50
Research paper
Doi 10.61892/stp202401050G
ISSN 2566-4484

Milan Gavrilović, University of Novi Sad, milangavrilovic@uns.ac.rs
Igor Ruskovski, University of Novi Sad, rus_igor@uns.ac.rs
Željko Bugarinović, University of Novi Sad, zeljkob@uns.ac.rs
Dušan Jovanović, University of Novi Sad, dusanbuk@uns.ac.rs
Miro Govedarica, University of Novi Sad, miro@uns.ac.rs

INTEGRATION OF RESUNET AND YOLO ALGORITHMS INTO A UNIFIED MODEL FOR OBJECTS DETECTION

Abstract

Automatic extraction of footprints of buildings from orthophotos is a challenge in the field of remote sensing data processing. The combination of image classification and object localization tasks in this research aims to develop a new model based on ResUNet and the YOLO algorithm. By applying the proposed model and publicly available data, a high level of building extraction success of 89% is achieved. Although there is potential to improve the results by introducing other types of data, the integration of these models represents a significant step towards improving the technology of automatic extraction of buildings from orthophotos.

Keywords: ResUNet, YOLO, neural networks, deep learning, building detectio.

ИНТЕГРАЦИЈА RESUNET И YOLO АЛГОРИТАМА У ЈЕДИНСТВЕНИ МОДЕЛ ЗА ДЕТЕКЦИЈУ ОБЈЕКАТА

Сажетак

Аутоматско издвајање отисака зграда са ортофото снимака представља изазов у области обраде података даљинске детекције. Комбинација задатака класификације слике и локализације објеката у овом истраживању има за циљ развој новог модела заснованог на *ResUNet*-у и *YOLO* алгоритму. Примјеном предложеног модела и јавно доступних података, постиже се висок ниво успјешности екстракције зграда од 89%. Иако постоји потенцијал за побољшање резултата увеђењем других типова података, интеграција ових модела представља значајан корак ка унапређењу технологије аутоматског издвајања зграда из ортофото снимака.

Кључне ријечи: ResUNet, YOLO, неуронске мреже, дубоко учење, детекција зграда

1. INTRODUCTION

Accurate and up-to-date information about buildings plays an essential role in many fields such as urban planning, environmental protection, real estate management, disaster risk assessment and many other fields [1], [2]. Such information is invaluable for modern society and city management, so it is not surprising that in recent years, with the rapid development of urban areas, automatic building extraction has become an important topic of scientific research [3], [4].

Remote sensing offers a large number of high-resolution sensors that record a wealth of information about objects on the ground. Therefore, remote sensing data provide significant support in the field of object classification and detection within urban areas [5]. The high spatial resolution of aerial and satellite images gives the possibility of distinguishing different objects in urban areas and enables the extraction of information about individual objects.

The manual method of collecting information about individual buildings, although it has high accuracy, is very expensive and time-consuming work, and cannot meet the requirements for quickly extracting and updating information about buildings in large areas. Using high-resolution images and modern image processing algorithms [6], [7], buildings can be accurately identified and classified based on their characteristics, such as roof shape, size, color, etc. However, this raises some new challenges in automatic object extraction due to the diversity of building features and complex environments [8]-[10]. First of all, buildings have significant differences in size, shape, height and function, and also have large variations in high resolution images caused by lighting, viewing angle, obscuration of the building by other objects and shadows [1]. Urban scenes consisting of spectrally similar objects such as roads, buildings and other artificial objects also make it difficult to accurately detect buildings [11].

Although this task has received a lot of attention in the scientific community, most approaches use additional data, such as point clouds, multispectral imagery, height information of objects and terrain (DSM, DEM), etc. This data, while important, is usually too expensive or unavailable for most cities around the world. Therefore, improving the accuracy and efficiency of automatic building extraction from high-resolution images is still a challenging task that is the focus of many researches [12], [13].

Traditional methods for extracting buildings from remote sensing images mainly involve image classification based on pixel features or image objects. Methods that rely on pixel features generally use the information of a single pixel for its classification. Initially, most studies used conventional machine learning techniques to process relevant objects by manually selecting features [3]. Machine learning methods such as K-Means [14], Support Vector Machines [15], Random Forest [16] and others were used for pixel analysis and classification. However, these methods ignore the relationship between neighboring pixels and do not use spatial information about objects. The results of this classification are prone to the influence of "salt and pepper", which results in blurred boundaries of separated buildings. Also, all these methods require prior knowledge, and their poor generalization capabilities due to manual feature selection may cause inaccurate results [8].

Detecting buildings from aerial and high-resolution satellite images is a complex task, partly due to the huge amount of data, i.e. of pixels to be processed, which often makes it challenging to handle large datasets in a fast and efficient manner, even with modern computing frameworks [17]. To deal with this challenge, new superpixel segmentation algorithms have been proposed and developed that group pixels to create contextually meaningful regions, leading to better processing efficiency while preserving important information [18]. However, the classification accuracy is highly dependent on the image segmentation results, and the segmentation scale is difficult to determine. Therefore, problems such as over- or under-segmentation are frequent occurrences, which can significantly affect the quality of the final result.

Due to the increase in computing power and the availability of large data sets, deep learning methods have emerged as successful tools for solving many tasks in the field of computer image processing. Deep learning has a strong generalization ability and the ability to efficiently express features [19]. It bridges the semantic gap, integrates feature extraction and image classification, and omits data pre-processing, such as image segmentation, through an end-to-end hierarchical construction method [5]. It can also automatically perform hierarchical feature extraction on massive raw data, reduce human labeling and reduce labor costs [5].

Deep learning, with convolutional neural networks (CNN) as a representative, is an automated artificial intelligence technique that has emerged in recent years, specialized in learning general patterns from large data sets, as well as exploiting the learned knowledge to solve unknown problems [20]. Deep learning has been successfully applied and rapidly developed in areas such as image classification, object detection, semantic segmentation, and instance segmentation.

Convolutional neural networks have a strong capacity to extract information from spatial context, and their automated learning mechanism allows for reuse [21]. All of the above, CNN is widely used on remote sensing images for object classification and detection.

Some of the key advantages of CNN-based image classification algorithms are that they provide solutions that offer greater generalization capabilities [22]. They also perform object-based classification, ie. they take into account features that characterize entire image objects, thus reducing the "salt and pepper" effect that affects conventional classifiers. CNNs can not only automatically extract features from raw image data, but also obtain semantic information level by level, which has resulted in great success in image classification tasks [23].

In 2015, Long et al. [24] proposed a fully convolutional network (FCN), the first end-to-end semantic segmentation method implemented in neural networks. Although FCN has achieved good results in building extraction, it does not consider the relationship between pixels. It also mainly focuses on global and ignores local features, resulting in poor results. However, most subsequent deep learning network models have been improved and innovated based on this model, i.e. many scholars have proposed some deepened and improved networks, such as U-Net [25], SegNet [26], etc. U-Net is one of the most commonly used network models for image segmentation tasks that belongs to one of the FCN variants. In recent years, many image segmentation algorithms have used U-Net as the original network model for segmentation.

In addition to the mentioned models for classification and segmentation, in the field of computer image processing there are algorithms for object detection that are also based on CNN. Object detection is the process of automatically finding and localizing objects of interest in images or videos. This process includes recognizing the presence of certain objects, as well as determining their locations and bounding boxes in the image or video. There are two basic types of algorithms for detecting objects in images using deep learning, namely two-level and one-level networks [21]. Two-level networks first identify potential regions that contain objects and then classify the image based on those regions. Examples of these algorithms include R-CNN (Region Based Convolutional Neural Networks) and Fast R-CNN. However, their disadvantage may be slower image processing. On the other hand, single-stage networks such as YOLO (You Only Look Once), SSD (Single Shot Detector), and similar ones perform object detection in a single step, which makes them more efficient than two-stage networks [27].

In the studies published so far that dealt with the classification of buildings and other objects on remote sensing images, a large number of different strategies are found. Jovanović et al. in the paper [12] propose a U-Net model for identifying changes in buildings in order to update the existing record of buildings. This study shows that the proposed model performed well in the identification of objects, but it also gives a lot of false positives, i.e. objects that do not belong to the building class are placed in that class. These results, using only very high-resolution images containing RGB and NIR bands, showed object identification accuracies ranging from 84% to 88%. In the paper [28], Chen et al. have proposed the Res2-Unet model to improve detection performance and generate accurate building boundaries. However, even this model is not able to distinguish individual roads from buildings. Kokeza et al. in their paper [29] test the possibility of using different publicly available datasets for training neural networks, and then test the ability of the model to generalize to the area of interest. The evaluation of the results showed that the models trained with publicly available datasets do not meet the required accuracy for updating cadastral maps in the study area. Much better results were achieved using orthophoto images, made from data obtained using UAV, for neural network training. Farajzadeh et al. in their work [30] investigate the ability of U-Net architecture with ResNet to extract building footprints from UAV-based orthophotos and digital surface models (DSM). Experiments highlight the effectiveness of height information for detecting and extracting building footprints with significant improvements in accuracy from 89% to 97%.

Donghang et al. [31] performed fast and accurate airport detection on remote sensing images using YOLO. Pham et al. in paper [32] introduce an improved single-stage detection model based on deep learning, called YOLO-Fine. This detector is designed to be capable of detecting small objects with high precision and high speed, enabling further real-time applications. Ma and others [27] implemented the detection of collapsed buildings using the YOLO algorithm. However, some results were characterized by incorrect bounding boxes.

Given the outlined challenges, the automatic extraction of building footprints, the outer surface of the building roof, from high-resolution orthophotos is one of the most challenging tasks in this field. Due to the aforementioned problems, when extracting objects, the tasks of image classification and localization of objects from the image must be combined, i.e. image classification is used to predict the class of an object in an image, and object localization is used to locate one or more objects present in an image and locate them using a bounding box. In response to these challenges, this paper proposes an enhanced building extraction method leveraging ResUNet and YOLO algorithms. In this proposed method, the ResUNet model acts as a feature extractor, while the YOLO algorithm is employed for object detection. The primary objective is to develop a novel model aimed at enhancing the efficiency and accuracy of object detection, thus advancing the technology of automatic building extraction from remote sensing data.

2. MATERIALS AND METHODOLOGY

2.1. METHODS

Within this case study, the authors proposed a workflow for the automatic extraction of buildings based on data generated on the principles of remote sensing. The suggested steps are shown in Figure 1.

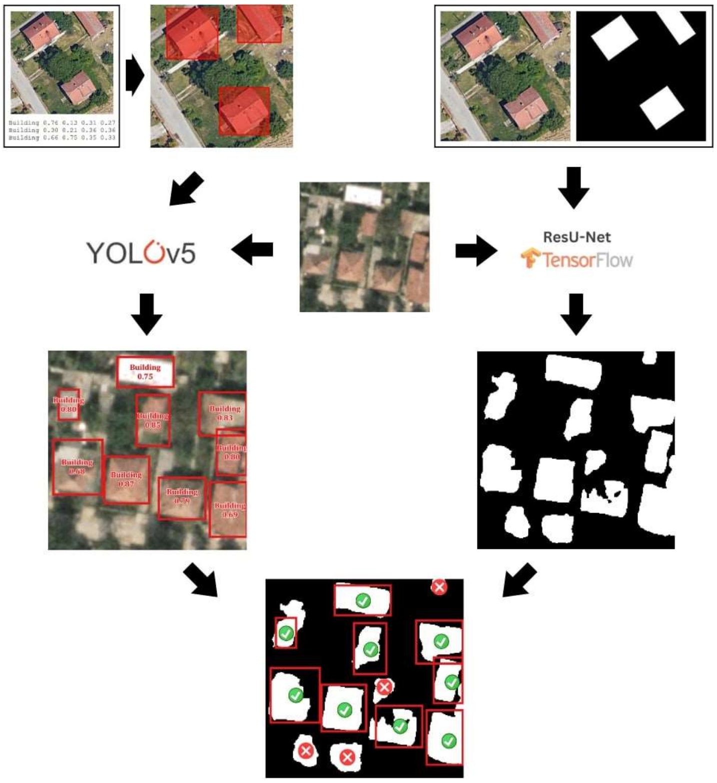

Building 0.76 0.12 0.31 0.27
Building 0.30 0.21 0.36 0.36
Building 0.66 0.75 0.35 0.33
YOLOv5
ResU-Net
TensorFlow
Building 0.75

Figure 1. Flow chart of research method

As the analysis of the literature found that ResUNet achieves much better results in the classification of satellite images than the U-Net architecture, this architecture was used as the basis for the classification of buildings in this paper. Since a certain number of false positives can be expected as a result, i.e. objects that are wrongly classified as buildings, in order to overcome this problem, a single-level YOLO network is used, which will be able to filter the obtained classification results in a fast and very simple way and thus localize only the objects of interest.

2.1.2. RESUNET

The problem of segmentation of satellite images represents a major challenge in remote sensing. In the last few years, algorithms based on convolutional neural networks have been developed with the aim of segmenting satellite images. The most common way of performing semantic segmentation is the use of convolutional neural networks because they achieve very good results, and one of the most famous architectures used is U-Net, which has a coder-decoder type structure. U-Net is a type of Fully Convolutional Network that was originally applied to medical image analysis, but later found application in many other fields, one of which is the classification of satellite images. U-Net is a special type of Fully Convolutional Network that merges low-level and high-level feature maps for better object localization.

Figure 2 is an illustration of the original U-Net architecture, with the downlink on the left and the uplink on the right. Max pooling operations reduce map sizes but increase the number of channels. In the expansive path, the sampling is followed by a convolution and the number of channels is halved so that the output has the same dimensionality as the input.

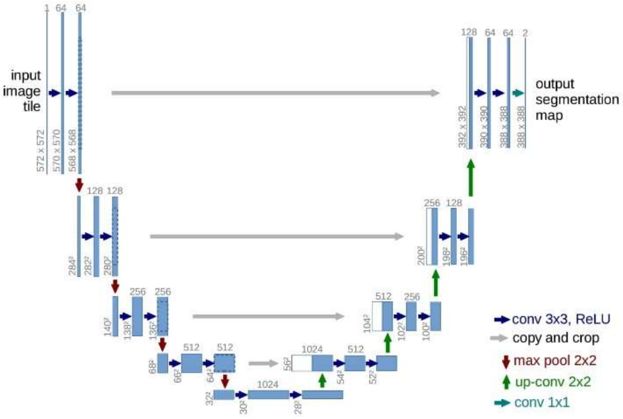


Figure 2. Original U-Net architecture [25]

ResUNet is a Deep Residual U-Net developed by Zhengxin Zhang et al. [33] for semantic segmentation. This architecture was originally applied to road extraction from high-resolution remote sensing imagery. Later, it found application in other areas such as segmentation of brain tumors, segmentation of human images and many others. The architecture of this model (Figure 3) consists of an encoding network, a decoding network and a bridge that connects these networks, just like U-Net. ResUNet is a fully convolutional neural network designed to achieve high performance with fewer parameters, and it represents an improvement on the existing U-Net architecture by taking advantage of both the U-Net architecture and Deep Residual Learning.

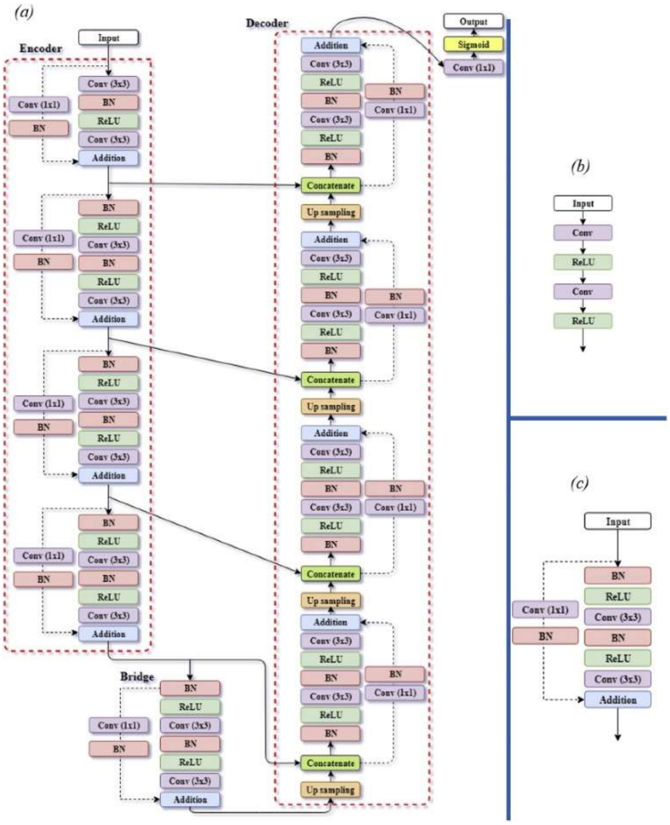


Figure 3. Architecture ResUnet [34]

Figure 3-b and Figure 3-c show the building block used in conventional U-Net and ResUNet, respectively. U-Net uses two convolutional layers (Conv) in the building block, each followed by a ReLU activation function. In the case of ResUNet, these layers are replaced by a residual block that uses batch normalization (BN), activation functions (ReLU) and convolutional layers (Conv). This combination of U-Net architecture and residual learning brings two advantages [33]:

- the residual unit facilitates network training,
- hopping connections within the network facilitates information dissemination without degradation.

2.1.3. YOU ONLY LOOK ONCE (YOLO)

The YOLO model is faster than the R-CNN family models, so it is commonly used in various real-time tasks, for example, object detection in videos. The YOLO model was first presented in 2015 by Redmon et al. [35]. R-CNN's key difference is that YOLO was the first to build a fast real-time object detector and it involves a single neural network trained end-to-end [36]. This algorithm differs from other object detection algorithms in that it "looks" at the image only once. The algorithm applies a single neural network to the entire image simultaneously predicting the probability that the object belongs to a certain class and the bounding boxes that determine its location in the image. Unlike two-stage detection network algorithms, YOLO treats target detection as a regression problem and simultaneously obtains target bounding boxes and the probability of object presence in the bounding box.

The boundary frame can be described by four descriptors:

- center of frame,
- frame width,
- height,
- a parameter that refers to the class to which the object belongs.

The version of YOLO used in this work is YOLOv5 developed by Ultralitics. The difference between YOLOv5 compared to previous models is that YOLOv5 uses a cross-level partial network (CSPNet) [37] as the backbone of the model and a path aggregation network (PANet) [38] as the backbone for feature aggregation. These new improvements provide better feature extraction. The architecture of the model is presented in Figure 4.

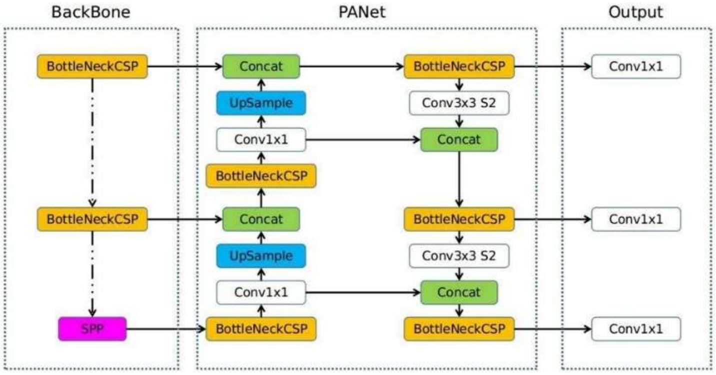


Figure 4. Overview of YOLOv5 architecture [36]

YOLOv5 provides different network models with different configurations and parameter sizes (Figure 5). It contains five different network models YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. The YOLOv5n model is the simplest, while the YOLOv5x model is the most complex. The larger the network, the more parameters that can be adjusted to get better performance, but this also means more time to train the model.

Figure 5. YOLOv5 different model sizes [36]

2.2. DATA COLLECTION

For a practical evaluation of the effectiveness and generalization performance of the model proposed in this paper, two publicly available datasets were used: the WHU aerial imagery dataset and the WHU satellite imagery dataset [39]. These datasets contain buildings of various types, shapes and sizes. The WHU aerial imagery dataset covers an area of 450 square kilometers in Christchurch, New Zealand and contains 187 000 buildings. The data set consists of 8 189 images with a resolution of 512×512 pixels, and the spatial resolution of the images is 0.3m [40]. Using this data, YOLO and ResUNet were trained, and testing of previously trained models was performed on orthophoto images of the area of interest, i.e. of the city of Novi Sad. These orthophoto images were generated using aerophotogrammetry methods with LEICA CAMERA RC 30 camera, with the longitudinal and transverse overlap of 60% and 25% respectfully.

3. RESULTS AND DISCUSSION

3.1. SEGMENTATION AND CLASSIFICATION

The input data for training the ResUNet model is composed of pairs of images (image and mask), as shown in Figure 6. Before training the ResUNet model, the input WHU data set is divided into two parts in the ratio of 80-20%. This step is necessary so that in the process of training the neural network, a quantitative assessment of the model can be obtained, which tells how well the network is trained to work with data that did not participate in the training.

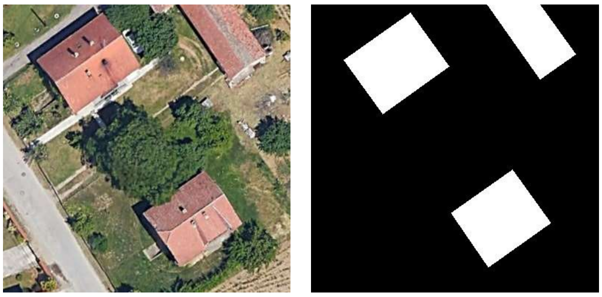


Figure 6. An example of a pair of input data used for training

In order to avoid overtraining the network, which can cause poor object identification, an early stop parameter was used in the training phase. Stopping the training is defined at the moment when the accuracy rating on the validation data starts to decrease, i.e. if in the three next epochs from the moment when the highest accuracy is reached, better results are not obtained, the training is interrupted, otherwise the training will go up to a maximum of 50 epochs. As a result of applying early stopping, it was obtained that the number of epochs required for training the ResUNet model is 20, while the accuracy of the model is 96.58%. In the following diagram (Figure 7) the curve of the model's accuracy rating obtained in the process of training the neural network can be seen.

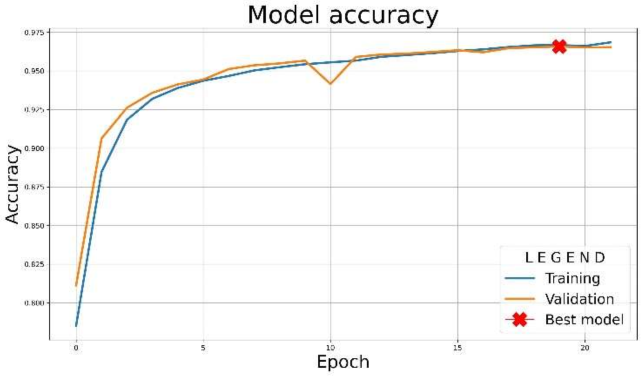


Figure 7. ResUNet model accuracy

After the completion of the training phase, the trained building identification model was applied to a new data set (orthophotograph of Novi Sad) in order to examine the possibility of knowledge transfer. As the model is applied to the input data, the result are raster images representing probability maps with a range of (0, 1), with 0 as the lowest probability of a building's existence and 1 as the highest probability of a building's existence. The next step represents the definition of the probability threshold value, based on which the buildings will be identified. The threshold value used in this paper is 0.5. Although it is expected that the slightly lower limit value chosen in this way will give many more false positives (objects that are not buildings and are classified in that class), it was deliberately chosen in order to cover all objects with certainty, i.e. in order to avoid the occurrence of false negatives (objects that exist but have not been identified), while false positives will be removed by integrating these results with the YOLO results. The results obtained using the ResUNet model on the test area are shown in Figure 8.

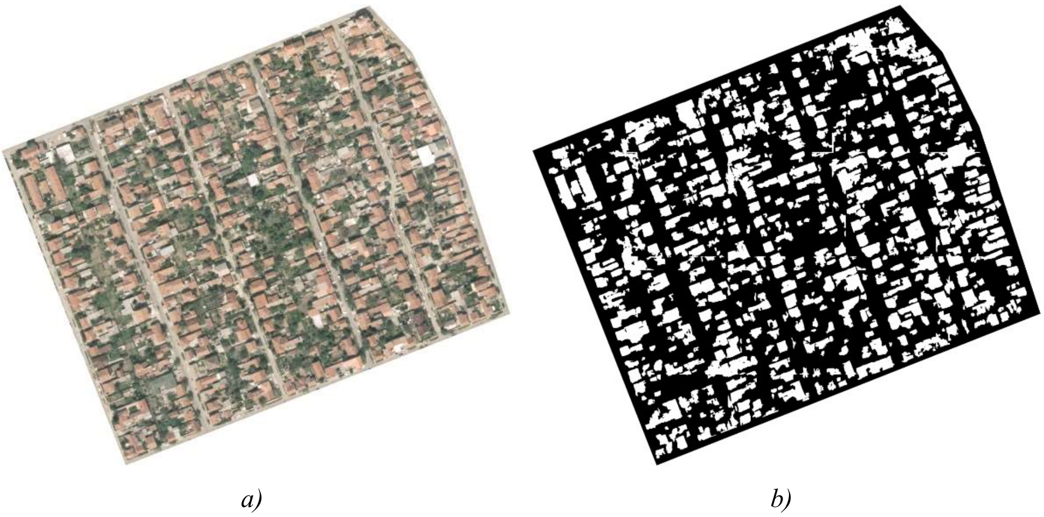
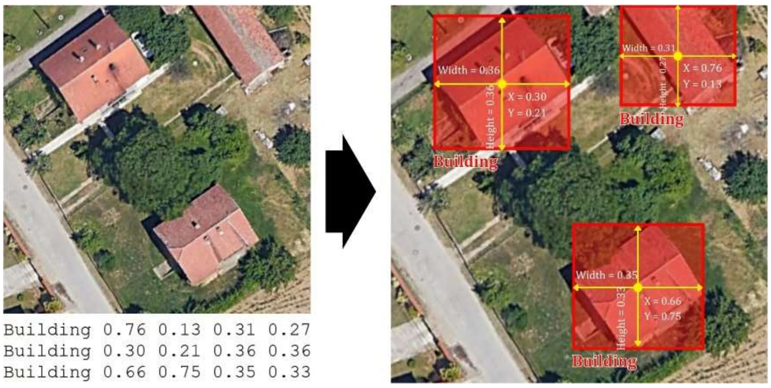


Figure 8. Orthophoto of parts of Novi Sad (a) and ResUNet model results (b)

From the results obtained using the ResUNet model (Figure 8-b), it can be seen that there is a large number of false positives, which was expected since a slightly lower threshold value was chosen for the identification of buildings, as previously explained. However, large number of false positives should not be seen as a problem since the goal was to completely include all objects on the analyzed area not regarding the potential occurance of false positives that will be filtered through YOLO algorithm.

3.2. OBJECT DETECTION

In order to detect buildings using the YOLO algorithm, the WHU data had to be adapted to the YOLO algorithm. What YOLO requires as training input is a dataset containing images and the coordinates of the bounding boxes of the objects in that image. The file structure with bounding boxes consists of five values that respectively refer to the class of the object, the coordinates of the center of the box, the width and the height of the bounding box, as shown in Figure 9.



```
Building 0.76 0.13 0.31 0.27
Building 0.30 0.21 0.36 0.36
Building 0.66 0.75 0.35 0.33
```

Figure 9. An example dataset for training the YOLO algorithm

During the training of the YOLOv5l network model, the maximum number of epochs is set to 50, with the fact that in the end the model from the epoch in which the highest accuracy is achieved is used for object detection. As with the ResUNet model, the division of the input data set into two parts in the ratio of 80-20% was used, in order to obtain the accuracy of the model during training. Performance was evaluated based on Precision, Recall and mAP (mean Average Precision) metrics when IoU (Intersection over Union) was 0.5 (50%) and 0.95 (95%). The next figure (Figure 10) shows graphs of metric curves during training.

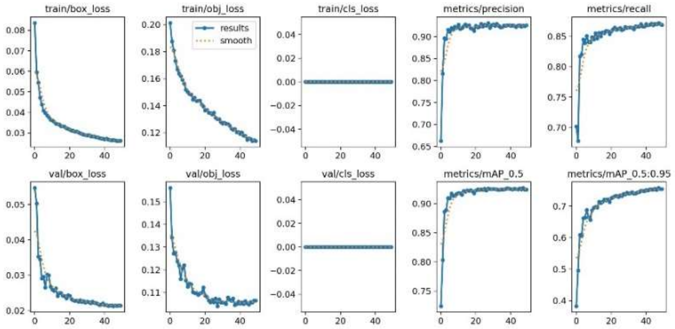


Figure 10. Graph of Precision, Recall, and mAP as YOLO training progresses

After evaluation, the model with the best accuracy obtained in 47 epochs and had a Validation Precision score of 0.93, Recall score of 0.87, as well as mAP score of 0.93 and 0.76 for @0.5IoU and @0.95IoU, respectively. This result confirms the effectiveness of this approach in correctly predicting objects. After completing the training phase, the trained building detection model was applied to the same data set as ResUNet (orthophotograph of Novi Sad). The results of object detection using the previously trained YOLO model in the test area are shown in the following figure.

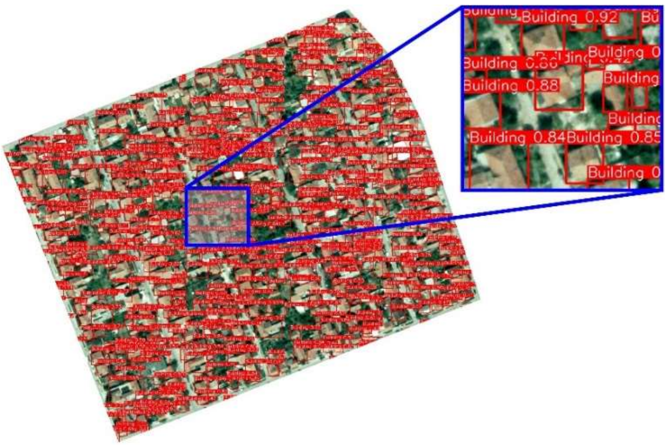


Figure 11. Building detection using the YOLO algorithm

The results of building detection using the YOLO algorithm (Figure 11) show a very high relative accuracy (percentage of the existence of objects within the boundary frames), but as with the ResUNet model, there are a certain number of false positives.

3.3. INTEGRATION RESUNET AND YOLO

The solution to the aforementioned problems regarding the presence of false positives in the results of segmentation, classification, and object detection lies in merging these results before drawing final conclusions. As can be concluded from the individual results (Figure 8 and Figure 11), while YOLO demonstrates the capability to detect objects more accurately (with a significantly lower false detection rate), it lacks the ability to differentiate their shapes like ResUNet, which yields a considerable number of false positives. Consequently, to retain the most crucial information from both models by integrating the results, it becomes feasible to enhance the accuracy of object detection and extract their shapes and positions more precisely.

An example of a correctly mapped building is given in Figure 12-detail A, where it can be seen that the building is correctly detected and that the boundaries of the object are preserved in the segmentation process. By crossing the results, the large number of false positives present in the ResUNet results is eliminated (Figure 12-detail B). As said YOLO gives much less false positives, so eliminating ResUNet polygons that are not inside the bounding box solves this problem. In this way, 81 polygons were removed (polygons detected only by ResUNet). False positives from YOLO results are also eliminated in the same way, i.e. all bounding boxes that do not contain ResUNet polygons are removed (Figure 12-detail C). In this way, 14 polygons were removed (polygons detected only by YOLO). The problems encountered by this methodology are given in Figure 12-detail D and E. In the first case (detail D) both models detected the existence of an object in a place where it is not present on the ground, while in the second case (detail E) neither ResUNet nor YOLO were able to detect the existence of the object. The reason for the appearance of these errors may lie in the differences, both in the shape and size of the objects in the images, as well as in the brightness and angle of the recording of the WHU data set for training and the orthophoto plan of Novi Sad used for testing. The number of false negatives, such as detail E in Figure 12, in the analyzed area is 38 objects, while the number of false positives, such as detail D in Figure 12, is 40 objects. These results can be considered successful, given that knowledge transfer was applied due to the lack of training data in the analyzed area, so it can be said that the proposed model copes well with new data.

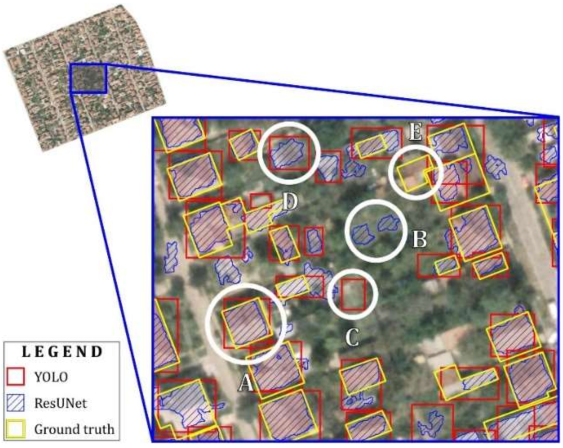


Figure 12. Analysis of the obtained results

One of the major problems this method is faced with is shown in Figure 13-detail F. Namely, in the case of the existence of buildings located next to each other, none of the used models is able to identify those buildings as separate entities, i.e. two individual buildings. Therefore, solving this problem as well as the problem with irregularly mapped edges of buildings, which is particularly pronounced in irregularly shaped buildings (Figure 13-detail G), is imposed as the primary goal of improving the proposed model.

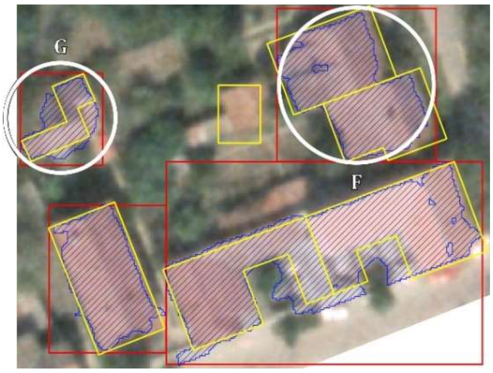


Figure 13. Problems in the results of the proposed model

In the end, it can be concluded that out of 367 buildings located in the analyzed area, 327 buildings were correctly identified. Accordingly, the percentage of success of the proposed methodology for the detection and extraction of buildings of 89% represents a very good basis for obtaining results quickly.

4. CONCLUSION

Traditional methods have proven to be insufficiently fast and efficient when extracting buildings from high-resolution images. With the development of deep learning technology, such as convolutional neural networks (CNN), new opportunities have opened up for automatic object detection and classification on remote sensing data. Models such as ResUNet and YOLO represent advanced techniques that have achieved high accuracy in object segmentation and detection. However, despite these advances, existing models face challenges in extracting buildings, such as shadows, different textures, and orientation of buildings.

The integration of ResUNet and YOLO represents a new approach that combines the strengths of both models to overcome these challenges. The ResUNet model, based on the idea of fully convolutional networks (FCN), has the ability to extract spatial information from the image, while the YOLO algorithm enables fast and efficient object detection. The combination of these models enables a better understanding of the image context and a more accurate detection of buildings even in complex scenes.

One of the goals of this paper was to examine the use of knowledge transfer, i.e. training models with publicly available data, and applying such trained models to new data. This approach significantly speeds up the time of obtaining results because there is no need for prior preparation of training data, and also solves the problem of availability of training data. In both cases (ResUNet and YOLO) it was shown that knowledge transfer is an applicable method that gives satisfactory results. As it can be concluded from the results shown (Figure 8 and Figure 11), in both cases there is a certain number of false positives that need to be eliminated and a significantly smaller number of false negatives. The very fact that the number of false negatives is very small indicates that knowledge transfer is applicable for this kind of object classification and detection.

Therefore, by using these two models together, one can eliminate false positives from the results and thus improve the accuracy, in such a way that the information about the shape and dimensions of the objects will be obtained using ResUNet, and for the verification of those results and the elimination of noises, the results of the YOLO model will be used. Based on the results thus obtained, the authors come to the conclusion that the extraction of buildings was performed with a high accuracy of 89%, especially if one takes into account the speed of obtaining results, which was obtained thanks to the use of publicly available data for training.

If necessary, the accuracy of the model could be increased by using some more publicly available data for training, in order to cover as many possible scenarios and diversity in the type, size, shape and texture of building roofs as possible. In this way, the problem of false negatives could be solved, i.e. not recognizing objects that exist on the ground. Another important source of information would be the DSM, which would significantly improve the results, but up-to-date data on the heights of objects are usually not available.

The issues encountered by the proposed model include the inability to identify adjacent buildings as separate units and difficulties in extracting building edges, resulting in the appearance of fuzzy edges. The process of building edge extraction is influenced by various factors, such as the complex appearance of the object, the capturing angle, glare, shadows, etc. The integration of ResUNet and the unique YOLO model for object detection represents a significant advancement in the technology of automatic building extraction from remote sensing data. Furthermore, these aforementioned challenges will guide future research directions and represent crucial issues that need to be addressed in order to achieve fully automated object mapping using orthophoto imagery.

ACKNOWLEDGMENTS

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad” (No. 01-3394/1).

LITERATURE

[1] J. Li, X. Huang, L. Tu, T. Zhang, and L. Wang, “A review of building detection from very high resolution optical remote sensing images,” *GIScience Remote Sens.*, vol. 59, no. 1, pp. 1199–1225, Dec. 2022, doi: 10.1080/15481603.2022.2101727.

[2] X. Huang and Y. Wang, “Investigating the effects of 3D urban morphology on the surface urban heat island effect in urban functional zones by using high-resolution remote sensing data: A case study of Wuhan, Central China,” *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 119–131, Jun. 2019, doi: 10.1016/j.isprsjprs.2019.04.010.

[3] B. Sariturk and D. Z. Seker, “Comparison of residual and dense neural network approaches for building extraction from high-resolution aerial images,” *Adv. Space Res.*, vol. 71, no. 7, pp. 3076–3089, Apr. 2023, doi: 10.1016/j.asr.2022.05.010.

[4] Y. Zhu, Z. Liang, J. Yan, G. Chen, and X. Wang, “E-D-Net: Automatic Building Extraction From High-Resolution Aerial Images With Boundary Information,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 4595–4606, 2021, doi: 10.1109/JSTARS.2021.3073994.

[5] D. You *et al.*, “EfficientUNet+: A Building Extraction Method for Emergency Shelters Based on Deep Learning,” *Remote Sens.*, vol. 14, no. 9, Art. no. 9, Jan. 2022, doi: 10.3390/rs14092207.

[6] S. Vasavi, H. Sri Somagani, and Y. Sai, “Classification of buildings from VHR satellite images using ensemble of U-Net and ResNet,” *Egypt. J. Remote Sens. Space Sci.*, vol. 26, no. 4, pp. 937–953, Dec. 2023, doi: 10.1016/j.ejrs.2023.11.008.

[7] A. Ghaznavi, M. Saberioon, J. Brom, and S. Itzerott, “Comparative performance analysis of simple U-Net, residual attention U-Net, and VGG16-U-Net for inventory inland water bodies,” *Appl. Comput. Geosci.*, vol. 21, p. 100150, Mar. 2024, doi: 10.1016/j.acags.2023.100150.

[8] Y. Liu *et al.*, “ARC-Net: An Efficient Network for Building Extraction From High-Resolution Aerial Images,” *IEEE Access*, vol. 8, pp. 154997–155010, 2020, doi: 10.1109/ACCESS.2020.3015701.

[9] Q. Tian, Y. Zhao, K. Qin, Y. Li, and X. Chen, “Dense feature pyramid fusion deep network for building segmentation in remote sensing image,” in Seventh Symposium on Novel Photoelectronic Detection Technology and Applications, SPIE, Mar. 2021, pp. 1361–1366. doi: 10.1117/12.2587144.

[10] B. Swan, M. Laverdiere, H. L. Yang, and A. Rose, “Iterative self-organizing SCEne-LEvel sampling (ISOSCELES) for large-scale building extraction,” *GIScience Remote Sens.*, vol. 59, no. 1, pp. 1–16, Dec. 2022, doi: 10.1080/15481603.2021.2006433.

[11] X. Huang and L. Zhang, “A Multidirectional and Multiscale Morphological Index for Automatic Building Extraction from Multispectral GeoEye-1 Imagery,” *Photogramm. Eng. Remote Sens.*, vol. 77, no. 7, pp. 721–732, Jul. 2011, doi: 10.14358/PERS.77.7.721.

[12] D. Jovanović, M. Gavrilović, D. Sladić, A. Radulović, and M. Govedarica, “Building Change Detection Method to Support Register of Identified Changes on Buildings,” *Remote Sens.*, vol. 13, no. 16, p. 3150, Aug. 2021, doi: 10.3390/rs13163150.

[13] H. Ye, S. Liu, K. Jin, and H. Cheng, “CT-UNet: An Improved Neural Network Based on U-Net for Building Segmentation in Remote Sensing Images,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp. 166–172. doi: 10.1109/ICPR48806.2021.9412355.

[14] T. Celik, “Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and k-Means Clustering,” *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009, doi: 10.1109/LGRS.2009.2025059.

[15] J. Inglada, “Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features,” *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 3, pp. 236–248, Aug. 2007, doi: 10.1016/j.isprsjprs.2007.05.011.

[16] Y. Dong, B. Du, and L. Zhang, “Target Detection Based on Random Forest Metric Learning,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, no. 4, pp. 1830–1838, Apr. 2015, doi: 10.1109/JSTARS.2015.2416255.

[17] A. Benchabana, M.-K. Kholladi, R. Bensaci, and B. Khaldi, “Building Detection in High-Resolution Remote Sensing Images by Enhancing Superpixel Segmentation and Classification Using Deep Learning Approaches,” *Buildings*, vol. 13, no. 7, Art. no. 7, Jul. 2023, doi: 10.3390/buildings13071649.

[18] B. Mao, B. Li, and J. Sun, “Large Area Building Detection from Airborne Lidar Data using OSM Trained Superpixel Classification,” in *2019 Seventh International Conference on Advanced Cloud and Big Data (CBD)*, Sep. 2019, pp. 145–150. doi: 10.1109/CBD.2019.00035.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[20] L. Luo, P. Li, and X. Yan, "Deep Learning-Based Building Extraction from Remote Sensing Images: A Comprehensive Review," *Energies*, vol. 14, no. 23, Art. no. 23, Jan. 2021, doi: 10.3390/en14237982.

[21] J. Gao, Y. Chen, Y. Wei, and J. Li, "Detection of Specific Building in Remote Sensing Images Using a Novel YOLO-S-CIOU Model. Case: Gas Station Identification," *Sensors*, vol. 21, no. 4, Art. no. 4, Jan. 2021, doi: 10.3390/s21041375.

[22] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020, doi: 10.1016/j.isprsjprs.2020.01.013.

[23] W. Zhang *et al.*, "Combining Deep Fully Convolutional Network and Graph Convolutional Neural Network for the Extraction of Buildings from Aerial Images," *Buildings*, vol. 12, no. 12, Art. no. 12, Dec. 2022, doi: 10.3390/buildings12122233.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, doi: 10.48550/ARXIV.1505.04597.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[27] H. Ma, Y. Liu, Y. Ren, and J. Yu, "Detection of Collapsed Buildings in Post-Earthquake Remote Sensing Images Based on the Improved YOLOv3," *Remote Sens.*, vol. 12, no. 1, Art. no. 1, Jan. 2020, doi: 10.3390/rs12010044.

[28] F. Chen, N. Wang, B. Yu, and L. Wang, "Res2-Unet, a New Deep Architecture for Building Detection From High Spatial Resolution Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 1494–1501, 2022, doi: 10.1109/JSTARS.2022.3146430.

[29] Z. Kokeza, M. Vujasinović, M. Govedarica, B. Milojević, and G. Jakovljević, "Automatic building footprint extraction from UAV images using neural networks," *Geod. Vestn.*, vol. 64, no. 04, pp. 545–561, 2020, doi: 10.15292/geodetski-vestnik.2020.04.545-561.

[30] Z. Farajzadeh, M. Saadatseresht, and F. Alidoost, "AUTOMATIC BUILDING EXTRACTION FROM UAV-BASED IMAGES AND DSMs USING DEEP LEARNING," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. X-4-W1-2022, pp. 171–177, Jan. 2023, doi: 10.5194/isprs-annals-X-4-W1-2022-171-2023.

[31] Y. U. Donghang, Z. Ning, Z. Baoming, G. U. O. Haitao, and L. U. Jun, "Airport detection using convolutional neural network and salient feature," *Bull. Surv. Mapp.*, vol. 0, no. 7, p. 44, Jul. 2019, doi: 10.13474/j.cnki.11-2246.2019.0216.

[32] M.-T. Pham, L. Courtrai, C. Friguet, S. Lefèvre, and A. Baussard, "YOLO-Fine: One-Stage Detector of Small Objects Under Various Backgrounds in Remote Sensing Images," *Remote Sens.*, vol. 12, no. 15, Art. no. 15, Jan. 2020, doi: 10.3390/rs12152501.

[33] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018, doi: 10.1109/LGRS.2018.2802944.

[34] H. T. Kollmann, D. W. Abueidda, S. Koric, E. Guleryuz, and N. A. Sobh, "Deep learning for topology optimization of 2D metamaterials," *Mater. Des.*, vol. 196, p. 109098, Nov. 2020, doi: 10.1016/j.matdes.2020.109098.

[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2015, doi: 10.48550/ARXIV.1506.02640.

[36] D. Dlužnevskij, P. Stefanovič, and S. Ramanauskaitė, "Investigation of YOLOv5 Efficiency in iPhone Supported Systems," *Balt. J. Mod. Comput.*, vol. 9, no. 3, 2021, doi: 10.22364/bjmc.2021.9.3.07.

[37] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 8759–8768. doi: 10.1109/CVPR.2018.00913.

[38] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," 2019, doi: 10.48550/ARXIV.1911.11929.

[39] S. Ji, S. Wei, and M. Lu, “Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019, doi: 10.1109/TGRS.2018.2858817.

[40] J. Chang *et al.*, “Multi-Scale Attention Network for Building Extraction from High-Resolution Remote Sensing Images,” *Sensors*, vol. 24, no. 3, Art. no. 3, Jan. 2024, doi: 10.3390/s24031010.