# Visual Grounding in Remote Sensing Images

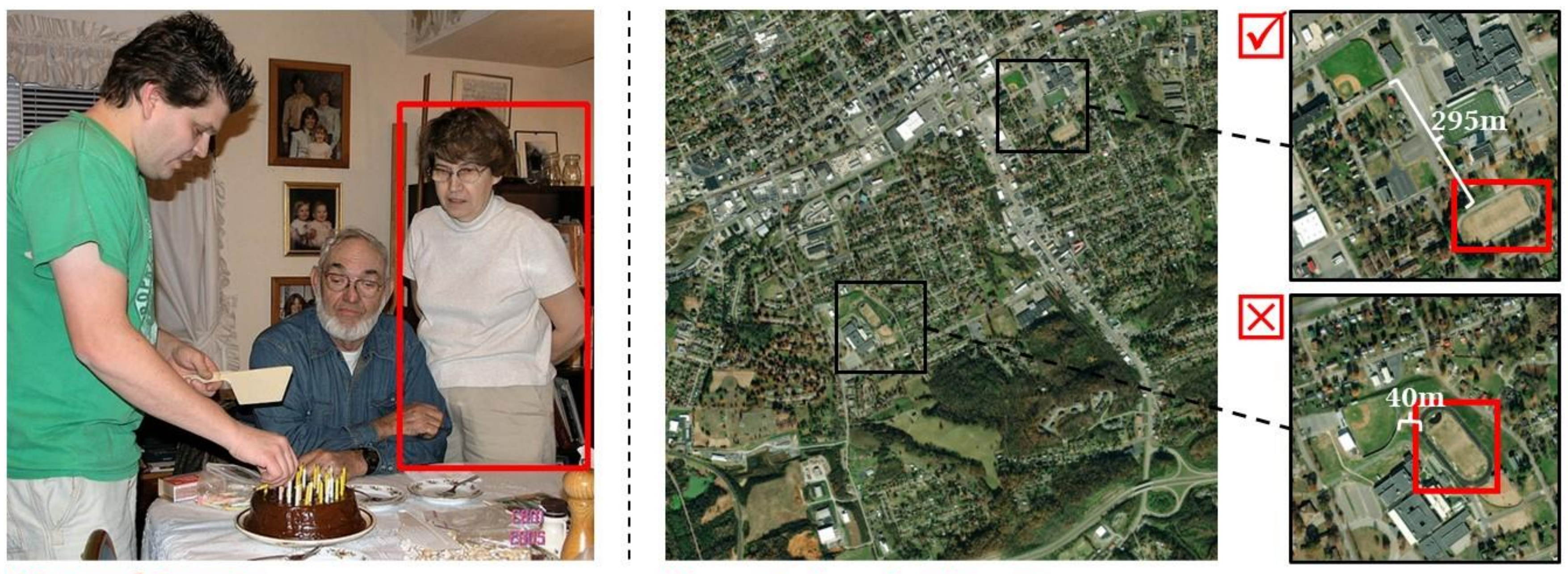
Yuxi Sun[1], Yunming Ye[1], Yifang Ban[2], Xutao Li[1]

1 Harbin Institute of Technology, Shenzhen, China
2 KTH Royal Institute of Technology, Stockholm, Sweden

## Abstract

We present a novel problem of visual grounding in remote sensing images. Compared with natural images, remote sensing images contain large-scale scenes and the geographical spatial information of ground objects (e.g., longitude, latitude). The existing method cannot deal with these challenges. In this paper, we collect a new visual grounding dataset, called RSVG, and design a new method, namely GeoVG.

## Introduction



**Natural Image**
**Expression:** lady right

**Remote Sensing Image**
**Expression:** Find a ground track field, located approximately 295 meters southeast of a baseball field. The baseball field is next to a road.

- Remote sensing images (right) are geospatial images containing longitude and latitude, but natural images (left) are not. Correspondingly, the language expressions for remote sensing objects contain complex numerical geospatial relations such as distance.
- Remote sensing images are large-scale scenes and target objects are inconspicuous. However, natural images are small-scale scenes and target objects are generally salient.

To address the two drawbacks, we design a novel visual grounding model to find a ground object from a large-scale remote sensing image.

## Objective

The task aims to locate the particular objects in a remote sensing image by a natural language expression.

### (1) Image Collection

RSVG mainly contains satellite and aerial images, which are collected from multiple sensors with multiple resolutions.

### (2) Expression Generation

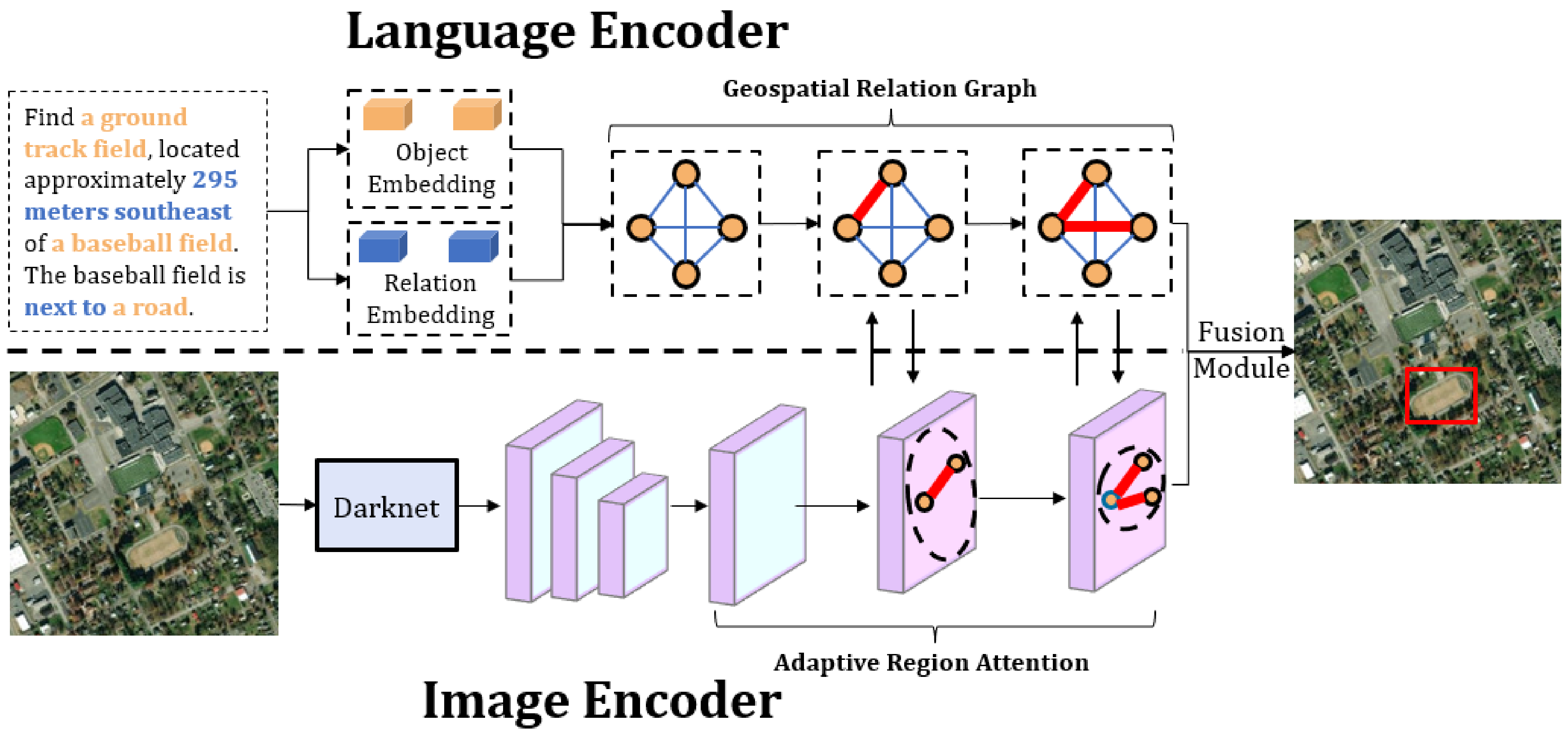
Expressions consist of ground objects and their spatial relations.

- Ground objects include baseball field, basketball court, tennis court, football field, etc.
- Spatial relations contain topological relations, directional relations, distance relations, etc.

| Exemplar templates | Expression examples |
|---|---|
| The <attr0> <obj0> <rel0>. | The largest baseball field. |
| The <att0> <obj0> <rel0> a/an <att1> <refobj1>. | The swimming pool located approximately 158 meters southeast of a tennis court. |
| The <att0> <obj0> <rel0> a/an <att1> <refobj1> and a/an <att2> <refobj2>. | The baseball field in the middle of a blue build and a track field. |
| The <att0> <obj0> <rel0> a/an <att1> <refobj1>. Within <rel1> of the <att1> <refobj1>, there are a/an <att2> <refobj2> and a/an <att3> <refobj3>. | The baseball field located approximately 200 meters southeast of a tennis court. Within 1000 meters of the tennis court, there are a blue build and a track field. |

Table 1. Examples of referring expression templates

## Method

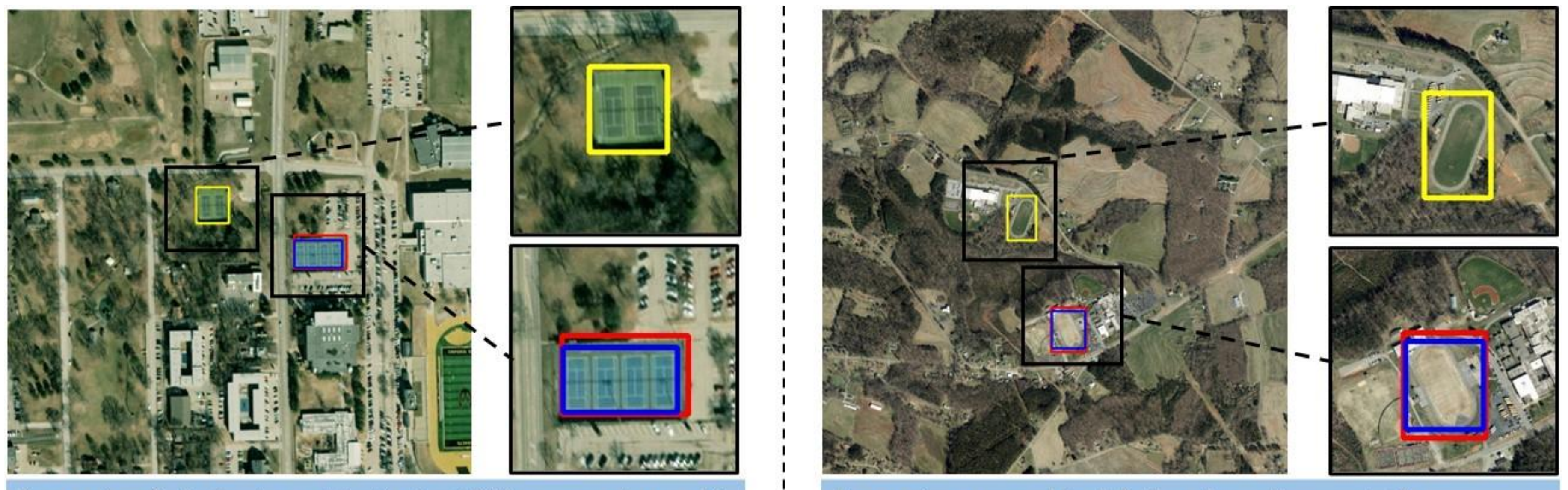

Language Encoder
Find a ground track field, located approximately 295 meters southeast of a baseball field. The baseball field is next to a road.
Object Embedding
Relation Embedding
Geospatial Relation Graph
Fusion Module
Darknet
Adaptive Region Attention
Image Encoder

- The language encoder learns numerical geospatial relations and represents a complex expression as a geospatial relation graph.
- The image encoder learns large-scale remote sensing scenes with adaptive region attention.
- The fusion module fuses the text and image feature for visual grounding.

## Results

| Method | Test | | Validate | |
|---|---|---|---|---|
| | acc@0.5 | acc@0.25 | acc@0.5 | acc@0.25 |
| OneStage | 30.15 | 34.06 | 30.06 | 34.47 |
| LBYLNet | 32.19 | 35.21 | 31.64 | 36.47 |
| ReSC | 51.18 | 57.05 | 53.96 | 58.45 |
| **GeoVG** | **59.40** | **64.95** | **58.20** | **65.11** |

## Discussion



**Expression:** find a tennis court that is about 2070 square meters, and it is located approximately 102 meters northwest of a ground track field.

**Expression:** a ground track field on the southernmost side.

**Red:** Ground-truth. **Blue:** Ours. **Yellow:** Baseline.

## Conclusion

We present a novel visual grounding method in remote sensing images and construct a new visual grounding dataset to evaluate the proposed method.