International Journal of
*Geo-Information*

*Article*

# Semantics-Driven Remote Sensing Scene Understanding Framework for Grounded Spatio-Contextual Scene Descriptions

**Abhishek V. Potnis *, Surya S. Durbha and Rajat C. Shinde**

Centre of Studies in Resources Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra 400076, India; sdurbha@iitb.ac.in (S.S.D.); rajatshinde@iitb.ac.in (R.C.S.)
* Correspondence: abhishekvpotnis@iitb.ac.in

**Abstract:** Earth Observation data possess tremendous potential in understanding the dynamics of our planet. We propose the Semantics-driven Remote Sensing Scene Understanding (Sem-RSSU) framework for rendering comprehensive grounded spatio-contextual scene descriptions for enhanced situational awareness. To minimize the semantic gap for remote-sensing-scene understanding, the framework puts forward the transformation of scenes by using semantic-web technologies to Remote Sensing Scene Knowledge Graphs (RSS-KGs). The knowledge-graph representation of scenes has been formalized through the development of a Remote Sensing Scene Ontology (RSSO)—a core ontology for an inclusive remote-sensing-scene data product. The RSS-KGs are enriched both spatially and contextually, using a deductive reasoner, by mining for implicit spatio-contextual relationships between land-cover classes in the scenes. The Sem-RSSU, at its core, constitutes novel Ontology-driven Spatio-Contextual Triple Aggregation and realization algorithms to transform KGs to render grounded natural language scene descriptions. Considering the significance of scene understanding for informed decision-making from remote sensing scenes during a flood, we selected it as a test scenario, to demonstrate the utility of this framework. In that regard, a contextual domain knowledge encompassing Flood Scene Ontology (FSO) has been developed. Extensive experimental evaluations show promising results, further validating the efficacy of this framework.

**Keywords:** remote sensing scene understanding; semantics-driven; grounded natural language scene descriptions; spatio-contextual; Scene Knowledge Graphs; flood ontology; semantic web; GeoSPARQL; Resource Description Framework (RDF); Semantic Web Rule Language (SWRL)

check for updates

**Citation:** Potnis, A.V.; Durbha, S.S.; Shinde, R.C. Semantics-Driven Remote Sensing Scene Understanding Framework for Grounded Spatio-Contextual Scene Descriptions. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 32. https://doi.org/10.3390/ijgi10010032

Received: 31 October 2020
Accepted: 11 January 2021
Published: 13 January 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

In recent years, the adoption of remote sensing across a wide spectrum of applications has increased rapidly. With an increase in the number of satellites launched over the last few years, there has been a deluge of Earth Observation (EO) data. However, the rate of data exploration largely lags behind the rate at which the EO data are being generated by these remote-sensing platforms [1]. The remote-sensing imagery captured by these platforms has great potential in understanding numerous natural, as well as manmade, phenomena. This remains largely unexplored, primarily due to the sheer volume and velocity of the data. This calls for a need for innovative and efficient ways to rapidly explore and exploit EO data. The research problem of empowering machines to interpret and understand a scene as a human has been gaining lots of attention in the remote-sensing community.

The area of remote sensing scene understanding for information mining and retrieval, scene interpretation including Land Use Land Cover (LULC) classification and change detection among numerous other applications has evolved significantly over the years. Most of the research on this paradigm has been focused on the problem of information mining and retrieval from remote sensing scenes. Earlier works [2,3] focused on the problem of scene identification and retrieval by comparing Synthetic Aperture Radar (SAR) data of different scenes by using a model-based scene inversion approach with Bayesian inference. These works propose and discuss mathematical models for extraction of low-level physical characteristics of the 3D scenes from the 2D images. Reference [4] introduced information fusion for scene understanding by proposing the mapping of the extracted primitive features to higher-level semantics representing urban scene elements.

There has been significant research in the area of Image Information Mining for Remote-Sensing Imagery that has led to the development of numerous Information Image Mining (IIM) frameworks in the last couple of decades. The GeoBrowse system [5] is one of the earliest IIM systems for remote-sensing imagery. Developed on the principles of distributed computing, the system used an object-oriented relational database for data storage and information retrieval. The Knowledge-driven Information Mining (KIM) system [6,7] was built over References [2,3], and Reference [4] proposed to use Bayesian networks to link user-defined semantic labels to a global content-index generated in an unsupervised manner. The Geospatial Information Retrieval and Indexing System (GeoIRIS) [8] identified specialized descriptor features to extract and map particular objects from remote sensing scenes, with support for querying by spatial configuration of objects for retrieval of remote-sensing-scene tiles. The PicSOM system [9] applied self-organized maps to optimize retrieval of images based on the content queried by the user. It also proposed supervised and unsupervised change-detection methodologies in addition to detecting manmade structures from satellite imagery. The Intelligent Interactive Image Knowledge Retrieval (I3KR) [1] system proposed the use of domain-dependent ontologies for enhanced knowledge discovery from the Earth Observation data. It used a combination of supervised and unsupervised techniques to generate models for object classes, followed by the assignment of semantic concepts to the classes in the ontology, achieved automatically by description-logic-based inference mechanisms. The Spatial Image Information Mining Framework [10] inspired by the I3KR [1] focused on the modeling of directional and topological relationships between the regions in an image and the development of Spatial Semantic Graph. The SIIM also proposed a Resource Description Framework (RDF)-based model for representing an image, associating regions with classes and their relationships, both directional and topological, amongst themselves, along with the structural metadata, such as geographical coordinates and time of acquisition.

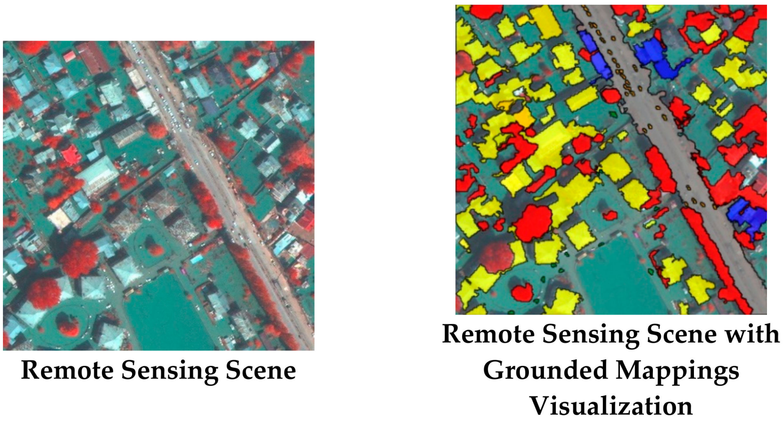
Each of the image information mining systems for remote sensing imagery mentioned above have strived to address the problem of effective retrieval of content-based information from huge remote sensing archives. Some of the recent semantics enabled IIM systems have also addressed the issue of the "semantic gap" between the low-level primitive features and high-level semantic abstractions in a remote sensing scene. However, the research problem of comprehensive understanding and interpretation of remote sensing scenes from a spatio-contextual perspective for man–machine interactions has still not been completely addressed.

Recently, a few research studies have focused on remote sensing scene captioning to interpret scenes into natural language descriptions. Reference [11] proposed the use of a deep multi-modal neural network consisting of a convolutional neural network to extract the image features followed by a recurrent neural network trained over a dataset of image–caption pairs to generate the single sentence text descriptions. The framework proposed in Reference [12] consists of two stages for the task of remote-sensing-image captioning: (1) Multi-Level Image Understanding, using Fully Convolutional Networks (FCNs), and (2) language generation, using a template-based approach. The first stage was designed to generate triplets from the images in the form of (ELM, ATR and RLT) representing the ground elements, their attributes and their relationships with other elements. These triples serve as input to the templating mechanism to generate appropriate natural language text descriptions. The Remote Sensing Image Captioning Dataset (RSICD) developed in Reference [13] consists of manually generated image–sentence pairs for 10921 generic remote sensing scenes of 224 × 224 pixels size. The developed dataset was evaluated over two methods of image captioning: (1) multi-modal approach [11] and (2) an attention-based approach proposed in Reference [13] that uses both deterministic and stochastic manner of attention. The Collective Semantic Metric Learning (CSML) framework [14] proposed the collective sentence representation corresponding to a single remote sensing image representation in the semantic space for a multi-sentence captioning task. With the objective of ensuring focus on the regions of interest, the Visual Aligning Attention model (VAA) [15] proposed a novel visual aligning loss function designed to maximize the feature similarity between the extracted image feature vectors and the word embedding vectors. The Retrieval Topic Recurrent Memory Network [16] proposed the use of topic words retrieved from the topic repository generated from the ground truth sentences at the training stage. In the testing stage, the retrieved topic words embedded into the network in the form of topic memory cells further control and guide the sentence generation process. Reference [17] addressed the problems of (1) focusing on different spatial features at different scales and (2) semantic relationships between the objects in the remote sensing image. It proposed the use of a multi-level attention module to account for spatial features at different scales and attribute graph-based graph convolutional network to account for the semantic relationship between the objects in the remote sensing image.

The remote sensing image captioning frameworks mentioned above have addressed the problem of captioning remote sensing scenes to natural language sentences. A few recent studies in this area have focused on the semantic visual relationships between the objects in the scenes. However, the research problem of generating detailed description paragraphs consisting of multiple natural language sentences, comprehensively describing a remote sensing scene, taking into account the spatio-contextual relationships between the objects, remains largely unexplored. Moreover, the generated sentences in the above-mentioned frameworks are not grounded to specific regions of the scene, and thus lack explainability. The term "grounded" in the scene description refers to the explicit mapping words or phrases in it to regions in the scene that it describes. This reinforces explainability and reliability of the scene description in its task of describing the scene in natural language.

Comprehensive, explainable and contextual interpretation of a remote sensing scene is of utmost importance especially in a disaster situation such as floods. During a flood occurrence, it is crucial to understand the flood inundation and receding patterns in context to the spatial configurations of the land-use/land-cover in the flooded regions. Moreover, the contextual semantics of the flood scene are also influenced by the temporal component. As time progresses, a flooded region may either shrink in size or grow, affecting the semantics of other regions that it spatially interacts with. Therefore, there is a dire need to develop approaches that can translate the real-world ground situation during or post disaster in a way that can be easily assimilated both by humans and machines, which can lead to a response that is well orchestrated and timely.

This paper addresses the problem of remote sensing scene understanding focusing specifically on comprehensive and explainable interpretation of scenes from a spatio-contextual standpoint for effective man–machine interactions. In that regard, the novel Semantics-driven Remote Sensing Scene Understanding (Sem-RSSU) framework was developed to generate grounded explainable natural language scene descriptions from remote sensing scenes. Figure 1 depicts comprehensive grounded spatio-contextual scene description as rendered by our proposed Semantics-driven Remote Sensing Scene Understanding (Sem-RSSU) framework for a remote sensing scene of urban floods. To the best of our knowledge, this research is the first of its kind to explore comprehensive grounded scene description rendering for remote sensing scenes using a semantics-driven approach.



**Remote Sensing Scene**

**Remote Sensing Scene with Grounded Mappings Visualization**

**Grounded Spatio-Contextual Scene Description in Natural Language:**
There is a road. There are accessibleBuildings along the road. There are floodedBuildings to the West and East direction of the road. There is traffic on the road. There are strandedVehicles to the West direction of the road. There is floodedVegetation to the West and East direction of the road.

**Figure 1.** Comprehensive grounded spatio-contextual scene description as rendered by our proposed Semantics-driven Remote Sensing Scene Understanding (Sem-RSSU) for a remote sensing scene of urban floods, depicting explainability by color-coded mapping of sentences to regions of interest in the scene.

The broad objective of this research is to transform a remote sensing scene to a spatio-contextual knowledge graph and further into explainable grounded natural language scene descriptions for enhanced situational awareness and effective man–machine interaction.

*Major Research Contributions*

Our major research contributions in this work are two-fold:

- We formalize the representation and modeling of spatio-contextual knowledge in remote sensing scenes in the form of Remote Sensing Scene Knowledge Graphs (RSS-KGs), through the development of Remote Sensing Scene Ontology (RSSO)—a core ontology for an inclusive remote-sensing-scene data product. We develop a contextual domain knowledge encompassing Flood Scene Ontology (FSO), to represent concepts that proliferate during a flood scenario.
- We propose and implement the end-to-end Semantics-enabled Remote Sensing Scene Understanding (Sem-RSSU) framework as a holistic pipeline for generating comprehensive grounded spatio-contextual scene descriptions, to enhance the user-level situational awareness, as well as machine-level explainability of the scenes.

In that regard, we propose (1) Ontology-driven Spatio-Contextual Triple Aggregation and (2) Scene Description Content Planning and Realization algorithms, to enable rendering of grounded explainable natural language scene descriptions from remote sensing scenes. We report and discuss our findings from the extensive evaluations of the framework.

The paper is structured as follows: Section 2 describes the proposed Semantics-driven Remote Sensing Scene Understanding (Sem-RSSU) framework. It presents in detail the various layers and components of the framework. Section 3 discusses the experimental setup, results with the evaluation strategies used to verify the efficacy of the framework. In Sections 4 and 5, we discuss and summarize the framework and conclude with future directions of this research.

## 2. Framework for Semantics-Driven Remote Sensing Scene Understanding

The core focus of the Semantics enabled Remote Sensing Scene Understanding (Sem-RSSU) framework is geared towards enabling enhanced situational awareness from remote sensing scenes. It intends to enable the users/decision makers to obtain increased understanding of the situation and help make better choices to respond to it prudently. It is well understood that the greater the amount of targeted information the better are the chances to positively react to the situation. However, the information needs to be highly contextual, easily understandable and pertinent to that particular situation. Otherwise, there is a danger of information overload leading to undesired consequences. Therefore, it is essential to develop approaches that can translate the real ground situation from remote sensing scenes in a way that can be easily understood and queried upon by man and machines alike, thereby leading to an appropriate and timely response in sensitive situations such as disasters.

Figure 2 depicts the system architecture of the proposed framework. The framework was logically divided into 6 layers: Data Storage layer, Scene Ingestion layer, Data Mediation layer, Semantic Enrichment layer, Natural Processing layer and the Rendering layer.

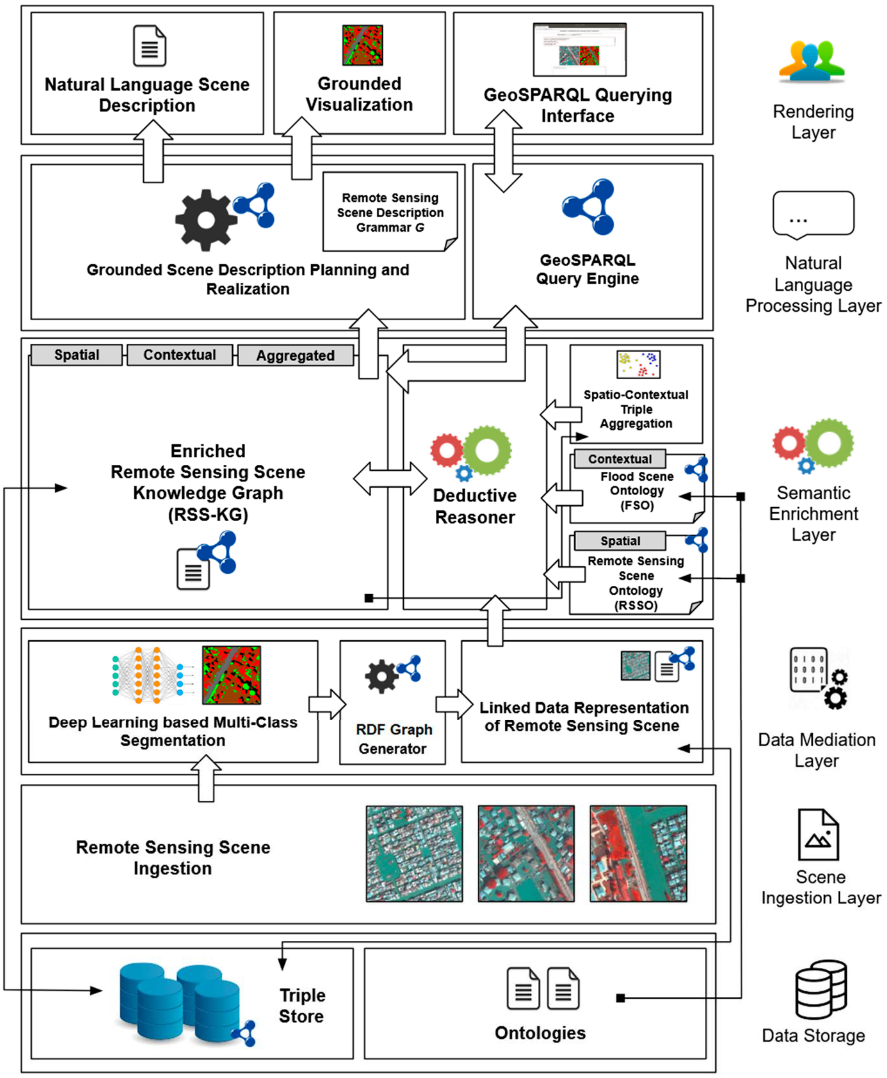

Natural Language Scene Description
Grounded Visualization
GeoSPARQL Querying Interface
Rendering Layer
Remote Sensing Scene Description Grammar G
Grounded Scene Description Planning and Realization
GeoSPARQL Query Engine
Natural Language Processing Layer
Spatial
Contextual
Aggregated
Enriched Remote Sensing Scene Knowledge Graph (RSS-KG)
Deductive Reasoner
Spatio-Contextual Triple Aggregation
Contextual
Flood Scene Ontology (FSO)
Spatial
Remote Sensing Scene Ontology (RSSO)
Semantic Enrichment Layer
Deep Learning based Multi-Class Segmentation
RDF Graph Generator
Linked Data Representation of Remote Sensing Scene
Data Mediation Layer
Remote Sensing Scene Ingestion
Scene Ingestion Layer
Triple Store
Ontologies
Data Storage

**Figure 2.** System architecture of the Semantics-driven Remote Sensing Scene Understanding (Sem-RSSU) Framework.

The Data Storage layer consists of a triple-store to store, retrieve and update the Remote Sensing Scene Knowledge Graphs (RSS-KGs) generated from the scenes. It also stores the ontologies—the Remote Sensing Scene Ontology (RSSO) and the contextual Flood Scene Ontology (FSO)—on a disk-based storage, for deductive reasoning. The Scene

Ingestion layer deals with ingesting Remote Sensing Scenes of interest into the framework for comprehensive, grounded and explainable scene description rendering.

## *2.1. Data Mediation*

The Data Mediation layer consists of (1) the deep-learning-based multi-class segmentation component that segments the ingested scene into land-use/land-cover regions and (2) the RDF graph generator component that transforms the segmented land-use/land-cover regions in the scene to a graph representation of the scene conforming to the proposed Remote Sensing Scene Ontology.

### 2.1.1. Multi-Class Segmentation

The remote sensing scene consists of multiple land-use/land-cover regions spatially interacting with one another. To infer higher level abstractions from the scene, it is essential to identify the primitive features and predict their land-use/land-cover. Each pixel in the scene is assigned a label of a land-cover, using a deep-neural-network approach. Some of the popular state-of-the-art deep-neural-network architectures based on the encoder–decoder architecture were experimented on for the urban flood dataset for this research.

The Fully Convolutional Network (FCN) [18] architecture popularized the use of end-to-end convolutional neural networks for semantic segmentation. The FCN first introduced the use of skip connections to propagate spatial information to the decoder layers and improve the upsampling output. U-Net architecture [19] built over the FCN, was first proposed for biomedical image segmentation and has proven to adapt well across a large spectrum of domains. It proposed a symmetric U-shaped architecture for encoding and decoding with multiple upsampling layers and using the concatenation operation instead of addition operation. The Pyramid Scene Parsing Network (PSPNet) [20] uses dilated convolution to increase the receptive field in addition to the use of the pyramid pooling module in the encoder to capture and aggregate the global context. The SegNet [21] architecture advocated storing and transmitting the max-pooling indices to the decoder to improve the quality of upsampling. The ResNet [22] architecture, a Convolutional Neural Network architecture, proposed the use of residual-blocks identity-skip connections to tackle the vanishing gradient problem encountered while training deep neural networks. The ResNet and VGG-16 [23] were used as backbone architectures for the FCN, U-Net, SegNet and PSP neural network architectures for experimenting over the urban flood dataset consisting of high-resolution remote sensing scenes captured during an urban flood event. Each of the architectures was implemented and evaluated over this dataset.

Both ResNet and VGG-16 as backbone architectures have shown promising results [24] for multi-class segmentation (also known as semantic Segmentation) for remote sensing scenes. The architectures for multi-class segmentation considered in Sem-RSSU were selected by considering their effectiveness and relevance for the task of segmentation and to limit the scope of the study. It must be noted that Sem-RSSU was structured to be modular and is thus amenable for use with any state-of-the-art deep neural approaches for multi-class segmentation.

### 2.1.2. RDF Graph Generator

The RDF Graph Generator component translates the land-use/land-cover regions from raster to a Resource Description Framework based graph representation. The semantic segmentation results (also known as classification maps) are vectorized into Well-Known Text (WKT) geometry representation based on the color labels assigned to the pixels. A predefined threshold for minimum pixels in a region to constitute an object in a knowledge graph, in this component, filters out stray and noisy pixel labels. This was implemented by using the shapely (https://pypi.org/project/Shapely) and rasterio (https://pypi.org/project/rasterio) libraries in Python. This process of vectorization is followed by encoding and rendering the WKT geometries in a string conforming to an RDF representation. The RDF graph representation of the remote sensing scene consists of triples corresponding to land-use/land-cover regions in the scene with their geometries and other spatial information stored in accordance with the GeoRDF ( https://www.w3.org/wiki/GeoRDF) standard by the W3C and the proposed Remote Sensing Scene Ontology (RSSO).

### 2.2. Semantic Enrichment Layer

Description Logic (DL) forms the fundamental building block of formalizing knowledge representation. It also forms the basis of the Web Ontology Language (OWL) used to construct ontologies. The Sem-RSSU framework proposes the formalization of remote sensing scene knowledge through the development of the Remote Sensing Scene Ontology (RSSO) and the contextual Flood Scene Ontology (FSO). The DL-based axioms discussed in this section were encoded in the proposed ontologies, to facilitate inferencing of implicit knowledge from the remote sensing scenes.

The Semantic Enrichment layer was structured in a multi-tier manner, to facilitate hierarchical semantic enrichment of the RDF graph representation of the remote sensing scenes to generate enriched Remote Sensing Scene Knowledge Graphs (RSS-KGs).

The Remote Sensing Scene Ontology (RSSO) enriches the RDF data by inferring spatial–topological and directional relationships between the land-use/land-cover regions, thereby facilitating generation of Remote Sensing Scene Knowledge Graphs (RSS-KGs). The KGs are further enriched with contextual concepts and relationships with the Flood Scene Ontology, a domain knowledge encompassing ontology that formalizes the concepts and relationships proliferating during the flood scenario. The enriched KGs are further aggregated contextually from a remote sensing scene description standpoint with the Spatio-Contextual Triple Aggregation algorithm.

An ontology-based deductive reasoner facilitates the enrichment of knowledge graphs at every tier. Figure 3 represents the multi-tier architecture of the Semantic Reasoning layer. The multi-tier approach of semantic enrichment in the form of spatial, contextual and aggregated knowledge enables (1) modularity and (2) extensibility, thus rendering the architecture amenable for scaling and integration with other data sources (e.g., GeoNames, UK Ordnance Survey, etc.) for other remote-sensing applications.

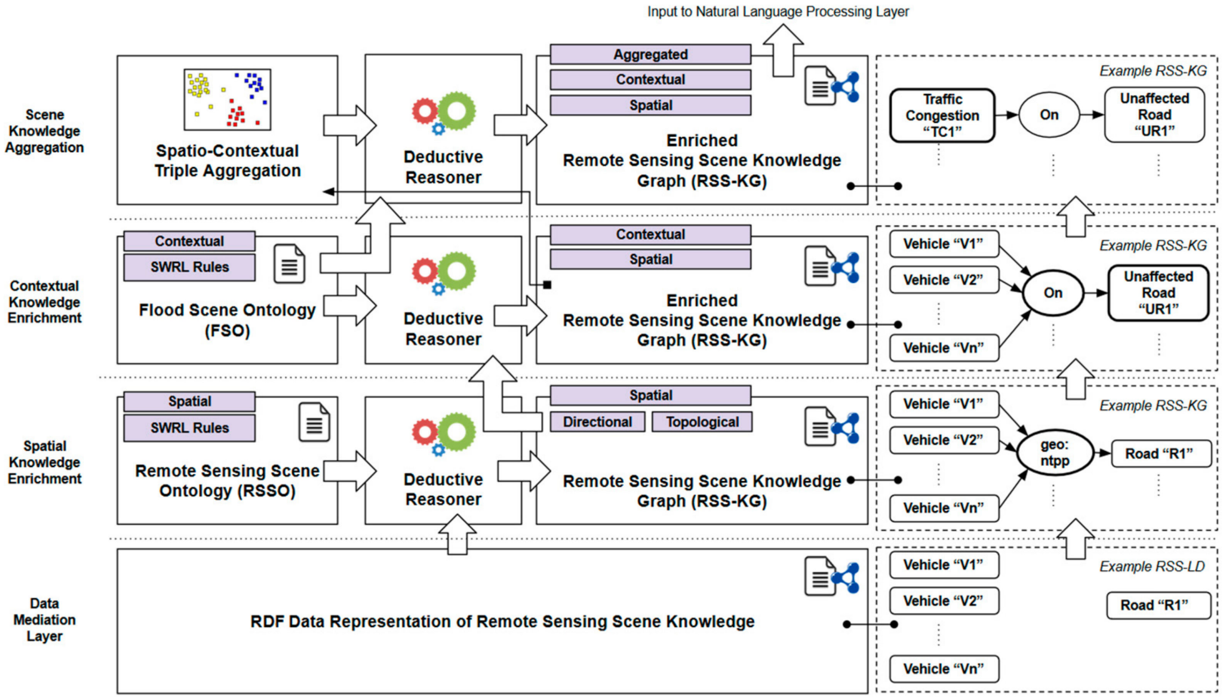


**Figure 3.** Multi-tier architecture of the Semantic Reasoning layer depicting semantic enrichment.

The figure also depicts an example of the hierarchical knowledge graph enrichment process as it propagates upwards to the Natural Language Processing layer. In the Data Mediation layer, the instances of vehicles and a road are represented using the Resource Description Framework (RDF) as a graph representation of the remote sensing scene. The graph representation in RDF form is propagated to the Spatial Knowledge Enrichment layer where the RSSO enriches and transforms it into a knowledge graph by inferring spatial–topological and directional relationships between the instances. The "geo:ntpp" Non-Tangential Proper Part topological relationship from RCC8 is inferred at this stage. The knowledge graph is further propagated upward for Contextual Enrichment using the FSO. In this layer, the contextual relationship "on" is inferred and the "road" instance is further specialized to a "unaffected road" class instance. Furthermore, the knowledge graph is aggregated using Spatio-Contextual Triple Aggregation for scene description rendering where the numerous vehicle instances "on" the road are aggregated and a "traffic congestion" class instance is inferred.

### 2.2.1. Semantic Data Modeling in Sem-RSSU

The Sem-RSSU framework formalizes the representation of remote sensing scenes in the form of knowledge graphs through the development of ontologies. It utilizes the developed ontologies for knowledge enrichment in a modular and hierarchical form to be amenable for integration and extension for other remote sensing applications.

#### Ontology Development for Spatial Semantic Enrichment

The Remote Sensing Scene Ontology (RSSO) was developed to translate the RDF data representation of remote sensing scenes to knowledge graphs by inferring spatial–topological and directional concepts and relationships between the identified regions. The ontology formalizes the semantics of a generic remote sensing scene captured by a remote sensing platform. To gauge the reliability of a remote-sensing data product and ascertain its origin, the metadata bundled with the data product plays a crucial role. In that regard, the RSSO defines classes, Object and Data Properties to model the metadata of a remote sensing scene to establish comprehensive data lineage. The Data Properties "hasGroundSamplingDistance", "hasProjection", enumerateBands", "hasResamplingMethod", "hasSpectralBands", "hasAcquisitionDateTime", etc., along with the specifics of the remote-sensing platform stored in Object Properties collectively model the metadata of the scene.

The Anderson Land-Use/Land-Cover Classification system was used as a reference to model the LULC classes in the RSSO. Figure 4a depicts the Classes Hierarchy, Data Properties and the Object Properties in the RSSO. The object properties were used to model and capture the topological and directional relationships between instances of different land-use/land-cover regions. Figure 4b depicts the visualization of an instance of class "scene" and "Scene256" captured by "GeoEye-01" which is an instance of class "satellite" that is a "RemoteSensingPlatform". A "scene" class instance has a relation with "region" class instances through the object property "hasRegions". Thus, a scene has multiple regions within it, with each of the regions having a LULC associated with it through the "hasLULC" Data Property. Figure 5 depicts a snippet of the Scene Knowledge Graph represented in the Resource Description Framework (RDF) form for a region "R40" in a remote sensing scene.

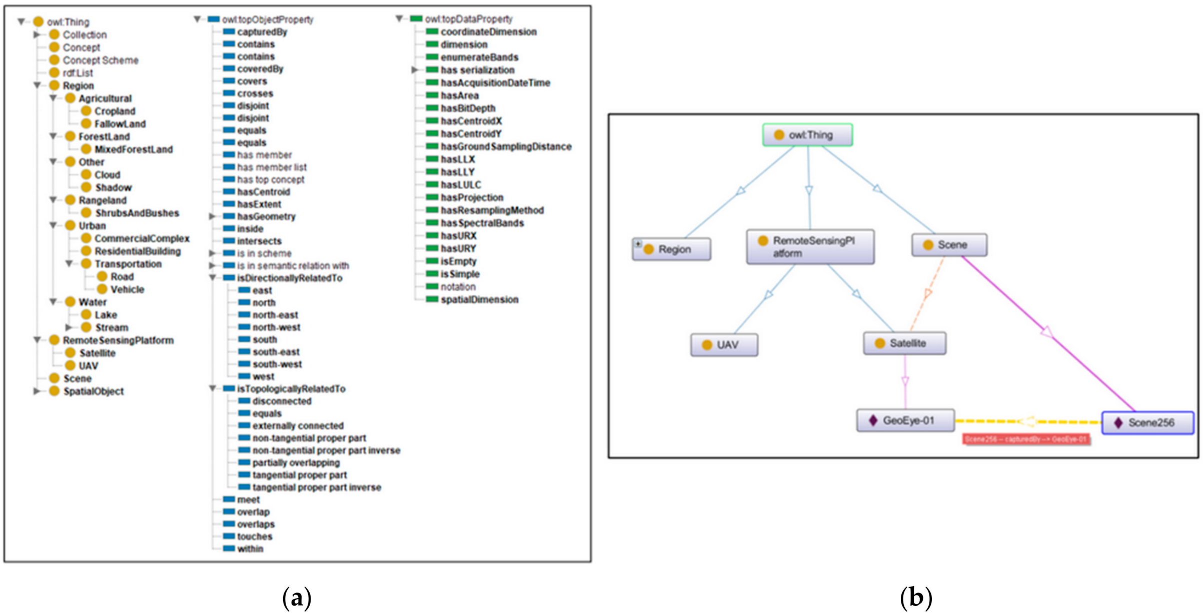
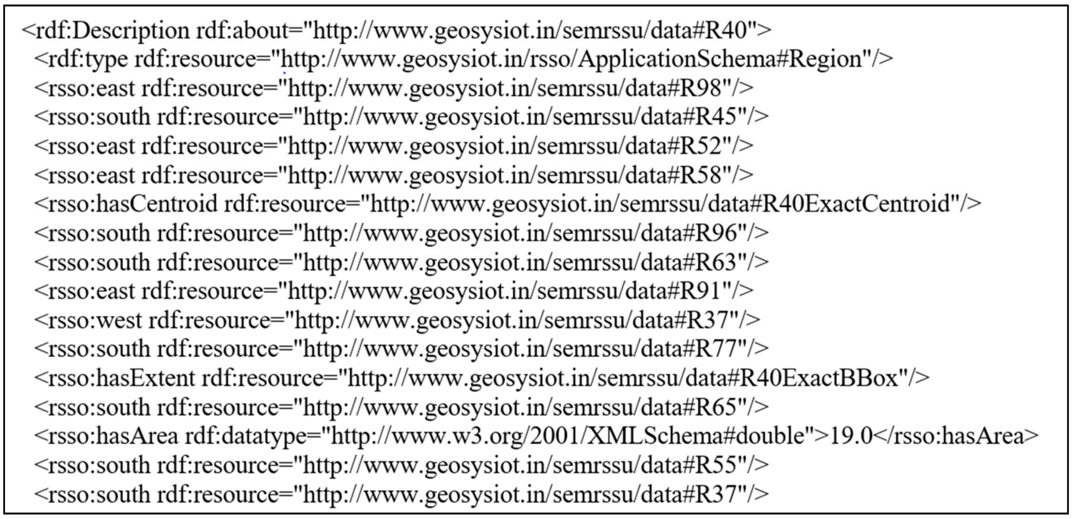


**Figure 4.** (**a**) Snapshot of the Classes Hierarchy, Object and Data Properties in the Remote Sensing Scene Ontology (RSSO) extending the GeoSPARQL ontology. (**b**) Visualization of a scene instance captured by the GeoEye-01 Satellite as modeled in the Remote Sensing Scene Ontology (RSSO).



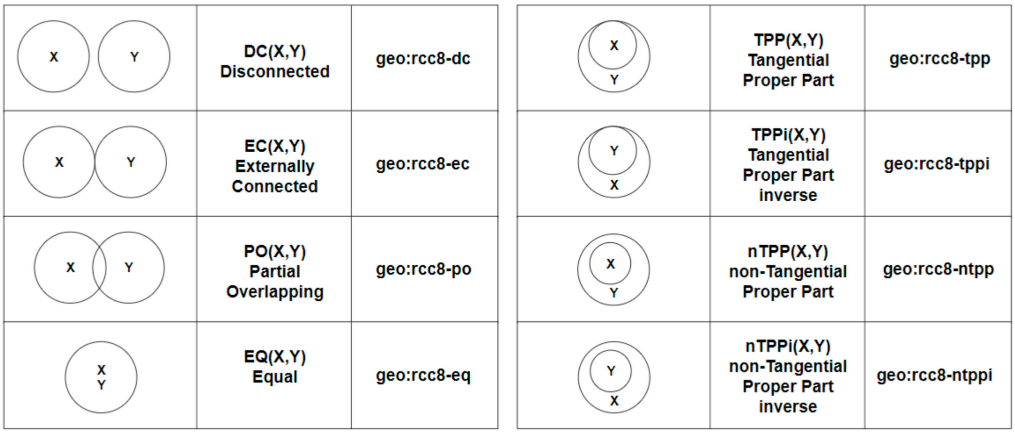
```
<rdf:Description rdf:about="http://www.geosysiot.in/semrssu/data#R40">
 <rdf:type rdf:resource="http://www.geosysiot.in/rsso/ApplicationSchema#Region"/>
 <rsso:east rdf:resource="http://www.geosysiot.in/semrssu/data#R98"/>
 <rsso:south rdf:resource="http://www.geosysiot.in/semrssu/data#R45"/>
 <rsso:east rdf:resource="http://www.geosysiot.in/semrssu/data#R52"/>
 <rsso:east rdf:resource="http://www.geosysiot.in/semrssu/data#R58"/>
 <rsso:hasCentroid rdf:resource="http://www.geosysiot.in/semrssu/data#R40ExactCentroid"/>
 <rsso:south rdf:resource="http://www.geosysiot.in/semrssu/data#R96"/>
 <rsso:south rdf:resource="http://www.geosysiot.in/semrssu/data#R63"/>
 <rsso:east rdf:resource="http://www.geosysiot.in/semrssu/data#R91"/>
 <rsso:west rdf:resource="http://www.geosysiot.in/semrssu/data#R37"/>
 <rsso:south rdf:resource="http://www.geosysiot.in/semrssu/data#R77"/>
 <rsso:hasExtent rdf:resource="http://www.geosysiot.in/semrssu/data#R40ExactBBox"/>
 <rsso:south rdf:resource="http://www.geosysiot.in/semrssu/data#R65"/>
 <rsso:hasArea rdf:datatype="http://www.w3.org/2001/XMLSchema#double">19.0</rsso:hasArea>
 <rsso:south rdf:resource="http://www.geosysiot.in/semrssu/data#R55"/>
 <rsso:south rdf:resource="http://www.geosysiot.in/semrssu/data#R37"/>
```

**Figure 5.** Snippet of the Scene Knowledge Graph represented in the Resource Description Framework (RDF) form for a region in the remote sensing scene.

Modeling Topological Relationships in RSSO

The RSSO builds over the OGC standardized GeoSPARQL ontology that formalizes representation of spatial objects and their topological relationships in the form of geospatial RDF data. The GeoSPARQL ontology enables the use of Geographic Markup Language (GML) based and Well-Known Text (WKT) Literals for representing geometries of spatial objects. It also defines vocabularies for topological relationships including Egenhofer [25] and RCC8 [26] relationships. The RSSO thus reuses the RCC8 relationships defined in the GeoSPARQL standard by referencing it with the prefix “geo”.

The RSSO defines topological relationships as depicted in Figure 6, for regions in remote sensing scenes and establishes an equality with its GeoSPARQL counterparts, using the owl:sameAs construct.



| | | | | | |
|---|---|---|---|---|---|
| X Y | DC(X,Y)<br>Disconnected | geo:rcc8-dc | X Y | TPP(X,Y)<br>Tangential<br>Proper Part | geo:rcc8-tpp |
| X Y | EC(X,Y)<br>Externally<br>Connected | geo:rcc8-ec | Y X | TPPi(X,Y)<br>Tangential<br>Proper Part<br>inverse | geo:rcc8-tppi |
| X Y | PO(X,Y)<br>Partial<br>Overlapping | geo:rcc8-po | X Y | nTPP(X,Y)<br>non-Tangential<br>Proper Part | geo:rcc8-ntpp |
| X Y | EQ(X,Y)<br>Equal | geo:rcc8-eq | Y X | nTPPi(X,Y)<br>non-Tangential<br>Proper Part<br>inverse | geo:rcc8-ntppi |

**Figure 6.** Region Connection Calculus 8 with corresponding GeoSPARQL topological vocabularies.

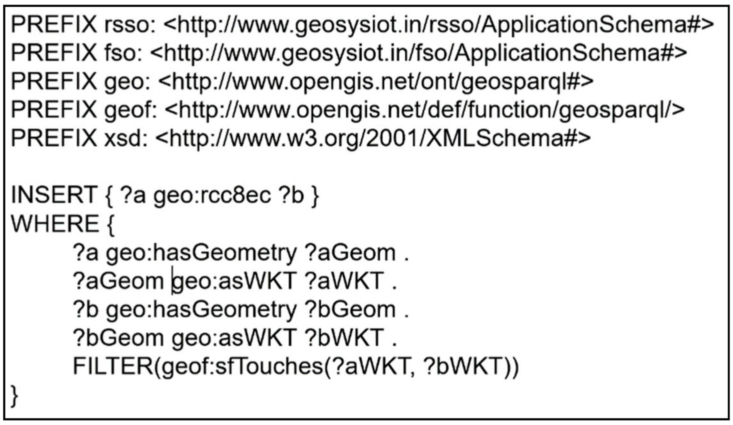
The following represents the formal expression of the externally connected topological relationship along with its definition in English, as modeled in RSSO:

> ***Externally Connected***: A *region a* is *externally connected* to *region b* if the *geometry* of *a touches geometry* of *b*. This is detected using the *WKT* representation of the *geometries* with the *sfTouches* predicate of GeoSPARQL.

$$\forall a \forall b\ (hasGeometry(a, aGeom) \wedge hasGeometry(b, bGeom) \wedge asWKT(aGeom, aWKT) \wedge asWKT(bGeom, bWKT) \wedge sfTouches(aWKT, bWKT) \rightarrow externallyConnected(a, b)) \tag{1}$$

The variables *a* and *b* in the expression are instances of the region class. The hasExactGeometry, sfTouches and externallyConnected are object properties defined in GeoSPARQL and Remote Sensing Scene Ontology (RSSO). The asWKT is a Data Property defined in the GeoSPARQL ontology. The expression depicts the conditions for the inferencing of the externallyConnected property between region instances *a* and *b*.

Deductive rule-based reasoning involves the use of a reasoner with access to an ontology containing the concepts and relationships defined in it. Rule-based reasoning specific to an ontology can be implemented by (1) encoding the rules in the form of Semantic Web Rule Language (SWRL) in the ontology and having the reasoner infer new triples using the rules or (2) using GeoSPARQL to query and infer new triples. The above query in Figure 7 depicts the GeoSPARQL-query-based implementation for inferring the externallyConnected topological relationship.



```
PREFIX rsso: <http://www.geosysiot.in/rsso/ApplicationSchema#>
PREFIX fso: <http://www.geosysiot.in/fso/ApplicationSchema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

INSERT { ?a geo:rcc8ec ?b }
WHERE {
        ?a geo:hasGeometry ?aGeom .
        ?aGeom geo:asWKT ?aWKT .
        ?b geo:hasGeometry ?bGeom .
        ?bGeom geo:asWKT ?bWKT .
        FILTER(geof:sfTouches(?aWKT, ?bWKT))
}
```

**Figure 7.** GeoSPARQL query depicting inferencing of externally connected topological relationship, using GeoSPARQL's *geof:sfTouches* construct.

Modeling Directional Relationships in RSSO

The RSSO models the four cardinal directions, North, South, East and West, and four intercardinal directions, North-East, North-West, South-East and South-West, for representing directional relationships between regions in a scene. It uses the projection-based approach [27], using half-planes for representing the 4 cardinal and 4 intercardinal directions. The relationships are formalized in the ontology using Semantic Web Rule Language (SWRL).

The directional relationship between two regions in a scene represent their spatial orientation with respect to one another. The two approaches [10] for computing the directional relationships between regions in the spatial domain are the (1) Minimum Bounding Rectangle (MBR)-based approach and (2) Centroid-based approach. The MBR based approach utilizes a pair of coordinates—Lower Left and Upper Right—for each region, to compute the directional relationship. The Centroid-based approach uses the coordinates of the Centroid of a region to compute the directional relationship.

It was experimentally found that for high resolution remote sensing scenes, due to the presence of numerous elongated and obliquely placed regions such as roads and buildings in the scenes, the MBR based approach generated erroneous directional relationships in contrast to the Centroid based approach using half-planes, which generated fairly good results. Thus, RSSO adopts the Centroid-based approach using half-planes from projections as depicted in Figure 8 for modeling directional relationships.

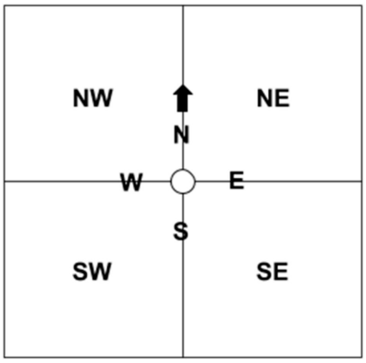


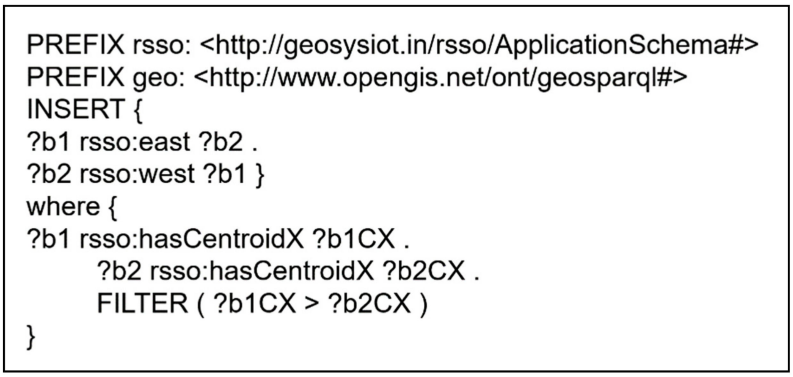
**Figure 8.** Projections-based approach using half-planes for cardinal and intercardinal directions, as modeled in RSSO.

The following represents the formal expression of the East and West directional relationship along with its definition in English, as modeled in RSSO:

> ***East Direction***: A *region a* is to the *East* of *region b* if the × coordinate of the centroid of *a* is greater than the × coordinate of the centroid of *b*. It also entails that *region b* is to the *West* of *region a*.

$$\forall a \forall b\ (hasCentroidX(a, aCX) \wedge hasCentroidX(b, bCX) \wedge greaterThan(aCX, bCX) \rightarrow East(a, b)) \wedge West(b, a)) \qquad (2)$$

The variables *a* and *b* in the expression are instances of the *region* class. The *hasCentroidX* is a Data Property defined in RSSO for storing the X coordinate of the Centroid of a region. The expression depicts the conditions for the inferencing of the *East* and *West* directions between *region* instances *a* and *b*. Similarly, expressions for all the cardinal directions were encoded as SWRL rules in the ontology. The following query in Figure 9, depicts the GeoSPARQL-query-based implementation for inferencing of *East* and *West* directional relationships. The East and West directions are defined as inverse properties of one another using the *owl:inverseOf* construct in RSSO.



```
PREFIX rsso: <http://geosysiot.in/rsso/ApplicationSchema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
INSERT {
?b1 rsso:east ?b2 .
?b2 rsso:west ?b1 }
where {
?b1 rsso:hasCentroidX ?b1CX .
      ?b2 rsso:hasCentroidX ?b2CX .
      FILTER ( ?b1CX > ?b2CX )
}
```

**Figure 9.** GeoSPARQL query depicting inferencing of East and West directional relationship.

**North-East Direction**: A region a is to the North-East of region *b* if a is to the North of *b* and a to the East of *b*.

$$\forall a \forall b\ (North(a, b) \wedge East(a, b) \rightarrow North\text{-}East(a, b)) \tag{3}$$

Consequently, the intercardinal directions were defined as the union of individual cardinal directions, as represented in Expression (3).

Ontology Development for Contextual Semantic Enrichment

The Flood Scene Ontology (FSO) introduced in Reference [28] was further improved and enriched as a part of this study. The FSO extends the RSSO and was conceptualized to consist of comprehensive domain knowledge of the flood disaster from the perspective of remote sensing scene understanding. The ontology was developed for contextual semantic enrichment of Scene Knowledge Graphs by defining context-specific concepts and relationships proliferating during the flood scenario.

The FSO builds over the RSSO and formalizes specialized classes that are intended to be inferred from remote sensing scenes of urban floods. The following are the formal expressions, along with their natural language definitions, for some of the specialized classes that were encoded as SWRL rules in the ontology. Moreover, some of their corresponding GeoSPARQL queries that can be used as an alternate implementation instead of SWRL rules have been depicted.

***Flooded Residential Building***: A *region* is termed as "*Flooded Residential Building*" as per FSO, if it is a *region* that has *LULC* as "*Residential Building*" and it is *externally connected* with at least one *region* that has *LULC* as "*Flood Water*".

$$\begin{aligned}&\forall a \forall b\ (hasLULC(a, \text{“ResidentialBuilding”}) \wedge hasLULC(b, \text{“FloodWater”})\\ &\wedge externallyConnected(a, b) \rightarrow hasInferredLULC(a, \text{“FloodedResidentialBuilding”})\\ &\wedge isA(a, FloodedResidentialBuilding))\end{aligned} \tag{4}$$

***Accessible Residential Building***: A *region* is termed as "*Accessible Residential Building*" as per FSO, if it is a *region* that has *LULC* as "*Flooded Residential Building*" and it is externally connected with at least one *region* that has *LULC* as "*Unaffected Road*". This class is envisaged to be of great importance from the perspective of disaster management specially to develop standard operating procedures (SOPs) for evacuations. Figure 11 depicts the corresponding GeoSPARQL query.

$$\begin{aligned}&\forall a \forall b\ (hasInferredLULC(a, \text{“FloodedResidentialBuilding”}) \wedge hasInferredLULC(b,\\ &\text{“UnaffectedRoad”})\ externallyConnected(a, b) \rightarrow hasInferredLULC(a,\\ &\text{“AccessibleResidentialBuilding”}) \wedge isA(a, AccessibleResidentialBuilding))\end{aligned} \tag{5}$$

***Unaffected Residential Building***: A *region* is termed as "*Unaffected Residential Building*" as per FSO, if it is a *region* that has *LULC* as "*Residential Building*" and it does not have an *Inferred LULC* as "*Flooded Residential Building*".

$$\begin{aligned}&\forall a \forall b\,(hasLULC(a, \text{"}ResidentialBuilding\text{"}) \land \neg\,(hasInferredLULC(b,\\ &\text{"}FloodedResidentialBuilding\text{"})) \to hasInferredLULC(a, \text{"}UnaffectedResidentialBuilding\text{"})\\ &\land isA(a, UnaffectedResidentialBuilding))\end{aligned} \tag{6}$$

***Stranded Vehicle***: A *region* is termed as "*Stranded Vehicle*" as per FSO, if it is a *region* that has *LULC* as "*Vehicle*" and it is *externally connected* with at least one *region* that has *LULC* as "*Flood Water*".

$$\begin{aligned}&\forall a \forall b\,(hasLULC(a, \text{"}Vehicle\text{"}) \land hasLULC(b, \text{"}FloodWater\text{"}) \land externallyConnected(a, b)\\ &\to hasInferredLULC(a, \text{"}StrandedVehicle\text{"}) \land isA(a, StrandedVehicle))\end{aligned} \tag{7}$$

Figure 10 depicts a snapshot of the classes formalized in the proposed Flood Scene Ontology (FSO) that extends the proposed Remote Sensing Scene Ontology (RSSO).

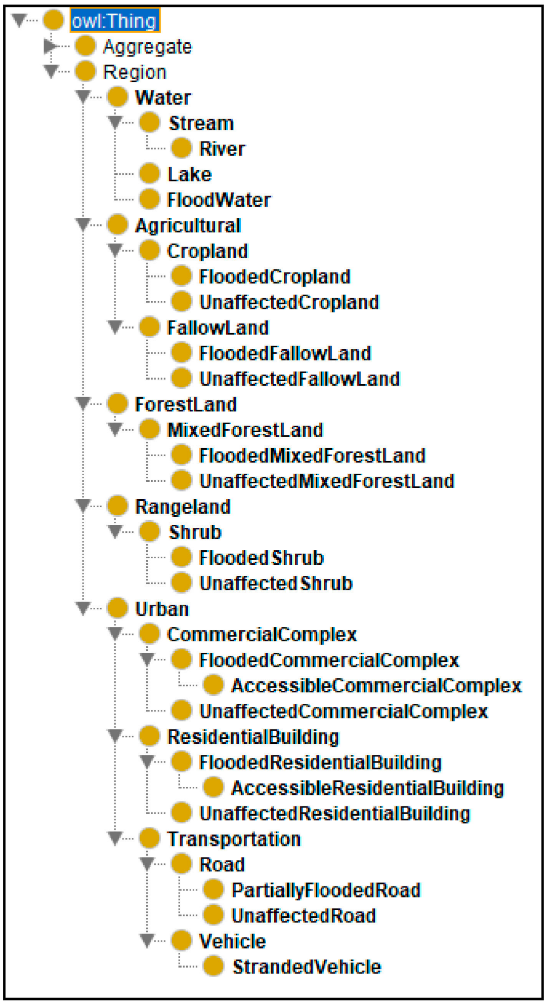
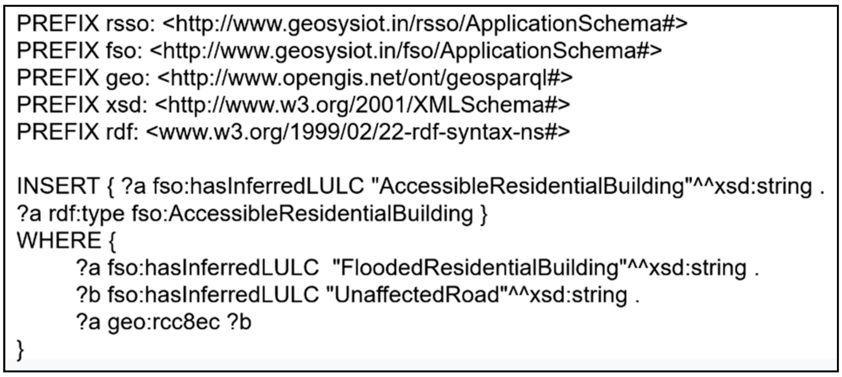


**Figure 10.** Snapshot of the Classes Hierarchy of the Flood Scene Ontology (FSO) that extends the Remote Sensing Scene Ontology (RSSO).



```
PREFIX rsso: <http://www.geosysiot.in/rsso/ApplicationSchema#>
PREFIX fso: <http://www.geosysiot.in/fso/ApplicationSchema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <www.w3.org/1999/02/22-rdf-syntax-ns#>

INSERT { ?a fso:hasInferredLULC "AccessibleResidentialBuilding"^^xsd:string .
?a rdf:type fso:AccessibleResidentialBuilding }
WHERE {
        ?a fso:hasInferredLULC  "FloodedResidentialBuilding"^^xsd:string .
        ?b fso:hasInferredLULC "UnaffectedRoad"^^xsd:string .
        ?a geo:rcc8ec ?b
}
```

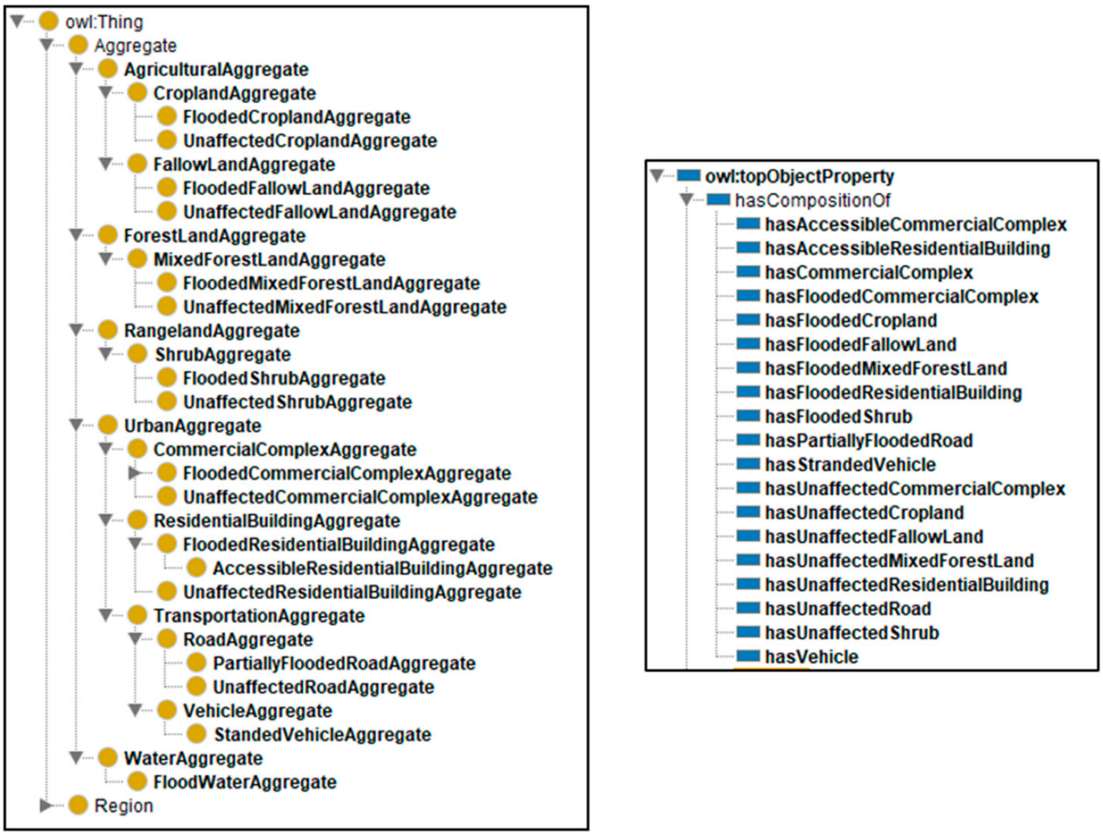
**Figure 11.** GeoSPARQL query depicting inferencing of "***AccessibleResidentialBuilding***" concept.

Spatio-Contextual Aggregation in Remote Sensing Scenes

A generic remote sensing scene consists of numerous individual regions depending on the scale and resolution of the scene, with each region belonging to a particular LULC class. A natural language scene description could entail to describe each of these individual regions. However, a scene description describing every single region in the scene in natural language would lead to severe information overload and would sabotage our primary objective of enhanced situational awareness through scene descriptions. Thus, there arises a need to aggregate regions in a way that conveys the necessary information pertaining to the scene to a user in a comprehensive yet concise manner. It is evident from the works of [29,30] in the area of human perception and psychology that we, humans, have a natural inclination towards grouping objects in scenes based on our understanding of their interactions (spatial and contextual) with one another. This phenomenon, termed as "perceptual grouping", was extended to regions in remote sensing scenes in the proposed Spatio-Contextual Triple Aggregation algorithm, to alleviate the issue of information overload. Reference [31] describes a methodology with a similar objective of discovering groups of objects in generic multimedia images for scene understanding. However, they propose a Hough-transform-based approach that automatically annotates object groups in generic multimedia images with bounding boxes. The proposed Spatio-Contextual Triple Aggregation algorithm groups triples in the semantically enriched Scene Knowledge Graphs to generate aggregates that reference multiple regions in a scene that are spatially and contextually similar.

The Flood Scene Ontology (FSO) defines aggregate classes on similar lines as the contextual classes defined as children of the "region" class discussed earlier. However, the instances of aggregate classes are intended to logically house multiple region class children instances through the "hasCompositionOf" object property.

Figure 12 depicts some of the object properties of the FSO. The "hasCompositionOf" object property has children properties for each of the corresponding class aggregates. For example, an instance of class "FloodedResidentialBuildingAggregate" would have multiple instances of "FloodedResidentialBuilding" connected through the object property—"hasFloodedResidentialBuilding". Thus, these object properties would help in mapping each of the aggregate instances to their component region instances. The "hasInferredAggregateName" Data Property allows the FSO to add contextually relevant names for the instance aggregates to use in the scene description. For example, multiple cars on a road can be aggregated and referenced as "traffic" or simply "vehicles" in the scene description. Such contextual knowledge from the perspective of remote sensing scene description was encoded in the FSO through SWRL rules.



- owl:Thing
  - Aggregate
    - AgriculturalAggregate
      - CroplandAggregate
        - FloodedCroplandAggregate
        - UnaffectedCroplandAggregate
      - FallowLandAggregate
        - FloodedFallowLandAggregate
        - UnaffectedFallowLandAggregate
    - ForestLandAggregate
      - MixedForestLandAggregate
        - FloodedMixedForestLandAggregate
        - UnaffectedMixedForestLandAggregate
    - RangelandAggregate
      - ShrubAggregate
        - FloodedShrubAggregate
        - UnaffectedShrubAggregate
    - UrbanAggregate
      - CommercialComplexAggregate
        - FloodedCommercialComplexAggregate
        - UnaffectedCommercialComplexAggregate
      - ResidentialBuildingAggregate
        - FloodedResidentialBuildingAggregate
          - AccessibleResidentialBuildingAggregate
        - UnaffectedResidentialBuildingAggregate
      - TransportationAggregate
        - RoadAggregate
          - PartiallyFloodedRoadAggregate
          - UnaffectedRoadAggregate
        - VehicleAggregate
          - StandedVehicleAggregate
    - WaterAggregate
      - FloodWaterAggregate
  - Region

- owl:topObjectProperty
  - hasCompositionOf
    - hasAccessibleCommercialComplex
    - hasAccessibleResidentialBuilding
    - hasCommercialComplex
    - hasFloodedCommercialComplex
    - hasFloodedCropland
    - hasFloodedFallowLand
    - hasFloodedMixedForestLand
    - hasFloodedResidentialBuilding
    - hasFloodedShrub
    - hasPartiallyFloodedRoad
    - hasStrandedVehicle
    - hasUnaffectedCommercialComplex
    - hasUnaffectedCropland
    - hasUnaffectedFallowLand
    - hasUnaffectedMixedForestLand
    - hasUnaffectedResidentialBuilding
    - hasUnaffectedRoad
    - hasUnaffectedShrub
    - hasVehicle

**Figure 12.** Snapshot of the Aggregate Children Classes and Object Properties in the Flood Scene Ontology that facilitate Spatio-Contextual Triple Aggregation.

The Sem-RSSU postulates the concept of Salient Region in a remote sensing scene. A salient region is defined as a region of significant importance from a spatio-contextual perspective in a scene. The concept of saliency proposed in Reference [32] was adapted to the remote sensing scene context in Sem-RSSU. The selection of the salient region in a scene depends on the saliency measures: (1) area it covers and (2) the LULC class it belongs to. A remote sensing scene may or may not have a salient region in it. A scene can have at most one salient region. The salient region is the primary region in the scene that is proposed to act as a reference to describe all the other regions in the scene. This facilitates planning and realization of the natural language scene description for the scene. In scenes that lack a salient region, the regions are aggregated with reference to the entire scene itself and the scene description is planned and realized accordingly.

The Salient Region Selection algorithm (Algorithm 1) facilitates the selection of the salient region in the scene. The algorithm filters regions in the Scene Knowledge Graphs based on the threshold value of area and the LULC set in the contextual Flood Scene Ontology to select the most salient region. The Sem-RSSU was designed to be modular, such that the contextual ontology (FSO in this case) houses the most relevant LULC and the threshold area value for the salient region, depending on the application. In the urban flood scenario from the perspective of remote sensing scene description, the “road” LULC was selected to be the most relevant class. Figure 12 depicts the snapshot of the aggregate classes and their corresponding object properties as formalized in the proposed Flood Scene Ontology (FSO).

**Algorithm 1** Salient Region Selection

```
Input:
    g: RS Scene Knowledge Graph
    fso: Contextual Ontology—Flood Scene Ontology
Output:
  sr: Salient Region
Constants:
     SalientRegionLULC ← LULC_Name—defined in Contextual Ontology—fso
     SalientRegionAreaThreshold ← Value—defined in Contextual Ontology—fso
function salientRegionSelection(g, fso)
    Read g
    Initialize SalientRegionCandidates: = []
    Initialize SalientRegionCandidateAreas: = []
    for Region in g:
         if Region.LULC == SalientRegionLULC and
              Region.Area > SalientRegionAreaThreshold then
              SalientRegionCandidates.append(Region)
                        SalientRegionCandidateAreas.append(Region.Area)
         else
              sr: = 0
              return sr
     maxArea: = max(SalientRegionCandidateAreas)
     maxAreainde: = SalientRegionCandidateAreas.index(maxArea)
     sr: = SalientRegionCandidates[maxAreaindex]
     return sr
end
```

Figure 13 depicts the visualization of a Spatio-Contextual Triple Aggregate (Algorithm 2) generated by the algorithm in the remote sensing scene. The algorithm aggregated all the regions with Inferred LULC as "FloodedResidentialBuilding" that are to the "West" direction of the salient region with Inferred LULC as "UnaffectedRoad". Thus, the regions visualized in red belong to a Spatio-Contextual Triple Aggregate with its "hasAggLULC" Data Property set to "FloodedResidentialBuildings". This mapping of aggregates to their individual composition regions with the "hasCompositionOf" object property facilitates grounded scene description rendering.

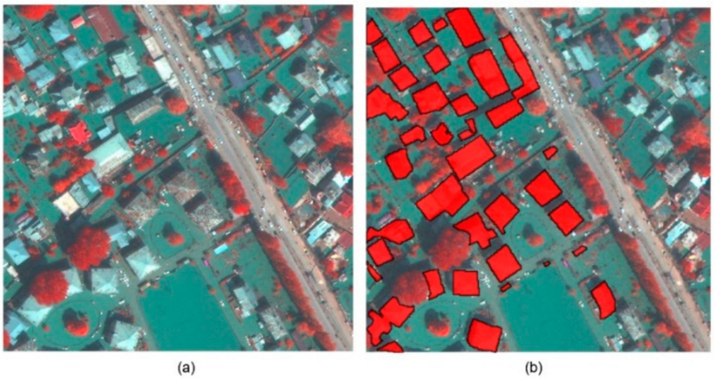


**Figure 13.** (**a**) Remote sensing scene captured during an urban flood. (**b**) Spatio-Contextual Triple Aggregate in the scene formed by aggregating spatial and contextual semantics—buildings (triples) that are flooded (contextual) and are to the West direction (spatial) of the road that is unaffected (contextual) by the flood.

**Algorithm 2** Spatio-Contextual Triple Aggregation

```
Input:
    g: RS Scene Knowledge Graph
    sr: Salient Region
    rsso: Spatial Ontology—Remote Sensing Scene Ontology
    fso: Contextual Ontology—Flood Scene Ontology
Output:
    $g'$: Enriched Scene Knowledge Graph with Spatio-Contextual Aggregates of LULC
    Regions
Constants:
    LeafNodes_LULC_Classes_List ← All Inferred LULC Class Names
    - defined in Contextual Ontology—fso
    LeafNodes_SpatialRelations_List ← All Inferred Spatial Relationships
    Topological and Directional (Object Properties)
     - defined in Spatial Ontology—rsso
    - defined in Contextual Ontology—fso
function spatioContextualTripleAggregation(g, sr, rsso, fso)
    Read g
    $g' := g$
    for RegionLULC in LeafNodes_LULC_Classes_List:
        for SpatialRelation in LeafNodes_SpatialRelations_List:
            Initialize SCAggregate = []
            for Region in $g'$:
                if Triple <Region, SpatialRelation, sr> in $g'$ and
                    Region.hasInferredLULC == Region.LULC then
                        SCAggregate.append(Region)
            if len(SCAggregate) > 1 then
                Insert Triple <sr, SpatialRelation, SCAggregate> into $g'$
                Insert Triple <SCAggregate, hasAggLULC, RegionLULC> into $g'$
                for Region in SCAggregate:
                    Insert Triple <SCAggretate, hasCompositionOf, Region> into $g'$
    return $g'$
end
```

### 2.3. Natural Language Processing Layer

Most state-of-the-art research [33–35] in image captioning deals with describing an image with a single sentence. This is feasible due to the fact that the images used in these studies are generic multimedia images and can be adequately summarized with a single sentence. Recent research [12,13,17] in the area of remote sensing image captioning deals with describing a scene in a single sentence, however such description is not comprehensive and does not convey the context of the scene in its entirety. Remote sensing scenes contain numerous objects of importance that spatially interact with each other, and thus cannot be comprehensively summarized in a single sentence. Moreover, in remote sensing scenes, the context of the event when the scene was captured plays a crucial role in the scene description. A simple existential description merely informing the existence of all the objects in a scene is undesirable too. Instead a detailed contextual description of the objects in a scene based on their directional and topological interaction with one another from a contextual perspective of the event is desirable. The Natural Language Processing (NLP) layer of the Sem-RSSU framework was designed to meet this objective of comprehensive explainable and grounded spatio-contextual scene descriptions in natural language.

The Scene Knowledge Graph enriched with spatial and contextual semantics and having been aggregated by the Spatio-Contextual Triple Aggregation algorithm serves as an input to the Natural Language Processing layer. The individual tasks, as identified by Reference [36], for a generic natural language generation system comprise (1) content determination, (2) document structuring, (3) aggregation, (4) lexicalization, (5) expression generation and (6) realization. The natural language generation for scene descriptions in

Sem-RSSU loosely follows this approach. The Semantic Enrichment layer in Sem-RSSU inherently performs the tasks of content determination and document structuring by generating and enriching the Scene Knowledge Graph. The Spatio-Contextual Aggregation algorithm in the Semantic Enrichment layer further aggregates the graph for scene description rendering, thus performing aggregation as one of the NLG tasks. The tasks of Lexicalization, Expression Generation and Realization are performed as a part of the NLP layer in Sem-RSSU through the Grounded Scene Description Planning and Realization (GSDPR) algorithm (Algorithm 3).

The proposed Grounded Scene Description Planning and Realization (GSDPR) is the primary algorithm that is geared towards describing the regions in the scene from a spatio-contextual perspective with the salient region as a reference, using a template-based approach. The orientation of the salient region is initially determined with Salient Region Orientation Detection algorithm (Algorithm 4). It provides the primary algorithm with the information whether the salient region is oriented in the North-South or East-West direction. This is determined using the coordinates of the bounding box of the salient region geometry. This information is crucial while describing other regions in the scene with reference to the salient region.

The GSDPR algorithm initially checks for the existence of the salient region (SR) in the Scene Knowledge Graph. On detection of the SR, it detects its orientation. Depending on the orientation of the SR, it further proceeds to describe the Spatio-Contextual Triple Aggregates (determined in the Semantic Enrichment layer) against the directional orientation of the SR. If the algorithm detects that a Scene Knowledge Graph lacks a salient region, then it proceeds ahead in a similar fashion considering the entire scene geometry as a salient region and describing the regions with reference to the scene itself. For example, with a salient region "UnaffectedRoad" oriented in the "North-South" direction, All the other regions such as "FloodedResidentialBuildings", "StrandedVehicles", "FloodedVegetation", etc., are described in reference to the "UnaffectedRoad" with the "East-West" direction. However, the triple aggregates whose geometry intersects with the geometry of salient region, are described along the direction of the SR orientation. For example, "Vehicles" or "Traffic" on the "UnaffectedRoad" oriented in the "North-South" direction, would be described in the "North-South" direction as well. These conditions, although specific, seem to generalize well for remote sensing scenes. The "describe" function in the GSDPR algorithm uses a templating mechanism based on the proposed Remote Sensing Scene Description Grammar G. It returns a sentence with appropriate natural language constructs describing the input triples passed to it. The "VisualizeAndMap" function is responsible for the visualization of color-coded regions mapping to the sentences generated by the "describe" function. This is implemented by visualizing the geometries of individual regions belonging to Spatio-Contextual Triple Aggregate mapped by the "hasCompositonOf" object property in the Scene Knowledge Graph. Thus, atomically describing the Spatio-Contextual Triple Aggregates and mapping their constituent regions aids in rendering grounded natural language scene descriptions.

**Algorithm 3** Grounded Scene Description Planning and Realization

```
Input:
    g: RS Scene Knowledge Graph
    rsso: Spatial Ontology—RS Scene Ontology
    fso: Contextual Ontology—Flood Scene Ontology
Output:
    sd: List of Natural Language Sentences as Scene Description
    groundedRegions: List of Grounded Regions Geometries corresponding to generated Natural
    Language Scene Description
Constants:
   LeafNodes_LULC_Classes_List ← All Inferred LULC Class Names
                     - defined in Contextual Ontology—fso
    EWDir ← ["East", "West"]
                     - defined in Spatial Ontology—rsso
   NSDir ← ["North", "South"]
                     -defined in Spatial Ontology—rsso
function SceneDescriptionPlanAndRealization(g, rsso, fso, rssao)
    Read g
    sr: = salientRegionSelection(g, fso)
    if len(sr) > 0 then
      //Case with 1 Salient Region
      srOrientation: = salientRegionOrientationDetection(g, sr)
      if srOrientation == 'NS' then
          alongSROrientation: = NSDir
          againstSROrientation: = EWDir
      else
          alongSROrientation: = EWDir
          againstSROrientation: = NSDir
      sd.append(describe(sr))
      groundedRegions.append(visualizeAndMap(sr))
      $g'$ = spatioContextualTripleAggregation(g, sr, rsso, fso)
      for SCTripleAggregate in $g'$:
        //Check if the SCTripleAggregate intersects with SalientRegion
        if ntpp(SCTripleAggregate.geometry, sr.geometry):
            direction: = alongSROrientation
        else:
            direction: = againstSROrientation
        for currentLULC in LeafNodes_LULC_Classes_List:
          //Check if SCTripleAggregate has currentLULC and is oriented in direction w.r.t. SR
          //If found then -
              sd.append(describe(SCTripleAggregate, direction, sr))
              groundedRegions.append(visualizeAndMap(SCTripleAggregate))
    else:
      //Case with 0 Salient Region
      sr: = "thisScene"
      $g'$ = spatioContextualTripleAggregation(g, sr, rsso, fso)
      for SCTripleAggregate in $g'$:
        for currentLULC in LeafNodes_LULC_Classes_List:
          //Check if SCTripleAggregate has currentLULC and is oriented in each of the Cardinal
              Directions w.r.t SR—Scene
          //If found then
            sd.append(describe(SCTripleAggregate, CardinalDirection, sr))
            groundedRegions.append(visualizeAndMap(SCTripleAggregate))
    end;
```

**Algorithm 4** Salient Region Orientation Detection

```
Input:
    g: RS Scene Knowledge Graph
    sr: Salient Region
Output:
    orientation: Salient Region Orientation
function salientRegionOrientationDetection(g, sr)
    Read g
    Get BoundingBox Coordinates—LLX, LLY, URX and URY from g
    Compute size of the Horizontal Side as horizontal and Vertical Side as vertical of the
    Bounding Box
    if horizontal > vertical then
        orientation = "EW"
    else
        orientation = "NS"
    return orientation
end
```

The algorithms developed in this research were implemented in Python. The RDFLib python library was used for facilitating the use of GeoSPARQL over the generated Scene Knowledge Graphs. Shapely, Descartes and Matplotlib libraries were used for rendering geometry visualizations for grounded scene descriptions.

The natural language scene descriptions generated by the NLP layer in Sem-RSSU conform to the following Remote Sensing Scene Description Grammar G. G is a Context-Free Grammar (CFG) in the research area of linguistics and is used to generate restricted languages. The grammar defines production rules specific to remote sensing scene description application context using the parameters T, N, S and R (Algorithm 5).

The natural language scene description of remote sensing scenes generated by G can be defined as follows:

$$\mathrm{L(G)} = \{x \in \mathrm{T}^* \mid \mathrm{S} \Rightarrow^* x\} \tag{8}$$

The grammar is invoked for every call to the "describe" function in the GSDPR algorithm. Thus, the natural language scene descriptions rendered by Sem-RSSU for remote sensing scenes can be derived and verified by parsing the proposed Remote Sensing Scene Description Grammar G.

The Rendering layer involves the generated Natural Language Scene Description in text form and its corresponding color-coded visualization depicting grounded mappings of the sentences in the description to the regions in the scene. This reinforces the explainability of the generated scene descriptions by Sem-RSSU. The figure below depicts the front-end of a web-based application implementing the Sem-RSSU framework to browse, preview, render grounded scene descriptions, query and visualize remote sensing scenes.

The web-based application depicted in Figure 14, uses Python with RDFLib and GraphDB triple-store at the back-end with REST based API calls originating from the user interactions and renders the scene descriptions, responses and visualizations over the web page.

**Algorithm 5** Remote Sensing Scene Description Grammar G

G = (T, N, S, R)

*T* is a finite alphabet of Terminals

*N* is a finite set of Non-Terminals

*S* is the Start Symbol and *S* ϵ *N*

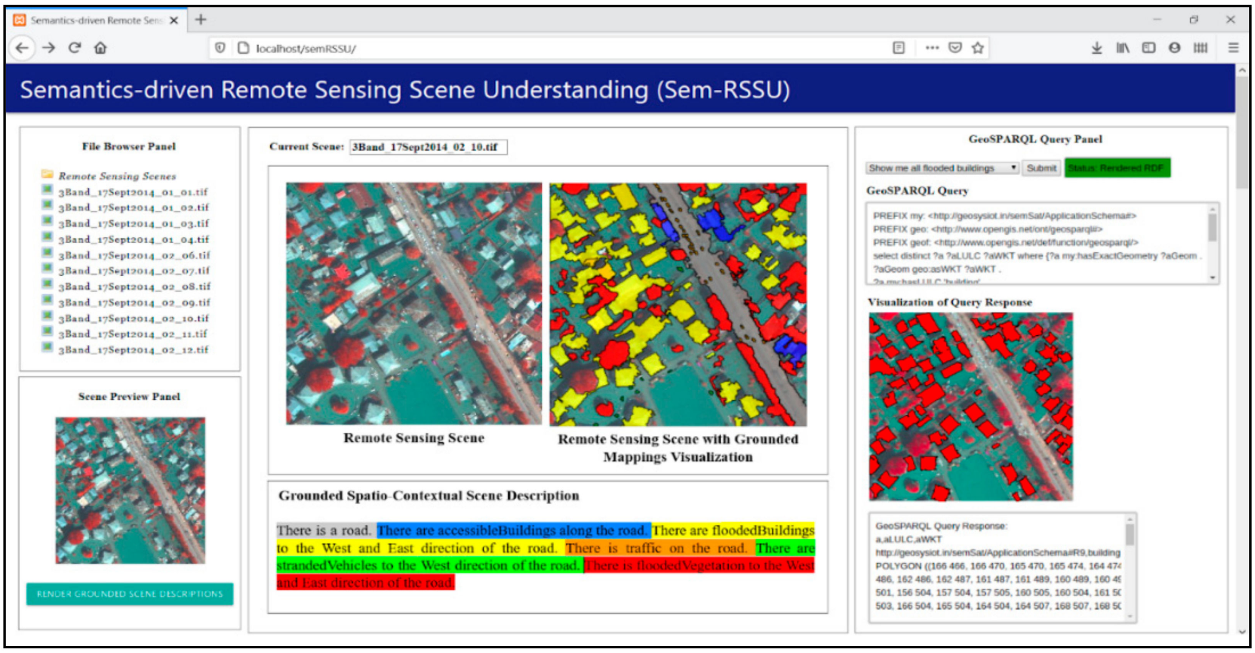
*R* is the finite set of Production Rules of the form $N \rightarrow (N \cup T)^*$

*T* = {*flooded buildings, accessible buildings, unaffected buildings, unaffected roads, vehicles, stranded vehicles, traffic, flooded vegetation, unaffected vegetation, road, scene, spread across, to the, of, on, east, west, north, south, there, the, a, an, is, are*}

*N* = {*ZeroSalientRegion, OneSalientRegion, DescribeSalientRegion, Pronoun, AuxiliaryVerb, Article, SalientRegion, DescribeSpatioContextualAggregate, SpatialReference, DirectionalReference*}

*S* = *DescriptionSentence*

```
R = {
<DescriptionSentence> → <ZeroSalientRegion> ∣ <OneSalientRegion>
<OneSalientRegion> → <DescribeSalientRegion> ∣ <DescribeSpatioContextualAggregate>
<ZeroSalientRegion> → <DescribeSpatioContextualAggregate>
<DescribeSalientRegion> → <Pronoun> <AuxiliaryVerb> <Article> <SalientRegion>
<DescribeSpatioContextualAggregate> → <Pronoun> <AuxiliaryVerb> <SpatioContextualAggregates> <SpatialReference> . <SalientRegion>
<SpatioContextualAggregates> → flooded buildings ∣ accessible buildings ∣ unaffected buildings ∣ unaffected roads ∣ vehicles ∣ stranded vehicles ∣ traffic ∣ flooded vegetation ∣ unaffected vegetation
<SalientRegion> → road ∣ scene
<SpatialReference> → spread across <Article> ∣ to the <DirectionalReference> of <Article> ∣ on <Article>
<DirectionalReference> → east ∣ west ∣ north ∣ south
<Pronoun> → there
<Article> → the ∣ a ∣ an
<AuxiliaryVerb> → is ∣ are
}
```



Semantics-driven Remote Sensing Scene Understanding (Sem-RSSU)

File Browser Panel

*Remote Sensing Scenes*
- 3Band_17Sept2014_01_01.tif
- 3Band_17Sept2014_01_02.tif
- 3Band_17Sept2014_01_03.tif
- 3Band_17Sept2014_01_04.tif
- 3Band_17Sept2014_02_06.tif
- 3Band_17Sept2014_02_07.tif
- 3Band_17Sept2014_02_08.tif
- 3Band_17Sept2014_02_09.tif
- 3Band_17Sept2014_02_10.tif
- 3Band_17Sept2014_02_11.tif
- 3Band_17Sept2014_02_12.tif

Scene Preview Panel

RENDER GROUNDED SCENE DESCRIPTIONS

Current Scene: 3Band_17Sept2014_02_10.tif

Remote Sensing Scene

Remote Sensing Scene with Grounded Mappings Visualization

Grounded Spatio-Contextual Scene Description

There is a road. There are accessibleBuildings along the road. There are floodedBuildings to the West and East direction of the road. There is traffic on the road. There are strandedVehicles to the West direction of the road. There is floodedVegetation to the West and East direction of the road.

GeoSPARQL Query Panel

Show me all flooded buildings ▾ Submit Status: Rendered RDF

GeoSPARQL Query

```
PREFIX my: <http://geosysiot.in/semSat/ApplicationSchema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
select distinct ?a ?aLULC ?aWKT where {?a my:hasExactGeometry ?aGeom .
?aGeom geo:asWKT ?aWKT .
?a my:hasLULC 'building'
```

Visualization of Query Response

```
GeoSPARQL Query Response:
a,aLULC,aWKT
http://geosysiot.in/semSat/ApplicationSchema#R9,building
POLYGON ((166 466, 166 470, 165 470, 165 474, 164 474
486, 162 486, 162 487, 161 487, 161 489, 160 489, 160 49
501, 156 504, 157 504, 157 505, 160 505, 160 504, 161 50
503, 166 504, 165 504, 164 504, 164 507, 168 507, 168 50
```

**Figure 14.** Web-based application front-end implementing the Sem-RSSU framework.

## 3. Experimental Setup and Results

The Sem-RSSU framework was designed and implemented to enhance the situational awareness from remote sensing scenes through the rendering of grounded spatio-contextual scene descriptions in natural language. Having realized the importance of enhanced situational awareness from remote sensing scenes during a dynamic disaster, such as a flood, we selected it as the test scenario for demonstrating the utility of Sem-RSSU. This section discusses the dataset and the results obtained including multi-class segmentation of remote sensing scenes and the grounded natural language scene description rendered by Sem-RSSU. It also discusses the different evaluation strategies employed to evaluate Sem-RSSU at different stages in the framework. The section concludes with a brief discussion of the results obtained with the framework, analysis, observations and scope of improvement for Sem-RSSU.

### *3.1. Data Description*

Multispectral and Panchromatic satellite imagery captured by WorldView-2, with a ground sampling distance of 0.5 m of the Srinagar area in India, during the floods of September 2014, was used for this work. Pansharpening operation was performed to fuse the two satellite products—panchromatic and multispectral—and the resultant imagery was selected to be used by Sem-RSSU. The satellite imagery spanning over an area of 25 sq. km on the ground was split into 512 × 512 pixel sized scenes to be used by Sem-RSSU framework for grounded spatio-contextual scene description rendering. Figure 15 depicts a couple of satellite scenes from the selected dataset. The training data required for multi-class segmentation of the scene, to be consumed by the neural network, were generated by annotating the satellite imagery.

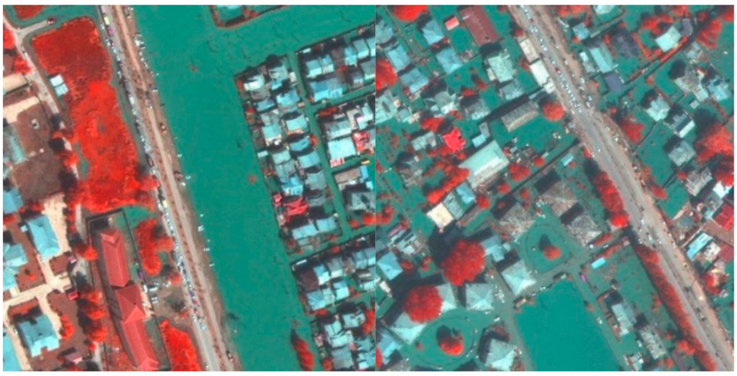


**Figure 15.** Remote sensing scenes depicting urban floods captured by WorldView-2 of Sringar, India, during the floods of September 2014.

The annotation of satellite imagery was performed manually in accordance with the principles of remote sensing image interpretation. A total of seven classes—"ResidentialBuilding", "Road", "Shrub", "Shadow", "FloodWater", "Vehicle" and "FallowLand"—were selected to be annotated for multi-class segmentation of the urban floods dataset. A collection of 150 annotated scenes was considered for this study. The data split of 70% and 30% was used for training and validation of the deep learning models. Figure 16 depicts a satellite scene from the dataset, with its corresponding annotated ground truth.

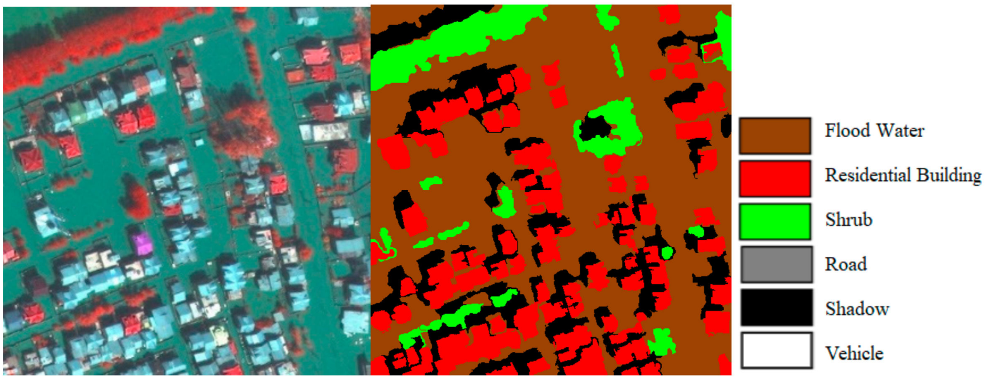


**Figure 16.** Urban flood satellite scene with its corresponding annotated ground truth.

## 3.2. Results and Evaluation

The experimental results and their corresponding evaluations for the different stages in the Sem-RSSU framework are discussed in this section.

### 3.2.1. Multi-Class Segmentation Results of Urban Flood Scenes

Different state-of-the-art deep-learning architectures were experimented with for multi-class segmentation of the urban floods dataset. The deep neural network models were re-trained from scratch over the annotated urban floods dataset without using any pre-trained model weights. The evaluation metrics of Intersection over Union (IoU)—class-wise, mean and frequency-weighted IoU—were used, in addition to the overall accuracy, to evaluate the models.

From the experiments, it was found that the SegNet, with ResNet as its backbone in the deep neural network architecture, produced the best overall accuracy, 89.74%, while SegNet with VGG-16 as its backbone produced the best mean IoU and frequency-weighted IoU of 0.5299 and 0.6702, respectively. Figure 17 depicts the result of multi-class segmentation as predicted by the SegNet with ResNet architecture.

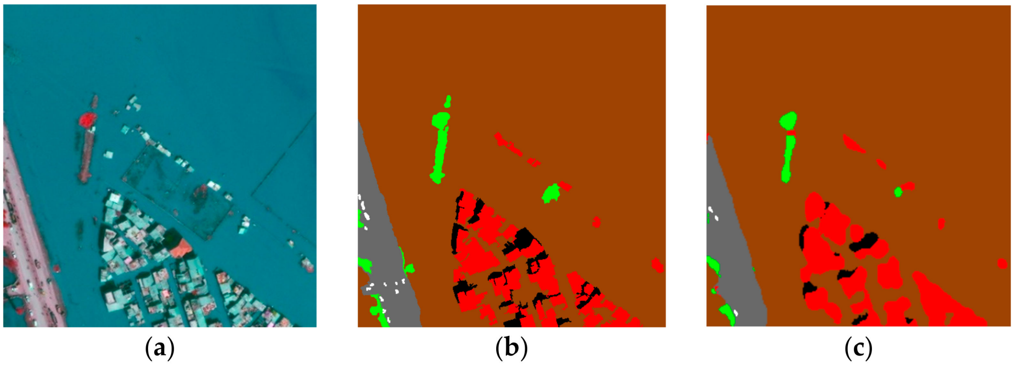


**Figure 17.** (**a**) Remote sensing scene of urban floods. (**b**) Manually annotated ground truth for the scene. (**c**) Multi-class segmentation result predicted by using SegNet with ResNet architecture.

From the consistent lower values of class-wise IoU in Table 1 for the "Vehicle" class for all the deep-learning architectures, it is evident that it is the most difficult to predict, given the small number of pixels for it in the training data. Similarly, from the consistent higher values of class-wise IoU for the "ResidentialBuildings" and the "FloodWater" classes, these classes are relatively easier for the deep neural networks to predict given their distinct spectral signature and relatively higher number of pixels in the training data. The neural network architectures was implemented in Python, using the Keras deep-learning library. The models were iteratively trained and validated over Nvidia Tesla P100 GPU.

**Table 1.** Class-wise and overall evaluation of the deep-learning-based multi-class segmentation over urban-floods remote sensing scenes.

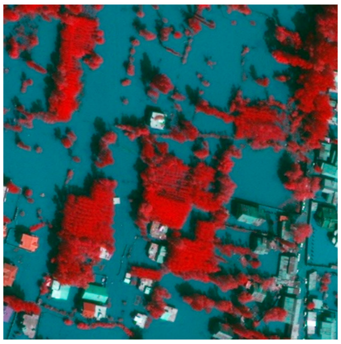
| Model | Class-Wise IoU | | | | | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Residential Building | Road | Flood Water | Shrub | Shadow | Vehicle | Fallow Land | Mean IoU | Frequency-Weighted IoU | Accuracy |
| U-Net with VGG | 0.421 | 0.102 | 0.406 | 0.646 | 0.39 | 0.0042 | 0.125 | 0.299 | 0.429 | 0.762 |
| U-Net with ResNet | 0.606 | 0.185 | 0.465 | 0.595 | 0.442 | 0.0006 | 0.028 | 0.332 | 0.502 | 0.8085 |
| FCN with VGG | 0.491 | 0.4756 | 0.5838 | 0.6152 | 0.3429 | 0.0055 | 0.2216 | 0.3908 | 0.52.8 | 0.7772 |
| FCN with ResNet | 0.4705 | 0.3556 | 0.5271 | 0.5733 | 0.2861 | 0.0081 | 0.009 | 0.3185 | 0.4711 | 0.7432 |
| PSPNet with VGG | 0.5801 | 0.5748 | 0.4836 | 0.3699 | 0.3149 | 0.0019 | 0.3328 | 0.3797 | 0.4739 | 0.8224 |
| PSPNet with ResNet | 0.5497 | 0.6154 | 0.634 | 0.5715 | 0.2798 | 0.0419 | 0.3716 | 0.4377 | 0.5507 | 0.8279 |
| SegNet with VGG | 0.6572 | 0.6415 | 0.7918 | 0.6669 | 0.3806 | 0.0018 | 0.5696 | **0.5299** | **0.6702** | 0.8625 |
| SegNet with ResNet | 0.5512 | 0.4232 | 0.5728 | 0.5936 | 0.3751 | 0.00129 | 0.1578 | 0.382 | 0.5304 | **0.8974** |

#### 3.2.2. Evaluation of Ontologies in Sem-RSSU

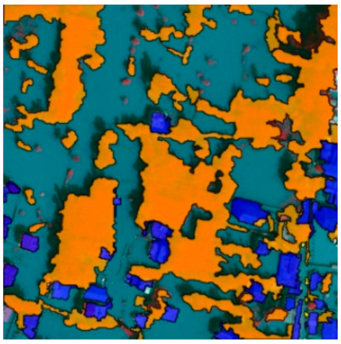
The Remote Sensing Scene Ontology (RSSO) and the Flood Scene Ontology were proposed and extensively used in the Sem-RSSU framework. The ontologies conform to the principle of clarity. They were designed for use in the domain of remote sensing scene understanding and use commonly used relevant terms for concepts and relationships in this domain. The ontologies conform to the principle of coherence. This is validated by visualization of inferred statements (region triples) as geometries in Sem-RSSU thus indicating correctness of the statements. The ontologies are extendible. This is reinforced by the fact that the Flood Scene Ontology itself extends the Remote Sensing Scene Ontology. The Flood Scene Ontology too can be extended and re-used for another application scenario. The ontologies exhibit minimum encoding bias. The ontologies were tested and found to be amenable for being consumed by different libraries for inferencing. The ontologies observe minimum ontological commitment, thereby enabling application developers to build, extend and reuse them in their applications. Thus, the ontologies were evaluated against and were found to satisfy the principles of ontology design proposed by Reference [37]. The ontologies were developed by using Protégé (https://protege.stanford.edu/) Ontology Editor.

#### 3.2.3. Grounded Spatio-Contextual Natural Language Scene Description Rendering

Figures 1 and 18 depict some of the natural language scene descriptions generated by Sem-RSSU, along with the remote sensing scenes and corresponding grounded visualizations. Sem-RSSU renders grounded mappings visualizations by color-coding the regions, to match their corresponding sentences in the scene descriptions.



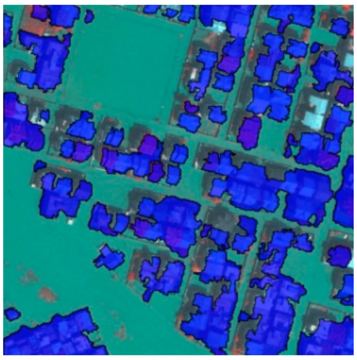
**Remote Sensing Scene**



**Remote Sensing Scene with Grounded Mappings Visualization**

**Grounded Spatio-Contextual Scene Description in Natural Language:** There are floodedBuildings spread across the scene. There are strandedVehicles to the South-East direction in the scene. There is floodedVegetation spread across the scene.

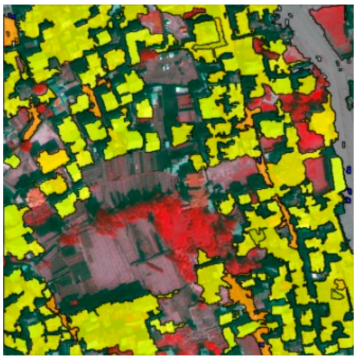
**Remote Sensing Scene**



**Remote Sensing Scene with Grounded Mappings Visualization**

**Grounded Spatio-Contextual Scene Description in Natural Language:** There are floodedBuildings spread across the scene.

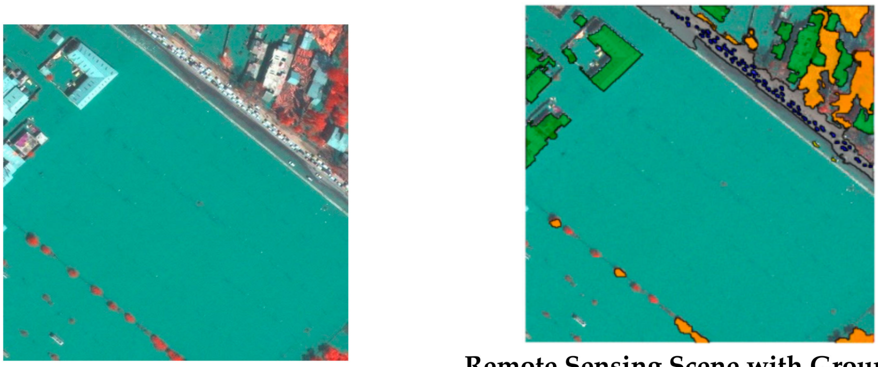
**Remote Sensing Scene**



**Remote Sensing Scene with Grounded Mappings Visualization**

**Grounded Spatio-Contextual Scene Description in Natural Language:** There is a road. There is traffic on the road. There are unaffectedBuildings to the West and East direction of the road. There are unaffectedRoads to the West direction of the road.

**Figure 18.** *Conts.*



**Remote Sensing Scene**

**Remote Sensing Scene with Grounded Mappings Visualization**

**Grounded Spatio-Contextual Scene Description in Natural Language:** There is a road. There is traffic on the road. There are strandedVehicles to the South direction of the road. There is floodedVegetation to the North and South direction of the road. There are floodedBuildings to the North and South direction of the road.

**Figure 18.** Remote sensing scenes with their corresponding grounded mapping visualizations and scene descriptions, as rendered by Sem-RSSU.

#### 3.2.4. Evaluation of Grounded Spatio-Contextual Natural Language Scene Descriptions

To the best of our knowledge, this research is the first of its kind in rendering comprehensive spatio-contextual scene descriptions from remote sensing scenes. Existing research studies in this area do not address this issue of comprehensive scene descriptions. In that regard, there is a lack of benchmark datasets involving remote sensing scenes with their corresponding comprehensive multi-sentence scene descriptions. Due to this shortcoming, this research validates the Sem-RSSU framework against a validation dataset of urban flood scenes with their corresponding manually transcribed scene descriptions. For comprehensive evaluation of the generated grounded natural language scene descriptions, they were evaluated by using a two-pronged strategy: (1) Automatic Evaluations, using widely accepted evaluation metrics, such as Bilingual Evaluation Understudy (BLEU) [38], Metric for Evaluation of Translation with Explicit Ordering (METEOR) [39], ROUGE_L [40] and Consensus-based Image Description Evaluation (CIDEr) [41]; and (2) Human Evaluations, using a set of proposed evaluation metrics to account for naturalness and quality in the scene descriptions.

##### Automatic Evaluations for Spatio-Contextual Natural Language Scene Descriptions

A corpus of remote sensing scenes from the urban flood dataset was selected for manually transcribing the scene description of scenes for the purpose of evaluation. This corpus of 50 remote sensing scenes and their corresponding manually generated scene description was used for Automatic Evaluation of the natural language scene descriptions generated by Sem-RSSU.

The Bilingual Evaluation Understudy (BLEU) computes a modified version of precision for quantifying the similarity between the machine generated text and reference text. The BLUE_1 to BLEU_4 refer to the n gram overlap between the machine generated text and the reference text. The Metric for Evaluation of Translation with Explicit Ordering (METEOR) computes the similarity between the texts based on harmonic mean of one-gram precision and recall. The Recall-Oriented Understudy for Gisting Evaluation (ROGUE_L) computes similarity between texts based on the longest common subsequence; while the Consensus-based Image Description Evaluation (CIDEr) quantifies the consensus of image captions by taking into account the precision and recall in addition to using the TF-IDF for every n-gram.

Table 2 depicts the evaluation metric scores for the scene descriptions generated by Sem-RSSU for the urban flood dataset. From the table, it is evident that the Sem-RSSU with Segnet consistently produces the highest scores for all evaluation metrics. This correlates well with the fact that the SegNet architecture produces the best accuracy and IoU for multi-class segmentation. Thus, it is evident that multi-class segmentation impacts the quality of the scene descriptions generated by Sem-RSSU.

**Table 2.** Automatic Evaluation metrics for natural language scene description rendering.

| | BLEU_1 | BLEU_2 | BLEU_3 | BLEU_4 | METEOR | ROUGE_L | CIDEr |
|---|---|---|---|---|---|---|---|
| **Sem-RSSU with UNet** | 0.3975 | 0.3635 | 0.3352 | 0.3087 | 0.2288 | 0.5169 | 0.1177 |
| **Sem-RSSU with FCN** | 0.6145 | 0.5723 | 0.5215 | 0.4812 | 0.4056 | 0.5945 | 1.1687 |
| **Sem-RSSU with PSPNet** | 0.6287 | 0.5793 | 0.5284 | 0.4798 | 0.3946 | 0.6428 | 1.4586 |
| **Sem-RSSU with SegNet** | **0.6395** | **0.5822** | **0.5343** | **0.4853** | **0.3987** | **0.6468** | **1.5776** |

Human Evaluation of Grounded Spatio-Contextual Natural Language Scene Descriptions

The Human Evaluation of the grounded spatio-contextual natural language scene descriptions was performed by using a scale-based approach, where human participants were asked to score the scene descriptions rendered by Sem-RSSU based on several parameters. For evaluating the scene descriptions comprehensively, the Human Evaluation was performed in two stages: (1) Subjective Evaluation of the Spatio-Contextual Natural Language Scene Descriptions, focusing on Natural Language Constructs and Readability, and (2) Subjective Evaluation of Grounded Mappings and Spatio-Contextual Information conveyed in the Scene Descriptions. The scores in the tables below were graded individually for the corpus of 50 remote sensing scenes by a group of seven remote-sensing researchers at IIT Bombay. The researchers are experienced and adept in the domain of remote sensing image interpretation and have full professional proficiency in English language. The scores graded by the human participants were averaged and scaled to range from 0 to 1 for uniformity.

Subjective Evaluation of the Spatio-Contextual Natural Language Scene Descriptions

The parameters [42] readability, accuracy, adequacy and relevance were selected for subjective evaluation of the scene descriptions. Readability accounts for the naturalness of the descriptions while accuracy or correctness, adequacy and relevance account for informativeness in the scene descriptions. From the Table 3, it is evident that the scene descriptions generated by Sem-RSSU were largely deemed as readable and accurate by the human participants, however the adequacy and relevance seemed fair to the participants.

**Table 3.** Human Evaluation metrics for natural language scene description rendering.

| | Readability | Accuracy | Adequacy | Relevance |
|---|---|---|---|---|
| **Sem-RSSU with SegNet** | 0.875 | 0.8 | 0.7 | 0.725 |

Subjective Evaluation of Grounded Mappings and Spatio-Contextual Information

The subjective evaluation of grounded mappings and spatio-contextual information led to the development of requirement specific parameters for evaluation of remote sensing scene descriptions. The generic evaluation parameters to gauge the quality of descriptions in the research area of natural language generation—adequacy and accuracy or correctness—were modified to suit the context of remote sensing scene descriptions. The proposed parameters with their definitions as displayed to the human participants are as follows:

**Grounding Correctness**: the accuracy or correctness of the mappings between the regions in the scene and the sentences in the scene description for the remote sensing scene.

**Directional Correctness**: the accuracy or correctness of the cardinal directions with respect to the reference region mentioned in the scene description for the remote sensing scene.

**Contextual Correctness**: the accuracy or correctness of the context conveyed by the scene description for the remote sensing scene.

**Topological Correctness**: the accuracy or correctness of the topological relations between the regions mentioned in the scene description for the remote sensing scene.

**Grounding Adequacy**: the adequacy of the mappings between the regions in the scene and the sentences in the scene description to comprehensively describe the remote sensing scene.

**Directional Adequacy**: the adequacy of the directional relations between the regions mentioned in the scene description to comprehensively describe the directions in the remote sensing scene.

**Contextual Adequacy**: the adequacy of the context mentioned in the scene description to comprehensively convey the context of the remote sensing scene.

**Topological Adequacy**: the adequacy of the topological relations between the regions mentioned in the scene description to comprehensively describe the topology of the remote sensing scene.

From the scores in Table 4, it is evident that grounding, context and topology were aptly conveyed in the scene descriptions generated by Sem-RSSU however the overall consensus in terms of directional relationships in the scene descriptions is fair as scored by the human participants.

**Table 4.** Human Evaluation metrics for grounded natural language scene description rendering.

| | Grounding Correctness | Directional Correctness | Contextual Correctness | Topological Correctness | Grounding Adequacy | Directional Adequacy | Contextual Adequacy | Topological Adequacy |
|---|---|---|---|---|---|---|---|---|
| Sem-RSSU with SegNet | 0.92 | 0.7 | 0.82 | 0.85 | 0.85 | 0.72 | 0.85 | 0.87 |

## 4. Discussion

The Sem-RSSU framework was meticulously evaluated by using relevant evaluation strategies at different stages of its operation. The multi-class segmentation of remote sensing scenes for the urban flood dataset in the Data Mediation layer of the framework was implemented by experimenting with different state-of-the-art deep neural network architectures. From the experiments, it was observed that the SegNet architecture with ResNet backbone produces the best results over the urban floods dataset in terms of accuracy and mean IoU. From the experiments, it was found that the Sem-RSSU framework is best suited for Very High Resolution remote sensing scenes considering the importance of effective multi-class segmentation in the framework. However, it was also noted that minor inaccuracies in the classification map generated by the segmentation component did not have a significant impact on the natural language scene descriptions, primarily due to the Spatio-Contextual Triple Aggregation algorithm, which groups multiple objects in a scene as “aggregates” based on their spatio-contextual relationship with the salient region. The ontologies Remote Sensing Scene Ontology and the Flood Scene Ontology, used extensively in Sem-RSSU, were evaluated in accordance with the principles of Ontology Design, as proposed by Reference [37]. The grounded spatio-contextual scene descriptions rendered by the Sem-RSSU were extensively evaluated by using a two-pronged strategy—Automatic Evaluations (Objective) and Human Evaluations (Subjective). The metrics BLEU, METEOR, ROGUE_L and CIDEr were employed for evaluation of the scene descriptions generated by Sem-RSSU with different deep neural network architectures for multi-class segmentation. It was observed that the Automatic Evaluation metrics consistently scored the highest for the SegNet with ResNet architecture, thus reinforcing the significance of multi-class segmentation. It should be noted that the Sem-RSSU framework was designed in a layered manner to be modular and can thus facilitate use of any state-of-the-art approach for multi-class segmentation for best results. The Human Evaluation of the rendered scene descriptions show promising results however they do highlight the minor imperfections in terms of adequacy and relevance specific to the directional relationships conveyed in the descriptions.

From the experimental analysis, it is understood that the directional relationships inferred by Sem-RSSU are imperfect in cases involving elongated and irregular regions due to the use of the centroid based approach in inferencing of directional relationships in the Semantic Enrichment layer. A more robust approach that takes into account the holistic geometries of regions needs to be researched on for inferencing of directional relationships. The Sem-RSSU at its core uses the Spatio-Contextual Triple Aggregation and the Grounded Scene Description Planning and Realization algorithms. The latter relies on a templating mechanism to generate natural language scene descriptions from the aggregated triples, the salient region and the directional information passed to it. Considering the limited number of natural language constructs necessary for comprehensively describing remote sensing scenes, a template-based approach seemed prudent for the development of Sem-RSSU framework. However, with Neural Machine Translation approaches being widely used, it would be intriguing to explore a neural approach to translate Scene Knowledge Graphs to natural language scene descriptions.

## 5. Conclusions and Future Directions

The Semantics-driven Remote Sensing Scene Understanding (Sem-RSSU) framework presented in this paper aims for enhanced situational awareness from remote sensing scenes through the rendering of comprehensive grounded natural language scene descriptions from a spatio-contextual standpoint. Although the flood disaster was chosen as a test scenario for demonstrating the utility of comprehensive scene understanding, Sem-RSSU can also be applied to monitoring other disasters, such as earthquakes, forest fires, hurricanes, landslides, etc., as well as urban sprawl analysis and defense-related scenarios, such as hostile surveillance in conflicted zones. It is envisaged that Sem-RSSU would lay the foundation for semantics-driven frameworks for natural language scene description rendering for comprehensive information dissemination for application scenarios such as disasters, surveillance of hostile territories, urban sprawl monitoring, etc., among other remote sensing applications.

The framework proposes the amalgamation of a deep learning and a knowledge-based approach, thereby leveraging (1) deep learning for multi-class segmentation and (2) deductive reasoning for mining implicit knowledge. The framework advocates the transformation of remote sensing scenes to Scene Knowledge Graphs formalized through the development of Remote Sensing Scene Ontology (RSSO). The ontology models the representation of a generic remote sensing scene in the form of knowledge graphs by defining concepts related to the scene's lineage and land-use/land-cover regions and the spatial relationships between them. The contextual Flood Scene Ontology developed as a part of this research defines concepts and relationships that are pertinent during a flood disaster in an urban landscape. The ontology thus demonstrates Sem-RSSU's adaptability to different application contexts. The Remote Sensing Scene Ontology (http://geosysiot.in/rsso/ApplicationSchema) and Flood Scene Ontology (http://geosysiot.in/fso/ApplicationSchema) have been published on the web, for reference. The framework proposes and implements the Spatio-Contextual Triple Aggregation and Grounded Scene Description Planning and Realization Algorithms to (1) aggregate Scene Knowledge Graphs for aiding in scene description rendering from a spatio-contextual perspective and (2) render mappings between regions in the scene and generated sentences in the scene description respectively. It also defines the Remote Sensing Scene Description Grammar that the rendered natural language scene descriptions conform to. The GeoSPARQL Query Interface of the framework enables querying and visualization over the inferred Scene Knowledge Graphs, thus allowing users to further explore and analyze the remote sensing scene. Extensive evaluation of individual components of the framework inspires confidence and demonstrates the efficacy of Sem-RSSU.

Although the approach for grounded natural language scene description rendering in Sem-RSSU produces fair results, it would be interesting to explore and compare with a neural approach to translate Scene Knowledge Graphs to natural language. In addition to the use of directional and topological relations, it would be beneficial to explore the use of inferred qualitative spatial relations, such as "near", "around" and "next to", in the future, to generate more natural scene descriptions. Moreover, the temporal component in natural language scene descriptions for depicting evolving remote sensing scenes has remained largely unexplored. The future directions of this research are (1) to explore neural approaches, to render natural language scene descriptions from Scene Knowledge Graphs; (2) to explore the use of inferred qualitative spatial relations, to improve the naturalness of the rendered scene descriptions; and (3) to explore the temporal component in Scene Knowledge Graphs, to aid in rendering natural language scene descriptions of rapidly evolving remote sensing scenes over time.

**Author Contributions:** Conceptualization, Abhishek V. Potnis and Surya S. Durbha; data curation, Abhishek V. Potnis; formal analysis, Abhishek V. Potnis; investigation, Abhishek V. Potnis; methodology, Abhishek V. Potnis, Surya S. Durbha and Rajat C. Shinde; project administration, Surya S. Durbha; resources, Abhishek V. Potnis, Surya S. Durbha and Rajat C. Shinde; software, Abhishek V. Potnis; supervision, Surya S. Durbha; validation, Abhishek V. Potnis; visualization, Abhishek V. Potnis; writing—original draft, Abhishek V. Potnis; writing—review and editing, Surya S. Durbha. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** The authors are grateful to the Google Cloud Research Credits Program for providing the credits that enabled the use of GCP for training and validating the deep-learning models.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Durbha, S.S.; King, R.L. Semantics-enabled framework for knowledge discovery from Earth observation data archives. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 2563–2572. [CrossRef]
2. Datcu, M.; Remote, G.; Data, S. Scene Understanding from SAR Images. In Proceedings of the 1996 International Geoscience and Remote Sensing Symposium, Lincoln, NE, USA, 31 May 1996.
3. Datcu, M.; Seidel, K.; Walessa, M. Spatial information retrieval from remote-sensing images. I. Information theoretical perspective. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 1431–1445. [CrossRef]
4. Quartulli, M.; Datcu, M. Information fusion for scene understanding from interferometric SAR data in urban environments. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 1976–1985. [CrossRef]
5. Marchisio, G.; Li, W.H.; Sannella, M.; Goldschneider, J.R. GeoBrowse: An integrated environment for satellite image retrieval and mining. *Int. Geosci. Remote Sens. Symp.* **1998**, *2*, 669–673. [CrossRef]
6. Datcu, M.; Daschiel, H.; Pelizzari, A.; Quartulli, M.; Galoppo, A.; Colapicchioni, A.; Pastori, M.; Seidel, K.; Marchetti, P.G.; D'Elia, S. Information Mining in Remote Sensing Image Archives—Part A: System Concepts. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 2923–2936. [CrossRef]
7. Daschiel, H.; Datcu, M. Information mining in remote sensing image archives: System evaluation. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 188–199. [CrossRef]
8. Shyu, C.-R.; Klaric, M.; Scott, G.J.; Barb, A.S.; Davis, C.H.; Palaniappan, K. GeoIRIS: Geospatial Information Retrieval and Indexing System—Content Mining, Semantics Modeling, and Complex Queries. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 839–852. [CrossRef]
9. Molinier, M.; Laaksonen, J.; Häme, T. Detecting Man-Made Structures and Changes in Satellite Imagery with a Content-Based Information Retrieval System Built on Self-Organizing Maps. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 861–874. [CrossRef]
10. Kurte, K.R.; Durbha, S.S.; King, R.L.; Younan, N.H.; Vatsavai, R. Semantics-Enabled Framework for Spatial Image Information Mining of Linked Earth Observation Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 29–44. [CrossRef]
11. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the IEEE CITS 2016—2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 6–8 July 2016. [CrossRef]
12. ZPan, B.; Zou, Z. Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [CrossRef]

13. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
14. Wang, B.; Lu, X.; Zheng, X.; Li, X. Semantic Descriptions of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1274–1278. [CrossRef]
15. Zhang, Z.; Zhang, W.; Diao, W.; Yan, M.; Gao, X.; Sun, X. VAA: Visual Aligning Attention Model for Remote Sensing Image Captioning. *IEEE Access* **2019**, *7*, 137355–137364. [CrossRef]
16. Wang, B.; Zheng, X.; Qu, B.; Lu, X. Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 256–270. [CrossRef]
17. Yuan, Z.; Li, X.; Wang, Q. Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning. *IEEE Access* **2020**, *8*, 2608–2620. [CrossRef]
18. Zhang, Y.; Qiu, Z.; Yao, T.; Liu, D.; Mei, T. Fully Convolutional Adaptation Networks for Semantic Segmentation. In Proceedings of the 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015. [CrossRef]
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Lecture Notes in Computer Science. Volume 9351, pp. 234–241. [CrossRef]
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
21. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
24. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181. [CrossRef]
25. Egenhofer, M.J.; Sharma, J.; Mark, D.M. A Critical Comparison of the 4-Intersection and 9-Intersection Models for Spatial Relations: Formal Analysis. In Proceedings of the 11th Auto-Carto Conference, Minneapolis, MN, USA, 30 October–1 November 1993.
26. Randell, D.A.; Cui, Z.; Cohn, A.G. A Spatial Logic based on Regions and Connection. In Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning, San Francisco, CA, USA, 25–29 October 1992. [CrossRef]
27. Frank, A.U. Qualitative spatial reasoning: Cardinal directions as an example. *Int. J. Geogr. Inf. Syst.* **1996**, *10*, 269–290. [CrossRef]
28. Potnis, A.V.; Durbha, S.S.; Kurte, K.R. A Geospatial Ontological Model for Remote Sensing Scene Semantic Knowledge Mining for the Flood Disaster. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 5274–5277. [CrossRef]
29. Ben-Av, M.B.; Sagi, D. Perceptual grouping by similarity and proximity: Experimental results can be predicted by intensity autocorrelations. *Vis. Res.* **1995**, *35*, 853–866. [CrossRef]
30. Han, S.; Humphreys, G.W.; Chen, L. Uniform connectedness and classical gestalt principles of perceptual grouping. *Percept. Psychophys.* **1999**, *61*, 661–674. [CrossRef]
31. Li, C.; Parikh, D.; Chen, T. Automatic discovery of groups of objects for scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012. [CrossRef]
32. Falomir, Z.; Kluth, T. Qualitative spatial logic descriptors from 3D indoor scenes to generate explanations in natural language. *Cogn. Process.* **2017**, *19*, 265–284. [CrossRef]
33. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164. [CrossRef]
34. Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. BabyTalk: Understanding and Generating Simple Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2891–2903. [CrossRef] [PubMed]
35. Karpathy, A.; Li, F.-F. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 664–676. [CrossRef] [PubMed]
36. Reiter, E.; Dale, R. Building applied natural language generation systems. *Nat. Lang. Eng.* **1997**, *3*, 57–87. [CrossRef]
37. Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum.-Comput. Stud.* **1995**, *43*, 907–928. [CrossRef]
38. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002.
39. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007.
40. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, 25–26 July 2004.

41. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [CrossRef]
42. Evans, R.; Grefenstette, E. Learning Explanatory Rules from Noisy Data. *J. Artif. Intell. Res.* **2018**, *61*, 1–64. [CrossRef]