Aerial Images Segmentation with Graph Neural Network

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

Keywords: Deep learning, Semantic segmentation, Graph convolutional network, Vectorization, Remote sensing images.

Abstract

The development of remote sensing platforms and sensors, as well as the improvement of remote data processing tools and methods, create new opportunities for automatic updating of maps. Currently, aerial photographs serve as the main source for automatic map updates due to their accessibility and significant informational value. One of the core elements for image to maps transition is accurate image segmentation. Nowadays, machine learning methods demonstrate the best results in task of image segmentation. At its core, maps represent information about a certain area in a vector form, that not only contains visual information about area, but also reflects some relations between objects in the map. This quality makes a map more convenient for human perception than an aerial photograph (raster image). This study addresses the problem of accurate aerial image segmentation with taking the advantages of using graph neural network as the more adequate model of map structure. We use graph neural network for retrieving semantic and vector information about a captured area from its aerial image. The developed framework at first phase utilizes visual transformer for retrieving deep features from the input aerial image. The graph neural network then performs clustering of the extracted deep features to obtain semantic segmentation of the image. 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

Timely updating of maps is very important for the proper operation of a wide range of organizations and services. With the recent progress in means for acquiring remote sensing data and striking advances in methods of its processing, the performance of automatic map updating algorithms reached very impressive level. The main improvements in the automatic updating of maps are based on a large amount of collected remote sensing data and on applying machine learning methods for data analysis.

High quality of aerial image segmentation is a key point in accurate map generating. Applying of machine learning methods for image semantic segmentation provided significant improvements in the quality of aerial image segmentation, and now deep learning neural network models demonstrate the state-of-the-art performance for this task. High results of deep neural network models in many machine vision and data analysis task are based on a huge amount of annotated data.

To reduce time and resources for data annotation some approaches were proposed such as semi-supervised (Assran et al., 2021) or weakly-supervised (Ren et al., 2020) techniques. Utilizing of an attention mechanism allowed to develop self-supervised methods for image segmentation. Thus, the use of self-distillation with no labels approach (Caron et al., 2021) in task of Vision Transformer training showed that self-supervised Vision Transformer aggregates explicit information about the semantic segmentation of an image as features, surpassing as supervised Vision Transformers, as convolution neural networks in this task. These extracted deep features contain significant semantic information, that can be used in task of semantic segmentation.

At its core, maps represent information about a certain area in a vector form, that not only contains visual information about area, but also reflects some relations between objects in the map. This quality makes a map more convenient for human perception than an aerial photograph (raster image).

The better the model reflects the essential features of the studied process (object), the better results it gives at the research stage. Recently introduced graph neural networks demonstrate the state-of-the-art performance on graph-structured data, in such domains as natural language processing (Document Classification, Text Generation, Question Answering, Sentiment Analysis), Bioinformatics (Protein-Protein Interaction Prediction, Genomic Sequence Analysis, Drug Discovery), traffic analysis and forecasting (Jiang and Luo, 2022).

In this study we apply graph neural network for retrieving semantic and vector information about a captured area from its aerial image. The developed framework at first phase utilizes visual transformer for retrieving deep features from the input image. The graph neural network then performs clustering of the extracted deep features to obtain semantic segmentation of the image.

To train and evaluate the developed framework, a special dataset is collected and annotated. It contains 10 thousand aerial photographs representing various types of objects taken in different years and seasons.

The main contributions of the presented study are the following:

- the framework for accurate aerial image segmentation based on graph neural network,
- the special dataset containing aerial images of various landscapes and land-use territories,
- results of evaluation of the developed framework and baselines on the created dataset.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-W1-2024-1-2024 | © Author(s) 2024. CC BY 4.0 License.

2. Related work

2.1 CNN for semantic segmentation

Recent advancements in deep learning (LeCun et al., 2015, Li et al., 2020c, Emelyanov et al., 2024) have greatly improved the deep semantic segmentation network (DSSN) for semantic segmentation of remote sensing (RS) images (Basaeed et al., 2016, Ouyang and Li, 2021) when compared to traditional methods like random forest (RF), decision trees (DT), and support vector machines (SVMs) (Camps-Valls et al., 2013). The fully convolutional network (FCN) was proposed by (Long et al., 2015) by adding deconvolution layers to the convolutional neural network (CNN), which became a new development in end-to-end semantic segmentation.

U-Net (Ronneberger et al., 2015) used skip connections to utilize multiscale information as the encoder-decoder architecture representative. U-Net achieved good results by combining low-level detail and high-level semantic information with skip connections, which strengthened the feature maps. Additionally, SegNet (Badrinarayanan et al., 2017) used the index of max pooling in the encoder to conduct nonlinear upsampling in the decoder.

Recently, there have been numerous advancements in DSSN-based RS image semantic segmentation within the field of RS. Several studies have utilized FCN for the task of semantic segmentation in remote sensing images (Wurm et al., 2019, Sherrah, 2016). (Kampffmeyer et al., 2016) introduced a new DSSN for the purpose of mapping urban land cover. (Wang and Li, 2020) utilized an ensemble multiscale residual deep learning approach inspired by the U-Net structure for building extraction. (Audebert et al., 2016) utilized a variants network of SegNet and incorporated multicore convolutional layers to efficiently gather predictions across different scales.

(Zhang et al., 2017) suggested using the DMSMR network to enhance segmentation performance. (Pan et al., 2018) used a detailed segmentation network to label objects in high-resolution aerial images. In order to use multisensor data (Kniaz, 2018, Knyaz et al., 2024), such as thermal imagery, both the RGB image and the multimodal data are combined to provide more information for DSSN. (Marmanis et al., 2016) suggested using a Siamese network to process images and DSM data, and included edge detection and semantic segmentation in the upgraded version. (He et al., 2020a) add edge information to DSSN to improve the segmentation results. Previous knowledge is applied to semantic segmentation of remote sensing images. (Alirezaie et al., 2019) used U-Net for quick and precise pixel classification, along with a knowledge-based post-processing step.

DSSN uses deep features to show the category of each pixel, highlighting the significance of features in semantic segmentation. To get better features, the network's expressiveness needs to be improved. Like the human visual system, the attention mechanism boosts important features and reduces less important features. In the channel domain, features are chosen based on their importance in the channel dimension. (Hu et al., 2018) proposed an squeeze-and-excitation (SE) block that adjusts feature responses across channels by modeling channel interdependencies. Spatial domain attention assigns different weights to pixels based on their positions to introduce spatial context. The U-net attention model (Oktay, 2018) utilizes an attention gate module to control the significance of features in various spatial locations. A Semantic Segmentation Network called SCAttNet (Li et al., 2020a) was introduced for RS image segmentation. It utilized a convolutional block attention module (CBAM) (Woo et al., 2018) with spatial and channel attention.

2.2 Graph convolutional network

The objects scattered on the ground are interconnected in various ways, forming a graph. In this graph, nodes represent the objects, while the edges represent their spatial relationships, like being nearby, overlapping, or apart.

While DSSNs have excelled in processing Euclidean data, their performance falls short when handling graph data, which exists in a non-Euclidean space. Graph convolutional neural network (GCN) has clear advantages in extracting features from irregular graph data by applying deep learning techniques. Through graph convolution, GCN is able to excel in analyzing graph data. Graph convolution utilizes edge connections to aggregate node information and create new node representations. This allows GCN to effectively capture the dependency relationship among graph nodes.

These benefits have fostered the advancement of studies involving graph analysis (Gori et al., 2005). (Welling and Kipf, 2016) developed a layered propagation graph model that utilized convolution in spectral space to directly process graph data. NN4G (Niepert et al., 2016) implemented graph convolution in the spatial domain by directly collecting information from neighboring nodes. (Li et al., 2019) successfully developed Deep-GCN by incorporating ideas from CNN, such as residual connections and dilated convolutions, to address the issue of gradient vanishing in the original GCN model, which was confined to shallow layers.

When aggregating the neighbor information, the attention mechanism allows Graph attention network (GAT) (Veličković et al., 2017) to calculate the weight of each neighbor node in relation to the central node. GCN focuses more on spatial relations than similarity weights, unlike GAT. (Lu et al., 2019) suggested utilizing the precise location of each pixel in the image for semantic segmentation through a pixel-based GCN model, which was initialized by an FCN. While each pixel contains its own local position, it fails to accurately depict ground objects and disregards the importance of spatial relationships.

To address multilabel aerial image scene classification, (Li et al., 2020b) presented a CNN-GCN framework that mines both the object information and topological relationships among multiple objects. CNN's abstract features are useful for scene classification, but pixel-level semantic segmentation entails providing specific details for categorizing each pixel.

3. Material and Method

With the aim to develop a data-driven method for accurate aerial image segmentation, we collected and annotated a special dataset designed for the tasks of aerial imagery segmentation, vectorization and change detection. This dataset was then used to train and test the proposed framework.

3.1 Dataset

The developed algorithm was trained on a purpose-built dataset for aerial imagery segmentation, vectorization and change detection tasks (SVAI dataset). SVAI dataset is a novel large-scale

Figure 1. Example images from the Segmentation and Visualization Aerial Images dataset. Images in columns 1 and 3 were obtained from Bing and Google satellites in 2018 and 2024, respectively; columns 2 and 4 show changes (destroyed objects are shown in red, and newly appeared objects are shown in blue).

```
<node id="625043" lat="42.5276604" lon="1.5690867" version="3" timestamp="2011-08-05T00:16:14Z" changeset="0"/>
<node id="625050" lat="42.5299751" lon="1.5721059" version="5" timestamp="2016-11-28T08:42:28Z" changeset="0">
    <tag k="highway" v="crossing"/>
```

Figure 2. An example of code for describing a node.

remote sensing dataset for building change detection, segmentation and vectorization. SVAI consists of 2000 pairs of Bing and Google Earth satellite images with very high resolution (VHR, 0.34 m/pixel) and a size of 512x512 pixels. These images with a time span of 6 years feature significant land-use changes, especially growth of buildings and roads.

SVAI covers different building types, such as single-family houses, small garages, and large warehouses. It primarily focuses on building-related changes, including building growth (change from soil, grass, or building under construction to new built-up areas) and building decay, as well as the emergence of new road routes.

The images used are annotated by remote sensing imagery interpretation experts using binary labels (1 for change and 0 for unchanged). Each sample in the dataset is annotated by one annotator and then double-checked by another to obtain high-quality annotations. A fully annotated SVAI contains over 40,000 individual building and road change instances.

The images in SVAI were obtained during surveys of the border of the Tula and Moscow regions, namely the villages of Kostino, Parshino, Lukyanovo, Verkhneye Romanovo, etc. Figure 1 shows example images from the dataset. The Bing satellite images were taken in mid-2018, and the Google Earth satellite images were taken in mid-2024.

For the vectorization task, SVAI was supplemented with data from OpenStreetMap. OpenStreetMap (literally ”open street map”), abbreviated OSM, is a non-profit web mapping project aimed at creating a detailed, free and open-source geographic map of the world by a community of participants — Internet users. To create maps, data from personal GPS trackers, aerial photography, video recordings, satellite images and street panoramas provided by some companies, as well as the knowledge of the person drawing the map, are used.

OpenStreetMap uses the wiki principle when creating a map. Each registered user can make changes to the map. The project data is distributed under the terms of the free Open Database License. The data obtained from OpenStreetMap provides a vector format of streets and buildings, which is necessary for vectorization. The sample data from the dataset intended for the tasks of aerial images vectorization is shown in Figure 3. We use open data from OpenStreetMap to initially annotate images for task of vectorization.

The original data file in OSM format is an XML format. The OSM file contains a set of object such as nodes, ways, relations and tags.

Nodes are points with a unique identifier and a pair of coordinates. Nodes can be independent objects (with descriptive tags), and also be part of ways and relations. The sample code describing a node is shown in Figure 2.

Ways are a set of nodes. It can be independent objects (with descriptive tags), and also be part of relations. The sample code describing a way is shown in Figure 4.

Relations can contain nodes, ways and other relations. The sample code describing a relation is shown in Figure 5.

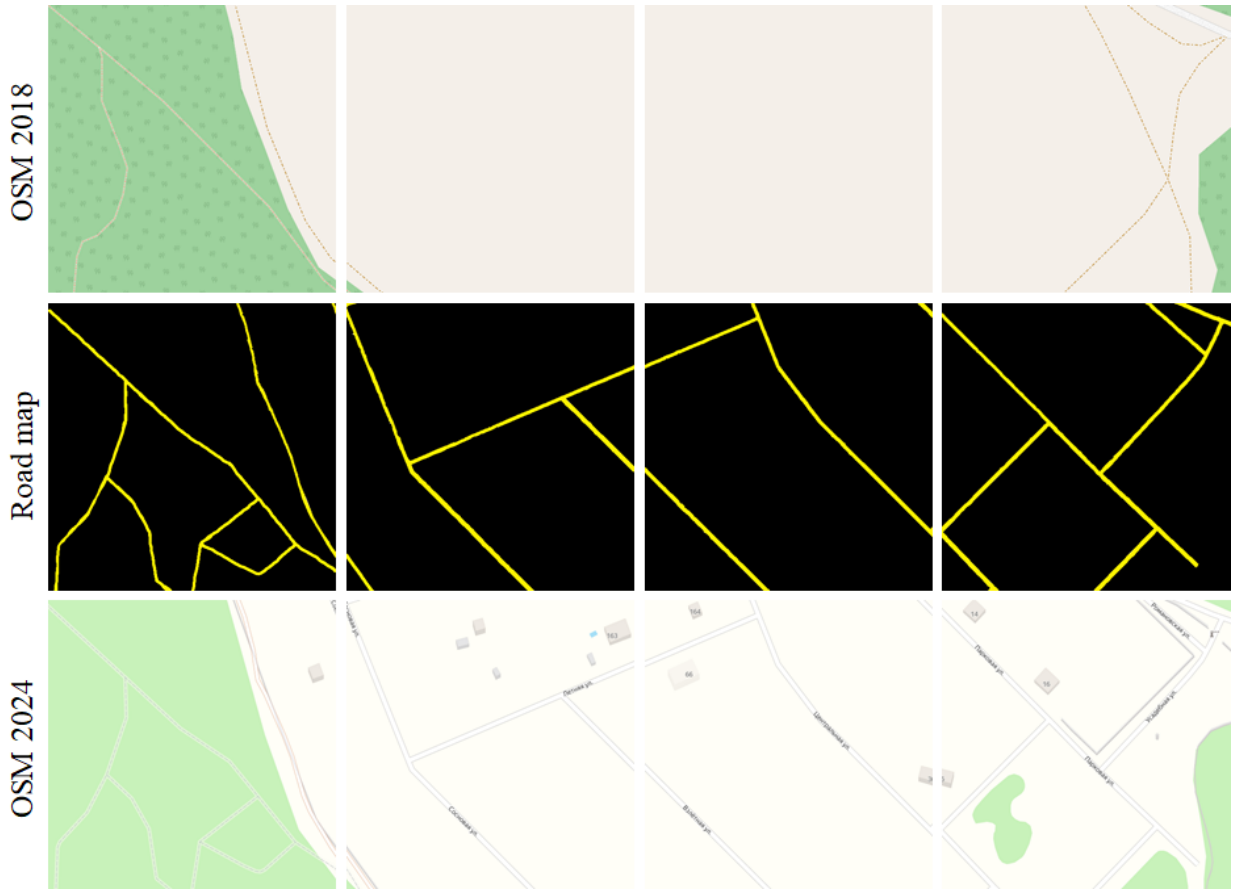

Figure 3. Example images from the Segmentation and Visualization Aerial Images dataset. The top and bottom rows are OSM images from 2018 and 2024, respectively, and the middle row is changes to the roadmap during this period of time.

```
<way id="8880955" version="4" timestamp="2018-08-18T14:17:53Z" changeset="0">
    <nd ref="64796309"/>
    <nd ref="64796310"/>
    <tag k="waterway" v="riverbank"/>
```

Figure 4. An example of code for describing a way.

```
<relation id="7380238" version="1" timestamp="2017-07-07T11:39:49Z" changeset="0">
    <member type="way" ref="6181313" role="from"/>
    <member type="node" ref="51390142" role="via"/>
    <member type="way" ref="489466870" role="to"/>
    <tag k="restriction" v="no_u_turn"/>
    <tag k="type" v="restriction"/>
</relation>
```

Figure 5. An example of code for describing relation.

3.2 Framework architecture

The proposed framework firstly exploits visual transformer to extract deep object features from the input aerial image basing on attention mechanism. Then a graphical neural network is applied to cluster these deep features into homogeneous segmented areas. The proposed architecture is shown in Figure 6.

3.2.1 Deep Features Retrieving To retrieve deep features from aerial image we apply Vision Transfomer (Dosovitskiy et al., 2020) trained with DINO (Caron et al., 2021).

DINO framework has the overall structure similar to modern self-supervised approaches (Caron et al., 2020, Grill et al., 2020, He et al., 2020b). It exploits the advantages of knowledge distillation (Hinton et al., 2015) to improve the performance in several tasks of computer vision such image segmentation, object detection, video instance segmentation, etc. This feature allows DINO framework to identify different areas in the image, that can be then used for accurate segmentation.

Vision Transfomer divides input image I into k^2 patches of size

$$s = \frac{h}{k} \times \frac{w}{k}, \tag{1}$$

where h and w are the height and width of the image I respectively. For the token embedding dimension t, the output feature vector Z has a size of $s \times t$.

The samples of DINO attention maps from multiple heads of Vision Transformer are shown in Figure 7.

These attention maps contain significant information about an image that can be used for segmentation. If to consider the similarity between image patches as a graph, then the problem of the image segmentation can be formulated as graph-cut task. Such technique was applied by a number of authors (Shi and Malik, 2000, Bansal et al., 2004, Bagon and Galun, 2011) with different approaches for finding optimal the graph-cut.

3.2.2 Image Graph Representation Let represent an image as an undirected graph $G = \{\mathcal{V}, \mathcal{E}\}$ with node set $\mathcal{V}$ and edge set $\mathcal{E}$, where each node corresponds to the image area. Than the task of image segmentation can be formulated as the task of grouping the nodes basing on their similarity, or, other word, to divide the graph into a sets of disjointed homogenious parts.

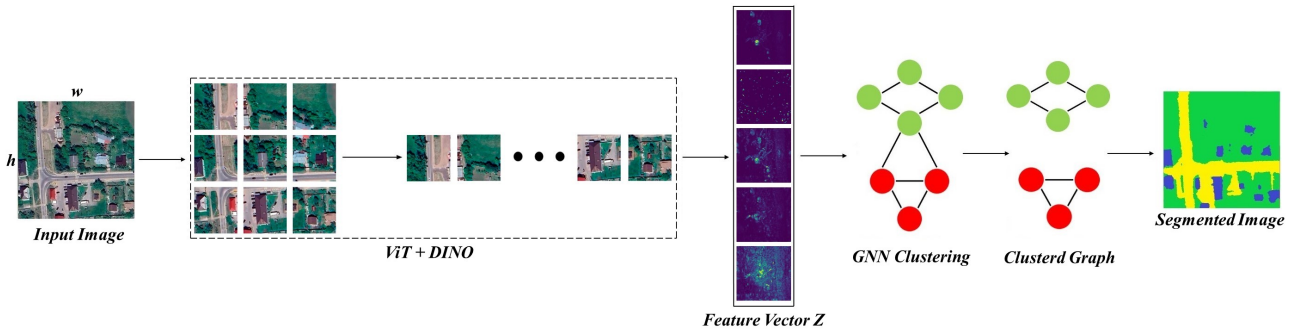

Figure 6. The framework architecture. Firstly, the pre-trained visual transformer retrieves deep features from the input aerial image basing on attention mechanism. Secondly, graphical neural network performs clustering of these deep features into homogeneous segmented parts.

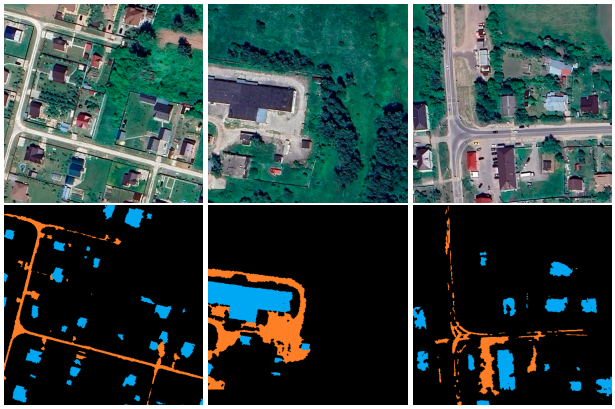

Figure 7. Attention maps from multiple heads of ViT-S/8 for `[CLS]` token query.

This task can be formulated as follows (Aflalo et al., 2023). It is required to divide the graph into the k disjoint sets $A_1, A_2 ... A_k$ such that $\cup_i A_i = \mathcal{V}$ and $\forall_{j \neq i} A_i \cap A_j = \emptyset$. This partition can be expressed as a binary matrix $S \in \{0,1\}^{n \times k}$ where $S_{ic} = 1$ if $i \in A_c$.

As a criterium of correct partitioning we consider the partition that maximizes the number of connections within the partition, and minimizes the connections between the partitions. The number of the connections between a part A and a part B of a given graph is given by the cut $cut(A, B)$ of the graph G:

$$cut(A, B) = \sum_{u \in A, v \in B} w(u, v). \tag{2}$$

From this follows the definition of a *normalized cut* $Ncut(A, B)$ of the graph G:

$$Ncut(A, B) = \frac{\sum\limits_{u \in A, v \in B} w(u, v)}{\sum\limits_{i \in A, j \in \mathcal{V}} w(i, j)} + \frac{\sum\limits_{u \in A, v \in B} w(u, v)}{\sum\limits_{i \in B, j \in \mathcal{V}} w(i, j)}, \tag{3}$$

So, to perform image segmentation using graph approach we have to define how to establish the similarities between image areas. We characterize the correspondence between image areas by $n \times n$ matrix W, the elements w_{ij} representing the similarity between image area i and image area j, $i, j = 1 \ldots n$. For these purpose we construct similarity matrix W using output deep feature vector Z from Vision Transformer.

$$W = zz^T \in \mathbb{R}^{k \times k}. \tag{4}$$

3.2.3 Graph Neural Network Clustering Let $\hat{X}$ be the matrix of node representations yielded by one or more layers of GNN convolution on a graph G with an adjacency matrix A.

Using the patch-wise correlation matrix from the extracted ViT features and a single-layer Graph Convolution Neural Network, we build a graph and then compute the cluster assignment of nodes using a multilayer perceptron (MLP) with softmax at the output layer:

$$\hat{X} = GNN(X, A, \Theta_{GNN}), \tag{5}$$

$$S = MLP(\hat{X}, \Theta_{MLP}), \tag{6}$$

where Θ_{MLP} and Θ_{GNN} are trainable parameters. The softmax activation of the multilayer perceptron guarantees that $s_{ij} \in [0, 1]$ and enforces the constraints $S1_K = 1_N$.

The Graph Convolution Neural Network is optimized using the normalized-cut relaxation proposed in (Bianchi et al., 2020).

The loss function is:

$$\mathcal{L}_{Ncut} = -\frac{Tr(S^T AS)}{Tr(S^T DS)} + \| \frac{S^T S}{\|S^T S\|_F} - \frac{\mathbb{I}_3}{\sqrt{3}} \|_F, \tag{7}$$

where $\| \cdot \|_F$ indicates the Frobenius norm and D is the degree matrix of A. The number 3 denotes the number of disjoint sets we aim to partition the graph into, and $\mathbb{I}_3$ is the identity matrix.

The first term of the objective function promotes the clustering of strongly connected components together, while the second term encourages the cluster assignments to be orthogonal and have similar sizes.

For segmenting an image into more than two clusters, the proposed technique is performed iteratively several times. At each phase the previously selected background is claustered into new background and new objects of given class by the developed technique.

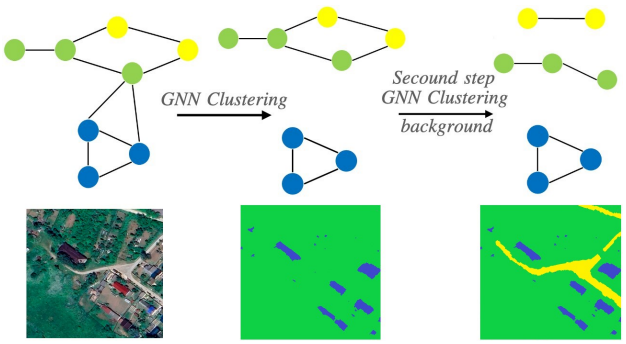

Figure 8. Iterative clustering.

The results of the iterative semantic segmentation of an aerial image is shown in Figure 8.

Such approach allows performing semantic segmentation for the given set of classes.

4. Results

The evaluation of our unsupervised segmentation was performed on developed SVAI dataset. We compare our unsupervised approach to state-of-the-art unsupervised methods. Mean Intersection-over-Union (mIoU) metric has been used as a measure of the accuracy of segmentation. The comparison of the numerical values of the mIoU for each of the methods is shown in Table 1.

Method	SVAI dataset
OneGAN (Benny and Wolf, 2020)	57.48
Voynov (Voynov et al., 2021) et al.	69.12
Spectral Methods (Melas-Kyriazi et al., 2022)	75.81
TokenCut (Wang et al., 2022)	72.36
Our method	**77.83**

Table 1. Values of the mIoU (mean Intersection-over-Union) metric on the developed dataset for various algorithms.

The qualitative results of segmentation by the proposed framework are shown in Figure 9.

The results of the framework evaluation on the SVAI dataset shows that the proposed technique demontstrates the state-of-the-art performance in the task of aerial images semantic segmentation.

5. Conclusion

The framework for accurate aerial image segmentation, based on graph neural network is developed.

We use graph neural network for retrieving semantic and vector information about a captured area from its aerial image. The developed framework at first phase utilizes visual transformer for retrieving deep features from the input aerial image. The graph neural network then performs clustering of the extracted deep features to obtain semantic segmentation of the image.

To train and evaluate the developed framework, a special dataset is collected and annotated. It contains more than 10k aerial

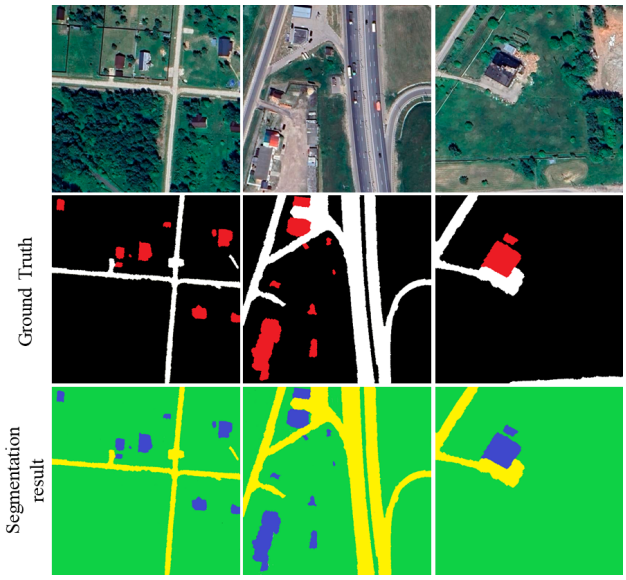

Figure 9. Semantic part segmentation. Top: original image; middle: ground truth; bottom: proposed method.

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.

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

Aflalo, A., Bagon, S., Kashti, T., Eldar, Y., 2023. DeepCut: Unsupervised Segmentation using Graph Neural Networks Clustering. *preprint arXiv:2212.05853*. https://arxiv.org/abs/2212.05853.

Alirezaie, M., Längkvist, M., Sioutis, M., Loutfi, A., 2019. Semantic referee: A neural-symbolic framework for enhancing geospatial semantic segmentation. *Semantic Web*, 10(5), 863–880.

Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Ballas, N., Rabbat, M., 2021. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8443–8452.

Audebert, N., Le Saux, B., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. *Asian conference on computer vision*, Springer, 180–196.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2481–2495.

Bagon, S., Galun, M., 2011. Large scale correlation clustering optimization. *arXiv preprint arXiv:1112.2903*.

Bansal, N., Blum, A., Chawla, S., 2004. Correlation clustering. *Machine learning*, 56(1), 89–113.

Basaeed, E., Bhaskar, H., Al-Mualla, M., 2016. Supervised remote sensing image segmentation using boosted convolutional neural networks. *Knowledge-Based Systems*, 99, 19–27.

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.

Bianchi, F. M., Grattarola, D., Alippi, C., 2020. Spectral clustering with graph neural networks for graph pooling. *International Conference on Machine Learning*, PMLR, 874–883.

Camps-Valls, G., Tuia, D., Bruzzone, L., Benediktsson, J. A., 2013. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE signal processing magazine*, 31(1), 45–54.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*.

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.

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., Knyaz, V. A., Kniaz, V. V., 2024. Extracting building outlines based on convolutional neural networks using the property of linear connectivity. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1-2024, 147–152. https://isprs-archives.copernicus.org/articles/XLVIII-1-2024/147/2024/.

Gori, M., Monfardini, G., Scarselli, F., 2005. A new model for learning in graph domains. *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.*, 2, IEEE, 729–734.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS*.

He, C., Li, S., Xiong, D., Fang, P., Liao, M., 2020a. Remote sensing image semantic segmentation based on edge information guidance. *Remote Sensing*, 12(9), 1501.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020b. Momentum contrast for unsupervised visual representation learning. *CVPR*.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. *preprint arXiv:1503.02531*.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jiang, W., Luo, J., 2022. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 117921.

Kampffmeyer, M., Salberg, A.-B., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1–9.

Kniaz, V. V., 2018. Conditional GANs for semantic segmentation of multispectral satellite images. L. Bruzzone, F. Bovolo (eds), *Image and Signal Processing for Remote Sensing XXIV*, 10789, International Society for Optics and Photonics, SPIE, 107890R.

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/.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *nature*, 521(7553), 436–444.

Li, G., Muller, M., Thabet, A., Ghanem, B., 2019. Deepgcns: Can gcns go as deep as cnns? *Proceedings of the IEEE/CVF international conference on computer vision*, 9267–9276.

Li, H., Qiu, K., Chen, L., Mei, X., Hong, L., Tao, C., 2020a. SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 18(5), 905–909.

Li, Y., Chen, R., Zhang, Y., Li, H., 2020b. A cnn-gcn framework for multi-label aerial image scene classification. *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 1353–1356.

Li, Y., Zhang, Y., Zhu, Z., 2020c. Error-tolerant deep learning for remote sensing image scene classification. *IEEE transactions on cybernetics*, 51(4), 1756–1768.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lu, Y., Chen, Y., Zhao, D., Chen, J., 2019. Graph-fcn for image semantic segmentation. *International symposium on neural networks*, Springer, 97–105.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., Stilla, U., 2016. Semantic segmentation of aerial images with an ensemble of CNSS. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, 3, 473–480.

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.

Niepert, M., Ahmed, M., Kutzkov, K., 2016. Learning convolutional neural networks for graphs. *International conference on machine learning*, PMLR, 2014–2023.

Oktay, O., 2018. Attention u-net: Learning where to look for the Pancreas. *arXiv preprint arXiv:1804.03999*.

Ouyang, S., Li, Y., 2021. Combining Deep Semantic Segmentation Network and Graph Convolutional Neural Network for Semantic Segmentation of Remote Sensing Imagery. *Remote Sensing*, 13(1). https://www.mdpi.com/2072-4292/13/1/119.

Pan, X., Gao, L., Marinoni, A., Zhang, B., Yang, F., Gamba, P., 2018. Semantic labeling of high resolution aerial imagery and LiDAR data with fine segmentation network. *Remote sensing*, 10(5), 743.

Ren, Z., Yu, Z., Yang, X., Liu, M.-Y., Lee, Y. J., Schwing, A. G., Kautz, J., 2020. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10598–10607.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 234–241.

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8), 888–905.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

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, C., Li, L., 2020. Multi-scale residual deep network for semantic segmentation of buildings with regularizer of shape representation. *Remote Sensing*, 12(18), 2932.

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.

Welling, M., Kipf, T. N., 2016. Semi-supervised classification with graph convolutional networks. *J. International Conference on Learning Representations (ICLR 2017)*.

Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., 2018. Cbam: Convolutional block attention module. *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wurm, M., Stark, T., Zhu, X. X., Weigand, M., Taubenböck, H., 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing*, 150, 59–69.

Zhang, M., Hu, X., Zhao, L., Lv, Y., Luo, M., Pang, S., 2017. Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images. *Remote Sensing*, 9(5), 500.