The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-4/W18-2025
Symposium on GeoSpatial Technologies: Visions and Horizons 2025 (GeoVisions2025), 8–10 October 2025, Çanakkale, Türkiye

Multi-scale scene graph generation for remote sensing imagery

Vladimir A. Knyaz[1,2], Vladimir V. Kniaz[1,2], Anton V. Emelyanov[1,2], Sergey Yu. Zheltov[2], Victor S. Aleksandrov[2]

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

Key Words: image vectorization, scatn graph generation, maps updating, convolutional neural networks.

Abstract

The map, as a way of representing geospatial data, is designed to reflect important information about the Earth as deeply and accurately as possible. To meet this requirement, maps are produced in different scales and different types, depending on the task being solved. Created by highly educated specialists, the map contains not only raw geospatial data, but also some high-level knowledge accumulated by people during the exploration of the Earth. The introduction of deep learning into the data analysis process has allowed the development of neural network models that can solve complex aerial image processing tasks, such as semantic image segmentation, object detection and recognition, and retrieving of semantic relations between objects in a scene. These advances created the background for moving to image (scene) understanding as a higher level of image analysis. The current study addresses to a problem of multi-scale scene graph generation from aerial images, similarly to creating maps of different scales.

1. Introduction

The map, as a way of representing geospatial data, is designed to reflect important information about the Earth as deeply and accurately as possible. To meet this requirement, maps are produced in different scales and different types, depending on the task being solved. Created by highly educated specialists, the map contains not only raw geospatial data, but also some high-level knowledge accumulated by people during the exploration of the Earth. We can say that the map is some kind of reflection of the human understanding of the scene.

The specifics of geospatial data are such that it may be required at different scales, depending on the application. Accordingly, maps are created at different scales, reflecting objects and their attributes in accordance with the application. This representation of geospatial data is very helpful in understanding the area under study.

Recent advances in data processing techniques, based on the collection and analysis of huge amounts of data, allow to move from processing data to understanding it.

Image classification (as the assignment of an image to one of the predefined classes) can be considered as the first step to image understanding, and then moving to the next level with labelling entities in the image (Li and Wang, 2003). But object detection and recognition is not enough to understand the scene, which should also include extracting semantic links between detected objects.

Deep learning methods of data analysis make it possible to solve with high quality such tasks as image classification, semantic segmentation, object detection and recognition, change detection, and other, that can be considered as preliminary stages of understanding an image (scene). For making the next step in image understanding it as necessary to extract semantic links between objects in the scene – to create so-called scene graph, reflecting hierarchy and relationships of the objects.

Scene graph is an abstraction of a scene, that operates with objects and relations between them (Johnson et al., 2015b). Standard form of presentation for the scene graph is triplet (a set of triplets) of the type <`subject - predicate - object`>. The example of scene graph for a scene with plain structure is shown in Figure 1.

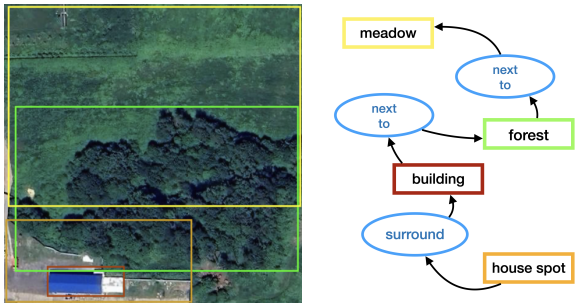

Figure 1. Scene graph for a scene with plain structure.

Scene graphs initially were introduced for analysis of ground-based images, and were applied for image captioning, visual question answering, and similar problems (Li et al., 2024), and deep learning techniques provided an impressive progress in solving these tasks. As to scene graph generation for aerial images, significantly less research has been conducted.

It should be noted, that scene understanding is also substantial for applications that use geospatial information (e.g. autonomous unmanned aerial vehicles, geo-information systems, environmental monitoring, and others).And almost all such applications work with a hierarchical data structure of various levels of detail.

This study examines the problem of generating a scene graph for input aerial images of various scales with the generation of a scene graph of scale and presentation level in accordance with the scale of the input image.

The main contributions of the study can be summarized as:

- the framework for multi-scale semantic scene graph generation that reflects the relationships between objects at different scales;

(a) Samples of maps of different scales

(b) Aerial images of the same scales

Figure 2. Examples of maps of different scales and aerial images of the same scale.

- the dataset for training and evaluating the proposed framework
- evaluation of the proposed framework in task of multi-scale semantic scene graph generation

2. Related work

2.1 Scene Graph Generation Research

Computer image understanding as a challenge to the scientific community was appeared along with digital image processing. At the earlier stages of studying this problem first attempts were made by developing hand-crafted method for image classification and object recognition (Torresani et al., 2010, Farhadi et al., 2010, Vishnyakov et al., 2015). Such methods were focused on automatic generating of image description (Torresani et al., 2010) or developing descriptors that serve for efficient classification and provide high performance in object recognition (Farhadi et al., 2010, Vishnyakov et al., 2015). Some studies developed techniques for establishing some definite kinds of relationships, such as relative location in a scene (Gould et al., 2008), or co-occurrence and appearance (Galleguillos et al., 2008).

But only with the development of modern deep learning methods, a fundamentally new level has been reached in solving the problem of image understanding. These data-driven techniques allow us to step from object detection and recognition to scene understanding with retrieving semantic information on relationships between objects. This semantic information is usually presented as a scene graph, that reflects objects presented in a scene and relationships between these objects (Johnson et al., 2015b). Scene graphs are widely used in such tasks of image analysis as visual question answering (Shi et al., 2019, Lee et al., 2019b), image captioning (Yang et al., 2019, Gu et al., 2019, Lee et al., 2019a).

First approaches in scene graph generation were based on two-stage processing (Lu et al., 2016a), beginning with object recognition in a scene and further combining the extracted objects and predicates to generate relationships using an objective function. Some improvements in the quality of scene graph generating were obtained via exploiting standard recurrent neural networks, that use message passing mechanism to refine the prediction iteratively (Xu et al., 2017).

Integration of object detection, image caption and scene graph generation modules in one multi-level scene description network (MSDN) (Li et al., 2017) allowed to obtain high performance in scene understanding due to passing rich information through three network models.

Applying the Transformer network models for scene understanding provided new improvement in solving this problem. Considering the image as a scene graph that reflects a scene as <subject - predicate - object> triplets and some global context allowed the RelTransformer (Chen et al., 2022) network achieving the state-of-the-art performance on two large-scale benchmarks. Using gated recurrent units (GRU) (Cho et al., 2014) in message passing after the tensor-based relational module (Hwang et al., 2018) gave some arising of the semantic relation accuracy. Also incorporating additional information about scene semantic (Gkanatsios et al., 2019, Yu et al., 2020) and statistics (Zhang et al., 2019) improved the quality of scene graph generation.

And while considerable attention has been paid to the problem of creating scene graphs for ground-based images, significantly less research has been conducted in the field of aerial photographs and satellite images. The works on image captioning and image representation (Wang et al., 2019, Shi and Zou, 2017) for remote sensing data were pioneer studies, that shift research focus from image processing to image understanding. Integration of fully convolutional U-Net network, and a long short-term

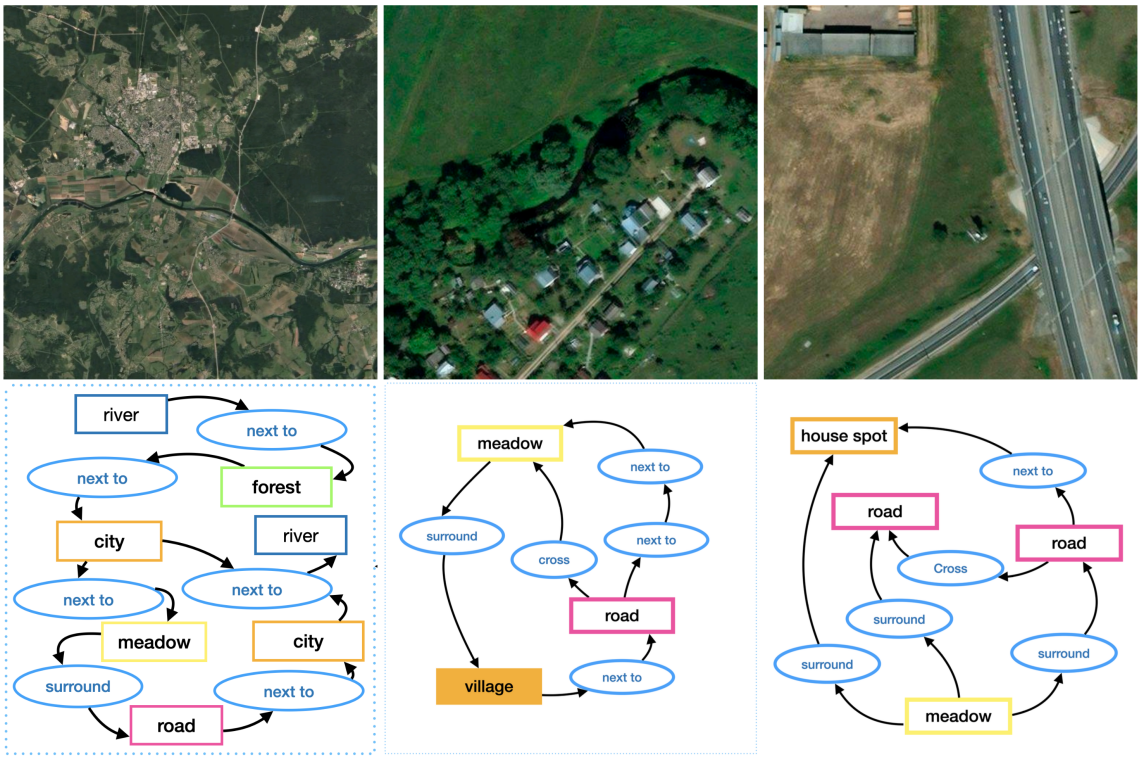

Figure 3. Example images from the Segmentation and Visualization Aerial Images dataset extended with relationship annotations.

memory network (LSTM) in multi-scale remote sensing image interpretation network (MSRIN) model (Cui et al., 2019) allowed to perform semantic segmentation of an image for detecting objects and to extract their relationships in one network model. Dilated convolution block included in the graph convolution network provides integration of multi-scale semantic information in multi-scale semantic fusion network (MSFN) (Li et al., 2021), improving the cognitive capacity of the model.

2.2 Scene Graph Generation Datasets

While currently there are many datasets for scene graph generation research in the field of ground based imagery, the number of datasets for scene graph generation problem in aerial and satellite imagery is noticeably smaller. This fact is one of the reasons of less number research on scene graph generation for aerial imagery. Despite the availability large datasets of ground-based images, that consist of thousands images and tens of triplet queries (Johnson et al., 2015a, Lu et al., 2016b, Krishna et al., 2017, Peyre et al., 2017), they could not be used for remote sensing scene graph generation because of differences in content of objects and relationships.

So the researcher in the field of remote sensing scene graph generation creates the intended datasets for this study. So, to develop and assess the multi-scale semantic fusion network (MSFN) (Li et al., 2021) the authors gathered and annotated the remote sensing scene graph dataset (RSSGD). It included annotations for object categories, their attributes, and the relationships between the objects, thus providing the data for training and evaluating the network model for scene graph generation.

One of the first dataset created for the problem of scene graph generation is Geospatial Relation Triplet Representation dataset (GRTRD) (Chen et al., 2021). It contains more than three thousands high-resolution (0.5 m/pixel) images, about 20 thousands annotation for objects and their geo-spatial relations. Several sets of geo-spatial (topological, orientation, and distance) relation triplets are available for each image.

Also the Remote Sensing Image Caption Dataset RSICD (Lu et al., 2017), containing about 40 categories for object and 16 for relationship, can be used for developing method of scene graph generation.

3. Materials and Methods

We use Graph Semantic Segmentation neural network model for Aerial Imagery (GSS-AI model), developed at the previous stage of the study (Emelyanov et al., 2024, Knyaz et al., 2025) as a starting point for developing simultaneous image semantic segmentation and multi-scale scene graph generation neural network model.

Scene graph G can be considered as a set of vertexes V, that represent image regions and edges E, that represent relationships between detected image regions. As the vertexes V, so the edges E are labelled by the categories objects O and the categories of the relationships R correspondingly.

Then the scene graph generation problem can be stated as to find the model $P(G|I)$, that predict the scene graph G for given image I:

$$P : I \rightarrow G \tag{1}$$

It can be written as a set of sub-tasks:

$$P(G \mid I) = P(V \mid I)P(E \mid V, I)P(R, O \mid V, E, I), \tag{2}$$

Namely:

1. to create the object region proposal $P(V|I)$,
2. to form objects' relationship proposal $P(E|V, I)$,
3. to label detected objects and relationships $P(R, O|V, E, I)$:

3.1 Framework for multiscale scene graph generation

So, the proposed multiscale scene graph generation (MSGG-AI) framework (Figure 4) includes three main blocks. Firstly, GSS-AI model, based on visual transformer and graph neural network, performs image semantic segmentation and generates object proposals (Emelyanov et al., 2024).

The GSS-AI network model uses 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.

To extend the developed framework for solving the task of multiscale scene graph generation it was modified by adding object classification block and multiscale relation retrieving block.

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.

This procedure allows to perform image semantic segmentation in self-supervising manner resulting in pixel-wise segmentation map. For generating the graph of the scene that reflect hierarchical structure of the scene the modified Graph R-CNN (Yang et al., 2018) network model is used. The connections between features of objects extracted at different scales allow retrieving the scene graphs of different levels of details, according to the scale of the image.

The relationship proposal network RePN (Yang et al., 2018) is used for predicting objects' relationship proposal $P(E \mid V, I)$ (the second term of Equation (2)). The RePN directly models the proposals in end-to-end manner.

For labeling the predicted scene graph (the third term of Equation (2)) the Gated Recurrent Unit (GRU) (Cho et al., 2014) is applied. It refines the scene graph in iterative mode.

3.2 Dataset

We use segmentation and vectorization of aerial imagery (SVAI) (Knyaz et al., 2025) dataset as a basis for creating the intended dataset for tasks of multi-scale scene graph generation. It was extending by adding aerial images of different scales and their annotations.

The SVAI (Aerial Image Segmentation and Vectorization) dataset is designed for aerial image analysis tasks, including change detection, segmentation and vectorization, as well as scene graph generation (Emelyanov et al., 2024). It contains 8,400 very high-resolution aerial photographs of various scenes taken at different times and using different sensors. During this study it has been extended by including about 2,500 aerial images of different scale corresponding to standard set of scales for maps. The SVAI data section for detecting changes contains two thousand pairs of images of the same scenes taken at different times and containing changes in the scene. The change detection section contains annotations for training and testing neural network models for change detection, and the annotations are binary masks marked with zero for unchanged areas and a non-zero value for changed ones.

To study the problem of generating multi-scale semantic scene graphs, the SVAI dataset was expanded by including images of different scales for the same area and annotating categories of objects and relationships at different scales.

Firstly, the objects shown in the images were divided into 32 classes, such as buildings, roads, rivers, bridges, as well as high-level categories such as factories, settlements, airports, etc. The classification classes were selected in accordance with the classification of topographic maps for further rapid adaptation to the task of updating maps.

Secondly, 16 categories of relationships were introduced, representing spatial topology, functional description and hierarchy of objects. They describe possible relationships between objects in the scene, such as nearby, distant, around, passing through, passing under, etc. to get a good initial approximation to the annotation triplets (<`subject - predicate - object`>) we use data from OpenStreetMap[1] resource.

4. Results

We evaluate of the proposed MSGG-AI framework in task of Phrase Detection (PhrDet) (Lu et al., 2016a) 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.

The evaluation has been carried out on the testing split of SVAI dataset and has demonstrated the performance of $R@100$ = 45.12 and $R@50$ = 38.97, being at the state of the art level.

5. Conclusion

The developed MSGG-AI framework for multi-scale semantic scene graph generation of aerial images allows to obtain scene graph representation at different scales according to standard

[1] `https://www.openstreetmap.org`

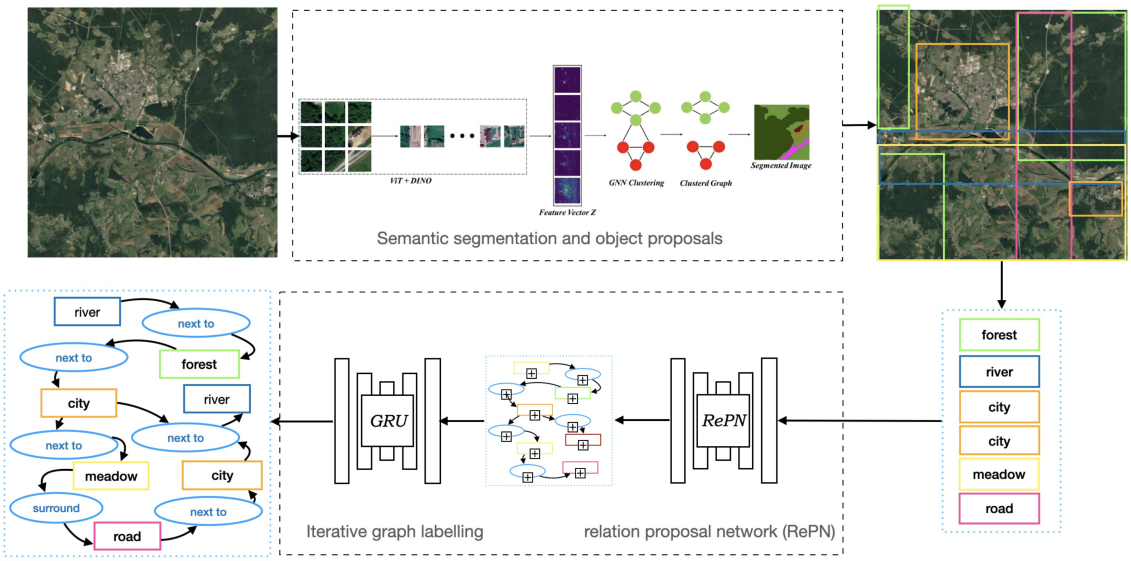

Figure 4. The MSGG-AI framework architecture. Firstly, the pre-trained visual transformer retrieves deep features from the input aerial image of different scales basing on attention mechanism, and graphical neural network performs clustering of these deep features thus creating object region proposals 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.

maps' scales. The MSGG-AI framework uses the Vision Transformer and graph neural network provides accurate image segmentation, and modified Graph R-CNN (Yang et al., 2018) network mode generates the graph of the scene that reflect hierarchical structure of the scene.

We trained the developed network model in task of multi-scale scene graph generation on training split of Segmentation and Vectorization of Aerial Imagery (SVAI) dataset, and then evaluated it on the testing split. The evaluation showed that deeper extraction of the scene graph structure allows to improve performance in terms of the $R@k$ metric for aerial imagery.

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/

The authors are grateful to Tatyana Vostatek for her insightful discussions and support.

References

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.

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.

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/.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D., 2010. Every picture tells a story: Generating sentences from images. K. Daniilidis, P. Maragos, N. Paragios (eds), *Computer Vision – ECCV 2010*, Springer Berlin Heidelberg, Berlin, Heidelberg, 15–29.

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.

Hwang, S. J., Kim, H. J., Ravi, S. N., Collins, M. D., Tao, Z., Singh, V., 2018. Tensorize, factorize and regularize: Robust visual relationship learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1014–1023.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., Fei-Fei, L., 2015a. Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3668–3678.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D., Bernstein, M., Fei-Fei, L., 2015b. 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., Emelyanov, A. V., Smirnov, E. R., 2025. Hierarchical Scene Graph Generation and Vectorization of Aerial Images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W9-2025, 161–167. https://isprs-archives.copernicus.org/articles/XLVIII-2-W9-2025/161/2025/.

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, J., Wang, J., 2003. Automatic Linguistic Indexing of Pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9), 1075-1088.

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., 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., 2016a. Visual relationship detection with language priors. *European conference on computer vision*, Springer, 852–869.

Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L., 2016b. Visual relationship detection with language priors. B. Leibe, J. Matas, N. Sebe, M. Welling (eds), *Computer Vision – ECCV 2016*, Springer International Publishing, Cham, 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.

Peyre, J., Laptev, I., Schmid, C., Sivic, J., 2017. Weakly-supervised learning of visual relations. *2017 IEEE International Conference on Computer Vision (ICCV)*, 5189–5198.

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.

Torresani, L., Szummer, M., Fitzgibbon, A., 2010. Efficient object category recognition using classemes. K. Daniilidis, P. Maragos, N. Paragios (eds), *Computer Vision – ECCV 2010*, Springer Berlin Heidelberg, Berlin, Heidelberg, 776–789.

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.

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.

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., 2018. 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. Auto-encoding 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*.

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.