Remote Sensing Image Captioning via Dual-Stream Fusion and Spatial Relation-Aware Encoding

Rong Chen, Guorui Ma, Lunjun Fan
State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China - crbrave@whu.edu.cn; mgr@whu.edu.cn; 2023186190087@whu.edu.cn

Keywords: Image Captioning, CLIP, Spatial Relation Awareness, Transformer, Remote Sensing

Abstract

Remote sensing image captioning (RSIC) aims to describe key objects in remote sensing images using natural language, with significant applications in disaster assessment, land-use identification, and scene understanding. Existing methods face two critical challenges: insufficient cross-modal alignment due to the domain gap between generic visual representations and remote sensing semantics, and inadequate spatial relation modeling among regions in complex scenes, which compromises the semantic precision and logical coherence of generated descriptions. To address these issues, this paper proposes the Dual-Stream Relation-Aware Transformer (DSRAT) for remote sensing image captioning. On the visual encoding side, multi-scale CNN features serve as the foundation, fused with domain-specific semantic priors from RemoteCLIP through a gated dual-stream fusion module to achieve adaptive alignment of multi-source visual information. Subsequently, a spatial relation-aware mechanism is introduced into the encoder self-attention, which explicitly encodes geometric relationships such as relative position, distance, and orientation between regions as attention biases, enhancing the model’s capability for structured representation of complex spatial layouts and multi-object interaction scenarios. Finally, adaptive weighted aggregation of multi-layer encoder outputs generates discriminative cross-modal memory representations for the decoder. Experiments on the RSICD and NWPU-Captions datasets demonstrate that DSRAT achieves state-of-the-art performance across six metrics on RSICD and all seven metrics on NWPU-Captions. In particular, DSRAT achieves a significant performance improvement of +14.45 CIDEr on NWPU-Captions compared to the state-of-the-art method, validating the effectiveness of the proposed approach.

1. Introduction

Remote sensing image caption generation requires models to not only perceive ground objects within images but also comprehend the relative positions, quantities, and functional relationships among targets, offering broad application prospects in rapid disaster assessment, land-use change narration, and remote sensing knowledge base construction (Lu et al., 2017). Unlike discriminative tasks such as object detection and semantic segmentation, RSIC simultaneously involves cross-modal semantic alignment and sequence generation, the difficulty of which is further exacerbated in the remote sensing domain due to scene complexity (Qu et al., 2016).

Early RSIC methods employed CNNs for global image feature extraction and RNNs or LSTMs for description generation, establishing the encoder–decoder framework (Qu et al., 2016; Lu et al., 2017). The subsequent introduction of attention mechanisms enabled models to dynamically focus on different image regions: Wang et al. (2019) learned a shared semantic space between images and sentences through semantic embedding. Liu et al. (2022) proposed MLAT, which aggregates features from each Transformer encoder layer via LSTM; Wang et al. (2022) proposed GLCM, integrating global semantic features with local region attention. To address multi-scale challenges, Cheng et al. (2022) proposed MLCA-Net to aggregate semantic features at different resolutions; Du et al. (2023) designed a deformable hierarchical Transformer for modeling foreground–background multi-scale information; Yang et al. (2024b) proposed HCNet, which combines hierarchical feature aggregation with cross-modal alignment and achieved notable progress on NWPU-Captions. The advent of large-scale vision–language pre-training has introduced a new research paradigm: Liu et al. (2024) proposed RemoteCLIP, which learns remote-sensing-specific image–text aligned representations through domain-adaptive contrastive pre-training; Meng et al. (2024) further incorporated CLIP’s global semantics as latent variables into a multi-scale grouping Transformer.

Despite the significant progress achieved by the aforementioned methods, the following two issues remain challenging:

The modality gap between visual features and remote sensing semantic priors. Mainstream methods directly feed visual features extracted by CNNs or Swin Transformer into the decoder, lacking domain-specific cross-modal semantic prior support. Some works have attempted to introduce generic CLIP features (Meng et al., 2024), but generic CLIP has limited semantic understanding of remote-sensing-specific land cover objects, and a significant modality bias persists between it and remote sensing visual features, leading to inaccurate vision–language alignment. More importantly, existing schemes for incorporating CLIP features into RSIC mostly employ feature concatenation or simple additive fusion—a shallow integration that fails to enable complementary enhancement between visual features and remote sensing semantic priors across multiple abstraction levels, thereby limiting the semantic discriminability of encoder outputs.

The absence of spatial relation modeling among regions during the encoding stage. Existing grid-feature-based encoding frameworks treat all spatial tokens as independent features, neglecting the inherent spatial topological structures among ground objects in remote sensing images (Du et al., 2023; Yang et al., 2024a). This deficiency causes the model to produce positional description errors or missing relational semantics when generating sentences involving multi-object spatial relationships, thereby reducing the semantic completeness and factual accuracy of descriptions.

To systematically address the two aforementioned problems, this paper proposes the Dual-Stream Relation-Aware Transformer (DSRAT). For Problem (1), a dual-stream remote sensing semantic fusion architecture is constructed, employing Swin-B combined with Feature Pyramid Network (FPN) to extract multi-scale visual features as the first stream, and in parallel introducing RemoteCLIP’s ViT-L/14 to extract domain-aligned semantic features as the second stream; the two feature streams are adaptively fused into a unified visual representation through a per-token learnable sigmoid gating mechanism. Unlike shallow feature concatenation schemes, the dual-stream architecture enables complementary enhancement between visual features and remote sensing semantic priors, significantly reducing the vision–semantic modality gap. For Problem (2), a relation-aware Transformer encoder is proposed, which explicitly constructs a spatial geometric relation graph among image regions on top of the standard self-attention framework and integrates relation embeddings into the attention computation of encoder layers, enabling the encoder to perceive the spatial topological structure among different ground object regions, thereby generating complete descriptive sentences with accurate positional and relational semantics.

The main contributions of this work are summarized as follows.

1) We propose DSRAT, a Dual-Stream Relation-Aware Transformer for remote sensing image captioning, which integrates a gated dual-stream fusion module and a spatial relation-aware encoder to simultaneously address the vision–semantic modality gap and the lack of explicit inter-region spatial relation modeling. DSRAT achieves state-of-the-art performance on both the RSICD and NWPU-Captions benchmarks.

2) We design GateFusion, a per-token sigmoid-gated dual-stream fusion module that adaptively integrates CNN multi-scale spatial features (Swin-B+FPN) with RemoteCLIP-LoRA domain-specific semantic priors, enabling complementary enhancement across multiple abstraction levels and effectively bridging the vision–semantic modality gap.

3) We introduce a Spatial Relation-Aware Transformer encoder featuring (RelMHA), which explicitly injects pairwise geometric relation embeddings (relative position, distance, and azimuth angle) into multi-head attention via per-head sigmoid-gated dual-path injection, together with adaptive multi-layer aggregation, providing structured and semantically rich visual memory for the caption decoder.

2. Methodology

2.1 Overall Structure

As illustrated in Figure 1, DSRAT comprises four principal modules: (1) a dual-source visual feature extraction module, employing a frozen Swin-B+Feature Pyramid Network (FPN) multi-scale CNN branch and a RemoteCLIP semantic branch fine-tuned via low-rank adaptation (LoRA) to provide hierarchical spatial features and rich semantic priors, respectively; (2) an adaptive gated fusion module, which integrates the two heterogeneous feature streams into a unified representation through a per-token learnable sigmoid gating mechanism; (3) a spatial relation-aware Transformer encoder, which injects explicit geometric relation encoding into multi-head attention and generates visual memory for the decoder through multi-layer adaptive weighted aggregation; and (4) a Transformer caption decoder, which autoregressively generates the target description sequence conditioned on visual memory, employing a weight-tying strategy to reduce parameter count.

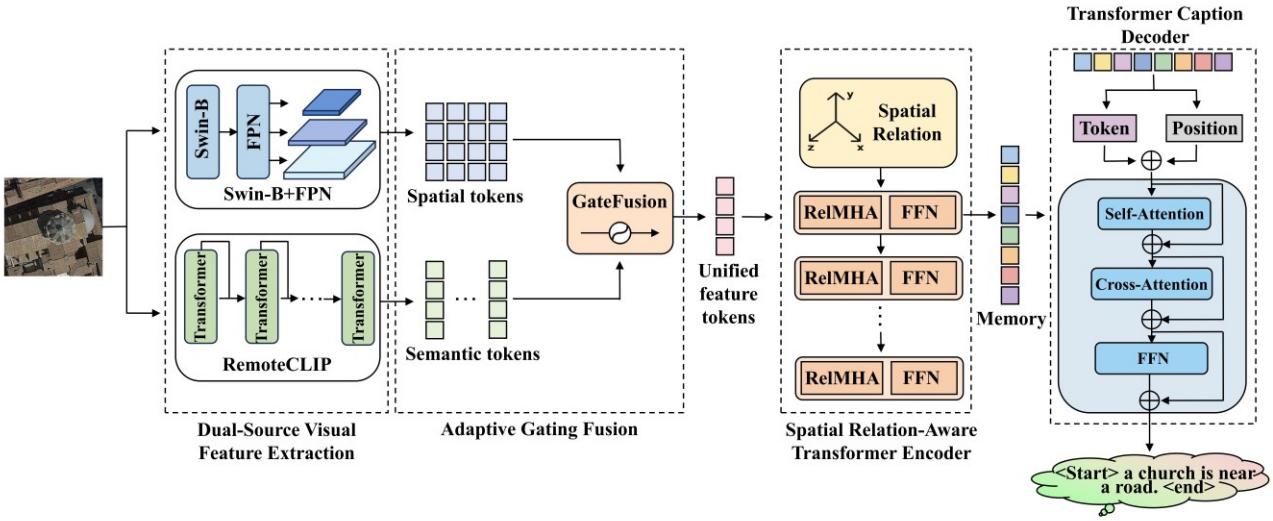

Figure 1. Overall architecture of DSRAT. Comprising Dual-Source Visual Feature Extraction, GateFusion module, 6-layer Spatial Relation-Aware Transformer Encoder with adaptive layer aggregation, and Transformer Caption Decoder.

2.2 Dual-Source Visual Feature Extraction

CNN-FPN Multi-Scale Feature Extraction Branch. Swin-B (Liu et al., 2021) pre-trained on ImageNet-22K serves as the backbone network. Given the limited scale of remote sensing annotation data, the backbone parameters remain entirely frozen during training to prevent overfitting. For a 256×256 input, Swin-B produces three-level feature maps at Stages 2, 3, and 4. Each level is projected to a unified dimension of 512 via 1×1 lateral convolutions, followed by local feature enhancement through a residual refinement module. A Feature Pyramid Network (FPN; Lin et al., 2017) fuses hierarchical semantics in a top-down manner:

$$p_i = \mathrm{Refine}(lat_i(c_i) + \mathrm{Up}(p_{i+1})), i \in \{4, 3, 2\} \quad (1)$$

where c_i denotes the feature map output from Swin-B Stage i, $lat_i(\cdot)$ is the 1×1 lateral convolution, Up $(\cdot)$ denotes bilinear upsampling, and Refine $(\cdot)$ is the residual refinement module. The three feature maps are aligned to 16×16 and fused via attention-gated weighted aggregation. Compared to simple concatenation, attention gating endows the model with the ability to adaptively adjust the contribution weights of each scale according to different ground object content—favoring fine-grained features in dense small-object scenarios and high-level semantic features for scene-level categorization.

RemoteCLIP-LoRA Semantic Branch. RemoteCLIP (ViT-L/14), pre-trained via contrastive learning on large-scale remote sensing image–text pairs, possesses strong cross-modal semantic alignment capability. However, full fine-tuning readily leads to overfitting. Therefore, this work adopts a Low-Rank Adaptation (LoRA) strategy (Hu et al., 2022), injecting low-rank increments only into the output projection matrices of the last four Transformer blocks:

$$W_{eff} = W_{frozen} + \frac{\alpha}{r}BA,\ A \in R^{r\times d},\ B \in R^{d\times r} \quad (2)$$

where $A \in R^{r\times d}$ and $B \in R^{d\times r}$ are the low-rank decomposition matrices with rank r = 8, scaling factor α = 16, and d = 1024 representing the attention layer dimension of ViT-L/14. B is initialized to zero to ensure that the increment is zero at the start of adaptation, guaranteeing training stability. This strategy introduces only approximately 66K additional trainable parameters, effectively adapting to the remote sensing domain distribution while fully preserving pre-trained semantic knowledge, outputting a semantic patch sequence.

2.3 Spatial Relation-Aware Encoder

Adaptive Gated Fusion. The two feature streams differ in both dimensionality and representation space, making simple concatenation insufficient for effective integration. This work designs a per-token sigmoid gated fusion mechanism that takes the concatenated linearly projected features from both streams as input, generates per-position fusion weights, and adaptively merges the two representations:

$$\alpha = \sigma(\mathrm{MLP}[c; v]),\ X_0 = \mathrm{LN}(\alpha\cdot c + (1-\alpha)\cdot v) \quad (3)$$

where c and v denote the linearly projected features from the two streams, σ $(\cdot)$ is the sigmoid activation function, and LN $(\cdot)$ denotes layer normalization. The gating network consists of two linear transformations with the output bias initialized to zero. This mechanism allows the model to dynamically balance the relative contributions of semantic priors and spatial details according to the ground object content, with the fused result serving as the encoder input.

Spatial Relation Encoding. The spatial topological structure of ground objects in remote sensing images is crucial for generating accurate descriptions, yet the absolute positional encoding of standard Transformers cannot explicitly model pairwise geometric relationships between tokens. For each pair of tokens on the 16×16 token grid (N = 256), this work constructs a four-dimensional geometric feature vector comprising normalized coordinate differences, Euclidean distance, and azimuth angle, which is then mapped to pairwise relation embeddings via a two-layer MLP. As illustrated in Figure 2, the relation embedding matrix is independently linearly projected to obtain a key-side component r_k and a value-side component r_v, which are injected into the attention score computation and value aggregation stages, respectively, through per-head learnable sigmoid gating, achieving dual-path enhancement of attention weights and context representations by spatial geometric relations.

Relation-Aware Multi-Head Attention (RelMHA). Following the relative position encoding framework of (Shaw et al., 2018), spatial relation embeddings are simultaneously injected into both the attention score computation and context aggregation stages. For the h-th attention head, the attention score is computed as:

$$\mathrm{score}_{ij}^{h} = \frac{q_i^h k_j^{h\top} + \sigma(\alpha_k^h) q_i^h (W_k r_{ij})^\top}{\sqrt{d}} \quad (4)$$

In context aggregation, the relation embeddings further participate as biases to the value vectors:

$$\mathrm{ctx}_i^h = \sum_j a_{ij}^h (v_j^h + \sigma(\alpha_v^h) W_v r_{ij}) \quad (5)$$

where r_{ij} is the pairwise spatial relation embedding between tokens i and j, W_k and W_v are the corresponding relation projection matrices, d is the attention head dimension, and a_{ij}^h denotes the normalized attention weight. α_k^h and α_v^h are per-head learnable scalars initialized to zero, so the gating values $\sigma(\alpha_k^h)$ and $\sigma(\alpha_v^h)$ equal 0.5 at the beginning of training, allowing the relation bias to participate with moderate weight, ensuring a smooth training start. The encoder stacks 6 layers in total, and the outputs of the last 3 layers are aggregated via learnable softmax weights to form the decoder’s visual memory, enabling the decoder to benefit from both mid-level structural features and high-level semantic representations.

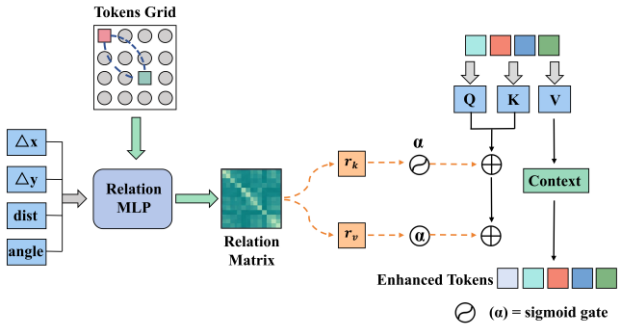

Figure 2. Illustration of the proposed RelMHA with spatial relation encoding and per-head sigmoid-gated injection.

2.4 Caption Decoder

The caption decoder, conditioned on visual memory, employs a 6-layer standard Transformer architecture (Vaswani et al., 2017) for autoregressive generation of the target description sequence. Each layer comprises three sub-modules in sequence: masked causal self-attention, visual memory cross-attention, and a feed-forward network. The masked self-attention ensures that step t attends only to historical tokens, preventing future information leakage; the cross-attention uses the current decoding state as queries and visual memory as keys and values, enabling dynamic visual feature addressing for language generation. The final output is processed through layer normalization and linear projection to yield the logit distribution over the vocabulary:

$$\mathrm{logits}_t = W_{emb}\ \mathrm{LN}(\mathrm{tgt}_t) \in R^V \quad (6)$$

where W_{emb} is the output projection matrix shared with the input word embeddings, and V is the vocabulary size. Weight tying between the output projection and input embedding matrices enforces consistency between input and output representation spaces while effectively reducing the parameter count. During inference, batch beam search with a beam size of 3 is employed.

2.5 Loss Function

The first training stage employs token-level cross-entropy loss for supervised training:

$$\mathcal{L}_{\mathrm{XE}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_{\theta}\left(y_t^* | y_{<t}^*; I\right) \quad (7)$$

where $\mathcal{L}_{\mathrm{XE}}$ is the cross-entropy loss, θ denotes model parameters, y_t^* is the ground-truth token at step t, $y_{<t}^*$ represents the preceding history tokens, and T is the sequence length. To mitigate decoder attention collapse, an attention coverage regularization term L_{reg} is introduced to penalize cross-attention distributions deviating from uniformity. The combined training objective is:

$$L_{train} = L_{XE} + \lambda L_{reg},\ \lambda = 0.01 \quad (8)$$

The second stage performs Self-Critical Sequence Training (SCST) reinforcement fine-tuning. Cross-entropy training suffers from two inherent drawbacks: exposure bias arising from the distributional gap between teacher forcing and autoregressive inference, and the fundamental inconsistency between the maximum likelihood objective and evaluation metrics such as CIDEr-D. Therefore, after cross-entropy convergence, SCST (Rennie et al., 2017) fine-tuning is conducted. For the same image, greedy decoding yields a baseline sequence, and random sampling generates a candidate sequence, defining the advantage estimate as the CIDEr difference. The SCST loss is:

$$L_{SCST} = -E_{\hat{y}\sim p_\theta}[A(\hat{y}) \cdot \sum_{t=1}^{T} \log p_\theta\left(\hat{y}_t \mid \hat{y}_{<t}, I\right)] \quad (9)$$

The advantage signal is detached from the computational graph and does not participate in gradient backpropagation. During the SCST fine-tuning stage, the learning rate is set to 1/10 of the cross-entropy stage, with validation-set CIDEr-D as the monitoring metric, employing an early stopping strategy with patience = 5 combined with learning rate decay for convergence control, ensuring that reinforcement learning gradients do not excessively perturb the converged language generation capability.

3. Experiment

3.1 Datasets

The proposed method is validated on two benchmark datasets: RSICD (Lu et al., 2017) and NWPU-Captions (Cheng et al., 2022). RSICD is collected from Google Earth, Baidu Maps, and other platforms, comprising 10,921 remote sensing images at 224×224 pixel resolution across 30 scene categories, with 5 human-annotated captions per image. NWPU-Captions is built upon the NWPU-RESISC45 scene classification dataset, containing 31,500 images at 500×500 pixels covering 45 scene categories with 157,500 caption sentences, making it one of the largest remote sensing image captioning datasets to date. Both datasets follow standard splits into training, validation, and test sets.

3.2 Evaluation Metrics

Seven automatic evaluation metrics are employed for comprehensive quality assessment. BLEU-1 through BLEU-4 (Papineni et al., 2002) measure text overlap by computing 1- to 4-gram precision between generated and reference descriptions. METEOR (Denkowski and Lavie, 2014) extends word-level matching with stemming and synonym matching, offering greater tolerance to linguistic variations. $\mathrm{ROUGE_L}$ (Lin, 2004) computes recall based on the longest common subsequence, reflecting description coverage completeness. CIDEr (Vedantam et al., 2015) uses TF-IDF weighting on n-grams to emphasize the capture of key land-cover vocabulary, exhibiting the highest correlation with human judgment and serving as the primary metric of this work.

3.3 Experimental Settings

All experiments are conducted on a single NVIDIA RTX 4090 GPU. The encoder employs a frozen Swin-B backbone with an FPN head, and only the output projection matrices of the last four blocks in the RemoteCLIP branch participate in training (LoRA rank r = 8, α = 16). Both the Transformer encoder and decoder stack 6 layers with model dimension d = 512, H = 8 attention heads, and feed-forward dimension 2048. Training proceeds in two stages: the first stage uses the Adam optimizer for cross-entropy pre-training with an initial learning rate of 5×10^{-4} and regularization weight λ = 0.01; the second stage fine-tunes with SCST at a reduced learning rate of 5×10^{-5}, using validation-set CIDEr-D as the monitoring metric with an early stopping patience of 5. Beam search with a beam size of 3 is used during inference.

3.4 Experimental Results

3.4.1 Comparison with Existing Methods.

Tables 1 and 2 present quantitative comparisons between DSRAT and existing methods on the RSICD and NWPU-Captions datasets. On the RSICD dataset, DSRAT surpasses all compared methods in BLEU-1 through BLEU-4, METEOR, and $\mathrm{ROUGE_L}$, achieving optimal performance across six metrics. Compared with the previous strongest baseline MLAT, BLEU-4 improves by 2.26%, METEOR by 1.55%, $\mathrm{ROUGE\_L}$ by 2.51%, and CIDEr by 11.18, reflecting the combined gains from dual-stream feature fusion and spatial relation encoding on description accuracy. On the NWPU-Captions dataset, DSRAT achieves optimal results across all seven evaluation metrics. Compared with the current state-of-the-art HCNet (Yang et al., 2024b), BLEU-4 improves by 3.19% and CIDEr by 14.45, validating the generalization capability and superiority of the proposed method on large-scale complex scenes.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
VLAD+RNN (Lu et al., 2017)	49.38	30.91	22.09	16.77	19.96	42.42	103.92
VLAD+LSTM (Lu et al., 2017)	50.04	31.39	23.19	17.78	20.46	43.34	118.01
mRNN (Qu et al., 2016)	45.58	28.25	18.09	12.13	15.69	31.26	19.15
mLSTM (Qu et al., 2016)	50.57	32.42	23.19	17.46	17.84	35.02	31.61
mGRU (Li et al., 2018)	42.56	29.99	22.91	17.98	19.41	37.97	124.82
CSMLF (Wang et al., 2019)	57.59	38.59	28.32	22.17	21.28	44.55	52.97
Sound-f-a (Lu et al., 2020)	59.35	45.11	35.29	28.08	26.11	49.57	**132.35**
Soft-attention (Xu et al., 2015)	65.13	49.04	39.00	32.20	26.39	49.69	90.58
MLAT (Liu et al., 2022)	66.90	51.13	41.14	34.21	27.31	50.57	94.27
DSRAT (Ours)	**69.06**	**53.46**	**43.54**	**36.47**	**28.86**	**53.08**	105.45

Table 1. Experimental results on the RSICD dataset. Bold indicate the best result.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
Multimodal (Qu et al., 2016)	72.50	60.30	51.80	45.50	33.60	59.10	117.90
Soft Attn (Lu et al., 2017)	73.10	60.90	52.50	46.20	33.90	59.90	113.60
Hard Attn (Lu et al., 2017)	73.30	61.00	52.70	46.40	34.00	60.00	110.30
FC-Att+LSTM (Zhang et al., 2019)	73.60	61.50	53.20	46.90	33.80	60.00	123.10
SM-Att+LSTM (Zhang et al., 2019)	73.90	61.70	53.20	46.80	33.00	59.30	123.60
MLCANet (Cheng et al., 2022)	75.40	62.40	54.10	47.80	33.70	60.10	126.40
RS-CapRet (Silva et al., 2024)	87.10	78.70	71.70	65.60	43.60	77.60	192.90
HCNet (Yang et al., 2024b)	89.54	82.51	76.58	71.68	47.42	80.53	209.27
DSRAT (Ours)	**91.23**	**84.97**	**79.60**	**74.87**	**47.74**	**81.58**	**223.72**

Table 2. Experimental results on the NWPU-Captions dataset. Bold indicate the best result.

3.4.2 Ablation Study.

This subsection conducts experiments on the RSICD dataset to validate the effectiveness of the gated fusion module, spatial relation-aware multi-head attention module, and adaptive aggregation module (Table 3). Experiments follow an incremental strategy: Baseline denotes the configuration with all three modules disabled; M1 enables the gated fusion module; M2 further enables the spatial relation multi-head attention module; the complete DSRAT model has all three modules activated. The alternative implementation for each disabled module is as follows: when the gated fusion module is disabled, only the CNN features are linearly projected without introducing CLIP semantic information; when the spatial relation-aware multi-head attention module is disabled, it is replaced by standard multi-head attention without spatial geometric relation biases; when the adaptive layer aggregation module is disabled, only the output of the single best encoder layer is used without multi-layer weighted fusion. The decoder structure remains identical across all variants to ensure fair comparison.

Effect of Gated Fusion. After introducing the GateFusion module (M1), BLEU-4 improves by 1.82% and CIDEr by 6.84 compared to Baseline—the most significant gain. This indicates that adaptively fusing RemoteCLIP semantic priors with CNN spatial features via per-token sigmoid gating effectively bridges the modality gap between remote sensing visual features and linguistic semantics, providing a more discriminative initial representation for subsequent encoding.

Effect of Spatial Relation-Aware Multi-Head Attention. Adding the RelMHA module on top of M1 (yielding M2) further improves BLEU-4 by 1.35% and CIDEr by 5.81. This validates the effectiveness of explicitly injecting inter-region geometric relation biases into the encoder self-attention, notably enhancing the model’s understanding of multi-object spatial topology.

Effect of Adaptive Layer Aggregation. The complete DSRAT model, adding AdaptiveAgg on top of M2, yields a further 0.44% BLEU-4 improvement and 2.03 CIDEr gain. The cumulative gains of the three modules relative to Baseline are +3.61 BLEU-4 and +14.68 CIDEr, with each module contributing independently and exhibiting complementary effects, validating the rationality of the overall architectural design.

Method	GateFusion	RelMHA	AdaptiveAgg	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
Baseline	✗	✗	✗	65.12	49.41	39.55	32.86	26.41	49.69	90.77
M1	✓	✗	✗	67.56	51.71	41.68	34.68	27.92	51.17	97.61
M2	✓	✓	✗	68.14	52.81	42.96	36.03	28.15	52.47	103.42
DSRAT	✓	✓	✓	**69.06**	**53.46**	**43.54**	**36.47**	**28.86**	**53.08**	**105.45**

Table 3. Ablation results on the RSICD dataset.

3.4.3 Visual Results.

Figures 3 and 4 show caption generation comparisons between DSRAT and Baseline (with gated fusion, spatial relation-aware multi-head attention, and adaptive layer aggregation disabled) on representative samples from the RSICD and NWPU datasets. On the RSICD dataset (Figure 3), Baseline exhibits notable deficiencies: in the first example, it erroneously describes a “residential area” as “both sides of the road were rows of houses,” losing the scene-type semantics; in the third column, it generates the grammatically incorrect “the ease of the road,” whereas DSRAT accurately identifies a “viaduct” and provides the spatial relation “near”; in the fourth example, DSRAT correctly describes the spatial topological relationship between “terminal” and “airport”; in the fifth example, DSRAT more precisely recognizes “basketball field” rather than the vague “playground.” On the NWPU-Captions dataset (Figure 4), Baseline completely overlooks the river in the third column, while DSRAT correctly generates “the river goes through bare and green farmland”; in the second column, DSRAT accurately counts “two planes” whereas Baseline only recognizes “an airplane”; in the fifth column, DSRAT comprehensively describes the architectural diversity of the industrial area. The visualization results demonstrate that dual-stream feature fusion contributes to improving the accuracy of scene semantic recognition, while spatial relation-aware encoding endows the model with greater logical consistency when generating descriptions involving multi-object positional relationships.

Base: on **both sides** of the **road** were rows of **houses**.
DSRAT: some **buildings** and many **green trees** are in a **medium residential area**.
GT-1: a row of trees is between two lines of cottages.
GT-2: a row of trees is between two lines of cottages.
GT-3: a row of trees is between two lines of cottages.
GT-4: several houses sits on the roadside in this green quiet community.
GT-5: several buildings and many green trees are in a medium residential area.

Base: the **residential buildings** on **both sides** of the **road** are **neatly arranged.**
DSRAT: many **buildings** and **green trees** are in a **dense residential area.**
GT-1: houses with gardens are built compactly in these ploygon blocks.
GT-2: houses with gardens are built compactly in these ploygon blocks.
GT-3: houses with gardens are built compactly in these ploygon blocks.
GT-4: in this residential area most houses have red roof.
GT-5: many red buildings and some green trees are orderly in a dense residential area.

Base: the **ease** of the **road** are **surrounded** by **dark green trees.**
DSRAT: many **green trees** are **near** a **viaduct** with some **cars**.
GT-1: two railways pass through a ring shaped viaduct.
GT-2: two railways pass through a ring shaped viaduct.
GT-3: two railways pass through a ring shaped viaduct.
GT-4: several cars are running on the flyovers and the road on the ground parralleled with railways.
GT-5: many green trees are near a viaduct with several cars.

Base: many of the **planes** were **docked** on the **wide**.
DSRAT: some **planes** are **near** a **terminal** in an **airport**.
GT-1: the two drop shaped terminal buildings are surrounded by planes on the tarmac which is adjacent to a runway with four oval lawns in it.
GT-2: the two drop shaped terminal buildings are surrounded by planes on the tarmac which is adjacent to a runway with four oval lawns in it.
GT-3: some planes stop in the drop shaped terminal and a wide runway.
GT-4: the c shaped termial building sits on the apron which is alongside the runways.
GT-5: some planes are near a building in an airport.

Base: many **buildings** and some **green trees** are **around** a **playground**.
DSRAT: a **basketball field** is **surrounded** by many **buildings**.
GT-1: every villa here are planned in a orderly way with a baksetball court in this block.
GT-2: a basketball field is surrounded by many buildings.
GT-3: a basketball field is surrounded by many buildings.
GT-4: a basketball field is surrounded by many buildings.
GT-5: a basketball field is surrounded by many buildings.

Figure 3. Examples of sentences generated by DSRAT on the RSICD dataset. Red and green words indicate mistakes and advantages compared to the baseline.

Base: there are several **houses among** the **lush woods**.
DSRAT: the **sparse residential** is on the **grass next** to some **trees** and a **lake**.
GT-1: The room is surrounded by trees.
GT-2: In a sparse residential area there is a gray house and several green trees, Surrounded by farmland, And a road.
GT-3: The sparse residential is on the grass next to some trees and a lake.
GT-4: Many trees are beside the sparse residential.
GT-5: A sparse residential area with a black building on the green meadow.

Base: **an airplane** was in the **open space beside** the grassland.
DSRAT: **two planes** are in the **clearing next** to the **grass**.
GT-1: Two adjacent planes parked on the open area and the lawn beside.
GT-2: Two red-tailed planes parked between the lawn and the concrete ground.
GT-3: Two planes of the same size parked neatly next to the buildings inside the airport.
GT-4: Two planes are parked on the open space next to the buildings and the grass.
GT-5: Two planes are in the clearing next to the grass.

Base: the **rectangular farmland** is **next** to some **trees**.
DSRAT: the **river goes through bare** and **green farmland.**
GT-1: The dark green river with a winding bank goes through the farmland.
GT-2: A curved green river with approximately two inner right angles, With green and tan mixed fields and green forests on both sides.
GT-3: The river goes through bare and green farmland.
GT-4: There are many trees on either side of the river.
GT-5: There are many yellow farmlands and some green trees near the river.

Base: there are several **houses among** the **lush woods**.
DSRAT: the **sparse residential** is on the **grass next** to some **trees** and a **road.**
GT-1: Here is a gorgeous home with a trail behind and a row of trees in front.
GT-2: In sparse residential areas there are several houses of different sizes and several green trees surrounded by farmland.
GT-3: The sparse residential is on the grass next to some trees and a road.
GT-4: A building in a sparse residential.
GT-5: A sparse residential area with a black building on the green meadow.

Base: the **railway** is **next** to some houses.
DSRAT: there are many **houses** of **different shapes** and **sizes** in the **industrial area.**
GT-1: The industrial area has some workshops of different colors and sizes and a road goes through the industrial area.
GT-2: Houses in industrial areas vary in size.
GT-3: In an industrial area there are factories and roads, And there is a parking lot for open-air items.
GT-4: There are many houses of different shapes and sizes in the industrial zone.
GT-5: There are some blue buildings and many containers on the industrial area.

Figure 4. Examples of sentences generated by DSRAT on the NWPU-Captions dataset. Red and green words indicate mistakes and advantages compared to the baseline.

4. Conclusion

This paper proposes DSRAT, a Dual-Stream Relation-Aware Transformer model for remote sensing image captioning that systematically addresses the two core challenges of the vision–semantic modality gap and inter-region spatial relation modeling. On the feature extraction end, the Swin-B+FPN multi-scale spatial feature stream and the RemoteCLIP-LoRA remote sensing semantic prior stream are adaptively complemented via per-token gated fusion, effectively bridging the gap between generic visual representations and the remote sensing semantic domain. On the encoder end, Relation-Aware Multi-Head Attention explicitly injects geometric relationships including relative position, distance, and orientation between regions into the attention computation, enhancing the model’s comprehension of multi-object spatial topological structures in complex scenes; adaptive multi-layer aggregation further fuses structural and semantic information across different encoding depths, providing richer visual memory for decoder generation. Experiments on the RSICD and NWPU-Captions benchmarks demonstrate that DSRAT achieves state-of-the-art results across six metrics on RSICD and all seven metrics including CIDEr on NWPU-Captions, and ablation studies further confirm the effectiveness and synergistic gains of the core modules.

Acknowledgements

This work was supported by the Guangxi Science and Technology Major Project (AA22068072).

References

Cheng, Q., Huang, H., Xu, Y., Zeng, P., 2022. NWPU-Captions dataset and MLCA-Net for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing 60, 1–19.

Denkowski, M., Lavie, A., 2014. Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT 2014). Association for Computational Linguistics, Baltimore, MD, USA, pp. 376–380.

Du, R., Cao, W., Zhang, W., Peng, J., 2023. From plane to hierarchy: Deformable transformer for remote sensing image captioning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16, 7704–7717.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models, in: Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022). OpenReview.net.

Li, X., Yuan, A., Lu, X., 2018. Multi-modal gated recurrent units for image description. Multimedia Tools and Applications 77, 29847–29869.

Lin, C.Y., 2004. ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. Association for Computational Linguistics, Barcelona, Spain, pp. 74–81.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). IEEE, Honolulu, HI, USA, pp. 2117–2125.

Liu, C., Zhao, R., Shi, Z., 2022. Remote-sensing image captioning based on multilayer aggregated transformer. IEEE Geoscience and Remote Sensing Letters 19, 1–5.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Lu, J., 2024. RemoteCLIP: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing 62, 1–16.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021). IEEE, Montreal, QC, Canada, pp. 10012–10022.

Lu, X., Wang, B., Zheng, X., Li, X., 2017. Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing 56, 2183–2195.

Lu, X., Wang, B., Zheng, X., 2020. Sound active attention framework for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing 58, 1985–2000.

Meng, L., Wang, J., Meng, R., Yang, Y., Xiao, L., 2024. A multiscale grouping transformer with CLIP latents for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing 62, 1–15.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). Association for Computational Linguistics, Philadelphia, PA, USA, pp. 311–318.

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

Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., 2017. Self-critical sequence training for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). IEEE, Honolulu, HI, USA, pp. 7008–7024.

Shaw, P., Uszkoreit, J., Vaswani, A., 2018. Self-attention with relative position representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018), Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, LA, USA, pp. 464–468.

Silva, J.D., Magalhães, J., Tuia, D., Martins, B., 2024. Large language models for captioning and retrieving remote sensing images. arXiv:2402.06475 [cs.CV].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Curran Associates, Inc., Long Beach, CA, USA, pp. 5998–6008.

Vedantam, R., Lawrence Zitnick, C., Parikh, D., 2015. CIDEr: Consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). IEEE, Boston, MA, USA, pp. 4566–4575.

Wang, B., Lu, X., Zheng, X., Li, X., 2019. Semantic descriptions of high-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters 16, 1274–1278.

Wang, Q., Huang, W., Zhang, X., Li, X., 2022. GLCM: Global–local captioning model for remote sensing image captioning. IEEE Transactions on Cybernetics 53, 6910–6922.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). PMLR, Lille, France, pp. 2048–2057.

Yang, C., Li, Z., Zhang, L., 2024a. Bootstrapping interactive image–text alignment for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing 62, 1–12.

Yang, Z., Li, Q., Yuan, Y., Wan, M., Gao, H., 2024b. HCNet: Hierarchical feature aggregation and cross-modal feature alignment for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing 62, 1–11.

Zhang, X., Wang, X., Tang, X., Zhou, H., Li, C., 2019. Description generation for remote sensing images using attribute attention mechanism. Remote Sensing 11, 612.