Hierarchical Scene Graph Generation and Vectorization of Aerial Images

V.A. Knyaz[1,2], V.V. Kniaz[1,2] , S.Yu. Zheltov[2], A.V. Emelyanov[1,2], E.R. Smirnov[2]

[1] Moscow Institute of Physics and Technology (MIPT), Moscow, Russia - (kniaz.va, kniaz.vv)@mipt.ru
[2] State Research Institute of Aviation Systems (GosNIIAS), Moscow, Russia - zhl@gosniias.ru

Key Words: image vectorization, scene graph generation, hierarchical representation, maps updating, convolutional neural networks.

Abstract

Vector representation of geodata is widely used in various application due to high density of information and the advanced level of information representation, introduced by the human operator while creating a map. We can say that a map is a vector representation of understanding a scene based on its image. Scene understanding can be considered at different levels of depth, beginning from image classification and semantic segmentation and completing with rich semantic relationships between objects and retrieving its hierarchy. With the progress in machine learning methods and tools for obtaining and processing large amounts of data a set of neural network models has been developed that demonstrate state-of-the art performance (humanlike and better) in image classification and image semantic segmentation tasks. After object detection and recognition, the next step in scene understanding is retrieving the relations between objects and their hierarchy. This problem is known as scene graph generation, and recently it received notable attention by the scientific community. The developed approach incorporates the information about the structural and functional relationships between objects in the image, which, on the one hand, improves the quality of segmentation through the use of new a priori data, and on the other hand, reduces the time spent by the operator on subsequent processing of the results of the neural network algorithm. To train and evaluate the developed framework, a special dataset is collected and annotated. It contains more than 10k aerial photographs representing various types of objects taken in different years and seasons. The evaluation results on the created dataset proved the state-of-the-art performance of the developed framework.

1. Introduction

Despite the rapid development of digital technologies in geospatial information systems, the vector representation of geodata is still effective and widely used. This phenomenon can be explained by the high density of information and the advanced level of information representation on vector maps, introduced by the human operator who created the map. We can say that a map is a vector representation of understanding a scene based on its image. Scene understanding can be considered at different levels of depth, beginning from image classification and semantic segmentation and completing with rich semantic relationships between objects and retrieving its hierarchy.

With the progress in machine learning methods and tools for obtaining and processing large amounts of data a set of neural network models has been developed that demonstrate state-of-the art performance (humanlike and better) in image classification and image semantic segmentation tasks. After object detection and recognition, the next step in scene understanding is retrieving the relations between objects and their hierarchy. This problem is known as scene graph generation, and recently it received notable attention by the scientific community.

Scene graph can be defined as a structured representation of a scene in the form of a set of nodes (objects) connected by edges reflecting their relations (Johnson et al., 2015). Scene graph (Figure 1) represents semantic links and interactions between entities in the scene (entities can be subject or object) in the form of <`subject - predicate - object`>. Scene graph is an abstraction, that gives a new level of the semantic understanding of an image, and currently widely used in such image analysis tasks. Among these tasks are image captioning (Nguyen et al., 2021), image retrieval (Johnson et al., 2015), visual question answering (Johnson et al., 2017), image generation (Ashual and Wolf, 2019) and similar.

The problem of scene graph generation (Li et al., 2024) has received a powerful impetus for development with advances in machine learning methods and tools for obtaining and processing large amounts of data. But the most of the studies addresses to scene graph generation for natural (not remote sensing) scenes. Meanwhile understanding of remote sensing images is also very important for updating maps and developing unmanned aerial systems. Moreover the semantic understanding of the remote sensing image provides higher performance in the tasks of change detection, semantic segmentation and vectorization of images by attracting additional information about the scene as a whole.

Remote sensing scenes usually have hierarchical structure that is reflected in the maps. Figure 1 shows the scenes with plain (Figure 1(a)) and hierarchical (Figure 1(b)) structures. Hierarchical representation more adequate reflects the scene structure and provides deeper understanding of the scene.

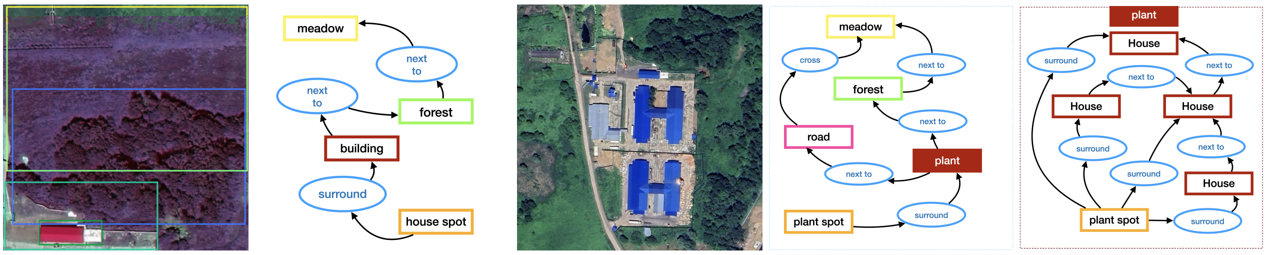
The developed approach incorporates the information about the structural and functional relationships between objects in the image, which, on the one hand, improves the quality of segmentation through the use of new a priori data, and on the other hand, reduces the time spent by the operator on subsequent processing of the results of the neural network algorithm.

The main contribution of the study are the following:

- the framework for hierarchical semantic scene graph generation and vectorization of aerial images

Publisher's note: Copernicus Publications has not received any payments from Russian or Belarusian institutions for this paper.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-161-2025 |



(a) Scene with plain structure

(b) Scene with hierarchical structure

Figure 1. Scenes with plain and hierarchical structures. In scene with hierarchical structure (b) the complex *plant* object is further represented as its interior graph structure.

- the dataset for training and evaluating the proposed approach
- evaluation of the proposed framework in task of hierarchical scene graph generation

2. Related work

2.1 Scene Graph Generation

The problem of image understanding has long attracted the interest of the scientific community (Li and Fei-Fei, 2007, Vishnyakov et al., 2015, Shu et al., 2015). Due to recent advances in the means and methods of obtaining, storing and processing large amounts of data, new machine learning methods have given a significant boost to this problem, and scene graph generation being the one of the core task of image understanding problem. While object detection and recognition techniques try to answer what objects present in the image, scene graph describes the relations between this objects.

Scene graphs, initially introduced for image retrieval problem, demonstrated their potential in other subtasks of scene understanding, such as image captioning (Yang et al., 2019, Gu et al., 2019, Lee et al., 2019a), visual question answering (Shi et al., 2019, Lee et al., 2019b), or image synthesis (Li et al., 2019, Talavera et al., 2019). If the first studies tried to extract relations of only some specific types (like spatial location in a scene) (Galleguillos et al., 2008, Gould et al., 2008), the more common statement and method of relationship identifying in an image (Lu et al., 2016) was proposed notably later. This work proposed two-stage approach for scene graph generation: at first, to detect objects in the image, and, secondly, to identify the relationships between these objects.

The proposed neural network model (Lu et al., 2016) is trained separately for objects and predicates, and than combines them for extracting multiple relationships in the image. The improvement in scene graph generation is obtained by using language priors from semantic word embeddings. It allows to improve the prediction of the relationship in terms of the likelihood. The authors demonstrated the improvements in image retrieval task due to understanding the relationships between objects.

Basing on this two-stage approach some improvements have been reached by using residual neural networks (Zellers et al., 2018, Xu et al., 2017, Li et al., 2017, Cong et al., 2020) with the global context, by applying standard RNNs to iteratively improve the prediction of the relationships involving the message passing approach.

The Graph R-CNN neural network model (Yang et al., 2018a) involves the attention mechanism for efficient retrieving the contextual information between objects and relationships. Such approach allowed the Graph R-CNN to show state-of-the-art performance on the Visual Genome dataset (Krishna et al., 2017).

Recently, with the invention of the Transformer networks (Vaswani et al., 2017), they were applied for retrieving the visual relationship and scene graphs generation. The RelTransformer (Chen et al., 2022) network model has outperformed the best baseline models on two large-scale visual relationship recognition benchmarks. RelTransformer network model considers each image as a fully-connected scene graph and represents the given scene in form of the relation-triplet and global-scene contexts. Efficient message passing from such scene representation to the target relation and integrated self-attention mechanism allowed notably improving the performance in visual relationship recognition task.

Some works involves semantic information (Cui et al., 2018, Gkanatsios et al., 2019, Yu et al., 2020) and statistics priors (Dai et al., 2017, Zhang et al., 2019) to improve the performance of scene graph generation.

It is worth noting that the most part of studies on scene graph generation and semantic image understanding addresses ground-based (or so-called natural) images. Due to the specificity of natural images and related tasks, the developed for this field scene graph generation techniques rather address to relationships of the action type such as ”walk”, ”seat”, ”drink”, etc. or state type such as ”lay”, ”wear”, etc. For the tasks of scene understanding in remote sensing imagery it is more important to retrieve spatial and structural relationships between objects such as mutual spatial position, spatial structure and hierarchy. This information allows to improve the quality of data processing in remote sensing tasks such as semantic image segmentation, change detection, image vectorization, etc.

Scene understanding in remote sensing imagery has received less attention due to necessity of creating large dataset required very time consuming manual annotation. First works addressed scene understanding in remote sensing imagery explored image captioning (Shi and Zou, 2017) and the image representation and the caption representation (Wang et al., 2019).

For simultaneous object detection and their relations retrieving in remote sensing images the multi-scale remote sensing image interpretation network (MSRIN) was proposed (Cui et al., 2019). The MSRIN is a parallel deep neural network that integrates the fully convolutional U-Net network, and a

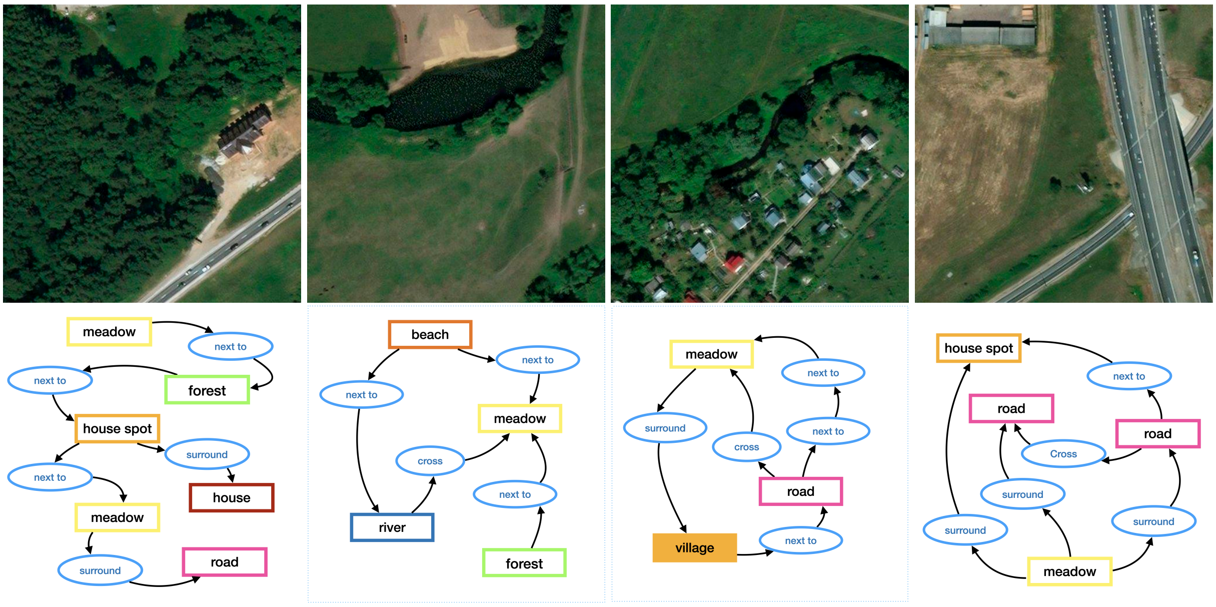


Figure 2. Example images from the Segmentation and Visualization Aerial Images dataset extended with relationship annotations.

long short-term memory network (LSTM). Such approached allowed simultaneously performing objects' semantic segmentation and identifying their relationships. To introduce the semantic meaning to spatial relationships produced by MSRIN, the multi-scale semantic fusion network (MSFN) (Li et al., 2021) introduced the dilated convolution block into a graph convolutional network for integrating and refining multi-scale semantic context. This architecture allowed to improve the cognitive ability of the MSFN. For developing and evaluating the proposed framework the authors created the remote sensing scene graph dataset (RSSGD), containing a set of objects, their attributes, and relationships.

Besides the RSSGD dataset there are not so many datasets designed for the tasks of scene understanding in remote sensing imagery. Among these are the Remote Sensing Image Caption Dataset RSICD (Lu et al., 2017), containing about 40 categories for object and 16 for relationship, and Geospatial Relationship Triplet Representation Dataset (GRTRD) (Chen et al., 2021) oriented for identifying of geographic objects and predicting their relationships. It includes twelve object classes and about 20k object and 20k relationship categories.

3. Materials and Methods

The proposed approach to simultaneous image semantic segmentation and scene graph generation is based on machine learning and uses the neural network model developed at the previous stage of the study (Knyaz et al., 2024, Emelyanov et al., 2024). The developed neural network model firstly applies visual transformer to extract deep features from the input aerial image. Then graph neural network carries out clusterization of these deep features resulting in semantic segmentation of the image. The developed network model was trained and evaluated on specially created SVAI (segmentation and vectorization of aerial imagery) dataset (Emelyanov et al., 2024).

To extend the developed framework for solving the task of scene graph generation it was modified by adding object classification block and relation retrieving block. The overview of the proposed framework is shown in Figure 3. For training the developed neural network model the SVAI (segmentation and vectorization of aerial imagery) dataset has also been modified. It was extending by including *about 39 object categories and 16 relationship categories* and annotations triplets <`subject - predicate - object`>.

3.1 Dataset

SVAI (Segmentation and Vectorization of Aerial Imagery) dataset is designed for the tasks of aerial imagery analysis, including change detection, segmentation and vectorization, and scene graph generation. Currently it includes 8400 very high resolution aerial images of different scenes obtained at different times and with different sensors. The change detection split of the SVAI dataset contains two thousand pairs of images of the same scenes acquired at various times and containing changes in the scene. The change detection split is annotated for training and testing change detection neural network models, the annotation being binary masks labelled by zero for unchanged regions, and by non-zero value for changed ones.

New annotations have been added to study the problem of scene graph generation. First, the objects represented in the images were classified into 24 classes, such as a building, a road, a river, a bridge, and a background. The classes for classification was chosen according the topographic map classification for further quick adaptation to the task of map updating.

Secondly, 12 categories of relations were introduced, representing the spatial topology and functional description of objects. They describe possible relationships between objects in a scene, such as adjacent, distant, around, passing through, passing under, etc. The samples of images and their annotations are shown in Figure 2.

To receive good initial approximation to annotation triplets (<`subject - predicate - object`>) data from OpenStreetMap resource (OpenStreetMap Foundation, 2026) was exploited.

The OSM offers three types of basic elements for conceptual modelling the real world. These are nodes, ways and relation.

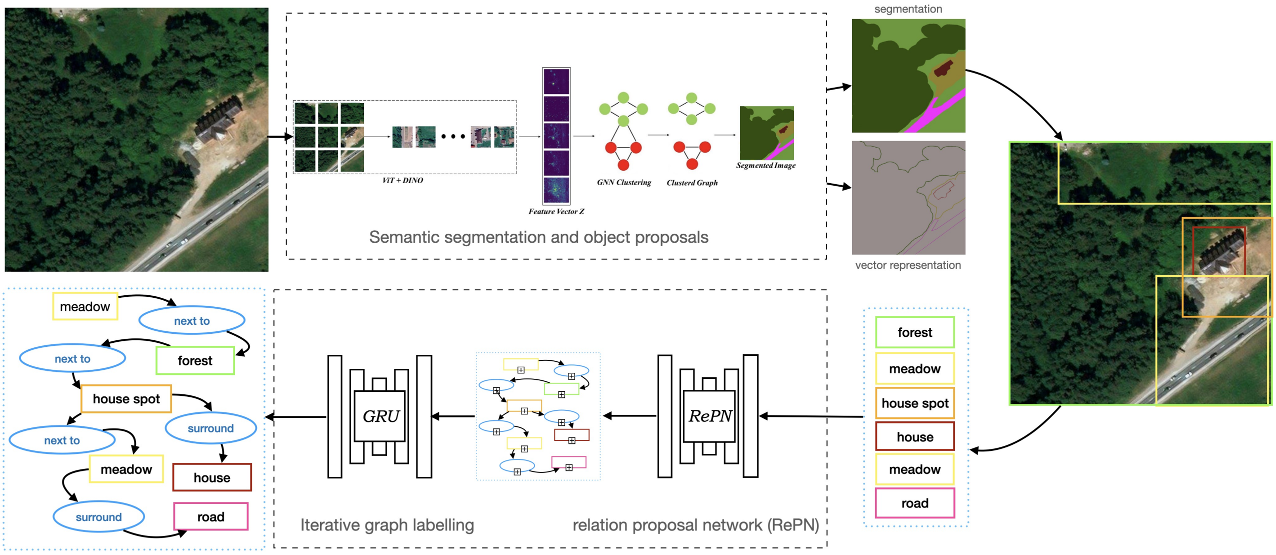


Figure 3. The HSGG-AI framework architecture. Firstly, the pre-trained visual transformer retrieves deep features from the input aerial image basing on attention mechanism, and graphical neural network performs clustering of these deep features thus creating object region proposals and a set of object's nodes and edges. Secondly, the relationship proposal network RePN. Finally, the graph labeling is performed iteratively to refine the scene graph.

Node defines the point in the space and is given by a pair of WGS84 coordinates and identifier.

Way defines a linear entity or a region boundary. *Way* is given by a set of nodes.

Relation defines the relationship between elements of the OSM data. These elements can be nodes, ways, or other relations.

We use OpenStreetMap data as initial annotation for images from SVAI dataset. Than this data was proved and edited to generate ground truth annotation for the task of scene graph generation.

3.2 Framework for hierarchical scene graph generation

In scene graph generation task, the scene graph G is represented as a set of vertexes V (image regions), edges E (relationships between image regions), and their labellings. For image I with a set V of the objects detected in the image I, and a set E of the edges (relationships), labels O correspond to objects and labels R denote correspond to relationships.

So the problem of scene graph generation can be formulated as the task of designing the model $P(G|I)$, that generate the scene graph S for given image I. This problem can be represented as three subtasks:

1. (1) object region proposal $P(V|I)$,
2. (2) objects' relationship proposal $P(E|V, I)$, and
3. (3) objects and relationships labelling $P(R, O|V, E, I)$:

Therefore, the generation of scene graph can be described as follows:

$$P(G \mid I) = P(V \mid I)P(E \mid V, I)P(R, O \mid V, E, I), \quad (1)$$

Accordingly, the proposed framework for generating hierarchical scene graphs and vectorizing aerial images HSGG-AI consists of three blocks that solve the whole problem. The outline of the proposes framework is shown in Figure 3.

3.2.1 Segmentation, vectorization and object region proposal.

We use the original graph semantic segmentation model for aerial images (GSS-AI) (Emelyanov et al., 2024) as a starting point for the framework for scene graph generation. The GSS-AI network model is used for scene semantic segmentation and object region proposal $P(V|I)$.

The GSS-AI network model utilizes attention mechanism to retrieve deep features, which then are aggregated in clusters by the graph neural network. Applying Vision Transfomer (Dosovitskiy et al., 2020) trained with DINO (Caron et al., 2021) allows to extract deep features in self-supervising mode resulting in attention maps.

Basing on this attention maps, the problem of image semantic segmentation is considered as the graph-cut task, the image being represented by an undirected graph $G = \{\mathcal{V}, \mathcal{E}\}$ with node set $\mathcal{V}$ and edge set $\mathcal{E}$. And the clusterization of the similar areas is performed using the similarity matrix W, whose elements w_{ij} are the similarities between image areas i and j, $i, j = 1 \dots n$ obtained from output feature vector of the Vision Transformer.

Considering the matrix W as a map of image areas similarities, the partitioning of the image is performed with normalized cut criterium, that requires maximizing interconnections within a partition and minimizing the number of partition-to-partition connections. For two parts A and B of a graph the *normalized cut* $Ncut(A, B)$ of the graph G is:

$$Ncut(A, B) = \frac{\sum\limits_{u \in A, v \in B} w(u, v)}{\sum\limits_{i \in A, j \in \mathcal{V}} w(i, j)} + \frac{\sum\limits_{u \in A, v \in B} w(u, v)}{\sum\limits_{i \in B, j \in \mathcal{V}} w(i, j)}, \quad (2)$$

This procedure allows to perform image semantic segmentation in self-supervising manner resulting in pixel-wise segmentation map.

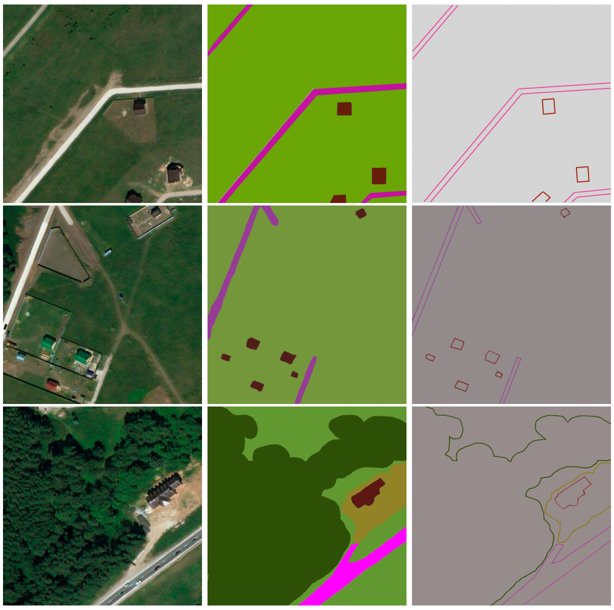


Figure 4. Sample of semantic segmentation and vectorization for the images from SVAI dataset

As a result, the initial graph $G = \{\mathcal{V}, \mathcal{E}\}$ representing the image, can be described by its adjacency matrix A with zero elements for edges eliminated by the graph-cut procedure (borders between clusters). The adjacency matrix A serves for generating the vector representation of the image that can be transformed into the map. Figure 4 shows an example of semantic segmentation and vectorization of images from the SVAI dataset after postprocessing the primary data to filter noise and outliers.

3.2.2 Objects relationships proposal and graph labelling. Remote sensing imagery has the specifics due to the need of analyzing data at different scales. This fact makes it imperative to tackle the data hierarchically, that is reflected in producing of maps at various scales and in applying the hierarchical classification of objects.

So, to take into account hierarchical structure of geospatial data we introduced hierarchical relationships such as ”belong to”, ”include” in annotation of the scene graph generation split of the SVAI dataset.

For the second term of Equation (1) we exploited the relationship proposal network RePN (Yang et al., 2018b) that directly models relationship proposals $P(E \mid V, I)$ and allows to perform learning in end-to-end mode.

The third subtask of the graph labeling was solved as iterative refinement of the scene graph (Xu et al., 2017), using the Gated Recurrent Unit (GRU) (Cho et al., 2014).

4. Results

The sample results of hierarchical scene graph generation by the proposed HSGG-AI framework are shown in Figure 5. We evaluate of the proposed HSGG-AI framework in task of Phrase Detection (PhrDet) (Lu et al., 2016) in terms of the $R@k$ metric, that considers the part of ground-truth relationship triplets (<`subject - predicate - object`>) among the top k most confident triplet predictions in an image.

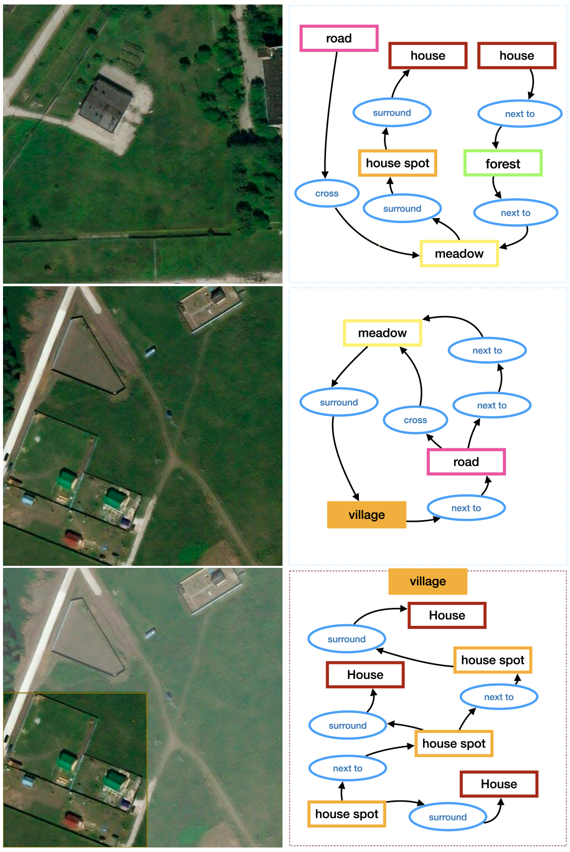


Figure 5. Qualitative results of aerial images processing for scene graph generation.

The evaluation has been carried out on the testing split of SVAI dataset and has demonstrated the performance of $R@100$ = 42.87 and $R@50$ = 36.02, being at the state of the art level.

We also evaluated our HSGG-AI framework in task of image segmentation on extended SVAI dataset in comparison with the state-of-the-art unsupervised baselines similar to the previous study (Emelyanov et al., 2024) in terms of mean Intersection-over-Union (mIoU) metric. Table 1 presents the numerical values of the mean Intersection-over-Union metric for each of the methods. Table 1 shows that adding information about hierarchical structure of the scene also contributes in overall segmentation accuracy.

Method	SVAI dataset
OneGAN (Benny and Wolf, 2020)	56.82
BigBigGAN (Voynov et al., 2021)	67.54
Spectral Methods (Melas-Kyriazi et al., 2022)	72.37
TokenCut (Wang et al., 2022)	74.74
GSS-AI (Emelyanov et al., 2024)	77.83
HSGG-AI (current study)	78.11

Table 1. Values of the mean Intersection-over-Union metric on the extended SVAI dataset for HSGG-AI and baselines.

5. Conclusion

The framework for hierarchical scene graph generation and vectorization of aerial images is developed. It uses the Vision Transformer and graph neural network for accurate image segmentation, vectorization, and object region proposal, and than generates the graph of the scene that reflect hierarchical structure of the scene using relationship proposal network (RePN) and Gated Recurrent Unit models.

Segmentation and Vectorization of Aerial Imagery (SVAI) dataset has been extended for training the developed framework in task of hierarchical scene graph generation. It was extended by annotations of object classes and relationship categories, including hierarchical annotation for the images.

The proposed HSGG-AI network model was evaluated on the testing split of the SVAI dataset in comparison with modern network model models. The evaluation showed that the developed model successfully competes with the baseline models, and that adding information about hierarchical structure of the scene contributes in overall segmentation accuracy.

6. Acknowledgements

The research was carried out at the expense of a grant from the Russian Science Foundation No. 24-21-00269,
https://rscf.ru/project/24-21-00269/

References

Ashual, O., Wolf, L., 2019. Specifying object attributes and relations in interactive scene generation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4561–4569.

Benny, Y., Wolf, L., 2020. Onegan: Simultaneous unsupervised learning of conditional image generation, foreground segmentation, and fine-grained clustering. *European Conference on Computer Vision*, Springer, 514–530.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, J., Agarwal, A., Abdelkarim, S., Zhu, D., Elhoseiny, M., 2022. Reltransformer: A transformer-based long-tail visual relationship recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19507–19517.

Chen, J., Zhou, X., Zhang, Y., Sun, G., Deng, M., Li, H., 2021. Message-Passing-Driven Triplet Representation for Geo-Object Relational Inference in HRSI. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5.

Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y., 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv preprint arXiv:1409.1259*. https://arxiv.org/abs/1409.1259.

Cong, Y., Ackermann, H., Liao, W., Yang, M. Y., Rosenhahn, B., 2020. Nodis: Neural ordinary differential scene understanding. *Proceedings of the European Conference on Computer Vision (ECCV)*, 636–653.

Cui, W., Wang, F., He, X., Zhang, D., Xu, X., Yao, M., Wang, Z., Huang, J., 2019. Multi-Scale Semantic Segmentation and Spatial Relationship Recognition of Remote Sensing Images Based on an Attention Model. *Remote Sensing*, 11(9). https://www.mdpi.com/2072-4292/11/9/1044.

Cui, Z., Xu, C., Zheng, W., Yang, J., 2018. Context-dependent diffusion network for visual relationship detection. *Proceedings of the 26th ACM international conference on Multimedia*, 1475–1482.

Dai, B., Zhang, Y., Lin, D., 2017. Detecting visual relationships with deep relational networks. *Proceedings of the IEEE conference on computer vision and Pattern recognition*, 3076–3086.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Emelyanov, A. V., Knyaz, V. A., Kniaz, V. V., Zheltov, S. Y., 2024. Aerial Images Segmentation with Graph Neural Network. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-2/W1-2024, 1–8. https://isprs-annals.copernicus.org/articles/X-2-W1-2024/1/2024/.

Galleguillos, C., Rabinovich, A., Belongie, S., 2008. Object categorization using co-occurrence, location and appearance. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 1–8.

Gkanatsios, N., Pitsikalis, V., Koutras, P., Maragos, P., 2019. Attention-translation-relation network for scalable scene graph generation. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D., 2008. Multi-class segmentation with relative location prior. *International journal of computer vision*, 80(3), 300–316.

Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., Wang, G., 2019. Unpaired image captioning via scene graph alignments. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10323–10332.

Johnson, J., Hariharan, B., Van Der Maaten, L., Hoffman, J., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R., 2017. Inferring and executing programs for visual reasoning. *Proceedings of the IEEE International Conference on Computer Vision*, 2989–2998.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D., Bernstein, M., Fei-Fei, L., 2015. Image retrieval using scene graphs. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3668–3678.

Knyaz, V. A., Kniaz, V. V., Zheltov, S. Y., Petrov, K. S., 2024. Multi-sensor Data Analysis for Aerial Image Semantic Segmentation and Vectorization. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1-2024, 291–296. https://isprs-archives.copernicus.org/articles/XLVIII-1-2024/291/2024/.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., Fei-Fei, L., 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision*, 123(1), 32–73. https://doi.org/10.1007/s11263-016-0981-7.

Lee, K.-H., Palangi, H., Chen, X., Hu, H., Gao, J., 2019a. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators. *arXiv preprint arXiv:1909.09953*.

Lee, S., Kim, J.-W., Oh, Y., Jeon, J. H., 2019b. Visual question answering over scene graph. *International Conference on Graph Computing (GC)*, 45–50.

Li, H., Zhu, G., Zhang, L., Jiang, Y., Dang, Y., Hou, H., Shen, P., Zhao, X., Shah, S. A. A., Bennamoun, M., 2024. Scene Graph Generation: A comprehensive survey. *Neurocomputing*, 566, 127052. https://www.sciencedirect.com/science/article/pii/S092523122301175X.

Li, L.-J., Fei-Fei, L., 2007. What, where and who? classifying events by scene and object recognition. *2007 IEEE 11th International Conference on Computer Vision*, 1–8.

Li, P., Zhang, D., Wulamu, A., Liu, X., Chen, P., 2021. Semantic Relation Model and Dataset for Remote Sensing Scene Understanding. *ISPRS International Journal of Geo-Information*, 10(7). https://www.mdpi.com/2220-9964/10/7/488.

Li, Y., Ma, T., Bai, Y., Duan, N., Wei, S., Wang, X., 2019. Pastegan: A semi-parametric method to generate image from scene graph. *Advances in Neural Information Processing Systems*, 32, 3948–3958.

Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X., 2017. Scene graph generation from objects, phrases and region captions. *Proceedings of the IEEE international conference on computer vision*, 1261–1270.

Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L., 2016. Visual relationship detection with language priors. *European conference on computer vision*, Springer, 852–869.

Lu, X., Wang, B., Zheng, X., Li, X., 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4), 2183–2195.

Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A., 2022. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8364–8375.

Nguyen, K., Tripathi, S., Du, B., Guha, T., Nguyen, T. Q., 2021. In defense of scene graphs for image captioning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1407–1416.

OpenStreetMap Foundation, 2026. OpenStreetMap Project. OpenStreetMap Foundation. osm.org (02 June 2025).

Shi, J., Zhang, H., Li, J., 2019. Explainable and explicit visual reasoning over scene graphs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8376–8384.

Shi, Z., Zou, Z., 2017. Can a machine generate humanlike language descriptions for a remote sensing image? *IEEE Transactions on Geoscience and Remote Sensing*, 55(6), 3623–3634.

Shu, T., Xie, D., Rothrock, B., Todorovic, S., Zhu, S.-C., 2015. Joint Inference of Groups, Events and Human Roles in Aerial Videos. *arXiv preprint arXiv:1505.05957*. https://arxiv.org/abs/1505.05957.

Talavera, A., Tan, D. S., Azcarraga, A., Hua, K.-L., 2019. Layout and context understanding for image synthesis with scene graphs. *IEEE International Conference on Image Processing (ICIP)*, 1905–1909.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 5998–6008.

Vishnyakov, B. V., Vizilter, Y. V., Knyaz, V. A., Malin, I. K., Vygolov, O. V., Zheltov, S. Y., 2015. Stereo sequences analysis for dynamic scene understanding in a driver assistance system. J. Beyerer, F. P. León (eds), *Automated Visual Inspection and Machine Vision*, 9530, International Society for Optics and Photonics, SPIE, 95300P.

Voynov, A., Morozov, S., Babenko, A., 2021. Object segmentation without labels with large-scale generative models. *International Conference on Machine Learning*, PMLR, 10596–10606.

Wang, B., Lu, X., Zheng, X., Li, X., 2019. Semantic Descriptions of High-Resolution Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, 16(8), 1274-1278.

Wang, Y., Shen, X., Hu, S. X., Yuan, Y., Crowley, J. L., Vaufreydaz, D., 2022. Self-supervised transformers for unsupervised object discovery using normalized cut. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14543–14553.

Xu, D., Zhu, Y., Choy, C. B., Fei-Fei, L., 2017. Scene graph generation by iterative message passing. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5419.

Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D., 2018a. Graph r-cnn for scene graph generation. *Proceedings of the European conference on computer vision (ECCV)*, 670–685.

Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D., 2018b. Graph r-cnn for scene graph generation. V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds), *Computer Vision – ECCV 2018*, Springer International Publishing, Cham, 690–706.

Yang, X., Tang, K., Zhang, H., Cai, J., 2019. Autoencoding scene graphs for image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10685–10694.

Yu, J., Chai, Y., Wang, Y., Hu, Y., Wu, Q., 2020. Cogtree: Cognition tree loss for unbiased scene graph generation. *arXiv preprint arXiv:2009.07526*.

Zellers, R., Yatskar, M., Thomson, S., Choi, Y., 2018. Neural motifs: Scene graph parsing with global context. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5831–5840.

Zhang, J., Shih, K. J., Elgammal, A., Tao, A., Catanzaro, B., 2019. Graphical contrastive losses for scene graph parsing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11535–11543.