This ACCV 2024 workshop paper, provided here by the Computer Vision Foundation, is the author-created version. The content of this paper is identical to the content of the officially published ACCV 2024 LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

RSSep: Sequence-to-Sequence Model for Simultaneous Referring Remote Sensing Segmentation and Detection

Ngoc-Vuong Ho*[1], Thinh Phan[1], Meredith Adkins[1], Chase Rainwater[1], Jackson Cothren[1], and Ngan Le[1]

[1]University of Arkansas, Fayetteville, AR, USA
*Corresponding author: vuongh@uark.edu

Abstract. Semantic segmentation in remote sensing images plays a crucial role in a wide range of geographic information applications. Despite the abundance of data, this field faces limitations due to the restricted set of categories and the inability of existing methods to accurately describe and localize individual or multiple objects within scenes. Addressing this challenge, the emerging fields of referring remote sensing image segmentation (RRSIS) and referring remote sensing object detection (RRSOD) have recently garnered attention. Both tasks, RRSIS and RRSOD, combine computer vision and natural language processing to localize objects based on a text query, with the outputs being segmentation masks and bounding boxes. Additionally, boundary information in remote sensing images, such as land-cover delineations, is crucial for segmentation tasks. To tackle this novel challenge, we introduce **RSSep**, a Sequence-to-Sequence model designed for simultaneous RRSIS and RRSOD. Unlike conventional approaches that use encoder-decoder blocks for pixel-level classification, our network leverages a sequence-to-sequence model to estimate polygonal boundaries, represented as sequences of vertices. Furthermore, we enhanced the network by improving the text encoder using both query and object noun features, employing the same architecture to extract these features. Our network is benchmarked on the recently introduced RRSIS-D dataset, notable for its extensive collection of image-caption-mask triplets across diverse scales and variations. Experimental results demonstrate the superiority of our method over existing techniques in both the RRSIS and RRSOD fields, underscoring its efficacy in semantic segmentation and object detection tasks in remote sensing imagery.

1 Introduction

Owing to the advancement and pervasiveness of satellites and aerial vehicles, remote sensing data collection has surged, leading to an increasing need for analyzing and understanding scene images. Remote sensing image segmentation (RSIS) has become one of the key tasks, applied in multiple fields such as urban planning [36], land resource management [22], environmental monitoring [23],

disaster monitoring [41], agricultural planning [48], street view extraction [14,42], land change detection [34, 40, 50], land cover classification [44], climate change studies [38], and deforested region monitoring [2], among others. Although current approaches in RSIS are proficient at identifying objects in a scene, they struggle to specify areas based on descriptions or to perceive the spatial and orientational features of objects. To address these limitations, Referring Remote Sensing Image Segmentation (RRSIS) has emerged. The goal of RRSIS is to take both an aerial image and a natural language query describing appearance, position, and direction, and return a pixel mask of the relevant objects or areas in the scene.

Unlike Referring Remote Sensing Image Segmentation (RRSIS), referring image segmentation (RIS) has been a well-known task for some time, with the majority of RIS approaches [18, 21, 24, 53] following an encoder-decoder framework where the encoder is responsible for extracting visual and language features using two separate networks and then aligning them through recurrent interaction [27], cross-modal attention [47], graph reasoning [20], or cross-attention from a Transformer [5], and subsequently, the decoder unravels the combined features and performs pixel-level classification, outputting the segmentation mask for the desired objects. Despite their success on multiple datasets, general RIS methods are sub-optimal when fully applied to specific data types like remote sensing images due to the aerial viewpoint reducing the noticeable discrepancies in color and appearance between objects and backgrounds, while the scales and sizes of objects vary greatly depending on the distance between the camera and the ground, resulting in weak contrast between object boundaries and backgrounds in low spatial resolution images, which often causes the predicted masks to appear smeared. To address these challenges, we adopt a sequence-to-sequence (Seq2Seq) framework to indirectly infer the segmentation mask where the input remains the same as in the encoder-decoder framework but the decoder performs a regression task and outputs the polygonal boundary instead, making this method better at recognizing object geometry and leading to more precise masks where the segmentation mask is converted into a sequence of polygon vertices (with unrestricted length) and a bounding box is also output by the Seq2Seq module, transforming each vertex into a coordinate embedding token, and in cases where multiple objects are queried, different sequences can be merged into a longer sequence and distinguished by separator tokens while the model learns to predict the next coordinate token based on the visual features, text features, and previous tokens, with the polygon sequence being extended iteratively until the end-of-sequence token is predicted, which allows our approach to be less susceptible to inconspicuous boundaries, scale variations, and omni-directional objects.

Our main contributions are summarized as follows:

– We present an effective RSSeq network for simultaneous RRSIS and RRSOD. Our RSSeq is built upon the Seq2Seq framework, modeling the object boundary as a sequence of vertices. Our network is designed not only to focus on the object boundary but also to handle an arbitrary number of objects.

– To effectively train the model to handle both boundary and segmentation mask tasks, our RSSeq model is trained using a combination of weighted L_1 and L_2 regression losses, cross-entropy (CE) loss, and $Dice$ loss. This comprehensive loss function aims to optimize the prediction of polygon vertices, vertex types, and segmentation masks.
– We benchmarked the proposed RSSeq on the newly introduced RRSIS-D dataset, demonstrating superior performance over all existing state-of-the-art methods in the RRSIS task.

2 Related work

2.1 Mask & Polygon-based Image Segmentation

Mask-based image segmentation has still been on the growth and the primary technique for object segmentation. Fully Convolutional Neural (FCN) [33] established the baseline for semantic segmentation field by repalce all fully-connected layers with convolution layers in classification network. For the purposed of accumulating multi-scale contextual information, DeepLab series [7] upgraded FCN with dilated convolutions. With the same intention, PSPNet [55] introduced the pyramid pooling operations. Latest work such as Mask2Former [9] utilized the end-to-end Transformer [3] encoder-decoder network and multi-scale high resolution features, deducing the each object mask from corresponding embedding query. Treating segmentation mask as set of polygon vertices is also considered because this task simulates how human annotates the mask. The boundary is refined or sequentially predicted until we reach the initial point. The early work [4] made use of the Recurrent Neural Network (RNN) and was extended by [1] with the application of graph neural network. Ling et al. [26] initiated with a circle and tried to deform it into the boundary. Done et al. [13] extended this task to spline curve prediction and did the multitasking training on edge detection and object segmention. PolyTransform [25] predicted the mask first and forwarded it as polygon type to deforming network for final polygon prediction.

2.2 Remote Sensing Image Segmentation (RSIS)

RSIS aims to segment and classify the objects such as building, vehicle, road or field on the earth surface from the aerial viewpoint. In the early period of deep learning application on this topic, FCN was the standard approach on many datasets [8]. The methods improved along with the development of deep learning segmentation model. ResUNet-a [12] combined U-Net with other CNN to eliminate the the problem of gradient disappearance and explosion. S-RA-FCN [35] enhanced the global contextual information by adding the spatial and channel relational reasoning modules. HMANet [37] proposing three attention modules to better obtain correlation features in space, channel and category. The efficiency of self-attention in transformer-based network set a new model trend in this field. Due to the low contrast between the foreground saliency and background noise,

RSSFormer [51] was designed with the Adaptive Transformer Fusion Module and Detail-aware Attention Layer. [45] introduced the a densely connected feature aggregation module for precise segmentation. While transformer-based methods are good at capturing long-range dependencies, intricate and tiny objects are still an obstacle for them.

2.3 Referring Image Segmentation (RIS)

RIS has been one of most active topics in the field of visual-language understanding and interaction. The main objective has been the fusion mechanism of visual and language features. Straightforward feature concatenation [43] was first implemented as the fusion operation. Chen et al. [6] followed this technique but applied recurrent refinement to polish the feature maps at different scales. Later works employed several types of attention mechanism [53] to model the visual-textual co-embeddings. CMPC [20] used graph-based reasoning to localize the image region that were highly related to the linguistic features of entity words and attribute words. BRINet [19] computed the relevance among each word and each image area in a bi-directional relationship modeling through vision and language-guided attention modules. As a result of the success of vision-language model such as CLIP [39], recent models [46] tried to transfer this rich knowledge to their fusion model. LAVT [52] replaced the complicated cross-modal decoder by early language-aware encoding module. PolyFormer [28] proposed the regression-based Tranformer decoder which directly output 2D coordinates from concatenated image feature and textual feature.

3 Methodology

3.1 Overall Methodology

Network Architecture: At the core of our approach is the idea of feature fusion between natural language processing and computer vision. Figure 1.b illustrates the network architecture: the inputs are the image and a text query, and the outputs are the polygon covering to the mask needed for segmentation and the bounding box for object detection as per the text description. Motivated by recent advancements in multimodal architectures such as CLIP [39], we use two separate encoder branches to extract visual and textual features from both the image and text prompt. The image encoder is based on Swin transformer [31], whereas the text encoder is based on BERT [10]. Both the visual and textual features are then concatenated by a fusion module before passing through a Seq2Seq network to obtain a sequence of vertexes of a polygon. To effectively handle floating-point coordinates, vertexes are passed through a regression network. Finally, the polygon will be converted to a segmentation mask.

Encoder: We use Swin Transformer $f_v(.|\theta_v)$, defined by weights θ_v, to extract visual feature F_v from a given image $I \in \mathbb{R}^{H\times W\times 3}$, i.e. $F_v = f_v(I|\theta_v)$. We select

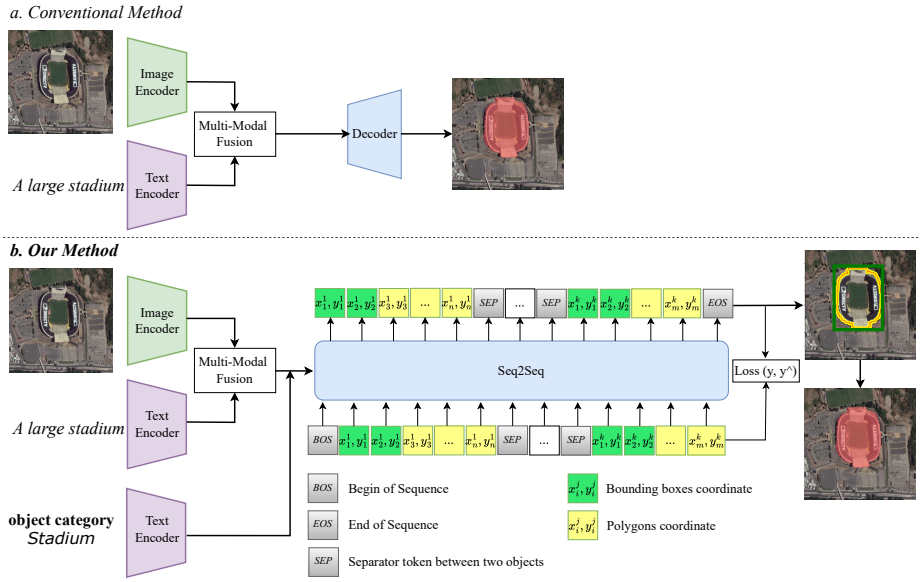

Fig. 1: Comparison between existing RRSIS approaches (top) and our proposed RSSeq (bottom). While conventional RRSIS methods directly generate a segmentation mask from a Decoder network, our RSSeq first produces a sequence of vertices and subsequently converts them into a polygonal segmentation mask.

the feature at stage-4, thus the feature $F_v \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_v}$. For the textual description extraction, we use BERT [10] $f_t(.|\theta_t)$, defined by weights θ_t, to extract textual feature from a given text prompt T, i.e. $F_t = f_t(T|\theta_t)$ and $F_t \in \mathbb{R}^{N \times C_n}$, where N is the number of words.

Fusion module: The visual feature F_v and textual feature F_t into new feature F_{vt}. To achieve this, the visual feature $F_v \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_v}$ is first flatten into 2D space $\hat{F}_v \in \mathbb{R}^{(\frac{H}{32} * \frac{W}{32}) \times C_v}$. Subsequently, both $\hat{F}_v$ and F_t are passed through two separate fully-connected (FC) layers to project them into the same embedding space. These projected features are then concatenated into F_{vl}, i.e., $F_{vl} = \Big[FC(\hat{F}_v), FC(F_t)\Big]$. Finally, F_{vl} is fed into a Transformer encoder consisting of multi-head self attention layer to object a visual-textual feature F.

Seq2Seq: The Seq2Seq takes the 2 inputs: (i) visual-textual feature F from the fusion module and (ii) input token, which represents the polygon vertexes. The input token of the module are define as format:

$$\begin{aligned}&[\texttt{<BOS>}, (x_1^1, y_1^1), (x_2^1, y_2^1), (x_3^1, y_3^1), ..., (x_n^1, y_n^1), \texttt{<SEP>}\\ &(x_1^2, y_1^2), ..., (x_m^1, y_m^1)\texttt{<SEP>}, (x_k^1, y_k^1), ..., \texttt{<EOS>}],\end{aligned} \tag{1}$$

Where (x_1^i, y_1^i) and (x_2^i, y_2^i) denote the bounding box coordinates, and $(x_3^i, y_3^i), ..., (x_n^i, y_n^i)$ represent the coordinates of the bounding polygons. The tokens <BOS> and <EOS> indicate the beginning and end of the tokens, while <SEP>

denotes the token used for separation between objects. With this definition, our model is capable of flexibly supporting multiple objects.

The prediction of vertex (x_t, y_t) at time step t depends on all the preceding tokens $(x_1, y_1), (x_2, y_2), .., (x_{t-1}, y_{t-1})$.

The model accommodates an arbitrary number of input tokens to handle an arbitrary number of objects. During training, we process based on the last token, while during inference, the procedure is based on the `EOS` token.

Decoder: To predict continuous coordinate values for vertex coordinates, we apply bilinear interpolation [15] into the decoder. Let $f(x, y)$ denote the coordinate embedding of (x, y). To capture the relations between the visual-textual feature F and the coordinate embedding $f(x, y)$, we apply N transformer decoder layers with multi-head cross-attention mechanism. Consequently, we obtain F^N as the output feature of the last decoder layer.

Prediction Heads: There are two output heads, i.e., token-based head and coordinate-based head. Both are built on the last feature F^N. The token-based head consists of a linear layer, which predicts the token type. Token types can be either coordinate token $((x_1^1, y_1^1), (x_2^1, y_2^1), (x_3^1, y_3^1), ..., (x_n^1, y_n^1))$, which is labeled as 0 or separate token `<SEP>`, which is labeled as 1, or ending token `<EOS>`, which is labeled as 1. The coordinate-based head is defined as a 3-layer feed-forward network (FFN). It aims to predict the 2D coordinates of the bounding box corner points $((x_1^i, y_1^i), (x_2^i, y_2^i)$ and polygon vertices $((x_3^i, y_3^i), ..., (x_n^i, y_n^i)$ for the reference object i.

Mask to Polygons Converter: The current available dataset provides masks only. Therefore, to train the model, we need to convert these masks into polygons. Given a mask $M \in \mathbb{R}^{H \times W}$, we obtain the contour, a set of points $(x, y) \in \mathbb{R}^2$, which are then converted into polygons. Due to the large number of points on the contour, we sample a subset of points, typically ranging from 100 to 200 points for each object. We select the top-left point as the starting point of the sequence $\{(x_i, y_i)\}_{i=1}^{P}$, $(x_i, y_i) \in \mathbb{R}^2$ in the clock-wise order, $P \in [100, 200]$. Finally, we construct the input tokens for the polygon coordinates as
`<BOS>`$(x_1^1, y_1^1), (x_2^1, y_2^1)...(x_n^1, y_n^1)$`<SEP>`$...$`<SEP>`$(x_2^k, y_2^k), (x_2^k, y_2^k)...(x_m^k, y_m^k)$`<EOS>`,
where k is the number of object, n and m are the number of token for each object. By using `<SEP>`, the number of objects and the number of tokens for each object are flexible.

Polygon to Mask Converter: The output of the network contains two components: a token-based head and a coordinate-based head. To convert from the polygon to a mask, we utilize both the token-type and coordinate outputs. By combining the outputs of both heads, we can obtain:
`<BOS>`$(x_1^1, y_1^1), (x_2^1, y_2^1)...(x_n^1, y_n^1)$`<SEP>`$...$`<SEP>`$(x_2^k, y_2^k), (x_2^k, y_2^k)...(x_m^k, y_m^k)$`<EOS>`

Then, we can separate the coordinate output into multiple bounding boxes and boundary vertices based on `<SEP>`. Finally, the boundary vertices are converted to a binary mask to be compared with the ground truth.

3.2 Loss function

Given an image I, a text prompt (referring description) T, and preceding tokens x_i, y_i, the model is trained to predict the next token x_t, y_t, its corresponding token type l, and the corresponding segmentation S. The prediction of the next token is guided by a combination of weighted $L1$ and $L2$ regression losses. The token type is is determined by the cross-entropy (CE) loss. The segmentation mask is based on $Dice$ loss.

$$\begin{aligned}\mathcal{L} =& L1\left((x_t, y_t), (\hat{x}_t, \hat{y}_t)\right) + L2((x_t, y_t), (\hat{x}_t, \hat{y}_t)) \\ &+ CE(l, \hat{l}) + Dice(S, \hat{S})\end{aligned} \tag{2}$$

The weighted $L1$ and $L2$ loss employs different weights to balance the importance of box coordinates and polygon coordinates, and is defined as follows.

$$\begin{aligned}L1\left((x_t, y_t), (\hat{x}_t, \hat{y}_t)\right) = & 0.1 \times L1\left((x_t, y_t), (\hat{x}_t, \hat{y}_t)\right)_{t=1,2} \\ &+ 0.9 \times L1\left((x_t, y_t), (\hat{x}_t, \hat{y}_t)\right)_{t>2}\end{aligned} \tag{3}$$

Similar weights are employed in the weighted $L2$ loss.

4 Experiments

4.1 Datasets and Metrics

RRSIS-D Dataset: In this work, we utilize the RRSIS-D dataset, which was made public at CVPR 2024, for our experiments. This dataset contains 17,402 image-caption-mask triplets, split into train/validation/test sets with 12,181/1,740/3,481 samples, respectively. The remote sensing images have various spatial resolutions ranging from 0.3 to 30.0 meters/pixel, with each image having a size of 800x800 pixels. The dataset comprises 20 categories: `ariplane`, `airport`, `basketball court`, `bridge`, `baseball field`, `chimney`, `dam`, `expressway service area`, `expressway toll station`, `golf field`, `ground track field`, `harbor`, `overpass`, `ship`, `stadium`, `storage tank`, `tennis court`, `train station`, `vehicle`, `wind mill`.

Metrics: In the experiments, we use overall Intersection-over-Union (oIoU), which is the overall ratio of intersection to union areas between predicted and ground truth masks, while mean Intersection-over-Union (mIoU) calculates the average accuracy for all the predicted and ground truth masks in pairs. Additionally, we use Precision@X (P@X) as an evaluation metric to evaluate precision based on IoU thresholds, reflecting the method's accuracy in object targeting.

4.2 Implementation Details

We trained the model on both the base and large versions of the Swin Transformer [31], with the language backbone based on BERT from Hugging Face's library [49]. The model was trained on an NVIDIA RTX 4090 TI for 100 epochs over 1 day. We initialized the learning rate to 0.00003 using Adam optimization. For better performance, we first trained RSSeq with RRSOD to obtain the initial checkpoint, making it aware of the global features. Then, we reloaded the checkpoint for the RRSOD task to train the multi-task RRSOD and RRSIS. In our experiments, we found that initializing the model with RRSOD resulted in better performance compared to training the multi-task setup from the start.

4.3 Comparison Results

Table 1 presents a quantitative comparison between our RSSeq model and existing state-of-the-art methods in referring segmentation. It is noteworthy that RMSIN [30] represents the latest advancement in RRSIS and introduces the RRSIS-D dataset. We showcase the performance of our RSSeq model with two backbone architectures: Swin-B and Swin-L. Both RSSeq-B and RSSeq-L outperform existing state-of-the-art methods, with RSSeq-L achieving superior performance across all metrics except for P@0.9. Further investigation into this metric will be included in future analyses.

Table 1: Quantitative comparison with state-of-the-art methods on **validation set of RRSIS-D dataset** [30]. Our proposed The best result is bold.

Methods	Venues	Visual Encoder	Language Encoder	Performance						
				P@0.5↑	P@0.6↑	P@0.7↑	P@0.8↑	P@0.9↑	oIoU↑	mIoU↑
RRN [24]	CVPR 2018	ResNet-101 [16]	LSTM [17]	51.09	42.47	33.04	20.80	6.14	66.53	46.06
CSMA [53]	CVPR 2019	ResNet-101	-	55.68	48.04	38.27	26.55	9.02	69.68	48.85
LSCM [21]	ECCV 2020	ResNet-101	LSTM	57.12	48.04	37.87	26.37	7.93	69.28	50.36
CMPC [20]	CVPR 2020	ResNet-101	LSTM	57.93	48.85	38.50	25.28	9.31	70.15	50.41
BRINet [18]	CVPR 2020	ResNet-101	LSTM	58.79	49.54	39.65	28.21	9.19	70.73	51.14
CMPC+ [29]	TPAMI	ResNet-101	LSTM	59.19	49.36	38.67	25.91	8.16	70.14	51.41
LGCE [54]	-	Swin-B [32]	BERT [11]	68.10	60.52	52.24	42.24	23.85	76.68	60.16
LAVT [52]	CVPR 2022	Swin-B	BERT	69.54	63.51	53.16	43.97	24.25	77.59	61.46
RMSIN [30]	CVPR 2024	Swin-B	BERT	74.66	68.22	57.41	45.29	**24.43**	78.27	65.10
RSSeq - B (Ours)	-	Swin-B	BERT	79.13	71.03	61.56	46.87	17.85	81.08	67.33
RSSeq - L (Ours)	-	Swin-L	BERT	**80.25**	**73.29**	**62.43**	**48.91**	17.87	**82.10**	**69.23**

Figure 2 visually illustrates a qualitative comparison between our RSSeq model and the runner-up, RMSIN [30]. While RMSIN only provides the segmentation mask, our RSSeq offers both the boundary and segmentation mask, along with the corresponding bounding boxes. By leveraging polygons to focus on boundaries, our model adeptly localizes objects, particularly along their edges.

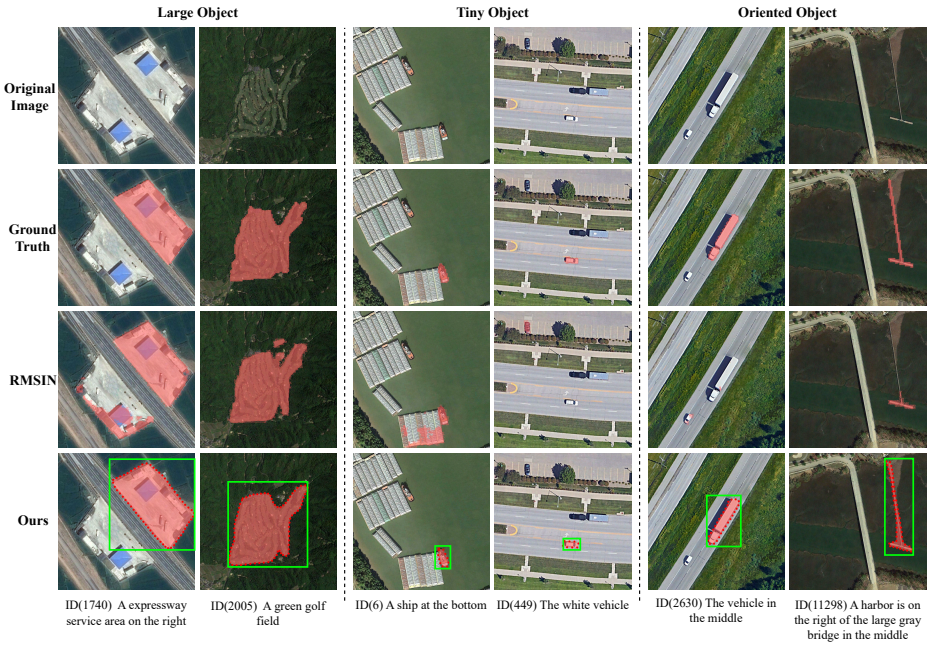

Fig. 2: Visualization comparison on the RRSIS-D Dataset [30]. From the top: 1^{st}: Original images with various types of objects of interest; 2^{nd}: Ground truth; 3^{rd}: Segmentation mask by RMSIN [30]; 4^{th}: Our proposed RSSeq, which simultaneously generates both segmentation and detection.

4.4 Ablation Studies

We further investigate the effectiveness of our proposed RSSeq by pre-training the model on the RMSIN-D dataset to obtain initial weights. Table 2 shows a comparison between two scenarios: one with and one without initializing weights on the RMSIN-D dataset across two different backbones.

Table 2: Ablation study of our proposed RSSeq on the **validation set of the RRSIS-D dataset** [30], comparing the performance with and without the pre-training procedure for initializing weights.

Language-Encoder	Visual-Encoder	Pre-train	Performance						
			P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	oIoU	mIoU
BERT	Swin-B	✗	75.31	67.09	57.60	43.16	17.50	79.26	64.26
		✓	**79.13**	**71.03**	**61.56**	**46.87**	**17.85**	**81.08**	**67.33**
	Swin-L	✗	79.63	71.21	59.17	44.71	21.49	78.20	68.18
		✓	**80.25**	**73.29**	**62.43**	**48.91**	**17.87**	**82.10**	**69.23**

To better illustrate how our RSSeq predicts polygon vertices at inference time, we visualize some vertices around objects, as shown in Figure 3, at different time steps.

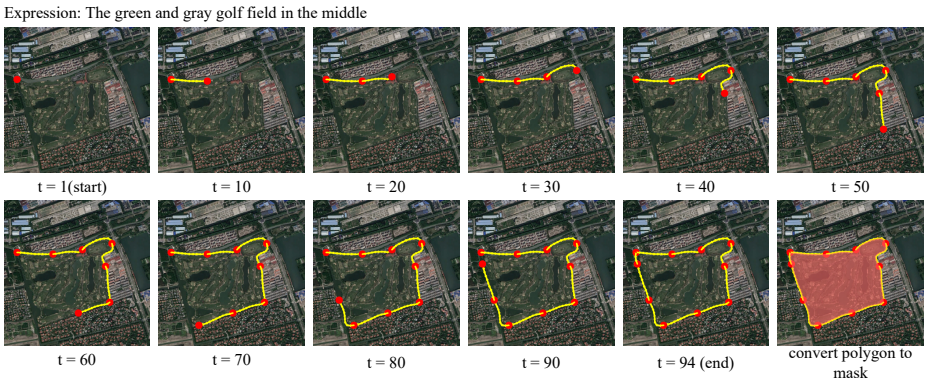

Fig. 3: Visualization of the steps during inference. The model begins with the `BOS` token. At step t = 1, it predicts the top-left vertex of the object. The model continues predicting until it reaches the `EOS` token. Finally, once the full polygon covers the object, the last step converts the polygon to a mask.

5 Conclusion

In this study, we introduce RSSeq, an end-to-end framework designed for referring remote sensing image segmentation (RRSIS) and object detection (RRSOD). RSSeq employs a multimodal approach within a sequence-to-sequence framework for multitask learning. By leveraging this architecture, RSSeq effectively segments object boundaries as sequences of vertices while supporting image segmentation for multiple objects. Additionally, RSSeq outperforms all existing state-of-the-art methods in RRSIS. We anticipate that this method can be extended to other remote sensing tasks, such as multi-label image segmentation and crop-type classification.

In future work, we plan to explore sequence-to-sequence methods for multi-label remote sensing image segmentation.

References

1. Acuna, D., Ling, H., Kar, A., Fidler, S.: Efficient interactive annotation of segmentation datasets with polygon-rnn++. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 859–868 (2018) 3
2. Andrade, R., Costa, G., Mota, G., Ortega, M., Feitosa, R., Soto, P., Heipke, C.: Evaluation of semantic segmentation methods for deforestation detection in the amazon. ISPRS Archives; 43, B3 **43**(B3), 1497–1505 (2020) 2
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 3
4. Castrejon, L., Kundu, K., Urtasun, R., Fidler, S.: Annotating object instances with a polygon-rnn. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5230–5238 (2017) 3
5. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 357–366 (2021) 2
6. Chen, D.J., Jia, S., Lo, Y.C., Chen, H.T., Liu, T.L.: See-through-text grouping for referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7454–7463 (2019) 4
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017) 3
8. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018) 3
9. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022) 3
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 4, 5
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019) 8
12. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing **162**, 94–114 (2020) 3
13. Dong, Z., Zhang, R., Shao, X.: Automatic annotation and segmentation of object instances with deep active curve network. IEEE Access **7**, 147501–147512 (2019) 3
14. Griffiths, D., Boehm, J.: Improving public data for building segmentation from convolutional neural networks (cnns) for fused airborne lidar and image data using active contours. ISPRS Journal of Photogrammetry and Remote Sensing **154**, 70–83 (2019) 2
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 6
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 8

17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997) 8
18. Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H.: Bi-directional relationship inferring network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 2, 8
19. Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H.: Bi-directional relationship inferring network for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4424–4433 (2020) 4
20. Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., Liu, L., Li, B.: Referring image segmentation via cross-modal progressive comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 4, 8
21. Hui, T., Liu, S., Huang, S., Li, G., Yu, S., Zhang, F., Han, J.: Linguistic structure guided context modeling for referring image segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 59–75 (2020) 2, 8
22. Kumar, S., Meena, R.S., Sheoran, S., Jangir, C.K., Jhariya, M.K., Banerjee, A., Raj, A.: Remote sensing for agriculture and resource management. In: Natural Resources Conservation and Advances for Sustainability, pp. 91–135. Elsevier (2022) 1
23. Li, J., Pei, Y., Zhao, S., Xiao, R., Sang, X., Zhang, C.: A review of remote sensing for environmental monitoring in china. Remote Sensing **12**(7), 1130 (2020) 1
24. Li, R., Li, K., Kuo, Y.C., Shu, M., Qi, X., Shen, X., Jia, J.: Referring image segmentation via recurrent refinement networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 2, 8
25. Liang, J., Homayounfar, N., Ma, W.C., Xiong, Y., Hu, R., Urtasun, R.: Polytransform: Deep polygon transformer for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9131–9140 (2020) 3
26. Ling, H., Gao, J., Kar, A., Chen, W., Fidler, S.: Fast interactive object annotation with curve-gcn. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5257–5266 (2019) 3
27. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1271–1280 (2017) 2
28. Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R.K., Mahadevan, V., Manmatha, R.: Polyformer: Referring image segmentation as sequential polygon generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18653–18663 (2023) 4
29. Liu, S., Hui, T., Huang, S., Wei, Y., Li, B., Li, G.: Cross-modal progressive comprehension for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 4761–4775 (2021) 8
30. Liu, S., Ma, Y., Zhang, X., Wang, H., Ji, J., Sun, X., Ji, R.: Rotated multi-scale interaction network for referring remote sensing image segmentation. arXiv preprint arXiv:2312.12470 (2023) 8, 9
31. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (October 2021) 4, 8

32. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 8
33. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015) 3
34. Marcos, D., Volpi, M., Kellenberger, B., Tuia, D.: Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. ISPRS journal of photogrammetry and remote sensing **145**, 96–107 (2018) 2
35. Mou, L., Hua, Y., Zhu, X.X.: Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. IEEE Transactions on Geoscience and Remote Sensing **58**(11), 7557–7569 (2020) 3
36. Netzband, M., Stefanov, W.L., Redman, C.: Applied remote sensing for urban planning, governance and sustainability. Springer Science & Business Media (2007) 1
37. Niu, R., Sun, X., Tian, Y., Diao, W., Chen, K., Fu, K.: Hybrid multiple attention network for semantic segmentation in aerial images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–18 (2021) 3
38. O'neill, S.J., Boykoff, M., Niemeyer, S., Day, S.A.: On the use of imagery for climate change engagement. Global environmental change **23**(2), 413–421 (2013) 2
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4
40. Samie, A., Abbas, A., Azeem, M.M., Hamid, S., Iqbal, M.A., Hasan, S.S., Deng, X.: Examining the impacts of future land use/land cover changes on climate in punjab province, pakistan: implications for environmental sustainability and economic growth. Environmental Science and Pollution Research **27**, 25415–25433 (2020) 2
41. Schumann, G.J., Brakenridge, G.R., Kettner, A.J., Kashif, R., Niebuhr, E.: Assisting flood disaster response with earth observation data and products: A critical assessment. Remote Sensing **10**(8), 1230 (2018) 2
42. Shamsolmoali, P., Zareapoor, M., Zhou, H., Wang, R., Yang, J.: Road segmentation for remote sensing images using adversarial spatial pyramid networks. IEEE Transactions on Geoscience and Remote Sensing **59**(6), 4673–4688 (2020) 2
43. Shi, H., Li, H., Meng, F., Wu, Q.: Key-word-aware network for referring expression image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 38–54 (2018) 4
44. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1. Curran Associates, Inc. (2021), https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4e732ced3463d06de0ca9a15b6153677-Paper-round2.pdf 2
45. Wang, L., Li, R., Duan, C., Zhang, C., Meng, X., Fang, S.: A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022) 4
46. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022) 4

47. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 2
48. Weiss, M., Jacob, F., Duveiller, G.: Remote sensing for agricultural applications: A meta-review. Remote sensing of environment **236**, 111402 (2020) 2
49. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Transformers: State-of-the-art natural language processing. In: Conference on Empirical Methods in Natural Language Processing (2019) 8
50. Xia, J., Yokoya, N., Adriano, B., Broni-Bediako, C.: Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6254–6264 (2023) 2
51. Xu, R., Wang, C., Zhang, J., Xu, S., Meng, W., Zhang, X.: Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation. IEEE Transactions on Image Processing **32**, 1052–1064 (2023) 4
52. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022) 4, 8
53. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10502–10511 (2019) 2, 4, 8
54. Yuan, Z., Mou, L., Hua, Y., Zhu, X.X.: Rrsis: Referring remote sensing image segmentation. arXiv preprint arXiv:2306.08625 (2023) 8
55. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017) 3