# INTISARI

Bangunan merupakan salah satu komponen tutupan lahan yang relatif cepat berubah karena berfungsi untuk berbagai hal oleh manusia sehingga data geospasial bangunan perlu diperbarui secara berkala. Saat ini, metode interpretasi visual pada citra satelit menjadi metode utama untuk menghasilkan data geospasial bangunan meskipun bergantung pada operator dan memakan waktu lama. Maka, penelitian ini mengusulkan metode segmentasi bangunan berbasis *deep learning* pada CSTRST untuk melakukan ekstraksi data bangunan yang lebih cepat dan dapat digunakan secara berkala. Secara umum terdapat dua jenis segmentasi citra pada *deep learning*, yaitu segmentasi semantik dan objek. Perbedaan utamanya adalah kemampuan memisahkan objek di kelas yang sama secara individu yang dapat dilakukan oleh segmentasi objek. Kemampuan tersebut karena penggunaan *bounding box* pada segmentasi objek yang tidak terdapat pada segmentasi semantik. Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi kinerja kedua tipe segmentasi tersebut di lokasi perkotaan di Indonesia.

Model *deep learning* yang digunakan adalah U-Net untuk segmentasi semantik dan Mask R-CNN untuk segmentasi objek membutuhkan bahan pembelajaran atau *training dataset* berisi pasangan *tile* citra dan labelnya (bangunan dan non-bangunan). Pada penelitian ini, label bersumber dari *ground truth* yang diperoleh dari digitasi secara manual dengan interpretasi visual. Proses *training model* dilakukan menggunakan *Google Colaboratory* dengan menggunakan *training dataset* dan *hyperparameter* yang sama terdiri dari *learning rate* = 0,001, *batch size* = 2, *epoch* = 70, *weight decay* = 0,0005, *momentum* = 0,9, dan *threshold* = 0,5. Bobot model dengan nilai *validation IoU* tertinggi digunakan untuk melakukan prediksi di enam lokasi tes. Lokasi 1 dan 2 (area pemukiman padat), lokasi 3 dan 4 (area bangunan berukuran besar dan kepadatan rendah), lokasi 5 (area terdapat objek non-bangunan yang menyerupai bangunan), serta lokasi 6 (area dengan sebagian bangunan tertutup vegetasi). Evaluasi terdiri dari dua, yaitu evaluasi akurasi melalui uji akurasi IoU dan *F1 score*, serta evaluasi geometri bangunan hasil segmentasi bangunan secara visual.

Berdasarkan nilai IoU, rata-rata akurasi kedua model adalah 0,687 dan 0,697 untuk U-Net dan Mask R-CNN. Secara umum, nilai IoU pada enam lokasi uji menunjukkan bahwa Mask R-CNN sedikit lebih unggul dibandingkan U-Net. Sementara itu, nilai rata-rata *F1 score* U-Net dan Mask R-CNN adalah 0,811 dan 0,813, dengan tren hasil yang serupa di setiap lokasi uji. Hasil dari dua evaluasi akurasi ini menunjukkan pola yang konsisten, yaitu akurasi segmentasi dipengaruhi oleh tingkat kompleksitas bangunan pada area pengujian. Meskipun secara rata-rata Mask R-CNN menghasilkan akurasi yang lebih tinggi dan unggul di empat lokasi uji dengan karakteristik bervariasi, U-Net menunjukkan performa yang lebih baik pada dua lokasi uji berkarakteristik pemukiman padat. Pada lokasi 1 dan 2, U-Net menghasilkan nilai IoU sebesar 0,612 dan 0,559 serta *F1 score* sebesar 0,761 dan 0,717, sedangkan nilai IoU model Mask R-CNN sebesar 0,510 dan 0,497 serta *F1 score* sebesar 0,677 dan 0,664. Berdasarkan nilai *precision*, model Mask R-CNN yang menggunakan *bounding box* memperoleh nilai yang lebih tinggi dibandingkan U-Net di seluruh lokasi uji, dengan selisih *precision* 0,038 hingga 0,128. Di sisi lain, model U-Net yang memberikan label pada tiap piksel mampu menjangkau lebih banyak bangunan, terutama di kawasan permukiman padat, sehingga memperoleh nilai *recall* yang lebih tinggi dengan selisih antara 0,011 hingga 0,169.

**Kata kunci** : Citra satelit, *deep learning,* U-Net, Mask R-CNN, indeks IoU, *F1 score*

# ABSTRACT

Buildings are one of the components of land cover that change relatively quickly because they are used for various purposes by humans, so building geospatial data needs to be updated regularly. Currently, the visual interpretation method on satellite images is the main method to generate geospatial data of buildings although it is operator-dependent and time-consuming. So, this research proposes a deep learning-based building segmentation method on CSTRST to perform faster building data extraction and can be used regularly. There are generally two types of image segmentation in deep learning, namely semantic and object segmentation. The main difference is the ability to separate objects in the same class individually which can be done by object segmentation. This ability is due to the use of bounding box in object segmentation which is not available in semantic segmentation. Therefore, this study aims to evaluate the performance of both types of segmentation in urban locations in Indonesia.

The deep learning models used are U-Net for semantic segmentation and Mask R-CNN for object segmentation, which require a training dataset containing image tile pairs and their labels (building and non-building). In this research, the labels are sourced from ground truth obtained from manual digitization with visual interpretation. The model training process was conducted using Google Collaboratory using the same training dataset and hyperparameters consisting of learning rate = 0.001, batch size = 2, epoch = 70, weight decay = 0.0005, momentum = 0.9, and threshold = 0.5. The model weight with the highest validation IoU value was used to make predictions at six test locations. Locations 1 and 2 (dense residential areas), locations 3 and 4 (large-sized and low-density building areas), location 5 (areas with non-building objects that resemble buildings), and location 6 (areas with partially vegetation-covered buildings). The evaluation consists of two, namely accuracy evaluation through the IoU accuracy test and F1 score, and evaluation of building geometry as a result of visual building segmentation.

Based on the IoU values, the average accuracy of the two models is 0.687 and 0.697 for U-Net and Mask R-CNN, respectively. In general, the IoU values at the six test locations show that Mask R-CNN is slightly superior to U-Net. Meanwhile, the average F1 score values of U-Net and Mask R-CNN were 0.811 and 0.813, with a similar trend of results at each test location. The results of these two accuracy evaluations show a consistent pattern, where the segmentation accuracy is affected by the level of building complexity in the test area. While on average, Mask R-CNN produced higher accuracy and excelled in the four test locations with varied characteristics, U-Net performed better in the two test locations with dense residential characteristics. In locations 1 and 2, U-Net produced IoU values of 0.612 and 0.559 and F1 scores of 0.761 and 0.717, while the Mask R-CNN model's IoU values were 0.510 and 0.497 and F1 scores of 0.677 and 0.664. Based on the precision values, the Mask R-CNN model using bounding boxes obtained higher values than U-Net in all test locations, with a difference in precision of 0.038 to 0.128. On the other hand, the U-Net model that labels each pixel is able to reach more buildings, especially in dense residential areas, thus obtaining higher recall values with a difference between 0.011 to 0.169.

**Keywords** : Satellite imagery, deep learning, U-Net, Mask R-CNN, IoU index, F1 score.