

DeepTriNet: A Tri-Level Attention Based DeepLabv3+ Architecture for Semantic Segmentation of Satellite Images

Tareque Bashar Ovi¹[0000-0001-6961-8894], Shakil Mosharrof²[0000-0002-7228-2823], Nomaiya Bashree³[0009-0008-6151-2356], Md Shofiqul Islam⁴[0000-0002-2597-4050], and Muhammad Nazrul Islam⁵[0000-0002-7189-4879]

¹⁻⁵Military Institute of Science and Technology (MIST)
Mirpur Cantonment, Dhaka-1216, Bangladesh

¹ovitareque@gmail.com ²shakilmrf8@gmail.com
³nomaiyabashree2002@gmail.com ⁴shafiqcseiu07@gmail.com
⁵nazrul@cse.mist.ac.bd

Abstract. The segmentation of satellite images is crucial in remote sensing applications. Existing methods face challenges in recognizing small-scale objects in satellite images for semantic segmentation primarily due to ignoring the low-level characteristics of the underlying network and due to containing distinct amounts of information by different feature maps. Thus, in this research, a tri-level attention-based DeepLabv3+ architecture (DeepTriNet) is proposed for the semantic segmentation of satellite images. The proposed hybrid method combines squeeze-and-excitation networks (SENNets) and tri-level attention units (TAUs) with the vanilla DeepLabv3+ architecture, where the TAUs are used to bridge the semantic feature gap among encoders output and the SENNs used to put more weight on relevant features. The proposed DeepTriNet finds which features are the more relevant and more generalized way by its self-supervision rather we annotate them. The study showed that the proposed DeepTriNet performs better than many conventional techniques with an accuracy of 98% and 77%, IoU 80% and 58%, precision 88% and 68%, and recall of 79% and 55% on the 4-class Land-Cover.ai dataset and the 15-class GID-2 dataset respectively. The proposed method will greatly contribute to natural resource management and change detection in rural and urban regions through efficient and semantic satellite image segmentation

Keywords: Attention mechanism· Deep learning· Satellite image· DeepLabv3+ · Segmentation

1 Introduction

Satellite image segmentation is required to monitor and evaluate land cover and land use. In other words, it is required for the management of natural resources.

Different methods are applied to analyse images and data acquired from remote sensing to offer land description and change detection in rural and urban areas. In many areas, including urban planning, a fundamental task for ecological environment conservation, vegetation monitoring, and even military reconnaissance, detailed information about land use or land cover is an invaluable resource [1], making precise satellite image segmentation a research aspect.

Traditional manual interpretation and classification of satellite images is the process of visually analyzing and categorizing satellite images based on human expertise and visual inspection, which can be time-consuming and error-prone. Again, semantic segmentation involves the implementation of pixel-level labels to images, which is a crucial task [2] in study of satellite images for land cover classification and urban planning.

The use of images in remote sensing (RS) applications has drawn more attention in recent years due to the recent expansion in satellite and aerial imagery availability. Achievements in computer vision, natural language processing, and audio processing have all been made possible by advances in machine learning (ML), larger benchmark datasets, and improved CPU capacity. Deep learning techniques, which have become the standard methods for remote sensing image landcover classification research, automatically extract low-level image features of objects from images by creating deep networks, then combine them into high-level abstract features to achieve higher classification accuracy [3]. Despite the availability of petabytes of freely available satellite imagery and a wide range of benchmark datasets for various RS tasks, the widespread success and popularity of machine learning—particularly of deep learning methods — has not yet been fully adopted to the RS domain. This is not to argue that there aren't any successful ML applications in RS, but rather that the promise of the nexus between both domains hasn't been entirely fulfilled.

Therefore, the objective of this research is to propose an effective deep-learning network using several attention techniques to reduce the semantic information gap between the encoder and decoder, and for allowing the accurate segmentation of satellite pictures. As outcomes, this research proposed a tri-level attention-based DeepLabv3+ architecture (DeepTriNet) for the semantic segmentation of satellite images. As such, the contribution could be summarized as: (i) integrating a self-supervised attention mechanism with vanilla DeepLabv3+ that combines channel-level, spatial-level, and pixel-level abstractions for better generalization of the pertinent information.(ii)Introducing Squeeze-and-Excitation(SE) in the decoder portion of the proposed architecture that explicitly models channel inter-dependencies and adaptively re-calibrates channel-wise feature outputs.(iii) Evaluating the robustness of the model by performing experiments on two datasets, one of which has 15 classes but is comparatively small in volume and the other one with only 4 classes but large in volume.

2 Related Works

Convolutional Neural Network (CNN) usage in the past years has shown promising results in automating satellite image segmentation tasks. Using FastFCN, Onim *et al.* [4] reported an accuracy of 93%, precision of 99%, recall of 0.98, and mIoU of 97% using the GID-2 dataset. Only 6 classes were used for their study.

Tong *et al.* [5] proposed CNN based land cover classification algorithm that uses deep learning, hierarchical segmentation, and multi-scale information fusion to classify satellite images using the GID-2 dataset. ResNet-50 was used for initial training purposes. The proposed method outperforms traditional methods such as colour histogram, grey-level co-occurrence matrix, and local binary patterns in land cover classification accuracy.

Yao *et al.* [6] used high-resolution remote sensing photos and two image datasets containing six common land cover classes to classify land use using a deep convolutional neural network method with an attention mechanism. With a Kappa coefficient of 0.91, the suggested model surpassed other models in terms of accuracy and metrics value, achieving an overall accuracy of 93.5%.

For LULC use using remote sensing data, a unique combined deep learning framework incorporating MLP and CNN models were implemented by Zhang *et al.* [7]. The framework's accuracy ranged from 85.5% to 92.3%, which was higher overall than previous techniques and less susceptible to smaller sample sizes. In order to separate land use and land cover classes from RGB satellite pictures, Nayem *et al.* [8] developed a semantic segmentation framework utilizing the FCN-8 algorithm. Their proposed architecture had an average intersection over union (IoU) of 84% and an average accuracy of 91.0% on the GID-2 dataset. Kang *et al.* [9] proposed a Multi-scale context extractor network for water-body extraction from high-resolution satellite images using the LandCover.ai dataset. Though 93.23% of mean IOU was reported in their study, only 1 class(water) was considered for their experiment. Lee *et al.* [10] performed a comparative analysis between various deep learning architectures using the LandCover.ai dataset and reported 88.4%, 91.4%, and 85.8% for SegNet, U-Net, and DeepLabv3+ respectively.

Recent literature shows that there were many attempts to semantically segment satellite images with the GID-2 dataset [4, 6, 8], using traditional end-to-end deep learning models. On the other hand, few networks were used to segment maps with other LandCover.ai datasets. But self-supervised attention mechanism-based satellite image segmentation on the GID-2 with 15 classes and LandCover.ai dataset with is yet to be tested.

3 Proposed Architecture

This section broadly introduces the theoretical concepts that includes

DeepLabv3+ The purpose of DeepLabv3+ is to assign semantic labels (such as a person, a dog, or a cat) to every pixel in the input image proposed by

Chen *et al.* [11]. The DeepLabv3+ model contains two phases: encoding and decoding. The encoding phase extracts the essential information from the image using a convolutional neural network (CNN) whereas the decoding phase reconstructs the output of appropriate dimensions based on the information obtained from the encoder phase. The decoder module was included to improve object boundary segmentation results. MobileNetv2, Xception, ResNet, PNASNet, and Auto-DeepLab are among the network backbones that Deeplab supports. It is used as a benchmark for semantic segmentation today.

Tri-Level Attention Unit (TAU) Tri-Level Attention Unit (TAU) proposed by Tanyir Mahmud *et al.* [12] combines three levels of abstraction—channel, spatial, and pixel—for greater generalization of the pertinent contextual information. It is a novel self-supervised attention technique. The channel attention (CA) mechanism functions from a wider perspective to highlight the corresponding channels that contain more information, the spatial attention (SA) mechanism focuses more on the local spatial regions that contain regions of interest, and finally, the pixel attention (PA) mechanism functions from the lowest level to assess the feature relevance of each pixel. Thus TAU unit module is frequently utilized throughout the DeepTriNet decoder network to recalibrate features and increase feature relevance. TAU architecture and mechanism is given in Fig.1 and Fig.2(a).

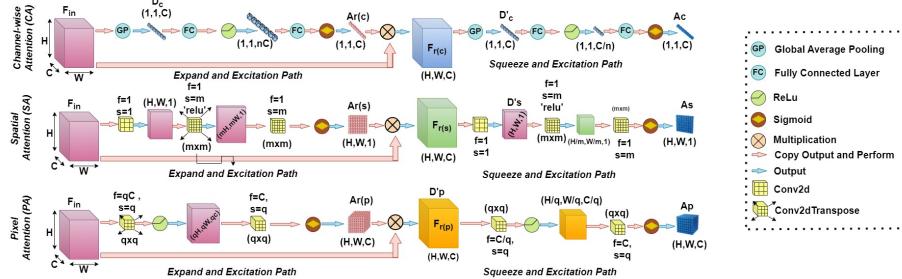


Fig. 1: Tri-Level Attention Unit Architecture

Squeeze-and-Excitation Networks(SENets) Squeeze-and-Excitation Networks (SENNets) provide a CNN building block that enhances channel interdependencies at essentially no computational cost. In CNN Before combining the data over all available output channels, the various filters will first locate spatial features in each input channel. All there is to know for now is that the network equally weights each channel when producing the output feature maps. By including a content-aware system to adaptively weight each channel, SENets aim to change this. In its simplest form, this may entail providing each channel with

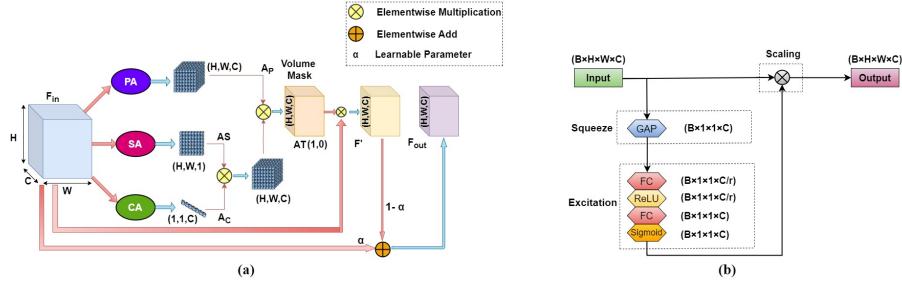


Fig. 2: (a) TAU Mechanism and (b) SENets Architecture

a single parameter and a linear scalar representing how relevant each one is. By condensing the feature maps to a single numerical value, they first gain a broad comprehension of each channel. As a result, a vector of size n is produced, where n is the number of convolutional channels. It is then input into a two-layer neural network, which produces an output vector of the same size. These n values can now be applied to the original feature maps as weights, sizing each channel according to its significance. This whole mechanism is visualised in Fig.2(b).

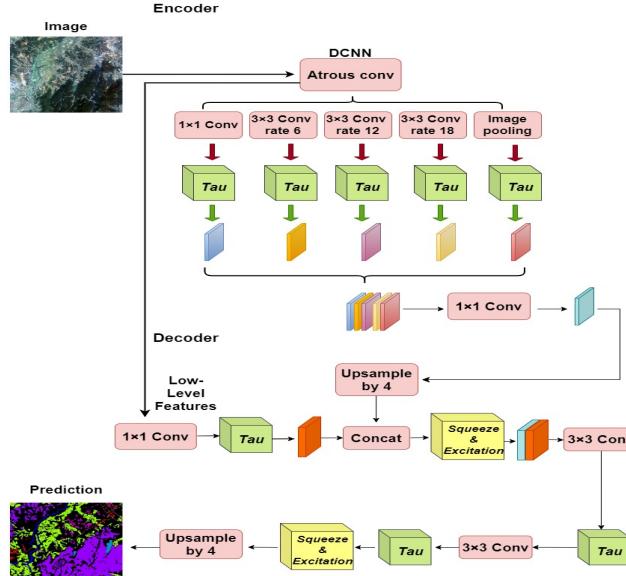


Fig. 3: Proposed Architecture

Proposed Architecture In this research both the Tri-Level Attention Unit (TAU) and the Squeeze-and-Excitation Networks (SENNets) have been integrated extensively with the vanilla DeepLabv3+ architecture. More specifically, the

TAU has been integrated with the Atrous Spatial Pyramid Pooling(ASPP) portion of the architecture, where each convolution has been passed through the TAU to extract more refined features. Squeeze-and-Excitation Networks (SENets) has been introduced at the decoder portion of this network. Detailed architecture of our proposed network is given in the Fig.3

4 Dataset

4.1 Dataset Description

Two open-source dataset LandCover.ai [13] and Gaofen-2 Image Dataset (GID-2) [14] are used in this study. LandCover.ai dataset has four classes: building (1), woods (2), water (3), and road (4). Again Gaofen-2 Image Dataset (GID-2) has mainly two portions: a large-scale classification set with six classes and a fine land-cover classification set with 15 classes. This work has used the latter part to verify the proposed architecture with more classes. An example of these two datasets is given in fig.4 and 5.

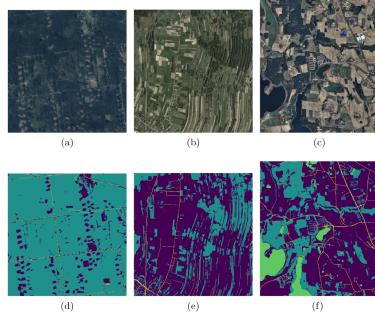


Fig. 4: Example of images from the LandCover.ai [13] dataset: (a–c) Input image and (d–f) Target image

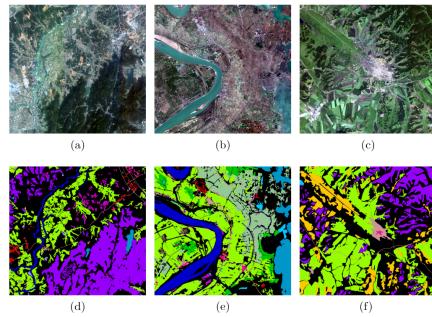


Fig. 5: Example of images from the Gaofen-2 Image Dataset (GID-2) [14]: (a–c) Input image and (d–f) Target image

4.2 Dataset Preprocessing

Tri-Level attention-based DeepLabv3+ requires the input image to be downsampled to $256 \times 256 \times 3$. This results in a significant loss of pixel information. To address and solve this issue, we followed the grid and patch method introduced in the work of Onim *et al.* [4]. Each of the original and annotated images was gridded to 728 sub-images of shape $256 \times 256 \times 3$. The sub-images were bounded to the grid and patch block to preserve their spatial position. Detailed patch-wise prediction of DeepTriNet on GID-2 and LandCover.ai dataset is shown in Fig.6 and 7 respectively.

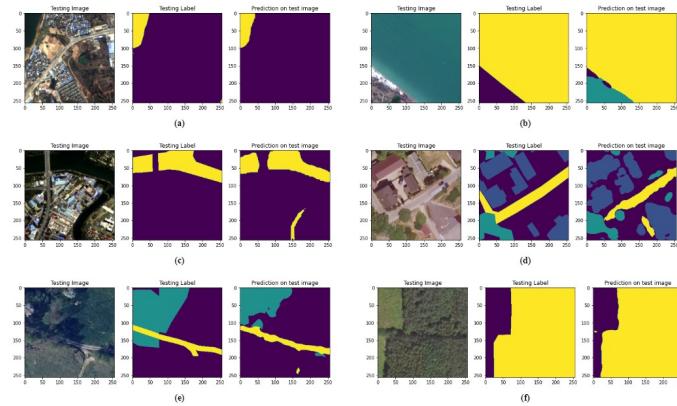


Fig. 6: Patch wise Prediction of DeepTriNet on GID-2 Dataset

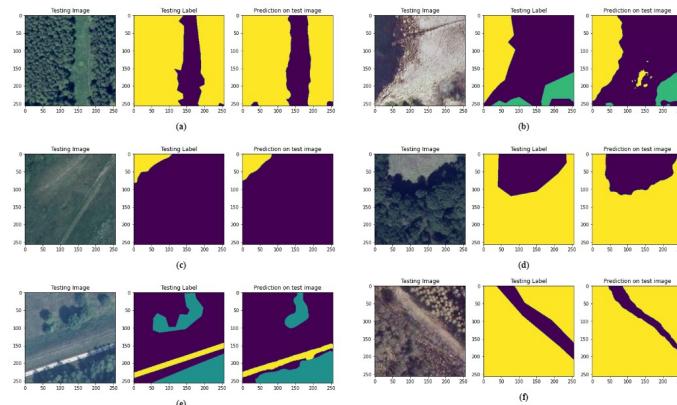


Fig. 7: Patch-wise Prediction of DeepTriNet on LandCover.ai Dataset

The final preprocessing step normalized the pixel values from -1 to 1 as the Equation (1). Here i, j represents image height and width respectively, $I_{i,j}$ is the original image and $I_{i,j}^n$ is the normalized image.

$$I_{i,j}^n = \frac{I_{i,j}}{127.5} - 1 \quad (1)$$

5 Methodology

5.1 Proposed Methods

Our research and evaluation workflow can be explained as follows.

- At first, both the training images and ground truths are subdivided into 784 images each by gridding the high-resolution satellite images.
- After that, these patches were normalized and then used for the training of the DeepTriNet.
- After the training, the prediction is done for each patch by the best saved weight, and after that smooth tiled prediction were done to reconstruct a high-resolution image. The generation of the evaluation metrics for each images follows through pixel-by-pixel calculation. The whole process is visualized in Fig.8

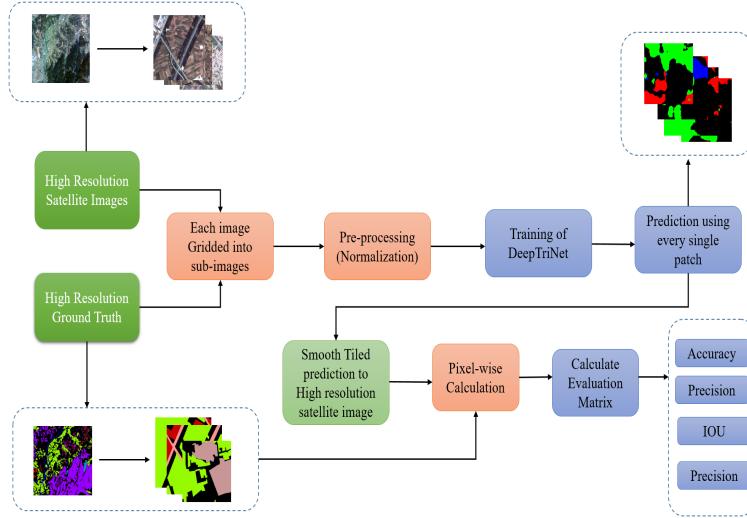


Fig. 8: Research Workflow

6 Result Analysis

Our proposed architecture has been evaluated based on the basis of accuracy, precision, recall and IoU.

Table 1: Performance Comparison with Existing Work

Year	Reference	Dataset	No of Classes	Pre processing	Network(Type)	Matrices
2022	Lee <i>et al.</i> [10]	LandCover.ai	4	grid and patch	DeepLabv3+(CNN)	Accuracy: 0.85 Precision: NA Recall: NA IoU: NA
2022	Lee <i>et al.</i> [10]	LandCover.ai	4	grid and patch	SegNet(CNN)	Accuracy: 0.88 Precision: NA Recall: NA IoU: NA
2019	Yao <i>et al.</i> [6]	ISPRS	6	single image	DeepLabv3+(CNN)	Accuracy: 0.91 Precision: 0.80 Recall: NA IoU: NA
2020	Nayem <i>et al.</i> [8]	Gaofen-2	6	grid and patch	FCN-8(CNN)	Accuracy: 0.91 Precision: NA Recall: NA IoU: 0.840
2022	Lee <i>et al.</i> [10]	LandCover.ai	4	grid and patch	U-net(CNN)	Accuracy: 0.91 Precision: NA Recall: NA IoU: NA
2020	Onim <i>et al.</i> [4]	Gaofen-2	6	grid and patch	FastFCN(CNN)	Accuracy: 0.93 Precision: 0.99 Recall: 0.98 IoU: 0.97
2021	Kang <i>et al.</i> [9]	LandCover.ai	1	grid and patch	Multi-scale context extractor with CNN	Accuracy: 0.98 Precision: NA Recall: NA IoU: 0.938
2023	Ours	Gaofen-2	15	grid and patch	DeepTriNet	Accuracy: 0.77 Precision: 0.68 Recall: 0.55 IoU: 0.58
2023	Ours	LandCover.ai	4	grid and patch	DeepTriNet	Accuracy: 0.98 Precision: 0.80 Recall: 0.79 IoU: 0.80

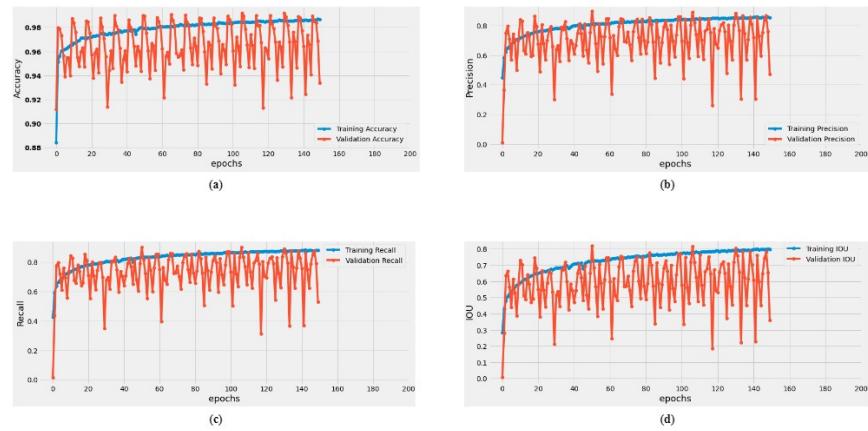


Fig. 9: Performance History of DeepTriNet on LandCover.ai Dataset

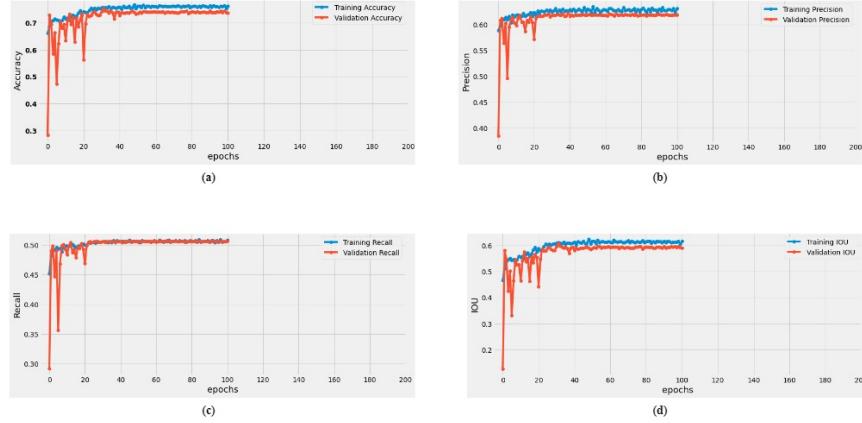


Fig. 10: Performance History of DeepTriNet on GID-2 Dataset

Fig.9 and Fig.10 show the training history of the proposed DeepTriNet model in terms of accuracy, precision, recall and IoU. It can be easily observed from both Fig.9 and Fig.10 that there were no overfitting while training the DeepTriNet, rather a perfect fitting curve for all the evaluation parameter has been shown. In Table.1 a comparative analysis between the existing literature and our proposed DeepTriNet architecture has been summarized.

7 Conclusion

In conclusion, DeepTriNet introduces a self-supervised DeepLabv3+ architecture based on tri-level attention for semantic segmentation of satellite images. The proposed model combines the vanilla DeepLabv3+ architecture with Tri-Level Attention Units (TAUs) and Squeeze-and-Excitation Networks (SENets) to increase segmentation efficiency and precision.

Experimental findings on the 4-class Land-Cover.ai dataset demonstrate that DeepTriNet outperforms conventional approaches, except for a method that only considers one class. Also presented are other measures, including accuracy, recall, and the F1 score. The performance of DeepTriNet for 15-class GID-2 dataset is noteworthy also comparing the data volume and diversity.

Better natural resource management and change detection in rural and urban regions are two uses of this technology. However, this approach has drawbacks, including the demand for a sizable amount of training data and computer power.

The potential of DeepTriNet in other domains, such as autonomous driving or medical imaging, may be investigated in subsequent research. Overall, the contributions of this model provide a viable method for precise and effective satellite picture segmentation, with potential applications in many different sectors.

References

1. N. Longbotham, C. Chaapel, L. Bleiler, C. Padwick, W. J. Emery, and F. Pacifici, “Very high resolution multiangle urban classification analysis,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 4, pp. 1155–1170, 2012.
2. “Key issues in image understanding in remote sensing [and discussion],” vol. 324, no. 1579, pp. 381–395, 1988. [Online]. Available: <http://www.jstor.org/stable/37940>
3. B. Neupane, T. Horanont, and J. Aryal, “Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis,” *Remote Sensing*, vol. 13, p. 808, 02 2021.
4. M. S. H. Onim, A. R. B. Ehtesham, A. Anbar, A. K. M. Nazrul Islam, and A. K. M. Mahbubur Rahman, “Lulc classification by semantic segmentation of satellite images using fastfcn,” in *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*, 2020, pp. 471–475.
5. C. Zhang, X. Pan, H. Li, S. Zhang, and P. Atkinson, “Joint deep learning: A novel framework for urban land cover and land use classification,” *GISRUK 2020*, pp. 1–5, 2020.
6. X. Yao, H. Yang, Y. Wu, P. Wu, B. Wang, X. Zhou, and S. Wang, “Land use classification of the deep convolutional neural network method reducing the loss of spatial features,” *Sensors*, vol. 19, no. 12, 2019. [Online]. Available: <https://www.mdpi.com/1424-8220/19/12/2792>
7. R. Zhang, X. Tang, S. You, K. Duan, H. Xiang, and H. Luo, “A novel feature-level fusion framework using optical and sar remote sensing images for land use/land cover (lulc) classification in cloudy mountainous area,” *Applied Sciences*, vol. 10, no. 8, 2020. [Online]. Available: <https://www.mdpi.com/2076-3417/10/8/2928>
8. A. B. S. Nayem, A. Sarker, O. Paul, A. A. Ali, M. A. Amin, and A. M. Rahman, “Lulc segmentation of rgb satellite image using fcn-8,” *ArXiv*, vol. abs/2008.10736, 2020.
9. J. Kang, H. Guan, D. Peng, and Z. Chen, “Multi-scale context extractor network for water-body extraction from high-resolution optical remotely sensed images,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102499, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0303243421002063>
10. S.-H. Lee and M.-J. Lee, “Comparisons of multi resolution based ai training data and algorithms using remote sensing focus on landcover,” *Frontiers in Remote Sensing*, vol. 3, p. 39, 2022.
11. L. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: <http://arxiv.org/abs/1706.05587>
12. Vol. 17, no. 9, pp. 6489–6498, sep 2021. [Online]. Available: <https://doi.org/10.1109%2Ftii.2020.3048391>
13. A. Boguszewski, D. Batorski, N. Ziembka-Jankowska, T. Dziedzic, and A. Zambrzycka, “Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery,” 2020. [Online]. Available: <https://arxiv.org/abs/2005.02264>
14. X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, “Land-cover classification with high-resolution remote sensing images using transferable deep models,” *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.