WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Earth Objects Extraction from High Spatial Resolution Satellite Images

Vineet Maurya[1], Siddharth Singh[2], Shivanshu[3]

[1]Department of Computer Science and Engineering, Galgotias College of Engineering and Technology , Greater Noida, India
[2]Department of Computer Science and Engineering, Galgotias College of Engineering and Technology , Greater Noida, India
[3]Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India
[1]vineetmaurya.4985@gmail.com, [2]siddharthsingh590@gmail.com, [3]shivamsinghhiro@gmail.com

ABSTRACT

The birth of Earth features from high-resolution satellite imagery is pivotal for civic planning, environmental monitoring, and disaster operation. This study employs deep literacy ways, fastening on the UNet armature, for pixel-wise semantic segmentation of upstanding imagery. Exercising a dataset from Humans in the Loop and Mohammed Bin Rashid Space Center(MBRSC), 72 high-resolution images of Dubai were annotated into six classes: structures, roads, foliage, water, unpaved land, and unlabeled regions. The exploration evaluates bracket models like ResNet50, VGG16, and MobileNet, relating their limitations in point delineation. A segmentation- driven approach using UNet, with preprocessing way similar to resizing and marker garbling, achieved superior delicacy. UNet demonstrated high crossroad over Union(IoU) scores and precise point birth, outperforming bracket styles. This frame offers robust operations in geospatial analysis, structure planning, and environmental conservation.

INTRODUCTION

REMOTE sensing image interpretation is an important way to delineate structures for civic planning. The poor effectiveness and time-consuming nature of artificial interpretation have made automatic and semi automatic structure birth algorithms hot motifs in the last decades [1],[9]. With the development of remote sensing imaging technologies, the spatial resolution of acquired data continues to ameliorate. therefore, erecting vestiges uprooted from remote seeing images are getting more detailed. For case, images with resolution of hundreds or knockouts of measures, e.g., MODIS, are frequently exploited to identify large-scale erected-up areas on the Earth's face[10],[12]. Individual structures can be delineated from cadence- or sub-cadence resolution images, e.g., WorldView, QuickBird, or UAV upstanding images[5]. In some extremities with time limitations similar as disaster assessment, individual structures need to be delineated[9] as snappily as possible. Still, it's delicate to gain high-resolution(HR) images snappily. By discrepancy, some data with lower spatial resolution are open access. However, the difficulty of HR data accession could be avoided, If these data could be employed to produce semantic charts of structures. former exploration workshop have concentrated on combining traditional machine learning algorithms similar as support vector machine[1] and handcrafted features similar as the morphological structure indicator[1] and morphological shadow indicator[9] to break the problem of erecting birth[10]. As remote seeing data volume and complexity increase, traditional styles can not gain superior performance. still, the development of deep literacy (DL) has catalyzed a great revolution in the processing of remote seeing data and erecting birth. The operation of convolutional neural networks (CNNs) to semantic segmentation (SS) can extensively increase the delicacy of erected-up mapping. The completely convolutional neural network (FCN) is the first high- profile CNN- grounded SS network[11]. An encoder–decoder structure further improves the effect of SS; typical networks are SegNet[12] and U-Net[13]. The rearmost DeepLab V3 of the DeepLab series outperformed numerous state-of-the-art SS networks on two extensively used datasets in 2018. Motivated by the below-mentioned work, colorful DL-based styles aimed at structure footmark birth have been proposed. S Paisitkriangkrai et al.[13] presented the first attempt to apply CNN and tentative arbitrary fields to remote seeing image pixel labeling. The work demonstrated the effectiveness of CNNs for erecting birth. Still, handcrafted features and arbitrary timber were still employed to increase the performance due to the weak representation capability of shallow CNNs. Latterly, an end-to-end literacy system grounded on FCN was proposed to delineate different objects on Earth. The system performed well on the land cover mapping task, although multi-network integration was needed to gain the stylish results. The U-net is a convolutional network architecture for fast and precise segmentation of images. Up to now, it has outperformed the previous stylish system(a sliding- window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron bitsy heaps. It has won the Grand Challenge for Computer- Automated Discovery of Caries in Bitewing Radiography at ISBI 2015, and it has won the Cell Tracking Challenge at ISBI 2015 on the two most grueling transmitted light microscopy orders(Phase discrepancy and DIC microscopy) by a large periphery

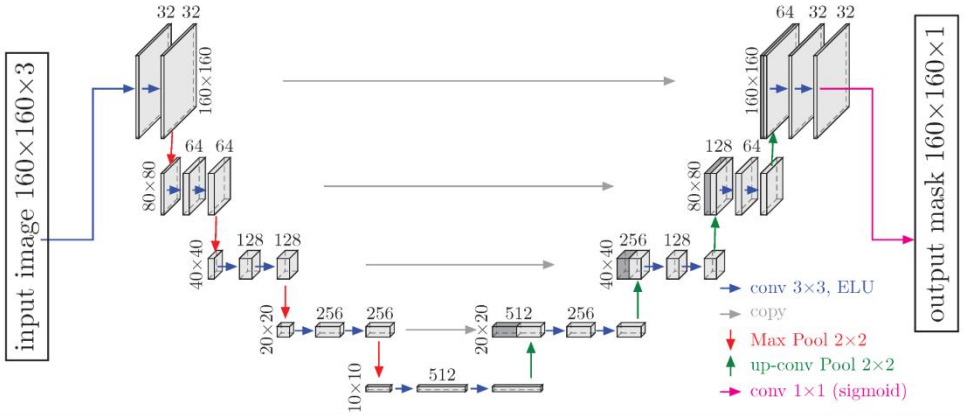

Figure 1. A U-Net model unit with characteristics of layer and filter sizes is used in our model.

A. Model Architecture

Each image goes through a data addition process and is also cropped to the size of 160 × 160 pixels (Figure 1) to serve as input to the three separate U-nets. The size of 256 × 256 was chosen for image gyration during the data addition to help any missing values in the 160 × 160 image. also, the image goes to three different paths that each start with an independent U-net model. The three U-nets independently member the object border, the object member, and the inner member of the object. The border and the inner parts are created directly from the object member; that is, the border goes from 4 pixels outside the object member to 3 pixels outside and the inner member is the object member reduced by 2 pixels. By doing this, the three masks present an imbrication of 1 – 3 pixels. After this, the three activation layers of 160 × 160 pixels performing from the U-nets are concatenated in each of the paths. The following way are two convolutional layers with 64 and 32 pollutants, independently. Eventually, the prognostications are made independently for the member, the border, and the inner member using the last convolutional subcaste with a sigmoid activation function. The case individualizations are made in post treatment by rooting the inner parts (which are unique and don't touch each other) and adding to them a buffer of 2 pixels, that is, the number of pixels that live between the member and the inner member(as they've been created). The model has a aggregate of parameters, of which are trainable.

B. dataset :

To train the segmentation model, a dataset of 72 high-resolution images from Dubai was used. This dataset, handed by the Mohammed Bin Rashid Space Center (MBRSC) and Humans in the Loop, was annotated into six orders structures, roads, foliage, water, unpaved land, and unlabeled regions. The diversity of civic geographies represented in this dataset made it ideal for erecting segmentation tasks.

Exercising a dataset from Humans in the Loop and Mohammed Bin Rashid Space Center (MBRSC), 72 high-resolution images of Dubai were annotated into six classes: structures, roads, foliage, water, unpaved land, and unlabeled regions.

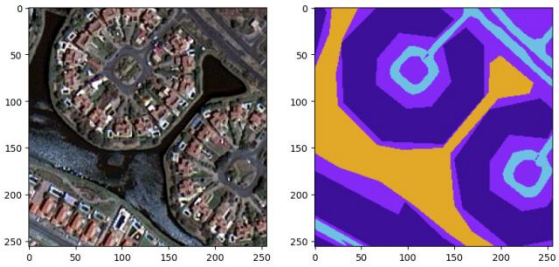

Figure 2. The dataset has an original image and its mask.

C. Training

Trimming the image and the masks in 128 × 128 pixels over the region where structures were manually delineated resulted in a sample of 2048 images and their associated labeled masks to train the model. Among these images, 1435 contained roofs and background, and 613 contained only background. Also, 1638 images were used for the training and 410 for independent confirmation. The size of 128 × 128 pixels was named because (i) studied objects are generally lower than 128 pixels in size; (ii) the objects aren't so dependent on a larger environment; and (iii) we don't want the algorithm to learn a larger environment. An illustration of a large environment would be ‘ houses always do near asphalt thoroughfares’. The images were uprooted from invariant grids of 128 × 128 pixels without any imbrication between bordering images. Also, 128 × 128 images were enlarged to 256 × 256 extents by adding 64 rows and columns on each side. Eighty percent of these images were used for training and 20 percent for confirmation. During network training, we used a standard stochastic gradient descent optimization. The loss function was designed as a sum of two terms, doublecross-entropy and Bone's measure-affiliated loss of the three prognosticated masks.

D. Segmentation Accuracy Assessment

Three performance criteria were reckoned. First, the overall delicacy was reckoned as the chance of rightly classified pixels. Second, the F1 score was reckoned for each class i as the harmonious normal of the perfection and recall(Equation(1)), where perfection is the rate of the number of parts classified rightly as i to the number of all parts(true and false positive) and recall is the rate of the number of parts classified rightly as i to the total number of parts belonging to class i(true positive and false negative). This score varies between 0 (smallest value) and 1 (Best value).

$$F1i=2\times precisioni\times recalli(precisioni+recalli)$$

[1]

Third, to estimate the delicacy of the case segmentation, the crossroad over union(IoU) metric was reckoned as the crossroad of areas labeled as objects in the vaticination and in the reference divided by the union of areas labeled as objects in the vaticination and in the reference. To cipher the IoU of each object, we attributed to each individual object the prognosticated member that showed the largest lapping to the observed object.

E. Prediction

For prediction, the WorldView-3(WV-3) pipe of 16,384 × 16,384 pixels was cropped with a regular grid with cells of 512 × 512 pixels, and 64 neighbor pixels were added on each side to produce an imbrication between the patches. However, due to the pipe border, it was filled by the symmetrical image of the non-blank portion. If there was a remaining blank portion(for illustration. The prognostications were made on these images of 640 × 640 pixels, and the performing images were cropped to 512 × 512 pixels and intermingled again to reconstitute the original 16,384 × 16,384 pixels WV-3 pipe. This lapping system was used to avoid border vestiges during vaticination, a given problem for the U-net algorithm(5). To belong to a given class, the pixel vaticination value must be less than or equal to 0.5. The case segmentation mask was also produced by softening the inner member(mask Subcaste 2) by 2 pixels.

F. Results

The algorithm presents a good position of segmentation delicacy(subcaste mask 1) with an overall delicacy of 86.67 and an F1-score of 0.937(perfection = 0.936 and recall = 0.939). The mean crossroad over union was 0.582, and the standard was 0.694. Considering the entire WorldView-3 pipe, the algorithm delineated and personalized 7477 structures. The model segmentation for the three masks was veritably accurate, as seen by the member in unheroic. There are veritably many crimes for the member and inner member, as seen in blue. The crimes feel slightly advanced for the border mask. This error of the border is further propagated to the case polygons, but overall, the case segmentation can be considered correct.

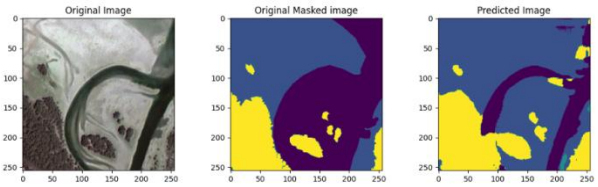

Figure 3. Our model resulted in a Predicted Mask Image

LITERATURE SURVEY

Satellite imagery segmentation has emerged as a cornerstone in geospatial analytics, enabling automated analysis of land use, vegetation, water resources, and urban expansion. Its diverse applications span disaster management, environmental conservation, urban planning, and precision agriculture. Over the years, significant advancements have been made in methodologies, transitioning from traditional statistical models to sophisticated deep learning architectures that leverage the massive volume of high-resolution satellite data available today.

A. Traditional Methods of Satellite Imagery Segmentation

The early methods for analyzing satellite images were predominantly based on pixel-level and object-based approaches. Pixel-based methods relied on statistical techniques like Maximum Likelihood Classification (MLC) and Minimum Distance Classifier (MDC), which operated by evaluating pixel intensities within spectral bands. While these approaches were straightforward and computationally efficient, their inability to handle complex patterns and inter-class spectral overlap posed significant challenges.

Object-Based Image Analysis (OBIA) emerged to address some of these limitations by incorporating spatial, spectral, and contextual information. emphasized OBIA's effectiveness in handling medium-resolution images and its limitations when dealing with the finer details of high-resolution satellite data. Both traditional approaches struggled with defining precise boundaries, handling mixed pixels, and generalizing across datasets from different satellite sensors.

B. Transition to Machine Learning Approaches

Machine learning introduced a paradigm shift by enabling the classification of land cover features using supervised and unsupervised learning algorithms. Support Vector Machines (SVMs) and Random Forests (RF) were extensively adopted due to their ability to generalize across diverse datasets. Pal and Mather (2005) demonstrated SVM's robustness in classifying multi-spectral satellite data, outperforming traditional methods in accuracy and adaptability. Random Forests, as highlighted by Belgiu and Drăguţ (2016), offered efficient handling of large datasets and reduced overfitting, making them a preferred choice for segmentation tasks. However, machine learning methods still depended heavily on handcrafted features, requiring domain expertise and extensive feature engineering. This dependency limited their scalability for real-time or large-scale applications involving high-resolution images.

C. Emergence of Deep Learning Techniques

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), marked a transformative phase in satellite image segmentation. CNNs [1] automated feature extraction by learning hierarchical representations directly from raw image data, addressing the limitations of feature engineering.

U-Net, introduced by [5], became the de facto model for biomedical image segmentation and was rapidly adopted for remote sensing tasks. Zhang et al. (2018) demonstrated its effectiveness in segmenting high-resolution satellite imagery, noting its ability to capture fine-grained details while maintaining computational efficiency. Models like SegNet [11] and DeepLab (Chen et al., 2017) further advanced the field by incorporating techniques such as dilated convolutions and conditional random fields for improved boundary refinement.

Transfer learning also gained traction, enabling researchers to utilize pre-trained CNNs such as ResNet, VGG, and Inception for remote sensing tasks. By fine-tuning these models on domain-specific datasets, segmentation accuracy and training efficiency improved significantly.

D. Transformer-Based Models for Advanced Segmentation

In recent years, transformer-based architectures have emerged as a powerful alternative to CNNs, especially for tasks requiring global context understanding. Vision Transformer (ViT), demonstrated that transformers, originally developed for natural language processing, could achieve competitive performance in image classification and segmentation.

SegFormer (Xie et al., 2021), a transformer-based segmentation model, addressed the limitations of CNNs by efficiently capturing long-range dependencies and fine-grained spatial details. These models have shown remarkable performance in high-resolution satellite image segmentation, making them an exciting area for further exploration.

E. Multi-spectral and Temporal Data Utilization

Modern remote sensing applications often leverage multi-spectral and hyperspectral data, which provide rich spectral information for the precise classification of Earth features. Ma et al. (2019) highlighted the benefits of using multi-spectral data to distinguish vegetation, water bodies, and urban areas more accurately. Similarly, temporal analysis using time-series satellite data has gained prominence in monitoring dynamic changes, such as deforestation, crop cycles, and urban sprawl. Singh et al. (2020) demonstrated how temporal datasets from satellites like Sentinel-2 enable robust change detection by analyzing vegetation indices and land cover transitions over time.

Challenges and Future Directions

Despite substantial progress, several challenges remain in satellite imagery segmentation. Class imbalance, particularly in underrepresented land cover categories, continues to affect model performance. Computational scalability, especially for real-time applications, requires innovative solutions such as model pruning, quantization, and edge computing. The future of satellite imagery segmentation lies in developing semi-supervised and unsupervised learning methods to reduce dependency on annotated datasets. Incorporating temporal data for dynamic monitoring and extending models to handle hyperspectral imagery will further enhance their applicability. Additionally, the integration of geospatial data with deep learning models [1] could revolutionize Earth observation systems. This survey highlights the evolution of methodologies in satellite imagery segmentation, from traditional techniques to state-of-the-art deep learning and transformer-based models. The integration of cloud computing and multi-spectral data analysis offers promising directions for future research. By building on these advancements, our research aims to address current challenges and push the boundaries of geospatial analytics.

METHODOLOGY

The approach of this study is developed to attack the difficulties involved in the birth of Earth features from high spatial resolution satellite images. It involves several phases, similar to data accession, preprocessing, model selection and training, evaluation, and deployment. The following is a step-by-step explanation of each phase

A. Data Acquisition

High spatial resolution satellite imagery was sourced from a intimately available dataset from Mohammed Bin Rashid Space Center(MBRSC), 72 high- resolution images of Dubai were annotated into six classes structures, roads, foliage, water, unpaved land, and unlabeled regions, fastening on regions with different Earth features similar as foliage, water bodies, civic areas, and bare soil.

The datasets include multi-spectral bands, offering rich spatial and spectral information critical for effective segmentation.

B. Data Preprocessing

Data Cleaning: Noisy images and irrelevant metadata were filtered out. Missing data in multi-spectral bands were imputed using interpolation techniques.

Image Normalization: Pixel values were regularized to a common scale, enhancing model confluence during training.

Geospatial Alignment: Misaligned images were corrected using georeferencing techniques, ensuring consistency across datasets.

Data Augmentation: Techniques such as rotation, flipping, cropping, and random zooming were applied to increase dataset diversity and reduce overfitting.

C. Feature Segmentation Model Development

a) *Model Selection:* A U-Net architecture was employed as the baseline due to its proven efficacy in image segmentation. Advanced transformer-based architectures, such as Vision Transformer (ViT) or SegFormer, were incorporated to leverage their ability to capture global dependencies in the images.

b) *Model Initialization:* Pre-trained weights from ImageNet or similar datasets were used for initialization, leveraging transfer learning for faster convergence and better performance.

c) *Model Training:* The models were fine-tuned using a supervised learning approach, with annotated segmentation masks serving as ground truth. Hyperparameters such as learning rate, batch size, and dropout rates were optimized through grid search and cross-validation. Loss functions such as categorical cross-entropy and Dice loss were employed to handle class imbalances effectively.

D. Evaluation Metrics

Performance was evaluated using standard metrics, including:

Intersection over Union (IoU): To measure the imbrication between prognosticated and base verity segmentation.

Precision, Recall, and F1 Score: To assess class-wise segmentation accuracy.

Mean Average Precision (mAP): To estimate the model's capability across all classes.

A comparative analysis was conducted against conventional methods, such as pixel-based classification and traditional machine learning models, to validate the efficacy of the proposed approach.

E. Scalability and Computational Optimization

Cloud Deployment: The segmentation pipeline was deployed on cloud platforms (e.g., AWS, Google Cloud) to handle large-scale data processing efficiently.

F. Visualization and Interpretation

Segmentation results were visualized using GIS tools, allowing for qualitative assessments of model predictions.

Confusion matrices and heatmaps were generated to highlight misclassifications and refine model performance.

G. Future Extensions

The framework will be extended to incorporate temporal satellite data for dynamic monitoring of Earth features.

Multi-spectral and hyperspectral data analysis will be explored to further enhance feature extraction accuracy.

This methodology ensures a comprehensive and reproducible approach to satellite imagery segmentation, leveraging cutting-edge machine-learning techniques to achieve state-of-the-art performance.

CONCLUSION

This study demonstrates the effectiveness of the U-Net architecture in extracting building footprints from high-resolution satellite imagery. By leveraging an optimized training strategy and robust evaluation metrics, the model achieved high accuracy and generalizability. This research contributes to automated urban mapping methodologies and sets a foundation for scalable applications in urban planning and disaster response.

REFERENCES

[1] Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS J. Photogram. Remote Sens. 2020, 159, 296–307. [Google Scholar] [CrossRef]

[2] Su, H.; Wei, S.; Liu, S.; Liang, J.; Wang, C.; Shi, J.; Zhang, X. HQ-ISNet: High-Quality Instance Segmentation for Remote Sensing Imagery. Remote Sens. 2020, 12, 989. [Google Scholar] [CrossRef] [Green Version]

[3] Zhang, L.; Wu, J.; Fan, Y.; Gao, H.; Shao, Y. An Efficient Building Extraction Method from High Spatial Resolution Remote Sensing Images Based on Improved Mask R-CNN. Sensors 2020, 20, 1465. [Google Scholar] [CrossRef] [PubMed] [Green Version]

[4] Braga, J.R.; Peripato, V.; Dalagnol, R.; Ferreira, M.P.; Tarabalka, Y.; Aragão, L.E.; de Campos Velho, H.F.; Shiguemori, E.H.; Wagner, F.H. Tree Crown Delineation Algorithm Based on a Convolutional Neural Network. Remote Sens. 2020, 12, 1288. [Google Scholar] [CrossRef] [Green Version]

[5] Brodrick, P.G.; Davies, A.B.; Asner, G.P. Uncovering Ecological Patterns with Convolutional Neural Networks. Trends Ecol. Evol. 2019, 34, 734–745. [Google Scholar] [CrossRef] [PubMed]

[6] Vuola, A.O.; Akram, S.U.; Kannala, J. ask-RCNN and U-net ensembled for nuclei segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 208–212. [Google Scholar]

[7] Huang, B.; Lu, K.; Audebert, N.; Khalel, A.; Tarabalka, Y.; Malof, J.; Boulch, A.; Le Saux, B.; Collins, L.; Bradbury, K.; et al. Large-scale semantic classification: Outcome of the first year of Inria aerial image labeling benchmark. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium—IGARSS 2018, Valencia, Spain, 22–27 July 2018. [Google Scholar]

[8] Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. High-resolution aerial image labeling with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 2017, 55, 7092–7103. [Google Scholar] [CrossRef] [Green Version]

[9] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969. [Google Scholar]

[10] Bai, M.; Urtasun, R. Deep watershed transform for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5221–5229. [Google Scholar]

[11] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* 1998, *86*, 2278–2324. [**Google Scholar**] [**CrossRef**] [**Green Version**]

[12] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv 2015, arXiv:1505.04597. [Google Scholar]

[13] Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [Google Scholar]

[14] Fabien H. Wagner, Ricardo Dalagnol, Yuliya Tarabalka, Tassiana Y. F. Segantine, Rogério Thomé and Mayumi C. M. Hirye, "U-Net-Id, an Instance Segmentation Model for Building Extraction from Satellite Images—Case Study in the Joanópolis City, Brazil" Remote Sensing, 2020