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

Domain-Specialized Multi-Agent Collaboration with Consensus-Aware Fusion for Remote Sensing Image Captioning

Chandrashekhar S. Pawar[1] , Dr. Ashwin Makwana[1]*

[1]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 , 388421, Gujarat, India, India

* Corresponding author's Email: ashwinmakwana.ce@charusat.ac.in

Abstract:
Now a days Remote sensing image captioning (RSIC) helps a lot in understanding complex satellite images by generating human-understandable textual descriptions. These textual descriptions useful in many domains such as land-use analysis, disaster management, urban planning, and environmental monitoring. For generating such descriptions, most of the current methods use single-agent encoder-decoder frameworks, which perform well to some extent but generally have trouble with domain generalization, semantic richness, and robustness in the varied and heterogeneous environments shown in remote sensing images. To overcome these restrictions, we presented MAC-CapNet (Multi-Agent Collaborative Captioning Network), an innovative architecture that utilizes the collaboration of many specialized agents, and each agent trained to concentrate on a certain land-use domain (e.g., urban, forest, agricultural, water bodies, barren land). These agents independently gives us candidate captions and confidence scores, which are then combined using a dynamically weighted meta-decoder that takes into account both contextual relevance and agent trustworthiness. The final caption is generated through a context-aware gating technique. The improvements we made include: (1) a multi-agent architecture designed specifically for RSIC; (2) a unique meta-decoder that dynamically combines agent predictions; and (3) numerous experiments evidence of higher performance on standard datasets. Evaluations carried out with RSICD and UCM-Captions demonstrate considerable gains in BLEU-4, METEOR, and CIDEr scores compared to cutting-edge approaches. The proposed MAC-CapNet was evaluated on widely used benchmark datasets for remote sensing image captioning. Across all experiments, it achieved better overall performance than the baseline models on BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE. We also carried out domain-wise comparisons, robustness experiments, scalability analysis, and statistical evaluation. Together, these results show that the framework performs reliably under different experimental settings and supports the design choices introduced in this work.
Keywords: Remote Sensing (RS), Satellite Image Processing, RS Caption generation, Land Cover Classification

I. Introduction

The Satellite revolving around the earth capture large amounts of data about the Earth, which includes information about recording cities as they grow, forests as they change, rivers as they shift, and farmlands as they evolve. This captured data (images) can give us useful information for various domains such as environmental monitoring [1], urban planning [2], disaster response [3], agriculture [4].

However, to extract such useful information from remote sensing imagery needs significant knowledge, and manual analysis, but it becomes impractical as data volumes grow.

Problem Statement

To Recognition Remote Sensing Images are very complex and variable from each other. Urban areas are well plan, which consists of human-made structures; agricultural regions display nearly similar views of repeating field layouts; forests can be classified by uneven textures; and water body have completely different visual features. In many RS images consists of more than one of these land-use categories.

The many RSIC methods are handling this variation of land-use categories with the help of homogenous captioning model. The captions generated using these methods are grammatically correct, but not covers all land-use categories. Like for example, industrial areas are sometimes referred to as ``buildings," whereas water regions are sometimes confused with agricultural land. This incorrect labeling is mostly done in RS images with multiple land-use categories.

Recent improvements in large-scale vision-language pretraining, such as BLIP and its associated models,

also advances done in computer vision and natural language processing [5,6], helps to generate more accurate captions. The traditional RS captioning methods majorly makes use of encoder-decoder architectures, and now a days these methods are improved with region-based attention [7] or these methods are trained using domain specific datasets such as RSICD [8]. But these methods are trained on generic concepts, and they do not include domain specialization procedures. So, by observing this we can say that improved pretraining alone not able to address the basic problem of domain heterogeneity in remote sensing.

II. Problem Statement

To overcome the limitation of homogenous captioning models, we introduce a novel approach of using specialized agents for each land-use category, instead of single prediction task. This idea is based on human analysis of complex situations, where experts from different fields work together to reach a final decision.

Based on this idea, we proposed MAC-CapNet (Multi-Agent Collaborative Captioning Network). Within MAC-CapNet we use a group of specialized agents, one agent for each specific land-use category, and each agent is fine-tuned only for that specific category-such as urban, agriculture, water, forest or barren land.

This allows each agent to produce a separate caption with confidence score that reflects the relevance of specific domain knowledge. To generate the final output, all these specialized agents must be able to communicate with each other; for that purpose, we developed the Cross-Agent Dynamic Routing (CADR) module. Using this CADR module, agents exchange semantic signals, resolve conflicts, and filter out redundancies before generating the final output. Next, the Consensus-Aware Meta-Decoder (CAMD) is responsible for generating the final caption. For this purpose, it combines the output of all agents considering both contextual agreement and agent reliability.

The major contributions of our work are listed below:

A Multi-Agent Paradigm: We are using multiple agents that specialize in one particular land-use category, and this approach We used to handle the inherent heterogeneity of remote sensing images [9].

Specialized Agent Design: We design a specialized agent for each specific land-use category and these agents are allowed to communicate with each other for exchanging common visual representation [10].

Semantic Routing Scheme: We use dynamic routing mechanism to allow communication between agents while also guaranteeing a consistent and non-redundant description of complex scenes.

Trust-Based Fusion: To ensure the generation of precise results, we suggested a meta-decoder which works on a consensus-aware approach, incorporating an explicit notion of “agent trust” and contextual alignment.

Rigorous Benchmarking: We demonstrate that MAC-CapNet consistently sets new performance standards across the RSICD dataset, validated through a comprehensive suite of statistical analyses, ablation studies, and qualitative reviews.

III. Rationale for Method Selection

The design of MAC-CApNet is based on three main pillars: specialization, collaboration, and interpretability. The main aim of the selection of this architecture is to address the unique challenges of remote sensing, where high-level abstraction is unable to capture the granular details.

The main focus of the proposed method is on domain specialization. As most of the satellite images are heterogeneous in nature, the singular/ monolithic model struggles to generalize across different land-use categories. So, to minimize this issue, we use specialized agents for different land-use categories. The specialized agents can more accurately represent the unique visual features of each domain.

However, only specialization is not sufficient for complex scenes that contain overlapping land-use categories. For this purpose, we include the cooperative technique using the Cross-Agent Dynamic Routing module.

This results in effective information exchange between agents and resolves uncertainty without having to pay the high computational costs associated with traditional dense networks.

Finally, we give top priority to robustness and interpretability by using our Consensus-Aware Meta-Decoder. Instead of considering fusion to be a ``black box," our system clearly assesses contextual agreement and agent reliability. This assures us that the generated final caption is not just a statistical average, instead is most relevant to the all-domain present in remote sensing image.

Literature Review

CNN–RNN based RSIC Methods

The RSIC methods developed in past were use CNN-RNN architecture, to integrate visual feature extraction and sequential language generation [11]. For generating meaningful caption for aerial imagery, encoder-decoder framework was used in [11]. And forming this encoder-decoder framework as a base, Anderson et al. developed attention-based extensions to focus on salient image regions [12].

Critical Comparison: Feature Representation: The CNN–RNN based RSIC Methods focus on global level

features only, not able to capture fine spatial details. Whereas Attention mechanisms capture local spatial details, but using fixed-resolution features. However, both approaches lack explicit multi-scale fusion.

Limitation and Gap: These methods are not able to handle multi-scale information or domain-specific variations such as rotation invariance or scale-adaptive feature extraction. Due to this limitation these methods are not able to describe complex scenes.

Attention and Transformer-based Methods

Transformer-based architectures [14, 15] are used previously to improve long-range dependency modelling in image captioning. By replacing self-attention with recurrent structures, these techniques improve contextual reasoning across textual and visual modalities.

Critical Comparison: Context Modelling: Transformer-based architectures capable of capturing contextual data between image regions and words. However, a clear distinction between domain-specific semantics is missing.

Limitation and Gap: The domain heterogeneity found in remote sensing images is not naturally addressed by transformers, despite the fact that they enhance sequence modelling. Generic captions that overlook minute semantic differences are frequently the result of a lack of explicit domain separation.

Graph-based and Relational Models

Graph-based approaches extend captioning models by relating objects and regions with each other [24, 22]. The representation of images as graphs helps to capture spatial and semantic dependencies more effectively.

Critical Comparison: Structural Modeling: Graph-based models provide a structured representation of relationships, which improves reasoning over spatial layouts. However, their performance depends heavily on the quality of object detection and graph construction.

Limitation and Gap: These methods often struggle in remote sensing scenarios where object boundaries are unclear or densely packed. Additionally, they do not explicitly incorporate domain-level semantics, limiting their effectiveness in multi-domain scenes.

Vision–Language Pretraining Models

BLIP and BLIP-2 [13], two recent developments in vision-language pretraining, have greatly enhanced captioning performance by utilizing extensive multimodal data. These models acquire robust cross-modal representations that perform well on a variety of tasks.

Critical Comparison: Generalization vs. Specialization: Due to their exposure to multiple datasets, pretrained models exhibit high performance. However, they are mostly trained on natural images, which differ significantly from remote sensing data in terms of scale,

viewpoint, and semantics.

Limitation and Gap: Without domain-specific adaptation, these models often produce captions that are semantically sound but not precise enough for distant sensing applications. This highlights the need for certain systems that can integrate domain knowledge.

Motivation for Multi-Agent Captioning

The drawbacks of existing approaches highlight a common issue: the inability to effectively express semantic variance in complex distant sensing contexts. Single-model architectures, whether CNN-based, transformer-based, or graph-based, seek to capture all semantics in a single representation, which usually leads to loss of detail or generalization issues.

Key Insight: Remote sensing images frequently contain multiple co-existing domains (e.g., urban, vegetation, water), each with distinct visual patterns. A single model may struggle to represent all these domains simultaneously.

Motivation: To address this, a multi-agent framework can be used, where each agent specializes in a particular domain and contributes to the final caption. This allows for more precise semantic modeling and better interpretability.

Research Gap: Existing RSIC methods lack mechanisms for domain-specific specialization and collaborative reasoning. This gap motivates the proposed MAC-CapNet, which introduces coordinated multi-agent caption generation to better handle complex, multi-domain scenes.

Discussion: Critical Comparison and Gap Analysis

Table 1 provides a critical comparison of MAC-CapNet with existing image captioning approaches. This comparison list outs the key limitations and the resulting research gaps, and all these gaps are related to semantic understanding, interpretability, and domain awareness for remote sensing imagery.

Table 2 presents a mapping between the listed research gaps, in existing work, and related contributions of

the proposed MAC-CapNet framework. This mapping clearly shows that proposed approach addresses the drawbacks of existing methods by using domain-specific specialization, structured semantic decomposition, and improved interpretability.

Methodology

Overview

In this section, we explain our proposed Multi-Agent Consensus Captioning Network (MAC-CapNet), used for remote sensing image captioning. The overview of the proposed MAC-CapNet framework is shown in Figure 1 and later elaborated using Figure 2 and 3. We introduced this MAC-CapNet framework to deal with two major concerns semantic heterogeneity and multi-domain nature of remote sensing images and we are able to address these concerns by dividing caption generation process into domain-specialized reasoning, inter-agent communication, and consensus-aware fusion, as shown in Figure 3.

As shown in Figure 2, the input image is processed by a shared visual encoder, followed by multiple Domain-Specialized Captioning Agents (DSCAs). The outputs are refined through Collaborative Agent Discussion and Routing (CADR) and fused using the Collaborative Agent Mixture Decoder (CAMD) with agent-level confidence weights γ_i to generate the final caption. Existing single-agent models or late fusion aggregation methods [9, 16], ignore domain-specific details or fuse output post-hoc. MAC-CapNet explicitly models both specialization (agents focus on distinct land cover domains) and collaboration (agents exchange information before generating the final caption) [10, 17]. This dual approach plays an important role in remote sensing image captioning, as this approach is capable of handling highly heterogeneous RS imagery. The architecture consists of three primary modules (illustrated in Figure 2):

Domain-Specialized Captioning Agents (DSCAs): This module consists of a set of specialized agents design for each specific domain. [10].

Cross-Agent Dynamic Routing (CADR): By using this module, agents share the information and refine the caption via agreement-based attention.

Consensus-Aware Meta-Decoder (CAMD): In this module, the refine output obtain from agents combine using trust-aware gated aggregation [13] to generate the final caption.

Overview of MAC-CapNet Architecture

Figure 2 shows the work flow of propose MAC-CapNet architecture. The MAC-CapNet accepts remote sensing images as an input I, and initially high-level visual features are extracted from input image using a pretrained vision encoder. These extracted features are then process by multiple Domain-Specialized Captioning Agents (DSCA), here every agent focus only on one specific land-use domain. Next, there is a need to share the contextual information among agents by allowing them to communicate with each other and we achieve this via a Cross-Agent Dynamic Routing (CADR) mechanism. Finally, the all-agent outputs are combined using Consensus-Aware Meta-Decoder (CAMD) to generate a coherent and semantically consistent caption.

Formally, the caption generation process is defined as:

$$C_{final} = \mathrm{CAMD}\left(\{c_i\}_{i=1}^{N}\right) \quad (1)$$

where c_i denotes the candidate caption produced by the $i-$ th agent and N is the total number of agents.

Above Equation (1) represents the final caption generation process. Each Domain-Specialized Captioning Agent produces a candidate caption c_i. These candidate captions are converted into semantic embeddings and passed to the Consensus-Aware Meta-Decoder (CAMD). The CAMD module evaluates contextual agreement and agent trust weights to combine the candidate captions and generate the final caption c_{final}.

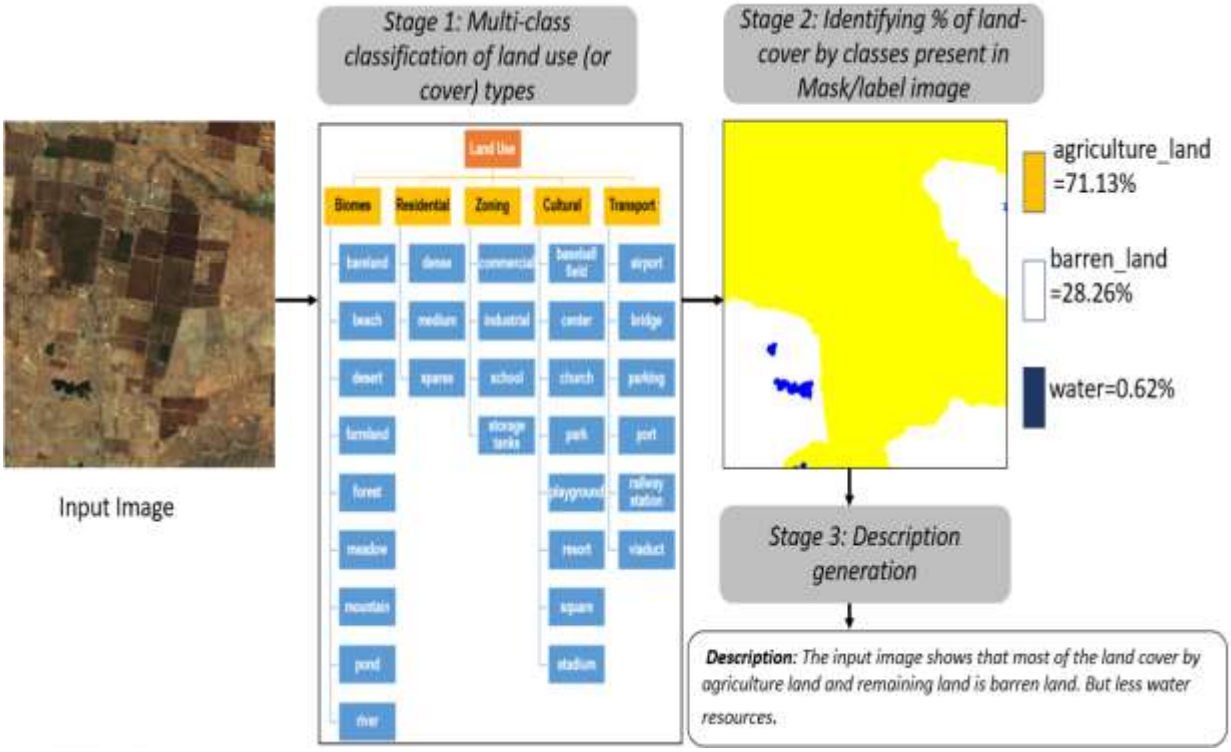


Fig. 1. Overview of the proposed MAC-CapNet framework.

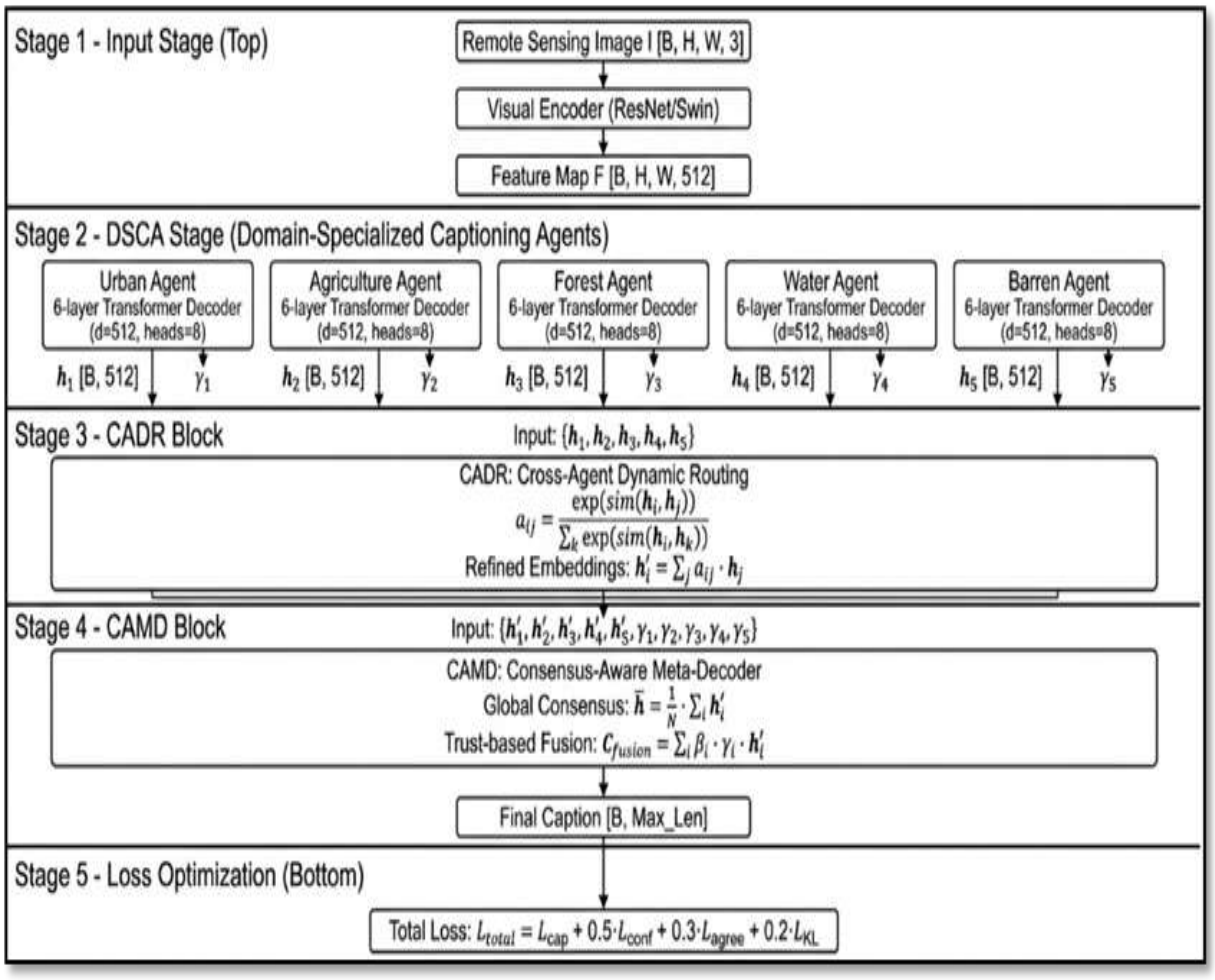


Fig. 2. Architecture of MAC-CapNet

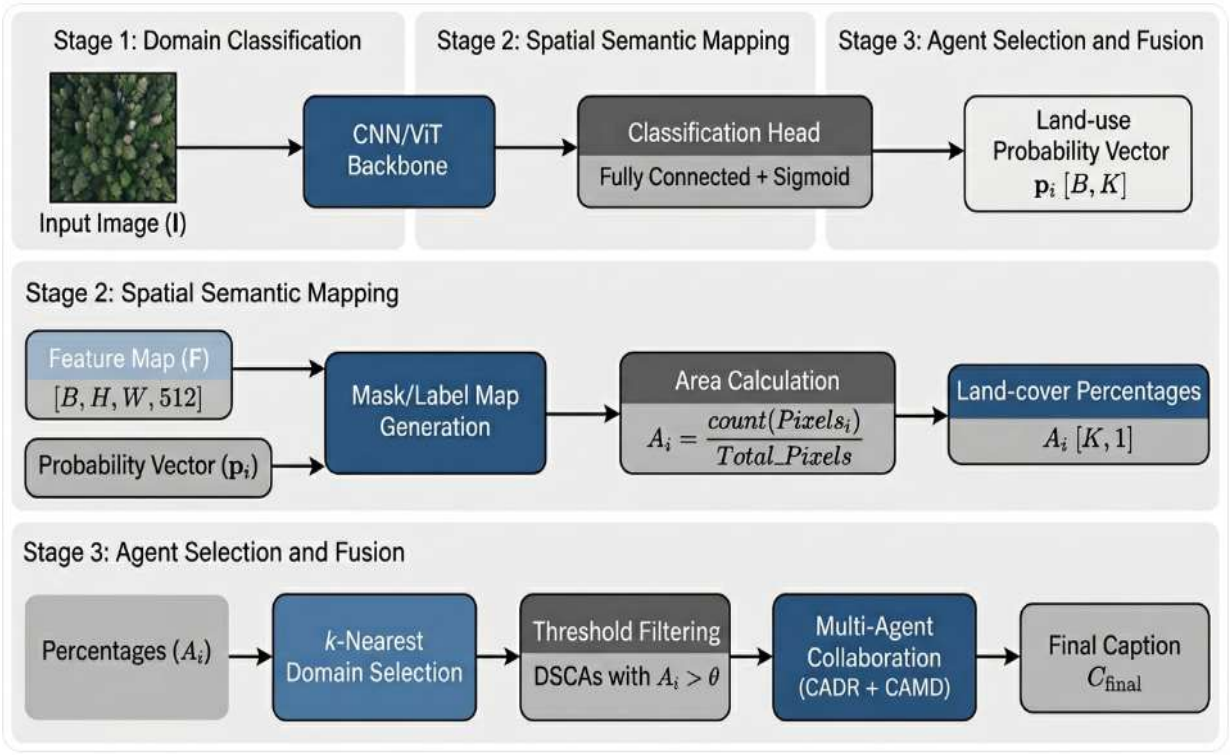


Fig. 3. Workflow diagram for the MAC-CapNet Domain Assignment and Inference process

Table 1. The critical comparison of existing captioning approaches and identified research gaps

Approach	Representative Works	Key Limitations	Research Gap
Attention-based Models	Bottom-Up & Top-Down [12]	Focus on salient regions but ignore global semantic structure; limited interpretability in complex scenes	Lack of explicit semantic decomposition for multi-object remote sensing scenes
RL-based Captioning	SCST [25]	Optimizes evaluation metrics but does not improve semantic understanding; unstable training	Absence of structured semantic reasoning despite improved metric scores
Remote Sensing Captioning	Lu et al. [11]	Domain-specific but relies on single-model architecture; limited diversity in generated captions	No mechanism for handling heterogeneous land-cover semantics
Vision-Language Pretraining	BLIP [13]	Strong generalization but computationally expensive; weak domain adaptation for remote sensing imagery	Lack of domain-aware specialization despite large-scale training
Transformer-based Models	GIT [14]	Captures long-range dependencies but acts as a black-box; limited interpretability	No explicit control over semantic focus or region-wise reasoning
Unified Multimodal Models	OFA [15]	General-purpose design sacrifices task-specific optimization; complex	Insufficient specialization for remote sensing

		training pipeline	captioning tasks
Ensemble Learning	Hansen & Salamon [16], Nguyen et al. [18]	Improves robustness but lacks semantic coordination between models	No intelligent routing or domain-specific collaboration among models
Multi-Agent Learning	Yang et al. [17]	Supports distributed decision-making but rarely applied to caption generation	Underexplored for structured caption generation and semantic fusion
Domain Adaptation	Ganin & Lempitsky [19], Gong et al. [21]	Reduces domain shift but does not enhance caption generation quality directly	Missing integration with captioning frameworks for semantic generation
CNN–GNN Hybrid Models	Liang et al. [24], Yan et al. [22]	Captures spatial relationships but focuses on classification/segmentation tasks	Lack of extension to language generation and descriptive modeling
General-Purpose Vision-Language Instruction Tuning	LLaVA [42]	Primarily designed for natural-image understanding and visual instruction following. Does not explicitly model remote sensing semantics or heterogeneous land-use regions. Lacks domain-specialized reasoning and collaborative expert interaction.	There is a need for domain-aware caption generation frameworks that explicitly capture heterogeneous land-use semantics and support collaborative reasoning among specialized captioning agents.
Vision-Language Alignment with Large Language Models	MiniGPT-4 [43]	Aligns visual features with a frozen LLM but lacks mechanisms for domain-specific feature specialization, semantic collaboration, and explicit modeling of multiple coexisting land-use categories.	Existing alignment-based frameworks require mechanisms that incorporate domain specialization and semantic collaboration for complex remote sensing scenes containing multiple land-use classes.
Remote Sensing Vision-Language Foundation Model	RS-LLaVA [44]	Adapts large vision-language models to remote sensing but employs a unified architecture without explicit domain-specialized expert collaboration. Computational cost remains relatively high, and semantic specialization across land-	Future remote sensing captioning frameworks should integrate domain-specialized collaborative agents with adaptive semantic routing and

		use categories is implicit.	consensus-aware fusion to improve interpretability, semantic completeness, and caption quality.
Unified Vision Foundation Model	Florence-2 [45]	Designed as a general vision foundation model supporting multiple vision tasks rather than specialized remote sensing image captioning. Does not explicitly model land-use relationships or collaborative caption generation.	Remote sensing image captioning requires architectures capable of exploiting domain knowledge and modeling semantic relationships among multiple land-use categories within a single scene.
Comprehensive Survey of Transformer-based and Multimodal Large Language Model (MLLM)-based Image Captioning	Next-generation Image Captioning: A Survey of Methodologies and Emerging Challenges from Transformers to Multimodal Large Language Models [46]	The survey comprehensively reviews transformer-based captioning models and Multimodal Large Language Models (MLLMs), highlighting challenges such as computational complexity, hallucination, grounding, evaluation limitations, long-tail object recognition, and insufficient semantic faithfulness. However, it primarily focuses on general image captioning and does not investigate remote sensing-specific caption generation, domain-specialized semantic reasoning, or collaborative multi-agent architectures for heterogeneous aerial scenes. (ScienceDirect)	There remains a need for remote sensing image captioning frameworks that explicitly model heterogeneous land-use semantics, support collaboration among domain-specialized agents, provide interpretable caption generation through adaptive trust-aware fusion, and efficiently balance caption quality with computational complexity. Such frameworks should also address scene heterogeneity and improve semantic consistency beyond conventional transformer and MLLM-based captioning approaches. (ScienceDirect)
Proposed MAC-CapNet	--	Addresses semantic specialization using domain-specific agents and consensus-based fusion	Bridges gap between semantic decomposition, interpretability, and caption

			generation

Table 2. The mapping of identified research gaps to MAC-CapNet contributions.

Identified Research Gap	Limitation in Existing Works	MAC-CapNet Contribution
Lack of domain-specific semantic understanding	Transformer-based models (e.g., [13], [14], [15]) learn generic representations without explicit domain awareness	Introduces domain-specific agents (e.g., urban, vegetation, water) for explicit semantic specialization
Absence of structured semantic decomposition	Attention-based methods [12] focus on regions but do not decompose scenes into meaningful semantic categories	Decomposes image understanding into multiple agents, each responsible for a specific semantic domain
Limited interpretability of caption generation	Deep models act as black-box systems with minimal transparency in decision-making	Provides interpretable outputs via agent-wise contributions and confidence (trust) scores
Inefficient handling of heterogeneous remote sensing scenes	Remote sensing captioning models [11] rely on single-model pipelines	Employs multi-agent collaboration to handle diverse land-cover types effectively
Lack of intelligent model collaboration	Ensemble methods [16], [18] combine outputs but lack semantic coordination	Uses consensus-based fusion with trust weighting for coordinated caption generation
Underutilization of multi-agent learning in captioning	Multi-agent frameworks [17] are not applied to structured caption generation tasks	Adapts multi-agent learning paradigm for caption generation with task-specific design
No integration of domain adaptation with captioning	Domain adaptation methods [19], [21] focus on feature alignment, not language generation	Implicitly incorporates domain-aware reasoning through specialized agents without separate adaptation modules
Lack of spatial-semantic relationship modeling in captions	CNN and GNN approaches [24], [22] focus on classification, not descriptive generation	Captures spatial-semantic relationships through coordinated agent reasoning and fusion
Optimization focused on metrics rather than semantics	RL-based methods [25] improve scores but not semantic richness	Enhances both semantic quality and interpretability alongside competitive metric performance
General-purpose vision-language model trained on natural images with limited remote sensing specialization.	Does not explicitly model heterogeneous land-use regions or domain-specific semantic interactions. Performance depends heavily on large-scale instruction tuning and computational resources.[42]	MAC-CapNet introduces domain-specialized captioning agents for urban, forest, agricultural, water, and barren regions, enabling fine-grained semantic understanding of heterogeneous remote sensing scenes while remaining considerably lighter than large foundation models.
Aligns visual features with a large language model for general multimodal reasoning but is not optimized for remote sensing image captioning.	Lacks mechanisms for domain-aware caption generation, agent collaboration, and explicit modeling of multiple coexisting land-use classes.[43]	MAC-CapNet performs collaborative multi-agent reasoning, allowing specialized agents to exchange semantic information before caption generation, leading to richer and more context-aware remote sensing descriptions.
Adapts a large vision-language model to remote	Uses a unified vision-language architecture without	MAC-CapNet introduces a collaborative multi-agent

sensing and supports captioning and VQA jointly. (MDPI)	explicit domain-specialized expert collaboration. Large model size increases computational requirements, and semantic specialization across land-use categories is not explicitly modeled.[44]	architecture where multiple domain-specialized captioning agents cooperate through CADR and CAMD. This design explicitly models complementary land-use semantics, improves interpretability through trust-weight estimation, and offers a modular architecture that can be extended with additional domain experts.
Strong visual representation learning across multiple vision tasks but not specifically designed for remote sensing caption generation.	General-purpose visual features may overlook domain-specific characteristics of aerial imagery and complex land-use relationships. [45]	MAC-CapNet explicitly incorporates remote sensing domain knowledge through dedicated agents and multi-label domain assignment, improving semantic representation of heterogeneous satellite scenes.
Existing transformer and MLLM-based captioning methods primarily target general image understanding and lack mechanisms specifically designed for heterogeneous remote sensing scenes.	Current approaches generally rely on unified caption generation pipelines without explicit domain specialization, collaborative semantic reasoning, or interpretable fusion strategies. They also suffer from high computational requirements, hallucination, limited semantic grounding, and inadequate handling of complex multi-land-use scenes.[46]	MAC-CapNet introduces a collaborative multi-agent framework consisting of Domain-Specialized Captioning Agents (DSCAs), a Cross-Agent Dynamic Routing (CADR) module for semantic interaction, and a Consensus-Aware Meta-Decoder (CAMD) for trust-aware caption fusion. Unlike general transformer or MLLM-based captioning models, the proposed framework explicitly models heterogeneous land-use categories, supports multi-label scene understanding, improves interpretability through agent collaboration and trust-weight estimation, and provides a scalable architecture tailored to remote sensing image captioning.

Formal Mathematical Formulation of MAC-CapNet

In this section, we describe theoretical foundation for the proposed MAC-CapNet framework, in terms of a unified mathematical form. The operation of architecture flow through a sequence of coordinated steps, which starts from domain-specific caption generation, followed by inter-agent semantic interaction, and finally a consensus-driven fusion mechanism which generates the final description.

The MAC-CapNet accepts remote sensing image I as an input, this input image use by visual encoder to extracts rich visual representations that capture both spatial and semantic information. This transformation is expressed as:

$$F = \text{Encoder}(I), \qquad F \in R^{H \times W \times d} \tag{2}$$

The extracted feature map F serves as a common visual foundation for all domain-specialized agents. Each agent is designed to focus on a particular land-use category and is implemented using a Transformer-based decoder. Conditioned on the shared visual features, each agent i generates a candidate caption ci in an autoregressive manner. The probability of generating the caption is defined as:

$$P(c_i \mid I) = \prod_{t=1}^{T} P\left(w_t \mid w_{\{<t\}}, F, \theta_i\right) \quad (3)$$

where w_t represents the token generated at time step t, and θ_i denotes the learnable parameters specific to the $i-th$ domain-specialized agent. This formulation allows each agent to model linguistic dependencies while grounding its predictions in domain-relevant visual patterns. In addition to generating textual descriptions, each agent produces a semantic embedding that captures the overall meaning of its caption, along with a confidence score that reflects the reliability of its prediction. These are defined as:

$$h_i = \text{Embed}(c_i), \qquad \gamma_i \in R^+ \quad (4)$$

While individual agents provide domain-specific insights, complex remote sensing scenes often require integrating knowledge across multiple domains. To facilitate this, MAC-CapNet introduces a Cross-Agent Dynamic Routing (CADR) mechanism that enables semantic interaction among agents. Given the set of embeddings {hi}N i=1, pairwise relationships between agents are computed using similarity-based attention:

$$\alpha_{i,j} = \frac{\exp\left(\text{sim}(h_i, h_j)\right)}{\sum_{\{k=1\}}^{\{N\}} \exp\left(\text{sim}(h_i, h_k)\right)} \quad (5)$$

where sim(·, ·) denotes cosine similarity. These routing weights determine how much information each agent receives from others. Using these weights, each agent refines its representation by aggregating information from all agents:

$$\tilde{h}_i = \sum_{j=1}^{N} a_{ij}\, h_j \quad (6)$$

This routing process allows agents to incorporate complementary semantic cues, reduce redundancy, and resolve potential inconsistencies in their individual predictions. Following inter-agent interaction, the refined representations are passed to the Consensus-Aware Meta-Decoder (CAMD), which performs the final aggregation. The fusion process begins by computing a global consensus representation across all agents:

$$\tilde{h} = \frac{1}{N} \sum_{j=1}^{N} \tilde{h}_i \quad (7)$$

This consensus vector acts as a reference point to evaluate the contextual relevance of each agent. Accordingly, contextual agreement weights are computed as:

$$\beta_i = \frac{\exp(\text{sim}(\tilde{h}_i, \bar{h})}{\sum_{k=1}^{N} \exp(\text{sim}(\tilde{h}_k, \bar{h})} \quad (8)$$

These weights quantify how well each agent aligns with the overall semantic consensus. To further incorporate reliability, the model combines these agreement weights with the learned confidence scores γi. The final fused representation is thus obtained as:

$$C_{fusion} = \sum_{i=1}^{N} \beta_i \cdot \gamma_i \cdot \tilde{h}_i \quad (9)$$

This formulation ensures that agents contributing more relevant and reliable information are given higher importance during fusion. Finally, the fused representation is decoded to generate the final caption:

$$C_{final} = \text{Decoder}(C_{fusion}) \quad (10)$$

Through this structured process, MAC-CapNet effectively integrates domain specialization, inter-agent collaboration, and trust-aware fusion. The final caption is therefore not derived from a single model but emerges as a result of coordinated reasoning across multiple specialized agents, guided by both semantic agreement and learned reliability.

Domain Specialization Strategy

To implement a Domain Specialization Strategy in MAC-CapNet, we design multiple domain-specialized agents, each focusing on a specific category such as urban, agriculture, water, barren land, or forest.

In this strategy, we do not depend on manually defined labels; instead, we derived domain information from the captions of datasets. Within the training process, semantic cues and keywords are extracted from the ground truth descriptions, and these cues and keywords are used to relate each image with one or more relevant domains. With this kind of training, the model learns domain assignment in weakly supervised manner.

Most remote sensing images consist of more than one land-use type. Due to this, we decided to avoid strict boundaries between domains. So, an image is processed by multiple agents instead of assigning it to a single category. Using this kind of soft assignment, model effectively capture complex, multi-domain scenes.

In MAC-CapNet architecture, the balance between efficiency and specialization is achieved with the combination of shared a common visual encoder (shared by all agents) and use of own decoder by each agent (the separate decoder enable agent to learn domain-specific language patterns and focus on relevant semantic details).

As our framework is modular, we can easily extend it. So, new agents can be included for additional domains without changing the existing structure. All these agents can communicate with each other through the cross-agent routing mechanism, which helps the model to understand the unseen or mixed domain scenes.

Overall, this design allows MAC-CapNet to remain flexible and interpretable, while effectively handling the diversity and complexity of real-world remote sensing images.

As shown in Figure 4, domain assignment is performed using weak supervision derived from caption semantics. Keyword-level cues are extracted from ground truth captions and mapped to predefined land-use categories, enabling soft assignment of images to multiple domain-specific agents. This design allows flexible multi-domain learning without requiring explicit manual labelling.

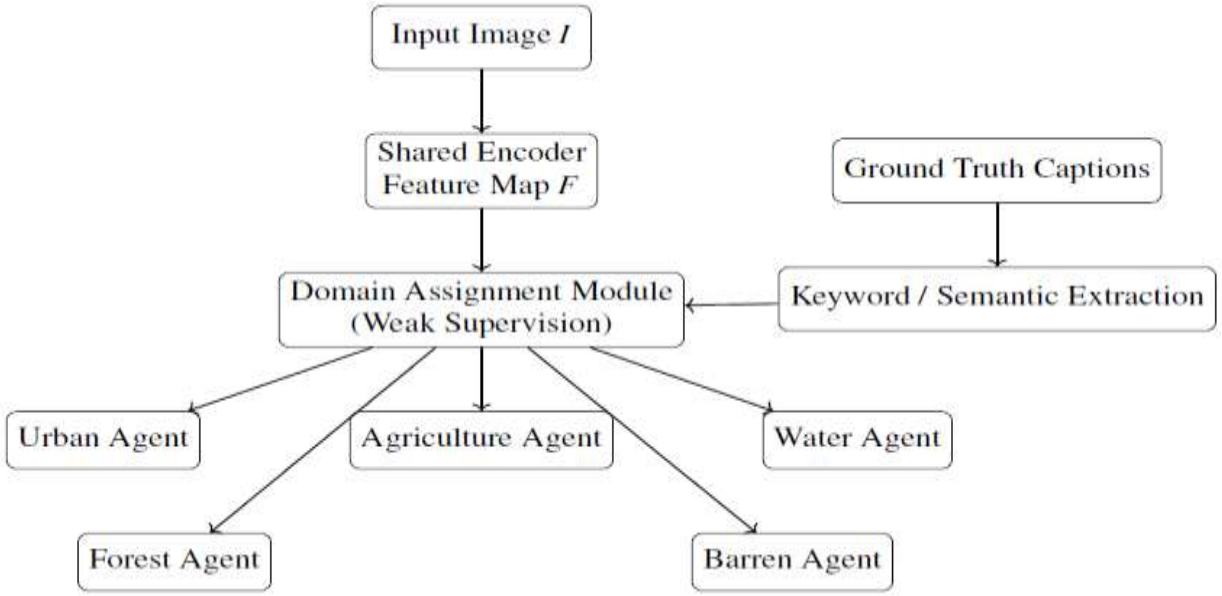


Fig. 4. Domain assignment strategy in MAC-CapNet. The input image is processed by a shared encoder, while semantic cues extracted from ground truth captions are used to assign the image to one or more domain-specialized agents using weak supervision.

Visual Feature Extraction

In order to extract visual feature, we use vision encoder, which capture both localized features and global context. And this vision encoder was initialized from a large-scale vision-language pretraining. After processing the input image, this encoder creates an extensive feature map:

$$F \in R^{H \times W \times d} \quad (11)$$

Here, d denotes the feature embedding dimension produced by the visual encoder. In our implementation, $d = 512$, which matches the hidden dimension used in the DSCA Transformer decoders to ensure consistent feature projection across agents.

This feature map becomes foundation for visual representation which is shared across the entire multi-agent suite. Here we clarify the scope of using pretraining: these high-level initial weights are beneficial to the visual encoder, but it is the only component which is taking the advantage of pretraining. The DSCA, CADR, and CAMD modules are built and trained from scratch for all the complexity of remote sensing captioning. We assure domain-specific reasoning and collaborative logic, use in MAC-CapNet, are trained right from geographical data by separating the pretraining to the encoder. During the initial stage of training, the visual encoder is initialized using a pretrained vision–language model. The purpose of using a pretrained encoder is to take advantage of the strong visual representations learned from large-scale datasets. However, during Stage I, the

parameters of this encoder are kept frozen. In this stage, only the domain-specific Transformer decoders of the Domain-Specialized Captioning Agents (DSCAs) are trained. By keeping the encoder fixed, the model preserves the general visual knowledge obtained during pretraining while allowing each agent to learn domain-specific caption generation patterns. In the final stage of training, a limited form of end-to-end fine-tuning is

performed. During this stage, the top layers of the visual encoder are partially unfrozen so that the model can slightly adapt its visual representations to the characteristics of remote sensing imagery. At the same time, most of the pretrained parameters remain unchanged to maintain stable and robust feature extraction. This gradual training strategy allows MAC-CapNet to benefit from transfer learning while still adapting effectively to the remote sensing captioning task.

Domain-Specialized Captioning Agents (DSCA)

Most remote sensing images are complex because they consist of various land-use categories, including urban infrastructure, agriculture, water bodies, forests, and barren land, and these complex scenes are a major challenge for monolithic models. If a single decoder is used to capture these different land-use categories, then it may result in "semantic dilution," where the model is not able to capture the minute details of any particular category. MAC-CapNet reduces this by using a group of Domain-Specialized Captioning Agents (DSCA), each agent acting as an expert for a particular land-use domain. These agents make use of a common visual backbone—such as ResNet-101 [26] or a Swin Transformer [27]. However, these agents use separate domain-specific decoders to transform these features into text [28].

The configuration for each agent includes:

A 6-layer Transformer architecture to capture deep linguistic dependencies.

A 512-unit hidden dimension for robust feature representation.

Multi-head attention mechanisms (8 heads) for both self- and cross-attention.

Standard position-wise feed-forward networks.

Domain Assignment Strategy

The training process required domain labels-, we extracted the domain labels from the land-use annotations given in the dataset. Each input image in the training process is forwarded to the agent (or agents) corresponding to its related land-use category, ensuring that every expert learns from relevant samples. At the inference stage, the system switches to a parallel processing mode: all agents analyze the image simultaneously to generate a set of complementary candidate captions. For each $i-th$ agent, the output consists of a candidate caption c_i and its corresponding sentence-level embedding, represented as:

$$e_i = \text{Embed}(c_i)$$

Model Capacity and Parameterization

A Transformer-based decoder architecture is used for designing each agent in MAC-CapNet method. Specifically, within our implementation, we have used a 6-layer Transformer decoder in each agent with a hidden dimension of 512 and 8 attention heads. This configuration helps the model to generate caption which reflects linguistic dependencies and visual–semantic relationships. Within this proposed model each agent approximately contains 38 million trainable parameters, when we combinely consider the embedding layers, multi-head attention modules, and position-wise feed-forward networks. Initially, when we started development, we used deeper decoders containing 8 to 12 Transformer layers. These models help to slightly improve captioning scores but with the cost of significant increase in training time and memory usage. As a practical decision, We opted for the 6-layer architecture, which offers adequate representational capacity while ensuring feasible computing demands for training the multi-agent framework.

Cross-Agent Dynamic Routing (CADR)

Every specialized agent in DSCA step generates an independent caption which may consist of redundancy or semantic inconsistency. To minimize/remove this redundancy or semantic inconsistency, there is need to exchange contextual information between these agents, at sentence level, for this purpose we introduced Cross-Agent Dynamic Routing (CADR) [29].

This routing mechanism allows agents to refine their semantic representations by incorporating information from other agents. As a result, redundant or contradictory captions are reduced before the final fusion stage. Given sentence embeddings from all agents, CADR computes attention-based routing weights:

$$\alpha_{i,j} = \frac{\exp\left(\text{sim}\left(\text{Embed}(c_i), \text{Embed}(c_j)\right)\right)}{\sum_{\{k=1\}}^{\{N\}} \exp\left(\text{sim}\left(\text{Embed}(c_i), \text{Embed}(c_k)\right)\right)} \quad (12)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

The routed representation for agent i is computed as:

$$\tilde{c}_i = \sum_{j=1}^{N} \alpha_{ij} \, \text{Embed}\left(c_j\right) \quad (13)$$

A single routing step is employed, as empirical analysis showed that iterative routing provides marginal

performance gains while increasing computational complexity. Compared to multi-head attention or graph attention networks, CADR operates at the sentence level with computational complexity O(N · d), where d denotes the embedding dimension.

Consensus-Aware Meta-Decoder (CAMD)

The Consensus-Aware Meta-Decoder (CAMD) responsible to generate the final caption. For this purpose it combines the output of all agents by considering both contextual agreement and agent reliability.

For each agent i, two weighting factors are defined:

Contextual Agreement Weight

$$\beta_i = \frac{\exp(\text{sim}(\tilde{c}_i, \bar{c})}{\sum_{k=1}^{N} \exp(\text{sim}(\tilde{c}_k, \bar{c})} \quad (14)$$

Where $\bar{c}$ denotes the mean routed embedding across all agents.

Agent Trust Weight: The trust weight of the i-th agent is computed using a learnable confidence estimation module:

$$\gamma_i = \sigma(W_t\, h_i + b_t) \quad (15)$$

where (h_i) denotes the routed semantic embedding produced by the agent, (W_t) and (b_t) are trainable parameters, and $(\sigma(.))$ represents the sigmoid activation function. The resulting trust score (γ_i in [0,1]) provides a normalized estimate of the agent's contribution to the final caption generation process.

During training, trust weights are optimized jointly with the overall captioning objective through end-to-end backpropagation. Rather than relying on manually defined reliability measures, the confidence estimation module learns to assign higher trust values to agent representations that consistently contribute useful semantic information for minimizing the caption generation loss. Conversely, representations with lower utility receive reduced trust scores.

This adaptive weighting mechanism enables the Consensus-Aware Meta-Decoder (CAMD) to dynamically emphasize informative agent representations while suppressing less relevant or noisy contributions. As a result, the fusion process becomes more robust to variations in scene complexity and domain-specific semantic ambiguity, leading to improved caption quality and contextual consistency

The routing of captions across agents is governed by agent-level confidence scores denoted as γ_i. These weights are computed at the sentence level and determine the contribution of each Domain-Specialized Captioning Agent.

In contrast, α_i represents the attention weights within the Transformer decoder, capturing token-level dependencies during caption generation.

The fused representation is computed as:

$$C_{fusion} = \sum_{i=1}^{N} \beta_i \,.\, \gamma_i \,.\, \tilde{h}_i \quad (16)$$

The CAMD then decodes the final caption C_{final} from C_{fusion} using a Transformer-based decoding process.

Training Objective and Optimization

The total loss combines caption accuracy, confidence supervision, agreement alignment, and diversity:

$$L_{total} = L_{cap} + \lambda_1 L_{conf} + \lambda_2 L_{agree} + \lambda_3 L_{KL} \quad (17)$$

where,

L_{cap}: Cross-entropy loss for final caption correctness [5, 6].

L_{conf}: Penalizes misaligned confidence scores [18].

L_{agree}: Encourages alignment among routed embeddings, stabilizing CADR [29].

L_{KL}: Promotes diversity among agent outputs to avoid mode collapse [30].

The loss weights are empirically selected based on validation performance. In our experiments, the hyperparameters are set as: $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$.

These values balance caption accuracy with confidence calibration, agent agreement, and output diversity L_{agree} is implemented as cosine similarity maximization between routed embeddings, and L_{KL} is the KL divergence between token-level probability distributions of agents.

Computational Complexity Analysis

The computational complexity of MAC-CapNet can be decomposed into three primary stages: visual feature extraction, domain-specialized caption generation, and cross-agent semantic collaboration.

Let (N) denote the number of Domain-Specialized Captioning Agents (DSCAs), (L) denote the caption sequence length, and (d) denote the hidden embedding dimension.

The visual encoder is shared among all agents. Therefore, the feature extraction cost is incurred only once and can be represented as:

$$O(E_v)$$

where (E_v) denotes the computational complexity of the visual backbone network.

Each DSCA employs an independent Transformer-based decoder for caption generation. Assuming a standard self-attention mechanism, the complexity of a single decoder is approximately:

$$O(L^2 d)$$

Consequently, the cumulative decoding complexity across all agents becomes:

$$O(NL^2 d)$$

The proposed Cross-Agent Dynamic Routing (CADR) mechanism performs semantic interaction among all participating agents. Since each agent exchanges information with every other agent, the routing operation requires approximately (N(N-1)) pairwise interactions. The resulting complexity can therefore be expressed as:

$$O(N^2 d)$$

The Consensus-Aware Meta-Decoder (CAMD) performs trust-aware fusion of the agent representations. Because the fusion operation aggregates information across all agents, its complexity scales linearly with the number of agents:

$$O(N_d)$$

Combining all components, the overall computational complexity of MAC-CapNet can be approximated as:

$$O(E_v + NL^2 d + N^2 d + Nd)$$

Since (N_d) is dominated by ($N^2 d$), the complexity can be simplified to:

$$O(E_v + NL^2 d + N^2 d)$$

Although the computational cost increases with the number of agents, the shared visual encoder eliminates redundant feature extraction, significantly reducing overhead compared with architectures that maintain separate visual backbones. Moreover, the DSCAs operate independently after feature extraction and can therefore be executed in parallel, making MAC-CapNet suitable for scalable deployment on modern GPU-based computing platforms.

Implementation Summary

The key implementation settings used in the proposed framework are summarized below:

Number of expert agents: 5 (Urban, Agriculture, Forest, Water, and Barren)

Vocabulary: Shared across all agents

Word embeddings: Shared across all agents

CADR: Applied at each decoding step

CAMD: Applied during sentence-level fusion only

IV. Design Rationale

The proposed framework comprises domain-specific agents, a routing mechanism, a meta-decoder, and a composite loss function. Their respective roles are listed below:

DSCA: Assigns a dedicated agent to each land-use category.

CADR: Facilitates information sharing among agents.

CAMD: Combines agent outputs to generate the final caption.

Loss Function: Incorporates terms for caption accuracy, confidence, agreement, and diversity.

The modular design also supports component-wise ablation analysis.

Distinction from Conventional Mixture-of-Experts Architectures

Although MAC-CapNet uses multiple specialized agents, its design is different from conventional Mixture-of-Experts (MoE) and ensemble learning methods.

In a typical MoE framework, a gating network decides which expert, or combination of experts, should contribute to the final prediction by assigning routing weights. The experts usually process the input independently, with little or no direct interaction during inference.

MAC-CapNet follows a different strategy. All Domain-Specialized Captioning Agents (DSCAs) contribute to the caption generation process instead of relying on a single selected expert. Each agent learns semantic representations related to its assigned domain, and these representations are then shared through the proposed Cross-Agent Dynamic Routing (CADR) mechanism. This interaction allows information learned by one agent to complement the knowledge of the others before the final caption is generated.

Furthermore, the Consensus-Aware Meta-Decoder (CAMD) performs trust-aware fusion by jointly considering contextual agreement and agent reliability. Unlike conventional MoE gating mechanisms that primarily rely on routing probabilities, CAMD incorporates semantic consensus among agents before

generating the final caption representation.

Therefore, MAC-CapNet is more appropriately characterized as a collaborative multi-agent captioning framework rather than a conventional MoE architecture. Its design integrates domain specialization, inter-agent semantic interaction, dynamic routing, and consensus-aware fusion within a unified caption generation process, making it particularly suitable for modelling the heterogeneous semantic content commonly observed in remote sensing imagery.

V. Experimental Setup

Dataset

We evaluate MAC-CapNet on following benchmark datasets:

RSICD: RSICD is selected due to its diversity in land-use categories and multi-caption annotations, which are essential for evaluating semantic richness. It consists of more thanc10,000 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 5, 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, the researchers have a resource to aid them in the remote sensing captioning endeavour thanks to this dataset. To ensure fair evaluation, we divided dataset in 3 sets as training, validation, and test sets. [8, 11].

UCM-Captions: UCM-Captions provides high-resolution aerial imagery with well-defined scene categories, enabling cross-dataset generalization analysis. 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 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 [31]. The example of this dataset is shown in Figure 6.

Domain-Specific RS Subset (DSRS): However, both datasets lack explicit domain annotations required for specialized learning. To address this limitation, we construct the Domain-Specific Remote Sensing Subset (DSRS), which enables controlled training of domain-specialized agents. The DSRS is constructed by clustering images into five land-use domains (urban, forest, agricultural, water, barren land) using semantic segmentation maps or metadata. DSRS is specifically used for training and evaluating he Domain-Specialized Captioning Agents (DSCAs).

The following Table 3, gives summary of datasets used for training and evaluation with details, as given below-

Table 3. Summary of datasets used for training and evaluation of MAC-CapNet.

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

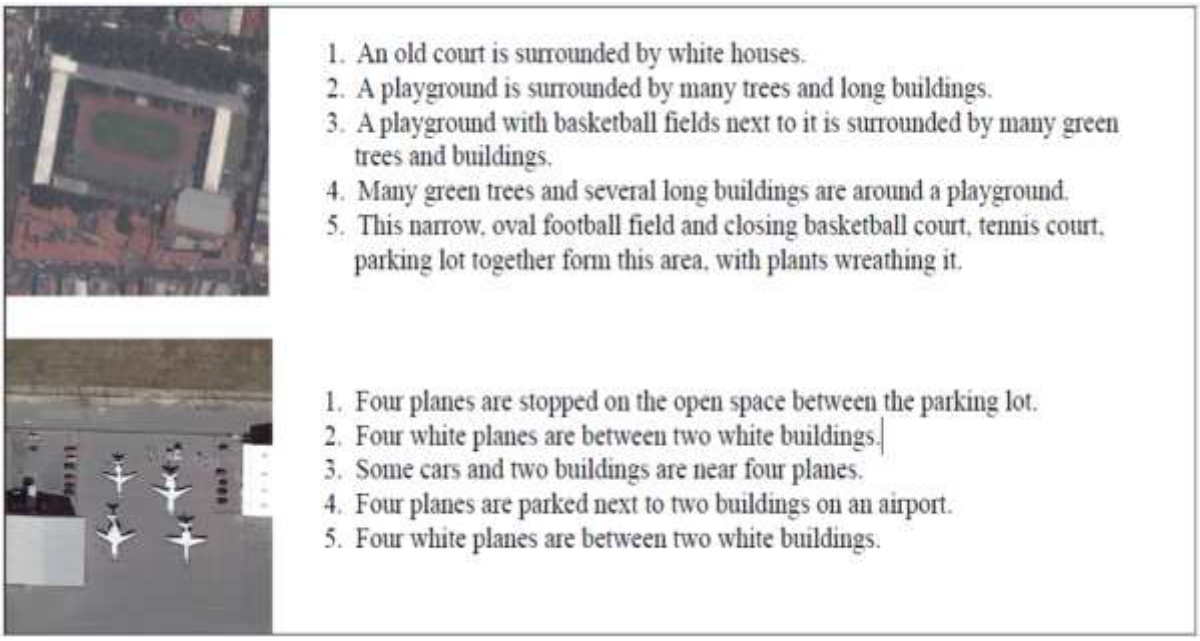


Fig. 5. Two examples in RSICD dataset

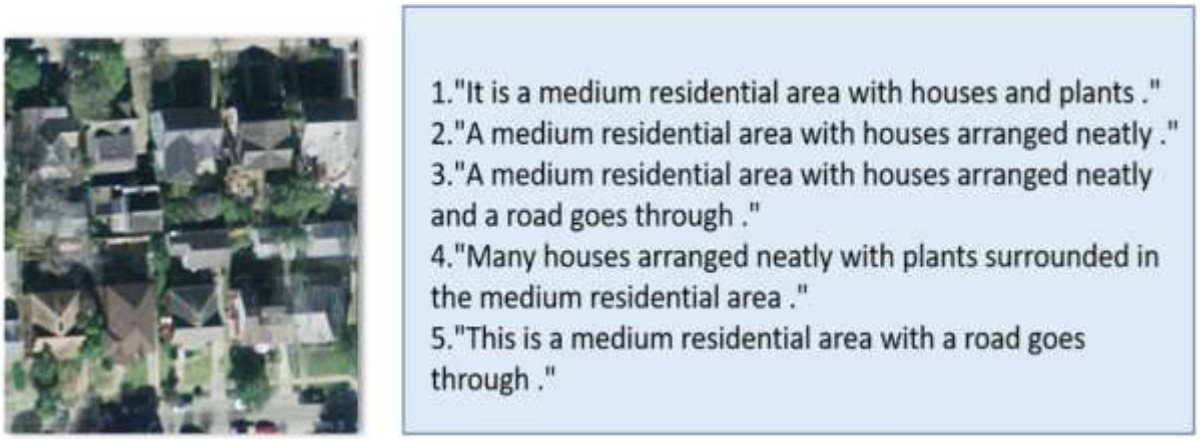


Fig. 6: Remote sensing sample and corresponding five captions extracted from the UCM-Captions dataset

VI. Domain Label Construction

Remote sensing images often contain more than one type of land-use within a single scene. For instance, an aerial image may show urban buildings, agricultural fields, rivers or lakes, and forest areas at the same time. Because of this diversity, assigning only one domain label to an image may not correctly represent the actual scene content. To address this challenge, we design a domain label construction strategy that supports the domain-specialized learning mechanism used in MAC-CapNet. The goal is to organize the training data in a way that allows different agents of the model to focus on specific land-use patterns. The proposed strategy is carried out in three steps. First, each image is assigned domain labels using a multi-label strategy. Next, a Domain-Specific Remote Sensing Subset (DSRS) is constructed to organize the training data according to domain categories. Finally, the dataset partitions are carefully maintained to ensure fair evaluation and to avoid any data leakage

between training and testing stages.

Multi-Label Domain Assignment

In real-world remote sensing scenes, multiple land-use categories often appear together. For example, an image might contain both agricultural fields and water bodies, or urban structures near forest regions. Because of this, assigning only a single domain label may ignore important information present in the scene. Therefore, instead of using a single-label approach, we adopt a multi-label domain assignment strategy, as shown in Figure 7. Using the available annotations, metadata, or scene descriptions from the benchmark datasets, each image is assigned to all relevant land-use categories present in that scene. The main domain categories considered in this work are:Urban, Agriculture, Forest, Water, and Barren land. If an image contains more than one domain,

it is associated with each of those domains. During the training process, such images contribute to the learning of multiple Domain-Specialized Captioning Agents (DSCAs). In this way, each agent can learn visual and semantic patterns related to its specific domain, even when the scene contains mixed land-use regions. This multi-label strategy helps preserve important semantic information that could otherwise be lost if images were forced into only one category.

Construction of the Domain-Specific RS Subset (DSRS)

After assigning domain labels, the next step is to organize the training data for domain-specialized learning. For this purpose, we construct a Domain-Specific Remote Sensing Subset (DSRS), as shown in Figure 8. It is important to note that DSRS is not a new dataset. Instead, it is a structured subset derived from the training portions of the benchmark datasets, namely RSICD and UCM-Captions. The goal of DSRS is to group images according to their domain categories so that each captioning agent can focus on learning domain-specific caption patterns. The construction of DSRS follows a simple process. First, all images from the training sets are examined based on their domain labels. Next, the images are grouped according to the land-use categories they belong to. Since a multi-label strategy is used, some images may appear in more than one domain group. Based on this grouping, five domain-specific subsets are created: Urban subset, Agriculture subset, Forest subset, Water subset and Barren land subset. These subsets are then used during the first stage of model training. In this stage, each Domain-Specialized Captioning Agent (DSCA) is trained using the images from its corresponding subset. As a result, each agent learns caption generation patterns that are more suitable for its specific land-use domain. To ensure that the domain labels used in the Domain-Specific Remote Sensing Subset (DSRS) are reliable, an additional verification step is incorporated during the construction process. Initially, domain labels are assigned to each image using the available metadata, scene descriptions, and annotations provided with the benchmark datasets. Because remote sensing scenes often contain multiple land-use categories, an image may be associated with more than one domain. In such cases, the multi-label assignment strategy described earlier is applied, allowing the same image to appear in multiple domain-specific subsets. After this first assignment, a verification is carried out, and this verification helps us to maintain the consistency and dependability of the domain labels. In this step, we analyse the easily noticeable visual content with domain categories assigned to it. The verification is done using caption keywords, scene descriptions, and annotation data from the dataset. When we obtain significant correlation between a certain land-use category and the semantic clues from the captions and comments, we kept that domain assignment.

Dataset Split Integrity

While constructing DSRS, it is important to maintain the integrity of the original dataset partitions. This ensures that the evaluation results remain fair and comparable with existing research.

To achieve this, the DSRS subsets are created only from the training portions of the RSICD and UCM-Captions datasets. The validation and test splits are not used during the DSRS construction process. Specifically, the datasets are used in the following way:

Training split: Used to construct DSRS and train the domain-specialized captioning agents.

Validation split: Used for hyperparameter tuning and early stopping during training.

Test split: Reserved strictly for final evaluation and never used during training.

By following this protocol, the model is always evaluated on images that were not seen during training. This prevents data leakage and ensures that the reported results reflect the true generalization ability of the proposed MAC-CapNet framework. Maintaining these original benchmark splits also allows the experimental results to remain directly comparable with prior remote sensing captioning studies.

VII. Evaluation Metrics

For performance evaluation of MAC-CapNet, we consider two types of metrics-Standard Captioning Metrics and Domain-Sensitive Metrics. Most of the remote sensing images are multi-domain in nature, so that standard metrics did not seem sufficient.

Standard Captioning Metrics: We consider BLEU-1 and BLEU-4 [32], METEOR [33], ROUGE-L [34], CIDEr [35], and SPICE [36], as these are widely reported in prior work. BLEU focuses on n-gram overlap and gives a reasonable estimate of fluency, although in practice it tends to reward surface-level similarity rather than true semantic correctness. This becomes noticeable in remote sensing images where multiple valid descriptions can exist for the same scene.

METEOR partially addresses this by allowing synonym and stem matching, which makes it more tolerant to variations in wording.

ROUGE-L measures the overlap between the generated and reference sentence structures. It is useful for assessing whether the generated caption follows a sequence that is similar to the reference caption.

CIDEr is especially suitable for remote sensing image captioning because it compares a generated caption with multiple human-written references [35]. Since a remote sensing image can often be described correctly in different ways, evaluating agreement across several reference captions provides a more reliable assessment

than relying on a single description.

SPICE evaluates captions by comparing their semantic content through scene graph representations rather than simple word matching [36]. This makes it useful for examining whether the relationships among objects are preserved in the generated caption. During our experiments, however, we found that strong scores on these standard metrics did not always guarantee correct interpretation of scene categories. For example, a caption could confuse agricultural land with open fields while still achieving a relatively high evaluation score.

Domain-Sensitive Metrics: To address this limitation, we introduced two additional measures that examine how well the proposed framework captures domain-specific information.

(1) Domain-Specific CIDEr (DS-CIDEr): Instead of reporting a single CIDEr score for the complete dataset, we calculate the score separately for each major land-use category, including urban, agricultural, forest, water, and barren-land scenes. This makes it easier to identify the strengths and weaknesses of the model in different environments and helps verify whether the domain-specialized agents learn distinctive semantic characteristics for their assigned domains.

(2) Agent Entropy: We also analyze the entropy of the fusion weights ((\alpha_i)) to understand how the Domain-Specialized Captioning Agents contribute during caption generation. Lower entropy indicates that the prediction is influenced mainly by one or a small number of agents, reflecting stronger specialization. Higher entropy indicates that the contribution is distributed across several agents. Our observations suggest that the best performance is obtained with a balanced level of entropy, where agents remain specialized while still contributing complementary information during the fusion process.

Evaluation Strategy: Caption quality cannot be fully described by a single evaluation metric, particularly for remote sensing images that often contain multiple valid semantic interpretations. For this reason, we evaluate MAC-CapNet from several perspectives. Standard captioning metrics are used to measure similarity with the reference captions, whereas the domain-sensitive metrics assess specialization and the behaviour of the collaborative multi-agent architecture. We also include a qualitative analysis (Section 5.5) to examine examples where numerical scores alone do not provide a complete picture. These examples are reviewed for domain coverage, contextual accuracy, and readability. Considering all of these evaluations together provides a more complete assessment of the proposed framework.

Baselines

MAC-CapNet is compared against classical and modern captioning models, summarized in Table 4.

Table 4. Benchmark models compared against MAC-CapNet.

Model	Description
Show-Tell (CNN–RNN) [5, 6]	Basic encoder–decoder baseline
Transformer-Cap [25]	Global transformer-based captioner
BLIP / BLIP-2 [13]	Bootstrapped vision–language pretraining, fine-tuned for captioning
GIT / GIT-2 [14]	Unified image–text transformer trained on web-scale corpora
VLM / mPLUG [37, 38]	Multimodal transformers with global-local fusion
MAC-CapNet (Ours)	Multi-agent domain-aware captioning with meta-decoder fusion

Training Strategy

To ensure that each component of MAC-CapNet learns its role effectively, we adopt a multi-stage training strategy consisting of independent agent training, routing learning, meta-decoder optimization, and final end-to-end refinement.

Stage I: Independent Training of DSCAs

Each Domain-Specialized Captioning Agent (DSCA) is trained independently on its respective land-use subset (urban, agriculture, water, forest, barren). All agents share a common visual backbone (e.g., ResNet-101 or Swin-T), while each maintains its own domain-specific language decoder.

DSCA training uses a standard cross-entropy captioning loss with scheduled sampling.

Early stopping is performed using CIDEr on the validation split of each domain.

Each trained DSCA produces caption proposals, semantic vectors, and a confidence estimation head.

Stage II: Training the CADR Module

With DSCAs frozen (or partially unfrozen in ablation studies), the Cross-Agent Dynamic Routing (CADR) module is trained.

CADR learns inter-agent agreement via attention coefficients and routed semantic vectors.

Only routing parameters, including the projection matrix W , are updated.

This stage enables CADR to model cross-agent consensus prior to meta-decoder training.

Stage III: Joint Training of the CAMD Module

The Consensus-Aware Meta-Decoder (CAMD), initialized from BLIP-2, is jointly trained with routed DSCA outputs.

DSCAs may remain frozen, partially trainable (decoder top layers), or fully trainable.

CAMD receives DSCA embeddings, confidence scores, routed vectors, and agreement weights.

The total loss combines caption loss, confidence supervision, agreement alignment, and KL-regularization.

Stage IV: End-to-End Fine-Tuning

Finally, the entire pipeline (DSCAs + CADR + CAMD) is fine-tuned end-to-end using a small learning rate to achieve:

globally optimized routing,

improved visual–semantic alignment,

calibrated confidence weights across agents.

Training Procedure

A training procedure consists of the following steps:

Step I: Forward Pass Through All DSCAs

Given an input image I, each $DSCA_i$ produces:

a caption proposal c_i,

a semantic vector h_i,

a confidence score γ_i representing sentence-level reliability of the caption generated by agent i

The confidence score γ_i is computed from the final hidden state of the agent decoder and reflects the overall reliability of the generated caption rather than token-level probabilities.

It is important to distinguish that γ_i denotes agent-level confidence weights used for routing, whereas αi refers to token-level attention weights within the Transformer decoder.

Step II: Cross-Agent Dynamic Routing (CADR)

CADR computes inter-agent attention weights ai j and routed vectors r_i , enriching each agent's representation with multi-agent consensus.

Step III: Meta-Decoder Fusion (CAMD)

CAMD generates:

contextual agreement attention β_i,

agent trust weights γ_i,

the fused semantic representation for the final caption decoder.

The final caption is then generated using the BLIP-2 based Transformer decoder.

Step IV: Loss Computation

The overall loss is:

$$L_{total} = L_{cap} + \lambda_1 L_{conf} + \lambda_2 L_{agree} + \lambda_3 L_{KL} \quad (18)$$

Step V: Parameter Update

Parameters for CAMD, CADR, and optionally DSCAs are updated using AdamW.

Step VI: Scheduler Step

A cosine-annealing scheduler with warm-up is applied to adjust the learning rate dynamically.

Hyperparameters, Hardware Setup, and Training Time

Table 5 summarizes the training configuration, model hyperparameters, and hardware setup used for all experiments.

Table 5. Training configuration and hyperparameter settings for MAC-CapNet.

Parameter	Setting
Optimizer	AdamW [39]
Learning Rate	5×10^{-5} (initial), 1×10^{-5} (fine-tuning)
Scheduler	Cosine decay with 5% warm-up [40]
Batch Size	64
Maximum Caption Length	30 tokens
Dropout	0.1
Weight Decay	10^{-4}
Framework	PyTorch [41]
Hardware	NVIDIA A100 GPUs (40 GB)
Training Time	8–12 hrs per DSCA; 20–24 hrs full model

Ablation Studies

Table 6 presents seven ablations evaluating module contributions. Key findings show that removal of CADR, CAMD, or agent diversity consistently reduces performance, validating the design choices [10, 29].

Table 6. Ablation study configurations.

Ablation ID	Description
A1: No Meta-Decoder	Use majority voting over agent captions (decision fusion only).
A2: Uniform Fusion	All confidence scores (c_i = 1) (removes adaptivity).
A3: Single-Agent Only	Use best-performing individual DSCA only.
A4: Homogeneous Agents	All agents share the same CNN backbone (no diversity).
A5: No Confidence Embedding	Remove confidence input to meta-decoder.
A6: Hard Attention	Use only the top-1 confident agent (no fusion).

Results and Discussion

With in this section, we present a comprehensive evaluation of the proposed MAC-CapNet framework. We first compare the proposed method with state-of-the-art remote sensing image captioning approaches using standard evaluation metrics. In next subsections, we examine the impact of using multi-agent approach by comparing it with standard vision–language pretraining methods. Lastly, we provide a closer look into the performance of MAC-CapNet through domain-specific evaluations, ablation studies, and qualitative comparisons.

VIII. Quantitative Comparison with State-of-the-Art Methods

We perform quantitative comparison of MAC-CapNet with State-of-the-Art methods using the RSICD dataset. Table 7 presents a comparison with CNN–RNN, attention-based, transformer-based, and vision–language pre-trained methods using BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE metrics.

The CNN–RNN and attention-based models report lower scores across most metrics. Transformer-based models show improved results, particularly on CIDEr and SPICE. Despite these improvements, noticeable differences remain between the generated captions and the reference descriptions, as reflected by the evaluation scores.

The effectiveness of large-scale pretraining demonstrated by the recent vision–language pretrained models, such as BLIP [13]. But as shown in Figure 9 and Figure 10, MAC-CapNet consistently exceeds these benchmarks values. One can notice the advantage of our proposed framework in CIDEr and SPICE scores, which is a noteworthy result as it shows greater level of propositional accuracy and semantic richness. Additionally, a uniform gain in BLEU-4, METEOR, and ROUGE-L validate that our collaborative approach improves semantic accuracy without losing proficiency in language. All these findings provide strong proof of the multi-agent collaborative approach, which help to bridge the gap between unprocessed visual inputs and detailed textual descriptions.

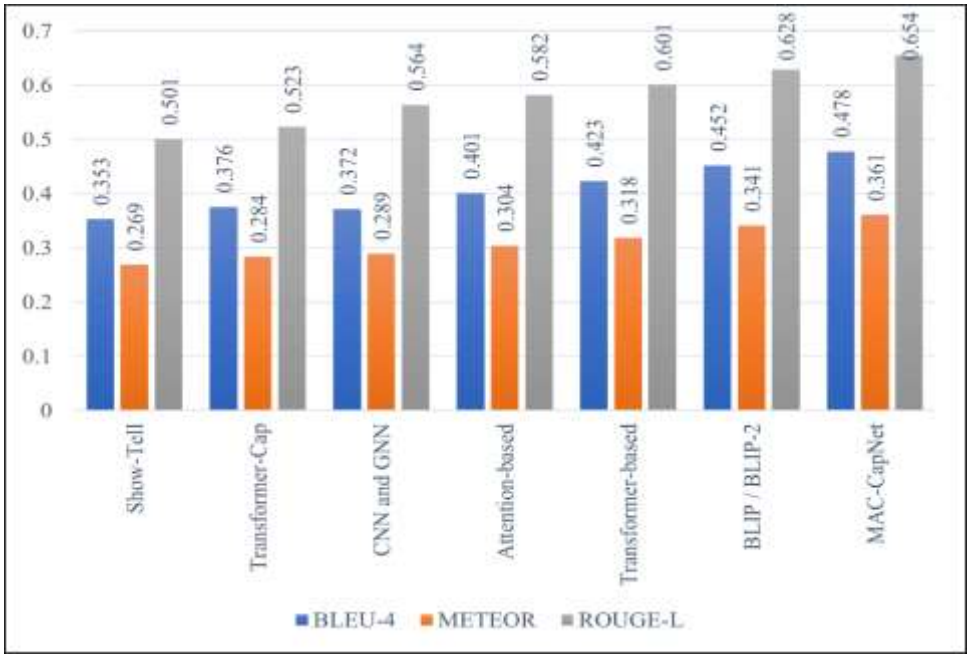


Fig. 9: Quantitative comparison across benchmarks (BLUE-4, METEOR, ROUGE-L)

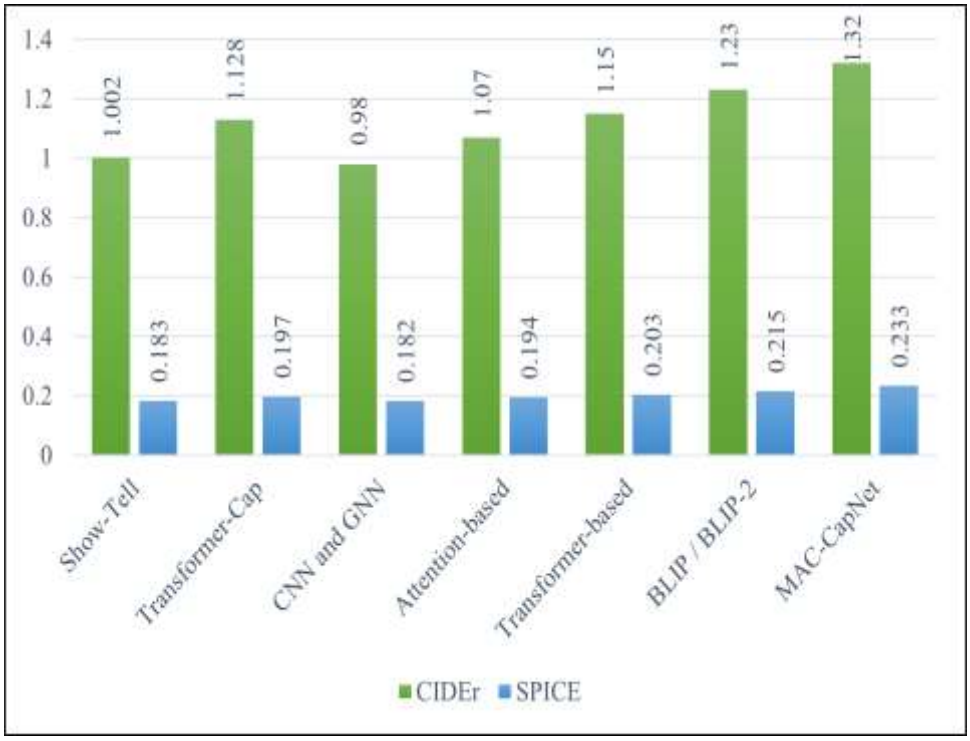


Fig. 10: Quantitative comparison across benchmarks (CIDEr and SPICE)

Table 7. Quantitative comparison across benchmarks

Model	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Show-Tell [5, 6]	0.353	0.269	0.501	1.002	0.183
Transformer-Cap [25]	0.376	0.284	0.523	1.128	0.197
CNN and GNN [24, 22]	0.372	0.289	0.564	0.980	0.182

Attention-based [12]	0.401	0.304	0.582	1.070	0.194
Transformer-based [14, 15]	0.423	0.318	0.601	1.150	0.203
BLIP / BLIP-2 [13]	0.452	0.341	0.628	1.230	0.215
MAC-CapNet (Proposed)	0.478	0.361	0.654	1.320	0.233

Domain-Specific Performance Analysis

All above result analysis are overall performance indicators, which may overlook the domain-specific behaviour. In this analysis, we examine captioning performance across different land-use categories such as urban, agricultural, forest, water, and barren land. Domain-aware DSCAs yield superior CIDEr scores across urban, forest, agricultural, water, and barren land domains (Figure 11), confirming the value of specialization [2, 4].

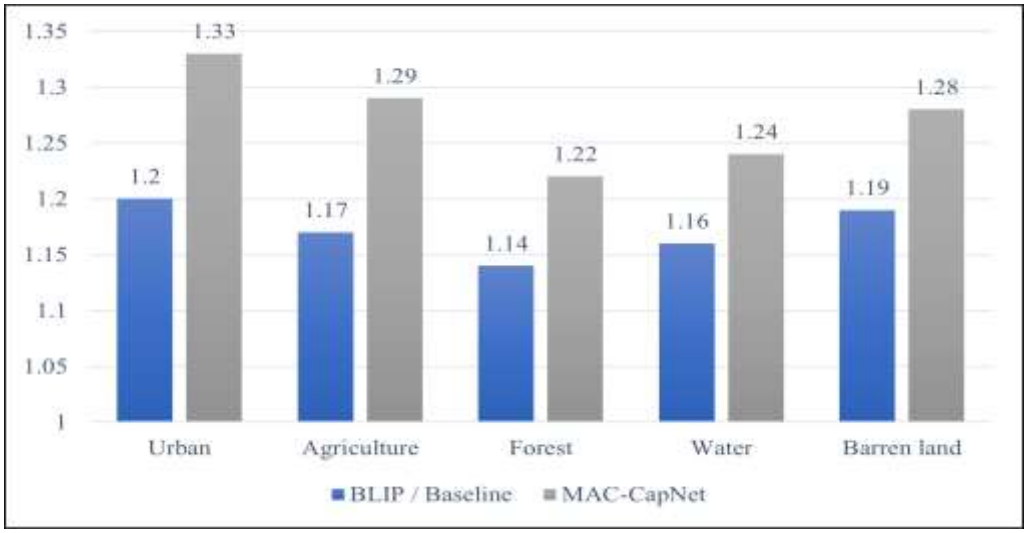


Fig. 11: Domain-specific CIDEr scores

As shown in Figure 11, MAC-CapNet produces better results over the baseline model in all domains. The noticeable improvements are obtained in urban and agricultural scenes, where images often contain multiple land-use patterns. These improvements shows that domain-specialized agents are able capture complementary semantic information, which help to generate more accurate and balanced caption for complex scenes.

The largest performance improvements are observed in urban and agricultural scenes. Urban and agricultural scenes usually contain a mixture of different land-cover and land-use elements, including roads, vegetation, water bodies, agricultural fields, and buildings. Because many of these features appear together in a single image, generating an accurate caption becomes more difficult for conventional models, which often fail to describe all important scene components.

MAC-CapNet performs better on these complex scenes because each domain-specialized agent focuses on a particular type of visual information while still exchanging semantic information with the other agents through the Cross-Agent Dynamic Routing (CADR) mechanism. This collaborative process helps combine complementary information from different domains, resulting in captions that describe the scene more completely and maintain better contextual consistency.

The performance gain is smaller for barren-land scenes. These images usually contain fewer objects and less variation in visual content, making them comparatively easier to describe. As a result, the additional interaction between domain-specialized agents has a smaller impact, although MAC-CapNet still achieves better results than the baseline model across all evaluated scene categories.Nevertheless, MAC-CapNet consistently outperforms the baseline across all evaluated scene categories, demonstrating the robustness and general applicability of the proposed collaborative domain-specialization framework.

IX. Ablation Findings

In this section, using ablation, we examine the contribution of each component of MAC-CapNet. Figure 12 shows that performance is reduce by removing or simplifying CADR (domain routing), CAMD (meta-decoder), or agents' diversity.

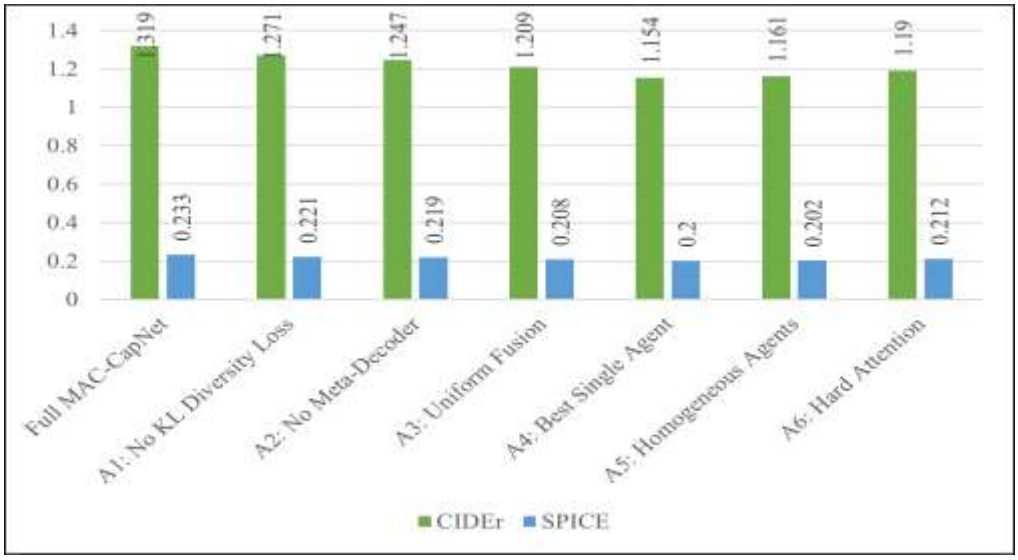


Fig. 12: Ablation findings for MAC-CapNet on CIDEr and SPICE

We noted the following findings:

Uniform fusion or single-agent reliance underperforms, showing the importance of confidence-guided adaptive fusion [18].

Agent diversity is crucial: heterogeneous backbones capture complementary features [26, 27].

Removing LKL reduces caption diversity and slightly degrades overall performance.

During the development of the model, different strategies for training the visual encoder were explored to understand how much adaptation is beneficial for remote sensing imagery. Figure 13 compares the three encoder training strategies considered in this study: a completely frozen encoder, partial fine-tuning of the top two layers, and full fine-tuning of the entire encoder. Among these settings, partial fine-tuning consistently produced the best performance across the evaluation metrics. Allowing only the upper layers of the encoder to adapt appears to help the model learn features that are more relevant to remote sensing imagery without losing the useful visual knowledge acquired during pretraining. In contrast, freezing the encoder limits its ability to adapt, while fine-tuning the entire encoder offers no clear advantage and introduces additional training complexity. Overall, partial fine-tuning provides a practical balance between domain adaptation and preservation of pretrained visual representations.

We measure caption diversity using distinct-n statistics (Distinct-1 and Distinct-2), which quantify the proportion of unique unigrams and bigrams in generated captions. Table 8 shows that including the KL-diversity loss increases both Distinct-1 and Distinct-2 scores, indicating that the proposed MAC-CapNet generates more diverse and less repetitive captions.

Table 8. Robustness comparison under noise and resolution degradation

Model	Distinct-1	Distinct-2
Without (L_{KL})	0.42	0.28
Full MAC-CapNet	0.51	0.35

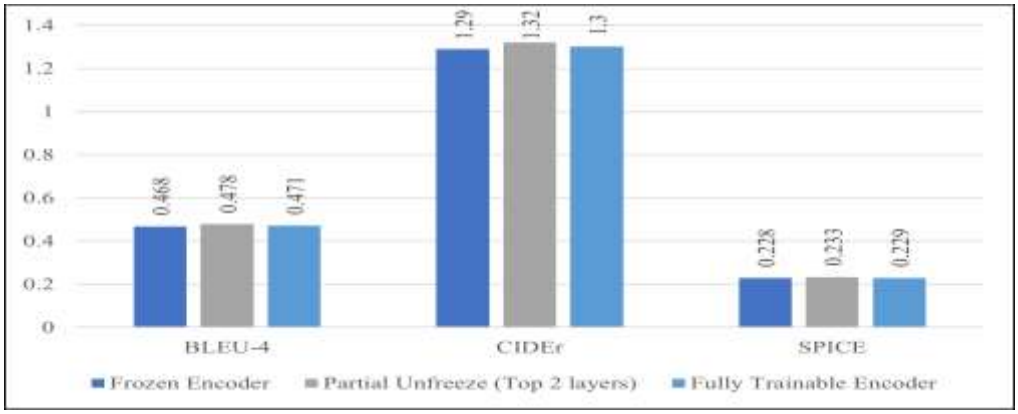


Fig. 13: Effect of encoder freezing strategy on captioning performance

Analysis of Agent Trust Weight Convergence

Within initial phase of training, all agents having similar trust weights and here model checks that which agent is suitable for which type of scene. As training goes on, model starts to update these trust weights. This update is depended upon effectiveness of agent and it is done by observing how many agents are continuously produces meaningful and relevant descriptions and how many are contributing less. This process helps to find out most valuable agents for the scene. As an example, when urban structures appear more in an image, the urban agent starts to gains more power while other agents continue to offer supporting data. This learning nature of model, results in more accurate and richer in context and meaning full final captions.

X. Interpretability Analysis

An important characteristic of MAC-CapNet is its ability to improve the transparency of the caption generation process through domain-specialized agents. Unlike conventional captioning models that produce a single latent representation, MAC-CapNet maintains separate semantic representations for each Domain-Specialized Captioning Agent (DSCA), enabling the contribution of individual agents to be analyzed during inference.

The trust weights generated by the Consensus-Aware Meta-Decoder (CAMD) provide an interpretable measure of the relative contribution of each agent to the final caption. Qualitative observations indicate that agents associated with the dominant scene characteristics generally receive higher trust weights, while other agents contribute complementary contextual information. For example, images containing predominantly urban structures tend to assign greater importance to the urban agent, whereas water-dominated scenes emphasize the water-specialized agent.

The entropy analysis presented in Section 4.3 further supports this behavior by quantifying the balance between specialization and collaboration. Moderate entropy values indicate that multiple agents contribute meaningful semantic information without excessive dominance by any single agent. In contrast, very low entropy would indicate over-reliance on a single agent, whereas very high entropy would suggest redundant or uniformly distributed contributions across agents.

Overall, these observations suggest that MAC-CapNet not only improves caption generation performance but also enhances the interpretability of the underlying decision process by exposing agent-level semantic contributions and adaptive trust assignments. This provides greater insight into how the collaborative multi-agent architecture integrates domain-specific knowledge to generate contextually coherent remote sensing image captions.

Qualitative Analysis and Failure Discussion

Figure 14, shows the different cases in which we can observe the performance of MAC-CapNet. Starting with multi-domain cases, it comes into notice that the baseline model only able to concentrate on major land-use category. But MAC-CapNet, catches most of the land-use categories and generate captions which are comprehensive. The improvement case shows the proof of enhancement in caption generation using MAC-CapNet. It is observed in example that MAC-CapNet able to note minute context of scene and relate it in description. The failure case shows where MAC-CapNet is not generating captions with better context information. This occurs when the fusion process gives less weight to weaker signals from particular domains. Collectively, this qualitative comparison shows that MAC-CapNet can handle a range of scenes and capture more semantics.

Although MAC-CapNet performs well on the evaluated datasets, it still has some limitations. One difficulty is handling remote sensing images with very complex scene compositions, especially when important objects occupy only a small part of the image. In such cases, features such as narrow roads, small water bodies, or

scattered vegetation may receive less attention, which can affect the completeness of the generated caption.

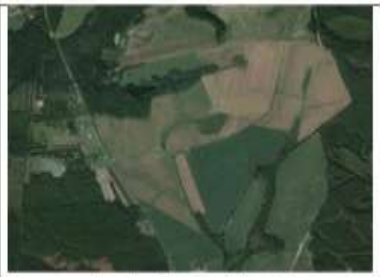
The framework also becomes more challenging to optimize when an image contains several semantic regions with similar visual importance. Since no single domain clearly dominates the scene, estimating the relative contribution of each domain-specialized agent becomes more difficult, which may reduce the effectiveness of the trust-weight assignment and the final fusion process.

In such cases, distinguishing the relative importance of different domain-specific representations becomes more difficult, which may reduce the effectiveness of the adaptive trust-weight estimation and consensus-aware fusion process.

Although the proposed collaborative multi-agent framework improves semantic representation through domain specialization and cross-agent interaction, further enhancements are possible. Future work will investigate adaptive attention mechanisms capable of better modelling fine-grained visual details, as well as multimodal foundation models and vision-language pre-training strategies to improve semantic reasoning and caption generation in highly heterogeneous remote sensing environments. Additionally, extending the framework to support a larger number of dynamically learned domain-specialized agents represents another promising direction for improving scalability and generalization.

Robustness Analysis

To check the robustness of our model, we test it on blurred images (either by introducing noise or by lowering their quality). This test is important because real-world remote images may contain such distortions in many cases. From Table 9, it is clearly observed that the baseline model and proposed model, both are not performing well. But we can notice that is how consistently MAC-CapNet outperforms the baseline in all of these conditions. MAC-CapNet outperforms than the baseline, is because of use of specialized agents and their ability to share visual representation. Due to these capabilities agents can still detect helpful cues when specific features are lost because of noise or low resolution. This robustness analysis shows that MAC-CapNet is more capable of managing difficult visual conditions.

		
Urban + Water	**Agriculture + Forest**	**Water + Barren**
GT: A river and several buildings are near a viaduct with many cars	**GT:** Many pieces of agricultural land are together	**GT:** Many bare lands are found on both sides of a curved river
BLIP: Road and water area	**BLIP:** Green land with trees	**BLIP:** Water body in open land
MAC-CapNet: Viaduct over water with road network	**MAC-CapNet:** Agricultural fields bordered by forest	**MAC-CapNet:** River through dry barren terrain

	
Mixed Urban	**Complex Scene (Failure)**
GT: A baseball field between two parallel roads is surrounded by many buildings	**GT:** Barelands are around a park with several buildings and a swimming pool
BLIP: Buildings and roads	**BLIP:** City with greenery
MAC-CapNet: Buildings with roads and green patches	**MAC-CapNet:** Urban + open land (misses water)

Fig. 14: Qualitative comparison of generated captions across different scene types. Ground truth (GT) captions are selected to best represent scene semantics. MAC-CapNet produces more detailed and context-aware descriptions compared to the baseline, while the failure case highlights limitations in capturing less prominent elements.

Table 9. Robustness comparison under noise and resolution degradation.

Condition	Model	BLEU-4	METEOR	CIDEr
Original	BLIP [13]	0.62	0.31	2.10
	MAC-CapNet	0.68	0.35	2.35
Gaussian Noise	BLIP [13]	0.55	0.28	1.85
	MAC-CapNet	0.61	0.32	2.10
Low Resolution (128 × 128)	BLIP [13]	0.50	0.26	1.70
	MAC-CapNet	0.57	0.30	1.95
Low Resolution (64 × 64)	BLIP [13]	0.44	0.23	1.45
	MAC-CapNet	0.51	0.27	1.70

Scalability and Computational Cost

We test the scalability of model by increasing the number of agents. The total computation of MAC-CapNet depend upon three components: inter-agent communication via routing and fusion, simultaneous caption synthesis by agents, and shared feature extraction. The cost of the visual encoder does not change because it is shared. The interaction between agents and the numerous decoders are the main sources of overhead. Computation expands in proportion to the number of agents, yet in reality, this growth is still manageable. Table 10 shows that MAC-CapNet takes somewhat more parameters and inference time than the baseline but with this additional cost MAC-CapNet able to provide quality captions. Furthermore, the modular design of model makes it easier to add new agent and expand it.

Table 10. Computational cost comparison.

Model	Params (M)	Inference Time (ms)	CIDEr
BLIP (Baseline) [13]	220	45	2.10
MAC-CapNet (3 Agents)	260	60	2.25
MAC-CapNet (5 Agents)	300	75	2.35

XI. Limitations

Although MAC-CapNet demonstrates consistent improvements across multiple benchmark datasets, several limitations remain that provide opportunities for future research.

First, the collaborative multi-agent architecture requires multiple Domain-Specialized Captioning Agents (DSCAs), resulting in a larger model size and higher computational complexity than conventional single-decoder captioning models. Although the shared visual encoder reduces redundant feature extraction and the agents can be executed in parallel, additional optimization may be required for deployment on resource-constrained edge devices or real-time remote sensing applications.

Second, the effectiveness of the proposed framework depends on the predefined domain specialization of the DSCAs. While the selected domains capture major semantic characteristics of remote sensing imagery, more complex scenes may involve overlapping or previously unseen semantic categories that are not explicitly represented by the current agent configuration. Developing dynamically adaptive or automatically learned domain-specialization strategies could further improve model flexibility and generalization.

Third, the experimental study is limited to the RSICD and UCM-Captions datasets. These datasets are commonly used for evaluating remote sensing image captioning methods, but they do not cover every type of scene encountered in real-world applications. Images collected from different geographic locations, seasons, sensors, and environmental conditions may present additional challenges that are not fully represented in the current evaluation.

The robustness experiments mainly examine the effects of image noise and reduced resolution. Other practical factors, such as cloud cover, atmospheric interference, changing illumination, seasonal differences, and variations between imaging sensors, were not considered in this study. Evaluating the proposed framework under these conditions would provide a more complete picture of its performance in real operational environments.

Future work will focus on addressing these limitations by exploring more flexible agent designs, validating the model on larger and more diverse datasets, and performing broader robustness experiments under a wider range of imaging conditions.

Cross-Dataset Generalization

We have considered RSICD dataset for training and the UCM-Captions dataset for testing and we did this to check the model capability to understand unseen data. The ability of MAC-CapNet, to share learned representations, overcome the lack of fin-tuning and it is shown in Table 11 that MAC-CapNet did better than the baseline. The use of domain-specialized agents and their cooperative interaction help model to understand complex scene.

Table 11. Cross-dataset Generalization Performance (Train: RSICD, Test: UCM).

Model	BLEU-4	METEOR	CIDEr
BLIP [13]	0.48	0.25	1.60
MAC-CapNet	0.54	0.29	1.85

XII. Conclusion

Here we confirm that our proposed MAC-CapNet approach is capable of handling the complexities of remote sensing imagery. It is possible because we used a multi-agent collaborative framework, this approach included a meta-decoder to perform confidence-weighted fusion, ensuring that the final output prioritizes the most reliable agent contributions.

This collaborative framework generates more accurate captions that describe all domains present within the image and are interpretable as well than previous approaches. Experimental results on RSICD demonstrate consistent improvements over strong baselines, achieving gains in BLEU-4 (0.452→0.478, +5.7%), METEOR (0.341→0.361, +5.9%), ROUGE-L (0.628→0.654, +4.1%), CIDEr (1.23→1.32, +7.3%), and SPICE (0.215→0.233, +8.4%). Domain-wise evaluation further shows notable CIDEr improvements across all categories, including urban (+10.8%) and agriculture (+10.2%), confirming the effectiveness of domain specialization. Ablation results highlight the importance of collaborative design, where removing the meta-decoder reduces CIDEr by 5.5%, and using a single agent leads to a 12.5% drop. Additionally, diversity improves with KL regularization (Distinct-1: +21.4%, Distinct-2: +25.0%).

In addition to improving caption quality, the proposed framework makes it easier to understand how the final prediction is formed. The use of domain-specialized agents, adaptive trust weights, and consensus-aware fusion allows the contribution of each agent to be examined during caption generation. The experimental results indicate that combining domain-specific knowledge with information exchange between agents helps produce captions that are more accurate and better aligned with the scene content. This advantage is most evident in remote sensing images that contain a mixture of different land-use categories.

The model also demonstrates robustness under degraded conditions, maintaining up to +0.25 CIDEr improvement over baselines, with statistically significant gains ($p < 0.05$). Finally, we conclude here that MAC-CapNet lays the foundation for the upcoming generation of intelligent remote sensing systems—capable of cooperative and context-sensitive scene understanding.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

This manuscript was created and written by Chandrashekhar; idea by Chandrashekhar; methodology by Chandrashekhar; software by Chandrashekhar; visualization by Chandrashekhar; project administration, Chandrashekhar and supervision by Dr. Ashwin.

References

[1] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao and L. Zhang, "Deep learning in environmental remote sensing: Achievements and challenges," Remote Sens. Environ., vol. 241, pp. 111716, 2020.

[2] P. Zhang, Y. Ke, Z. Zhang, M. Wang, P. Li and S. Zhang, "Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery," Sensors, vol. 18, no. 11, pp. 3717, 2018.

[3] Z. T. AlAli and S. A. Alabady, "A survey of disaster management and SAR operations using sensors and supporting techniques," Int. J. Disaster Risk Reduction, vol. 82, pp. 103295, 2022.

[4] M. Weiss, F. Jacob and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," Remote Sens. Environ., vol. 236, pp. 111402, 2020.

[5] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3156–3164.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in Proc. Int. Conf. Mach. Learn. (ICML), 2015.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 936–944.

[8] B. Zhao, "A systematic survey of remote sensing image captioning," IEEE Access, vol. 9, pp. 154086–154111, 2021.

[9] T. G. Dietterich, "Ensemble methods in machine learning," in Proc. Int. Workshop Multiple Classifier Systems, 2000, pp. 1–15.

[10] T. Rashid, M. Samvelyan, C. S. De Witt et al., "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 4292–4301.

[11] X. Lu, B. Wang, X. Zheng and X. Li, "Exploring models and data for remote sensing image caption generation," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 4, pp. 2183–2195, 2018.

[12] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 6077–6086.

[13] J. Li, D. Li, C. Xiong and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. Int. Conf. Mach. Learn. (ICML), 2022, pp. 12888–12900.

[14] Z. Wang et al., "GIT: A generative image-to-text transformer for vision-language tasks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022.

[15] P. Wang et al., "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in Proc. Int. Conf. Mach. Learn. (ICML), 2022.

[16] L. K. Hansen and P. Salamon, "Neural network ensembles," IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 10, pp. 993–1001, 1990.

[17] Y. Yang et al., "Mean field multi-agent reinforcement learning," in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 5571–5580.

[18] T. T. Nguyen, A. V. Luong, M. T. Dang, A. W.-C. Liew and J. McCall, "Ensemble selection based on classifier prediction confidence," Pattern Recognit., vol. 100, pp. 107104, 2020.

[19] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 1180–1189.

[20] H. Zhang, J. Xue and K. Dana, "Deep TEN: Texture encoding network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2896–2905.

[21] B. Gong, Y. Shi, F. Sha and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2012, pp. 2066–2073.

[22] J. Yan, S. Ji and Y. Wei, "A combination of convolutional and graph neural networks for regularized road surface extraction," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–13, 2022, Art. no. 4409113.

[23] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post flood scene understanding," IEEE Access, vol. 9, pp. 89644–89654, 2021.

[24] J. Liang, Y. Deng and D. Zeng, "A deep neural network combined CNN and GCN for remote sensing scene

classification," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 4325–4338, 2020.

[25] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross and V. Goel, "Self-critical sequence training for image captioning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1179–1195.

[26] S. Sabour, N. Frosst and G. E. Hinton, "Dynamic routing between capsules," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[27] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[28] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021.

[29] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.

[30] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. Int. Conf. Learn. Represent. (ICLR), 2014.

[31] X. Lu, B. Wang, X. Zheng and X. Li, "Exploring models and data for remote sensing image caption generation," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[32] K. Papineni et al., "BLEU: A method for automatic evaluation of machine translation," in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2002, pp. 311–318.

[33] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proc. ACL Workshop Intrinsic Extrinsic Evaluation Measures Mach. Translation Summarization, 2005, pp. 65–72.

[34] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Proc. ACL Workshop, 2004.

[35] R. Vedantam et al., "CIDEr: Consensus-based evaluation for image captioning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015.

[36] P. Anderson et al., "SPICE: Semantic propositional image caption evaluation," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016.

[37] Y. Zhang, C. Zhang, Y. Tang and Z. He, "Cross-modal concept learning and inference for vision-language models," Neurocomputing, vol. 583, pp. 127530, 2024.

[38] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu and Z. Cao et al., "mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections," arXiv preprint arXiv:2205.12005, 2022.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.

[40] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in Proc. Int. Conf. Learn. Represent. (ICLR), 2017.

[41] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 3319–3327.

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning," In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 1516, 2023, 34892–34916.

[43] Zhu, Deyao, et al. "Minigpt-4: Enhancing vision-language understanding with advanced large language models," International Conference on Learning Representations. Vol. 2024. 2024.

[44] Bazi, Yakoub, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani, "RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery," Remote Sensing 16, no. 9: 1477, 2024. https://doi.org/10.3390/rs16091477

[45] B. Xiao et al., "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 4818-4829, doi: 10.1109/CVPR52733.2024.00461.

[46] Huda Diab Abdulgalil, Otman A. Basir, "Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to Multimodal Large Language Models," Natural Language Processing Journal, Volume 12, 2025, pp. 100-159, ISSN 2949-7191, https://doi.org/10.1016/j.nlp.2025.100159.