Urban Land-use Features Mapping from LiDAR and Remote Sensing Images using Visual Transformer Network Model

Qinglie Yuan[a*]

[a.]School of Civil and Architecture Engineering, Panzhihua University, Panzhihua 617000, China-[*]yuanqinglie@pzhu.edu.cn

Keywords: LiDAR, remote sensing, Transformer, CNN, land-use classification.

Abstract

With the rapid development of science and technology in the acceleration of urbanization, it is important to achieve efficient and accurate monitoring and mapping of urban features. Traditional urban feature mapping methods often rely on a single data source, such as optical remote sensing images or LiDAR, which often encounter many challenges in complex urban environments, such as shading, occlusion, and land cover changes. LiDAR has relatively accurate three-dimensional spatial information, while remote sensing image has rich spectral information. Thus, the fusion of spatial-spectral features can improve the accuracy and robustness of automatic classification efficiency for urban feature mapping. Recently deep learning technology has achieved a profound impact on remote sensing data processing. However, some existing deep models have not effectively fused spatial-spectral information In addition, the lack of semantic information optimization could confuse classification, especially for some high spectral heterogeneity areas. Hence, this study proposed a visual Transformer model to achieve automatic mapping combined with LiDAR and remote sensing images. In addition, this study improved the global attention mechanism for adaptive enhancing spectral-spatial fusion. Finally, it is found that the proposed algorithm is generally better than other representative methods, and the classification accuracy using remote sensing data and LiDAR is improved. The proposed modules can improve the Kappa coefficient by 5%.

1. Introduction

Urban land mapping is crucial for urban planning, environmental management, and disaster response (Gómez-Chova et al. 2015). It involves the identification and classification of various land features within a city, as shown in Figure 1. Traditional methods of urban land mapping rely heavily on manual interpretation of aerial photographs and satellite images, which can be time-consuming and prone to human error. The advent of Light Detection and Ranging (LiDAR) technology and remote sensing images have revolutionized the field by providing high-resolution, three-dimensional data that can be automatically processed (Tuia et al. 2015).

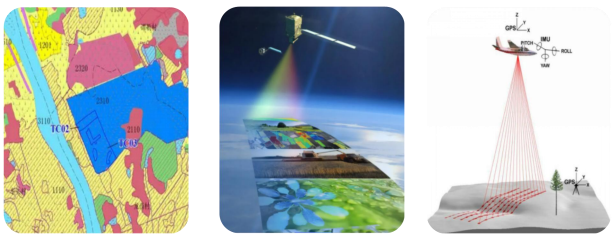


Figure 1. Urban land mapping and remote sensing platform.

With the rapid advancement of urbanization and the rapid development of remote sensing technology, the application of aerial remote sensing images and radar data in thematic mapping of urban land features has gradually received widespread attention (Yan et al. 2015). These two data sources have different characteristics and advantages. Aerial remote sensing images can provide rich surface information, while radar data has stronger penetration into surface structures. Integrating the two can obtain more accurate and comprehensive information on urban land features, providing important support for urban planning, urban management, disaster monitoring, and other fields.

The popularization and expansion of remote sensing technology, as well as the fusion of aerial remote sensing images and radar technology, have made significant progress in the research of urban land feature thematic mapping. Many domestic scholars and research institutions have conducted in-depth research on fusion algorithms, data processing, feature extraction, and other aspects, and have achieved a series of innovative results. With the launch of multiple high-resolution remote sensing satellites, image fusion technology based on deep learning has rapidly become a research hotspot. At this stage, various images such as remote sensing images were fused using image fusion methods, greatly improving the quality and application value of the images. For example, YUAN et al. (2021, 2024) applied Convolutional Neural Networks (CNN) to building extraction, resulting in improved accuracy.

The main advantages of a Visual Transformer (ViT) over CNN are to capture global context, higher parallelization processing efficiency, and wider adaptability (Yuan et al. 2021). The ViT can capture the dependency relationships between different positions in the input sequence through its self-attention mechanism. Regardless of the distance between two elements in the sequence, the Transformer can directly calculate their relationships, thus better understanding the global context. Wang et al. (2022) used the U-Net model combined with a visual Transformer, achieved good land use cover classification results.

In contrast, CNN uses convolutional kernels for local perception. Although it can expand the receptive field by increasing the number of layers and using larger convolutional kernels, it still tends to extract local features, making it difficult to directly capture global contextual information from long distances (Yao

et al. 2021). The self-attention mechanism of the Transformer can process all elements in the input sequence in parallel, which enables more efficient utilization of hardware resources such as GPU and TPU during training and inference, especially when dealing with long sequences. In contrast, although CNN's convolution operation can be parallelized, it requires layer-by-layer computation when processing sequential data, which may not be as efficient as the Transformer's parallel operation in some cases.

Transformer relies on self-attention mechanisms rather than convolution operations, making fewer assumptions about the structure of input data. This flexibility allows Transformer to adapt more widely to various types of data, including text, images, and time series. In contrast, CNN utilizes the local perception and weight-sharing properties of convolutional kernels to perform well in tasks with spatial local correlations, but may not be as flexible as Transformers when dealing with other types of data (Yuan 2024).

Aerial remote sensing images can provide rich texture and color information, which helps identify the appearance features and details of ground objects. However, in the presence of complex terrain and obstructions, it may be affected, resulting in incomplete or distorted information acquisition. As an active remote sensing system, LiDAR can penetrate cloud layers, provide surface physical and geometric feature information, and accurately measure the vertical spatial height information and horizontal spatial geometric structure information of ground objects. Although radar data performs well in some aspects, there may be data sparsity issues on targets with fewer textures.

Therefore, integrating aerial remote sensing images with LiDAR data can fully utilize their complementarity and improve the accuracy of urban land classification. The fused data can not only provide rich texture and color information but also obtain accurate three-dimensional spatial structure information, thus more accurately identifying and classifying urban land features. Specialized mapping can also provide more accurate and comprehensive data support for urban planning and management.

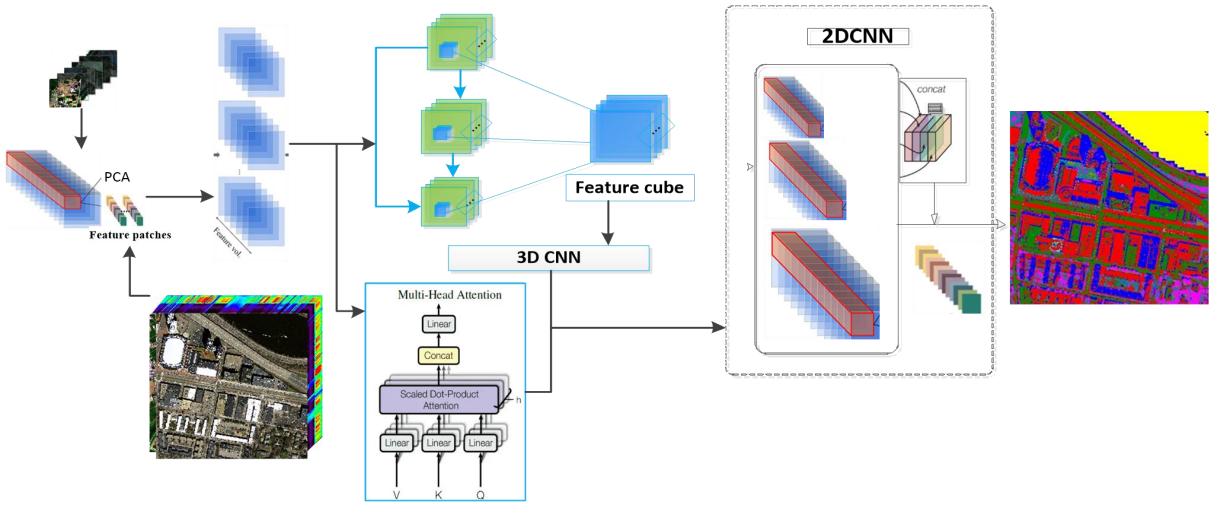


Figure 2. Network architecture.

2. Method

Remote sensing image segmentation methods based on CNN have been widely studied, most of which are based on u-net or its variants, and have achieved remarkable results in various tasks Although the CNN method performs well in terms of representation ability, due to the local nature of convolution operation, it has some limitations in capturing global and long-distance semantic information interaction, while ViT has made significant improvement in this regard.

In recent years, ViT has introduced a self-attention-based architecture to handle visual recognition tasks to better model long dependencies. Subsequently, many transformer variants have achieved great success in natural image recognition tasks, such as swing transformer, diet, PVT, tit, etc With the help of ViT 's excellent presentation ability, some researchers try to combine it with CNN or directly replace CNN to obtain better medical image segmentation results, such as TransUnet, Swin UNET, COTR. All these studies show that ViT can further improve its performance compared with CNN, and also point out that more attention should be paid to the development of ViT in the future However, although ViT has excellent presentation ability, it still needs a lot of data for identification tasks, and may even need more data than CNN Therefore, in the task of remote sensing image analysis with limited data, the effective fusion of CNN-ViT is important for the optimization of model structure.

Figure 2 shows the process of the proposed method. Firstly, the multispectral images are stacked with LiDAR data and input into a dual branch Transformer module for feature extraction. Then, a spectral feature encoder and a structural feature encoder composed of lightweight convolutional neural networks are used to extract spectral and structural features of multispectral and LiDAR, respectively. Among them, the spectral feature encoder adopts 3D convolution, and the structural feature encoder adopts 2D convolution. Finally, the features extracted by the Transformer module, spectral features extracted by the spectral feature encoder, and texture features extracted by the structural feature encoder are directly stacked and input into the classifier for classification to obtain the results.

2.1 Visual Transformer Network Model Architecture

Due to the powerful global information exchange capability of the Transformer, a dual branch Transformer module was developed for the interaction of spatial and channel ranges in

multi-source remote sensing images. The internal structure of the visual Transformer (ViT) module in this paper is shown in

Figure 3. The input of the dual branch Transformer module is a stack of multispectral and LiDAR. After linear projection, the encoded features can be used as features Q, K, and V for the next step of self-attention computation. After passing through self-attention layers and gated feedforward neural networks, it is used for multi-source remote sensing images.

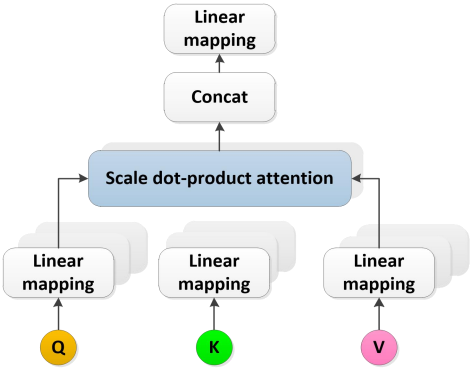


Figure 3. The structure of the ViT module.

2.2 CNN Local Feature Enhancement

Transformers tend to focus more on global information (i.e. semantic information of the entire image) while ignoring spectral information of multispectral images and structural information of LiDAR data. However, multi-source remote sensing image classification tasks often require the use of rich spectral information and fine ground structure information to achieve accurate classification. To address this issue, this paper proposes a lightweight convolutional neural network that extracts spectral information and geometric structure information separately, namely spectral feature extractor and structural feature extractor, which are used to extract spectral features from hyperspectral images and geometric structure features from LiDAR data, respectively. The former consists of three layers of Conv3D ReLU BN, used to extract spectral and texture features from texture-rich hyperspectral data, while the latter consists of three layers of Conv2D ReLU BN, used to extract height information and object geometry features from LiDAR.

3. Experiment And Result

3.1 Data Description

The dataset is derived from USGS. The dataset is an aerial multispectral remote sensing image data collected by imaging sensors in urban areas. That includes 4 bands with a spectral range of 0.363 micrometers to 1.018 micrometers and a spatial resolution of 0.5 meters. The entire image has undergone geometric and radiometric calibration. As shown in Figure 4. Through investigation and visual inspection using high-resolution color images, ground truth data of 8 categories were collected.

The sampling technique adopts a hierarchical system sampling. Based on hierarchical system sampling, a total of 34820 points were created from the entire sample, including buildings, forests, roads, cars, grasslands, sports fields, and bare land.

Figure 4. Study area overview.

3.2 Model training

In the experiment, the sample set is randomly divided into training sets and verification sets with a ratio of 7:3. During network training when the validation data set reaches 90%, the ViT layer is frozen, and fine-tuning training is conducted at the CNN layer. In the Keras framework, model compilation mainly completes the configuration of the loss function and optimizer. In the model compilation, the Adam algorithm is selected as the optimization function, and the multi-class cross entropy as the objective function is calculated for the prediction and the ground-truth.

On the dataset, the learning rate for 2000 generations is 0.001. Use backpropagation to minimize classification cross-entropy loss. Use batch normalization (BN) and 50% dropout to address overfitting issues. Accuracy indicators are used to evaluate experimental results, including overall accuracy (OA), Kappa coefficient, and F1 score.

3.3 Preprocessing

LiDAR data was obtained from the National Oceanic and Atmospheric Administration (NOAA) in the United States using LiDAR point cloud data (. las format), with acquisition accuracy including estimated point spacing of 0.35 meters, vertical accuracy of 0.5 meters, and horizontal accuracy of 0.36 meters, as shown in Figure 5. The LiDAR data and orthorectified images were converted to the same coordinate reference as the UTM area 19N, 1983 North American reference, and NAVD88 vertical reference. Vector labels refer to the Open Street Map and are interpreted through images and 3D point clouds.

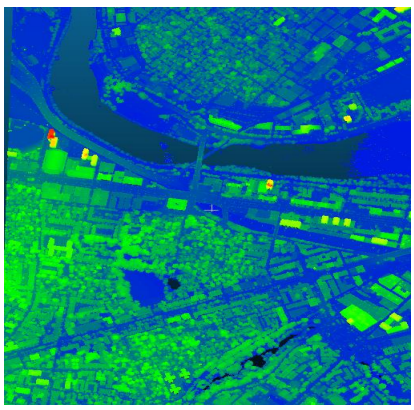
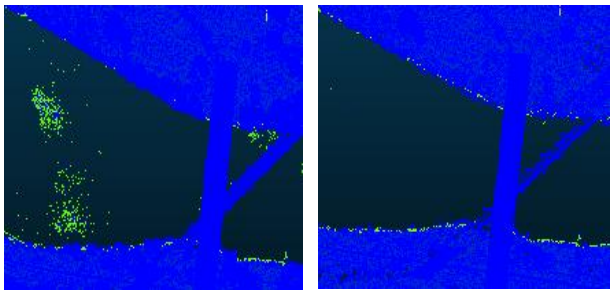


Figure 5. LiDAR point cloud data.

This article mainly uses the low-pass filtering algorithm in CloudCompare software to denoise point cloud data, as

illustrated in Figure 6. The experimental area consists of various types of features and complex scenes, such as roads, exposed soil, vegetation, lakes, rivers, and hills, as shown in Figure 4. Buildings present complex architectural structures and various scenes, such as large industrial areas, rural areas, densely populated residential areas, and suburbs of different heights. Some ground objects and buildings have similar textures and colors, such as floors, courtyards, sports fields, roads, and cars. The above factors pose challenges to automatic mapping.



(a) (b)

Figure 6. Denoising point cloud data. (a) Pre-denosie; (b) Post-denoise.

Digital Elevation Model (DEM) is a digital model used to represent the elevation information of the earth's surface. It divides the surface into regular grid cells, each cell stores an elevation value, to form a continuous elevation surface. DEM is a virtual representation of landforms, and it can derive contour lines, slope maps, and other related information. It can also be superimposed with Digital Orthophoto Map (DOM) or other thematic data to participate in the analysis and application of terrain, and DEM is also the basic data required for DOM production. In engineering construction, DEM can be used for earthwork calculation, intervisibility analysis, etc; In terms of flood control and disaster reduction, DEM is also an important tool for hydrological analysis. Digital Surface Model (DSM) is a digital model used to represent the three-dimensional shape of the earth's surface. It contains not only the elevation information of terrain but also the information of buildings, vegetation, and other surface objects. Therefore, DSM can be used in urban planning, landscape design, environmental simulation, and other fields.

As shown in Figure 7, the ENVI LiDAR software was used to generate DSM and DEM images from denoised point cloud data. Then, the band operation function in ENVI was used to generate nDSM images from DSM-DEM.

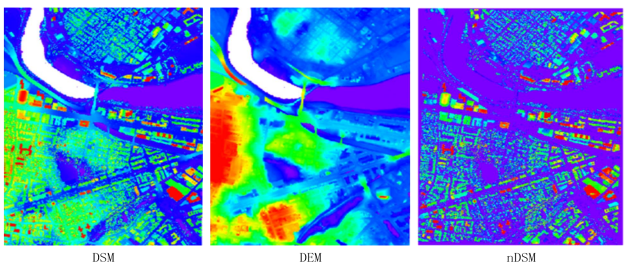


Figure 7. LiDAR digital products.

3.4 Accuracy Evaluation

As reported in Table 1, the proposed method combined with CNN and Transformer has achieved an improvement in classification performance. The experiment fused the original image and nDSM features to improve the Kappa coefficient by 5% compared to the support vector machine (SVM) and 4% compared to a random forest (RF). Compared to 3DCNN, the multi-head attention mechanism improves Kappa and F1 scores by 6.2%. Compared to 2DCNN, the proposed method improved the F1 score by 3.13 Kappa with 6.53%. This confirms that the fusion of CNN and Transformer can significantly improve prediction accuracy.

Table 1. Evaluation of Prediction Accuracy Using Different Models.

Method	Overall accuracy %	Kappa %	F1-score %	Precision %	Recall %
RF	98.11	94.23	95.56	96.23	95.32
SVM	98.57	93.36	94.63	95.12	93.17
2D-CNN	97.21	95.24	92.34	91.36	93.21
3D-CNN	91.23	92.17	92.47	97.24	95.34
Ours	98.65	98.37	98.87	98.56	98.15

The visualization results of the ablation experiment confirm the effectiveness of the improved structure. The 2D-CNN model has relatively low performance. The common classifiers of random forest and support vector machines perform better than 2D-CNN. In contrast, the proposed residual 2D-3D CNN has almost similar performance to random forests, support vector machines, and random forests.

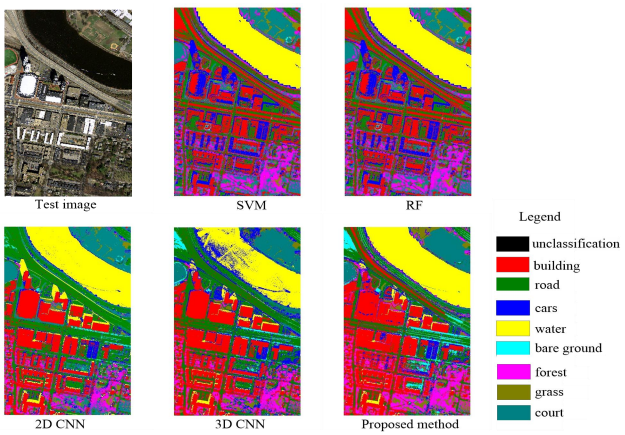


Figure 8. Prediction results using different models.

In this experiment, the accuracy of different land-cover types in the classification results was comprehensively analyzed. Conditional accuracy (user's accuracy) and recall (producer's accuracy) were used as key evaluation metrics, as illustrated in Figure 9. For building, the conditional accuracy was 79.31% and the recall was 63.89%, indicating that there were still misclassification results, and a considerable number of actual building pixels were not correctly classified. Road had a conditional accuracy of 71.43% and a recall of 89.29%, showing that although most actual road pixels were correctly classified, the classification results for users contained a relatively high proportion of misjudgments. For small cars, the conditional accuracy was 80.00% and the recall was 83.33%, with relatively better classification accuracy. Water bodies achieved a conditional accuracy of 96.55% and a recall of 100.00%, demonstrating excellent classification performance. The conditional accuracy of bare land was 75.00% and the recall was 100.00%, while forest land had a conditional

accuracy of 100.00% and a recall of 92.31%, and grassland had a conditional accuracy of 97.37% and a recall of 92.50%. The classification of the stadium was perfect, with both conditional accuracy and recall reaching 100.00%.

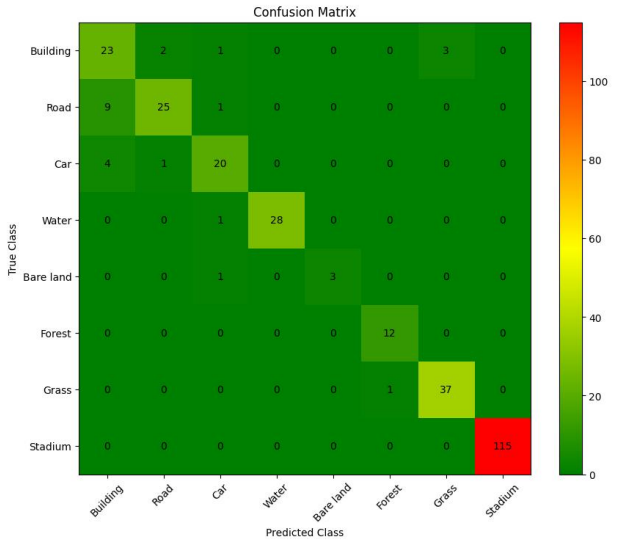


Figure 9. Confusion matrix for land use classification.

The differences in accuracy among various land-cover types can be attributed to several factors. The complexity of land-cover features is a significant factor. For instance, buildings and roads have complex features in remote-sensing images, and their spectral characteristics are easily affected by surrounding environments, causing misclassification. The small-sized cars are also easily interfered with the surrounding environment, and their spectral features are not stable. In contrast, water bodies have unique and stable spectral reflection features, which are easy to distinguish. In addition, the representativeness of sample data also affects the classification accuracy. If the training samples are not comprehensive or representative, the classifier's ability to identify certain land-cover types will be limited. Moreover, the unique properties of land-cover types themselves, such as the similarity of bare-land spectral characteristics to those of some surrounding land-cover types, also contribute to the accuracy differences. Stadiums, on the other hand, have distinct and regular features, making them easy to classify accurately.

Some data ablation experiments have been conducted to evaluate the effectiveness of multimodal data fusion. Remote sensing fusion images can provide feature information such as texture and color of land features, while LiDAR point cloud data can provide three-dimensional information such as geometric shape and height of land features. Integrating the two can provide a more comprehensive and accurate understanding of the distribution, characteristics, and attributes of urban land features. Comparing the final classification performance of the SVM and RF, as well as the influence of the near-infrared band on classification performance, the following conclusion can be drawn: under the same conditions, the classification performance of the RF classification algorithm is better than that of SVM.

In classification results using RF, the Kappa coefficients of the original image, original image+DSM, original image+DEM, and original image+nDSM were increased by 0.0494, 0.0812, 0.0588, and 0.1172, respectively, compared to SVM classification. The image classification performance in the near-infrared band is better than that in the RGB band. The Kappa coefficient of RGB+near-infrared+DEM classification using SVM increased by 0.1056 compared to the Kappa coefficient without near-infrared band, and the Kappa coefficient of random forest classification increased by 0.0134. The Kappa coefficient of RGB+near-infrared+nDSM classification using SVM increased by 0.0919 compared to that without the near-infrared band, and the Kappa coefficient of RF classification increased by 0.009.

Compared with the classification results of single high-resolution remote sensing data, the fusion of LiDAR and high-resolution remote sensing data has a significant improvement effect on urban land use classification. The classification effect after adding point cloud data is better than that of single remote sensing data. After adding DSM, DEM, and nDSM support vector machine classification to the original image, the Kappa coefficients increased by 0.0444, 0.0443, and 0.0263, respectively, compared to single remote sensing data. After adding DSM, DEM, and nDSM random forest classification to the original image, the Kappa coefficients increased by 0.0762, 0.0538, and 0.0941, respectively, compared to single remote sensing data.

4. Conclusion

The study constructs a network by combining Transformer and CNN to harness the full potential of deep learning for urban land mapping. The Visual Transformer Network Model represents a significant advancement in the field of urban land mapping, offering a powerful tool for network structure optimization. The proposed algorithm is generally better than other representative methods, and the classification accuracy using remote sensing data and LiDAR is improved. In future exploration, this study will further investigate the fusion of LiDAR and hyperspectral image features to improve the efficiency and predictive performance of algorithms for land use mapping.

Funding

This work was supported by the Municipal Guiding Science and Technology Plan Project of Panzhihua City, [NO. 2024ZD-S-82].

References

Gómez-Chova, L., Tuia, D., Moser, G. and Camps-Valls, G., 2015. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9), 1560-1584.

Tuia, D., Flamary, R., & Courty, N., 2015. Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 272-285.

Yan, W. Y., Shaker, A., & El-Ashmawy, N., 2015. Urban land cover classification using airborne LiDAR data: A review. *Remote sensing of environment*, 158, 295-310.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z. H., & Yan, S., 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *In Proceedings of the IEEE/CVF international conference on computer vision*, 558-567.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M., 2022. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 196-214.

Yuan, Q., & Mohd Shafri, H. Z., 2022. Multi-modal feature fusion network with adaptive center point detector for building instance extraction. *Remote Sensing*, 14(19), 4920.

Yuan, Q., & Xia, B., 2024. Cross-level and multiscale CNN-Transformer network for automatic building extraction from remote sensing imagery. *International Journal of Remote Sensing*, 45(9), 2893-2914.

Yao, J., Zhang, B., Li, C., Hong, D., & Chanussot, J., 2023. Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-15.

Yuan, Q., 2024. Multiscale global attention network with edge perceptron for automatic road extraction from remote sensing imagery. *IEEE Geoscience and Remote Sensing Letters*. 11 (21), 1-5.