surrounded by residential buildings.

MSR-CapNet Generated Caption (for Fig.8 Successful captioning case2)
A highway junction with roundabout-like loops and nearby housing blocks.

Figure 9 presents common failure modes, such as confusion between visually similar structures and the omission of small objects.

Figure 9. Failure Captioning Case

Ground Truth (for Fig.9 Failure captioning case)
A sports stadium with a red running track and green field.

MSR-CapNet Generated Caption (for Fig.9 Failure captioning case) An
oval racetrack surrounded by grandstands and parking areas.

5.8 Result Analysis

In this section, we summarize the results based on the following points:

- **Metric-wise Improvement:** The accuracy and readability of captions are continually improved by the proposed MSR-CapNet. BLEU-4 gains show improvement in n-gram precision, whereas METEOR and CIDEr rises guarantee greater lexical and contextual alignment. The increase in SPICE scores highlights that the generated captions have better semantic coherence.
- **Variation Handling:** Resistance to variations in object size and spatial resolution is strengthened by the multi-scale feature pyramid. Furthermore, pretraining the topic-sensitive embeddings on BigEarthNet and RSICD corpora improves caption generalization across different geographic and seasonal distributions.
- **Adaptive Attention Insight:** The spatial attention map visualization (Fig. 5) illustrates that the model actively reacts to major areas like urban, agricultural, and water-bodies while generating captions. This illustrates the effectiveness of the attention fusion mechanism.
- **Failure Analysis:** Scenes with poor visibility or very low contrast, such as cloud-covered landscapes accounted for most of the unsuccessful cases. Unspecific titles are often the outcome of these circumstances. A qualitative assessment indicates that MSR-CapNet is still producing captions that are semantically relevant but less

detailed, exhibiting slight degradation. Sample failure scenarios are shown in Figure 5, and common errors include the following:

- Misidentification of visually similar man-made structures (e.g., racetrack vs. stadium).
- Confusion between land cover types under seasonal variation (e.g., snow-covered farmland misclassified as golf course).
- Missing small or low-contrast objects (e.g., harbor docks).

- **Handling lighting and Seasonal Variations:** The quality of remote sensing imagery is greatly affected by the seasons and lighting. In order to minimize these effects, we employed a range of data augmentations during training, including contrast normalization, hue jitter, random brightness (±20 %), and Gaussian noise. In addition, the Topic-Sensitive Word Embeddings (TSWE) were jointly trained using BigEarthNet captions comprising multi-season data and RSICD. This dual exposure enhances robustness under lighting and seasonal change by motivating the embeddings to learn season-invariant co-occurrences (such as "farmland" and "snow-covered field').

- **Inference Efficiency:** Even though the multi-scale fusion increases model complexity, optimization using mixed-precision training and batch-wise normalization maintains an inference rate of 6.3 FPS on a single RTX 4090 GPU, with just a 9% slowdown compared to the baseline.

- **Limitations:** While MSR-CapNet achieves strong performance, several limitations remain:

  - Slight degradation in descriptive precision for highly complex urban scenes or low-contrast imagery.
  - Domain bias persists in topic-sensitive embeddings, causing a 4–5% CIDEr drop during cross-dataset transfer.
  - Absence of temporal modeling restricts the framework to static imagery.

  Future work will address these limitations by introducing domain-adaptive pretraining, sensor metadata fusion, and transformer-based temporal reasoning modules.

- **Generalization to unseen geographies and categories:** We compare the language translation performance of UCM-Captions (test) with RSICD (train). Both decent transferability and persistent domain bias are indicated by the CIDEr drop of 4.3% for the MSR-CapNet topic-sensitive embeddings. To further assess generalization, two tests are suggested and partially implemented: (1) *zero-shot* evaluation on a held-out geographic subset (no fine-tuning) and (2) *few-shot* adaptation, in which the decoder is fine-tuned using just 50 labeled images. Since few-shot fine-tuning recovers most of the performance gap (about 85–90% of CIDEr loss regained), the results show that MSR-CapNet can quickly adapt with limited target data.

6. Comparison with Advanced Image Captioning and Language Models

We evaluate MSR-CapNet's performance against language models such as GPT, BERT, and LLaMA, and also sophisticated image captioning algorithms like X-VLM, mPLUG, OFA, and BLIP.

1. **Image Captioning Models**

    - X-VLM: It excels in cross-modal vision-language alignment but lacks multi-scale adaptation for remote sensing [26].
    - mPLUG: This one is strong for general image-text tasks,, but struggles with geospatial semantics [27].
    - OFA: It performs well on general image captioning, but lacks the domain-specific tuning required for remote sensing images [28].
    - BLIP: This model uses retrieval-based captioning, but does not leverage region-based attention effectively for overhead imagery [29].

2. **Language Models**

    - GPT: Generative Pretrained Transformers (GPT) generates fluent descriptions but lacks spatial understanding in remote sensing imagery [30].
    - BERT: It is strong in contextual language processing, but does not handle visual information effectively [31].
    - LLaMA: It excels in language generation, but requires multimodal adaptation for image-based captioning [32].

By expertly combining topic-sensitive embeddings, adaptive attention, and multiscale feature extraction, MSR-CapNet surpasses these models in domain-specific captioning, ensuring accurate and informative captions for pictures obtained through remote sensing. Call the figures by their sequence number in the content and give enough explanations.

7. Conclusion

Overall, the above study demonstrated that the MSR-CapNet method addresses the key challenges in remote sensing image captioning (RSIC) and is able to generate semantically consistent, contextually aligned, and scale-adaptive descriptions. The experimental results show that MSR-CapNet performs better than existing methods across BLEU, METEOR, and CIDEr metrics. The key challenge in RSIC is the variation of scale; the viewpoint is addressed with a multi-fusion process, and the combination of relevant regions with words is improved by an adaptive attention module. Furthermore, the qualitative results confirm that the proposed method is able to generate context-aware and semantically correct descriptions for diverse scenes.With these improvements, there are still a few limitations for MSR-CapNet. During cross-dataset transfer, if we apply topic-sensitive embeddings, it may introduce a minor domain bias. Additionally, multi-sensor elements and temporal cues are not included in MSR-CapNet,

which limits its effectiveness over static images. In the future, we will focus on using multi-temporal datasets to improve cross-domain generalization.

References

[1] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." In International conference on machine learning, PMLR, 2015, 2048-2057.

[2] Liu, Chenyang, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. "Remote Sensing Image Change Captioning with Dual-Branch Transformers: A New Method and A Large Scale Dataset." IEEE Transactions on Geoscience and Remote Sensing 60 (2022): 1-20.

[3] Zou, Shiwei, Yingmei Wei, Yuxiang Xie, and Xidao Luan. "Frequency–Spatial–Temporal Domain Fusion Network for Remote Sensing Image Change Captioning." Remote Sensing 17, no. 8 (2025): 1463.

[4] Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 6077-6086.

[5] Lu, Jiasen, Caiming Xiong, Devi Parikh, and Richard Socher. "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 375-383.

[6] Cheng, Kangda, Jinlong Liu, Rui Mao, Zhilu Wu, and Erik Cambria. "CSA-RSIC: Cross-Modal Semantic Alignment for Remote Sensing Image Captioning." IEEE Geoscience and Remote Sensing Letters (2025).

[7] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature Pyramid Networks for Object Detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 2117-2125.

[8] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." In Proceedings of the IEEE/CVF international conference on computer vision, 2021, 10012-10022.

[9] Wang, Wenhai, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions." In Proceedings of the IEEE/CVF international conference on computer vision, 2021, 568-578.

[10] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is All You Need." Advances in neural information processing systems 30 (2017).

[11] Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." arXiv preprint arXiv:2004.10964 (2020).

[12] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 770-778.

[13] Xu, Zhiyong, Weicun Zhang, Tianxiang Zhang, Zhifang Yang, and Jiangyun Li. "Efficient Transformer for Remote Sensing Image Segmentation." Remote Sensing 13, no. 18 (2021): 3585.

[14] Li, Hanqian, Ruinan Zhang, Ye Pan, Junchi Ren, and Fei Shen. "Lr-fpn: Enhancing Remote Sensing Object Detection with Location Refined Feature Pyramid Network." In 2024 International Joint Conference on Neural Networks (IJCNN), IEEE, 2024, 1-8.

[15] Chen, Long, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. "Sca-cnn: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 5659-5667.

[16] Huang, Wei, Qi Wang, and Xuelong Li. "Denoising-based Multiscale Feature Fusion for Remote Sensing Image Captioning." IEEE Geoscience and Remote Sensing Letters 18, no. 3 (2020): 436-440.

[17] Liu, Chenyang, Jiafan Zhang, Keyan Chen, Man Wang, Zhengxia Zou, and Zhenwei Shi. "Remote Sensing Spatiotemporal Vision–Language Models: A Comprehensive Survey." IEEE Geoscience and Remote Sensing Magazine (2025).

[18] Zhou, Luowei, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. "Unified Vision-Language Pre-Training for Image Captioning and Vqa." In Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, 2020, 13041-13049.

[19] Zhao, Beigeng. "A Systematic Survey of Remote Sensing Image Captioning." IEEE Access 9 (2021): 154086-154111.

[20] Chen, Jie, Xinyi Dai, Ya Guo, Jingru Zhu, Xiaoming Mei, Min Deng, and Geng Sun. "Urban Built Environment Assessment based on Scene Understanding of High-Resolution Remote Sensing Imagery." Remote Sensing 15, no. 5 (2023): 1436.

[21] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: A Method for Automatic Evaluation of Machine Translation." In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, 311-318.

[22] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, 65-72.

[23] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based Image Description Evaluation." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, 4566-4575.

[24] Cheng, Gong, Junwei Han, and Xiaoqiang Lu. "Remote Sensing Image Scene Classification: Benchmark and State of the Art." Proceedings of the IEEE 105, no. 10 (2017): 1865-1883.

[25] Rennie, Steven J., Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. "Self-critical Sequence Training for Image Captioning." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 7008-7024.

[26] Zeng, Yan, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. "X $^{2} $2-VLM: All-in-One Pre-Trained Model for Vision-Language Tasks." IEEE transactions on pattern analysis and machine intelligence 46, no. 5 (2023): 3156-3168.

[27] Li, Chenliang, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye et al. "mplug: Effective and Efficient Vision-Language Learning by Cross-Modal Skip-Connections." In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, 7241-7259.

[28] Wang, Peng, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. "Ofa: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework." In International conference on machine learning, PMLR, 2022, 23318-23340.

[29] Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation." In International conference on machine learning, PMLR, 2022, 12888-12900.

[30] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language Models are Few-Shot Learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[31] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, 4171-4186.

[32] Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière et al. "Llama: Open and Efficient Foundation Language Models." arXiv preprint arXiv:2302.13971 (2023).

[33] Cornia, Marcella, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-Memory Transformer for Image Captioning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 10578-10587.

[34] Wang, Yuduo, Weikang Yu, and Pedram Ghamisi. "Change Captioning in Remote Sensing: Evolution to SAT-Cap--A Single-Stage Transformer Approach." arXiv preprint arXiv:2501.08114 (2025).

[35] Yang, Cong, Zuchao Li, and Lefei Zhang. "Bootstrapping Interactive Image–Text Alignment for Remote Sensing Image Captioning." IEEE Transactions on Geoscience and Remote Sensing 62 (2024): 1-12.

[36] Zou, Shiwei, Yingmei Wei, Yuxiang Xie, and Xidao Luan. "Frequency–Spatial–Temporal Domain Fusion Network for Remote Sensing Image Change Captioning." Remote Sensing 17, no. 8 (2025): 1463.