 ISSN:2302-8149
e-ISSN:2540-9719

Comparative Analysis of Instance Segmentation Models for House-Level Visual Socioeconomic Classification using Satellite Imagery

[1]David Cahyapratama, [2]Chairani Fauzi*
[1,2]Informatics Engineering, Informatics and Business Institute Darmajaya
[1,2]Jl. ZA. Pagar Alam No.93, Gedong Meneng, Kec. Rajabasa, Kota Bandar Lampung, Lampung 35141, Indonesia
*e-mail: chairani@mail.darmajaya.ac.id

(***received*:** 16 July 2026, ***revised*:** 28 July 2026, ***accepted*:** 29 July 2026)

Abstract

House-level visual socioeconomic information can provide a more detailed spatial understanding of residential areas and can support applied urban and commercial analysis. However, official socioeconomic data are often available only at broader administrative levels and may not capture variation among small residential clusters or individual houses. This study compares five instance segmentation models for detecting, segmenting, and classifying individual houses into lower, middle, and upper visual socioeconomic classes using high-resolution satellite imagery. The evaluated models are Mask R-CNN, Cascade Mask R-CNN, YOLO11n-seg, SOLOv2, and Mask2Former. The dataset consists of 1,209 training images with 7,052 annotations and 213 validation images with 1,353 annotations from residential areas in Banten and DKI Jakarta. Annotation reliability was assessed using 100 annotation pairs, producing a quadratic-weighted Cohen's kappa of 0.8077. Model performance was evaluated using COCO metrics for both segmentation masks and bounding boxes. The results show that Cascade Mask R-CNN achieved the highest observed overall validation performance among the tested configurations. Under the current experimental setting, it produced the strongest combination of object-localization and mask-segmentation metrics. These findings show that comparing multiple instance segmentation models can help identify a more suitable method for house-level visual socioeconomic classification. Unlike previous studies that generally perform area-level socioeconomic estimation or building extraction alone, this study compares multiple instance-segmentation approaches for expert-defined socioeconomic classification at the individual-house level. The resulting output can serve as a supplementary visual socioeconomic layer that complements demographic, accessibility, and commercial data in applied spatial and market-development analyses.
Keywords: Instance segmentation, Socioeconomic classification, Residential area, Satellite imagery, Model comparison.

1 Introduction

Detailed spatial information about the socioeconomic characteristics of residential areas is important for urban analysis, commercial planning, and shopping center market development analysis. In practice, socioeconomic conditions are often estimated using official statistical data, such as income, expenditure, population, and demographic indicators. However, these data are commonly available only at broader administrative levels and may not capture socioeconomic differences within smaller residential clusters or individual houses. This limitation can reduce the spatial detail needed to understand local residential characteristics.

High-resolution satellite imagery provides an additional source of information because it captures visible physical characteristics of residential areas. Housing size, settlement density, road access, roof condition, vegetation, and area regularity can reflect differences in residential conditions. Previous studies have shown that satellite imagery and machine learning can support socioeconomic mapping, poverty estimation, and welfare prediction, especially when detailed household survey data are limited [1], [2], [3]. However, many of these studies estimate socioeconomic conditions at the area level rather than identifying variation at the individual-house level.

 ISSN:2302-8149
e-ISSN:2540-9719

This creates an important research problem. Residential socioeconomic variation may occur even within the same small area, but area-level analysis may obsecure these differences. At the same time, detecting and classifying individual houses from satellite imagery is difficult because residential objects can be small, dense, irregular, visually similar, and located close to one another. Therefore, this study asks which instance segmentation model provides the most suitable validation performance for detecting, segmenting, and classifying individual houses into visual socioeconomic classes from high-resolution satellite imagery.

To answer this question, this study compares five instance segmentation models: Mask R-CNN, Cascade Mask R-CNN, YOLO11n-seg, SOLOv2, and Mask2Former. These models represent different instance segmentation approaches, including two-stage detection-based segmentation, cascade-based segmentation, one-stage YOLO-based segmentation, box-free segmentation, and transformer-based segmentation. The comparison is conducted using the same residential socioeconomic dataset so that model performance can be evaluated under a consistent experimental setting.

While previous studies have demonstrated the usefulness of satellite imagery for socioeconomic mapping, many of them estimate poverty, welfare, or living standards at broader spatial units rather than at the level of individual houses. In contrast, building extraction studies generally focus on detecting buildings, roofs, or footprints without linking each object to socioeconomic interpretation. This study addresses that gap by treating each house as an individual object and comparing several instance segmentation models to classify houses into lower, middle, and upper visual socioeconomic classes. This approach connects house-level segmentation with socioeconomic interpretation, rather than only producing area-level socioeconomic estimates or general building maps.

This study evaluates the comparative performance of Mask R-CNN, Cascade Mask R-CNN, YOLO11n-seg, SOLOv2, and Mask2Former using high-resolution residential satellite imagery from Banten and DKI Jakarta. The comparison is designed to identify the most suitable model for detecting, segmenting, and classifying individual houses based on visual socioeconomic class. The resulting house-level socioeconomic layer is expected to complement demographic, accessibility, and commercial data, including for applied spatial analysis and market development analysis.

2 Literature Review

Satellite imagery has increasingly been used as an alternative data source for socioeconomic analysis because it captures physical and environmental characteristics that are difficult to obtain from conventional statistical data alone. Built-environment features such as building size, housing density, road structure, roof condition, vegetation, and settlement regularity can provide indirect indications of socioeconomic differences. Previous studies have shown that satellite imagery and machine learning can support socioeconomic mapping and poverty estimation, especially when detailed household-level survey data are limited [1], [2]. Review-based research also shows that satellite imagery has been widely used for poverty prediction and welfare estimation in different geographic contexts [3].

Urban visual data can also strengthen the interpretation of neighborhood conditions. Street-view imagery has been used to analyze urban environments, accessibility, environmental quality, and social characteristics [4]. Multispectral imagery and convolutional neural networks have also been applied to estimate neighborhood context from visual urban data [5]. In addition, satellite and street-level imagery have been used to estimate urban inequality indicators, including income, overcrowding, and environmental deprivation [6]. These studies show that visual data can provide useful socioeconomic signals, but many of them still operate at neighborhood, grid, or area level rather than treating each house as an individual object.

The use of expert visual interpretation remains important in image-based socioeconomic mapping because socioeconomic conditions are not directly visible in satellite imagery. Human interpretation of satellite imagery can contain bias, while convolutional neural networks may capture visual patterns that are not always easy to explain manually [7]. This issue is relevant to residential socioeconomic classification because house class labels are based on visible indicators, such as roof condition, road access, housing density, building size, and surrounding environment, rather than direct income or expenditure measurement. Therefore, visual socioeconomic labels should be treated as interpreted indicators that require careful annotation and clear operational definitions.

 ISSN:2302-8149
e-ISSN:2540-9719

Instance segmentation is suitable for residential analysis because it can separate adjacent houses as individual objects. This is important in dense urban imagery, where houses may be close together, partially connected, or visually difficult to distinguish. Previous studies have shown that instance segmentation can be applied to large remote sensing imagery and can support building or roof segmentation tasks [8], [9]. Roof-level instance segmentation research also shows that object-level segmentation can be useful for separating individual built structures from imagery [10]. However, these studies mainly focus on physical object extraction, such as buildings, roofs, or footprints. They do not directly address the classification of individual houses into socioeconomic classes.

Remote sensing object detection also faces specific challenges that can affect residential socioeconomic classification. Object detection in remote sensing imagery is influenced by scale variation, object orientation, complex backgrounds, and differences in spatial resolution [11]. Small-object detection remains difficult because small objects contain fewer pixels and are more easily affected by background noise, occlusion, and image resolution [12]. Large-scale individual building extraction from satellite imagery is possible, but performance can still be affected by building size, density, and environmental variation [13]. These challenges indicate that model architecture can strongly influence the quality of house-level visual socioeconomic classification.

Mask R-CNN is widely used in instance segmentation because it produces bounding boxes, class labels, and segmentation masks for each detected object [14]. This structure is useful for residential analysis because the model can locate houses, classify them, and generate object-level masks. Nevertheless, single-model performance may be limited when objects are dense, small, or difficult to localize precisely. Therefore, relying on only one model may not be sufficient to determine the most suitable approach for residential socioeconomic classification.

Cascade Mask R-CNN extends the Mask R-CNN-based instance segmentation approach by using a cascade-based detection structure. The cascade framework improves detection quality through multiple stages trained with increasing Intersection over Union thresholds, allowing object hypotheses to be refined progressively before final prediction [15]. Cascade Mask R-CNN has also been applied to aerial image analysis and has shown better performance than Mask R-CNN for detecting greenhouse structures [16]. This structure is relevant for residential socioeconomic classification because better localization can support better segmentation, especially when houses are close to one another.

Other instance segmentation approaches provide different advantages. SOLOv2 removes dependence on bounding box detection by directly predicting instance masks using dynamic mask generation and Matrix NMS [17]. YOLO-based segmentation models represent one-stage approaches that are commonly associated with faster inference and practical deployment. Recent YOLO11-seg-based research also shows that YOLO segmentation architectures can be optimized for lightweight and efficient instance segmentation tasks [18]. Mask2Former represents a transformer-based universal segmentation architecture that uses masked attention and can be applied to semantic, instance, and panoptic segmentation tasks [19]. These models reflect different segmentation strategies, making them suitable for comparison in a residential socioeconomic classification task.

Table 1 Summary of previous studies

Author	Techniques Used	Advantages	Disadvantages
Arshad et al., 2023 [1]	CNN transfer learning using satellite imagery and night-time light proxy for socioeconomic mapping	Shows that satellite imagery can support socioeconomic estimation in developing countries	Focuses on area-level prediction, not individual-house segmentation
Agyemang et al., 2023 [2]	Ensemble CNN using ResNet50, ResNet50V2, and ResNet101 for rural poverty mapping	Uses high-resolution imagery and multiple data sources for poverty estimation	Predicts poverty at grid or area level, not house-level visual classes
Suel et al., 2021 [6]	Multimodal deep learning using satellite and street-level imagery for urban	Combines different visual data sources for socioeconomic	Focuses on grid-level inequality, not individual residential objects

 ISSN:2302-8149
e-ISSN:2540-9719

	inequality indicators	interpretation	
He et al., 2017 [14]	Mask R-CNN for object detection and instance segmentation	Provides bounding boxes, class labels, and segmentation masks for each object	General instance segmentation model
Oh et al., 2022 [16]	Cascade Mask R-CNN for greenhouse detection in aerial imagery	Shows that cascade-based segmentation can improve aerial object detection	Object type is greenhouse
Wang et al., 2020 [17]	SOLOv2 box-free instance segmentation using dynamic mask generation	Provides an alternative segmentation approach without box dependence	General model paper
Cheng et al., 2022 [19]	Mask2Former transformer-based segmentation using masked attention	Provides a modern transformer-based segmentation approach	General segmentation framework
Qiu et al., 2025 [18]	YOLO11-seg-based lightweight instance segmentation	Supports practical and efficient one-stage segmentation	Focuses on efficient segmentation,

Comparing several models is important because different segmentation architectures may respond differently to object density, object size, boundary complexity, and dataset characteristics. Previous classification research has shown that comparing multiple algorithms can help identify the most suitable method for a specific classification problem [20]. In this study, model comparison is necessary because the task is not only to detect houses, but also to generate segmentation masks and classify each house into a visual socioeconomic class. Therefore, the evaluation must consider both bounding box detection and mask segmentation performance.

Based on Table 1, previous studies show that satellite imagery, deep learning, and instance segmentation have been widely applied in socioeconomic mapping and remote sensing object detection. However, most socioeconomic studies estimate conditions at broader area or grid levels, while most segmentation studies focus on physical object extraction without linking each object to socioeconomic interpretation. In addition, few studies have compared multiple instance segmentation models for house-level visual socioeconomic classification using the same dataset. This study addresses these gaps by comparing Mask R-CNN, Cascade Mask R-CNN, YOLO11n-seg, SOLOv2, and Mask2Former for detecting, segmenting, and classifying individual houses into lower, middle, and upper visual socioeconomic classes. The study therefore positions instance segmentation as a supporting method for producing a house-level visual socioeconomic layer from satellite imagery.

3 Research Method

3.1 Research Design

This study uses a comparative experimental design to evaluate several instance segmentation models for house-level visual socioeconomic classification. The research focuses on detecting, segmenting, and classifying individual houses from high-resolution satellite imagery into three visual socioeconomic classes: lower, middle, and upper. The models compared in this study are Mask R-CNN, Cascade Mask R-CNN, YOLO11n-seg, SOLOv2, and Mask2Former.

The comparison was conducted using the same frozen training and validation dataset to ensure that all models were evaluated under a consistent data condition. The general workflow consisted of research area selection, satellite image preparation, visual socioeconomic annotation, image tiling, annotation format conversion, model training, validation evaluation, and comparative analysis. The workflow is shown in Figure 1.

 ISSN:2302-8149
e-ISSN:2540-9719

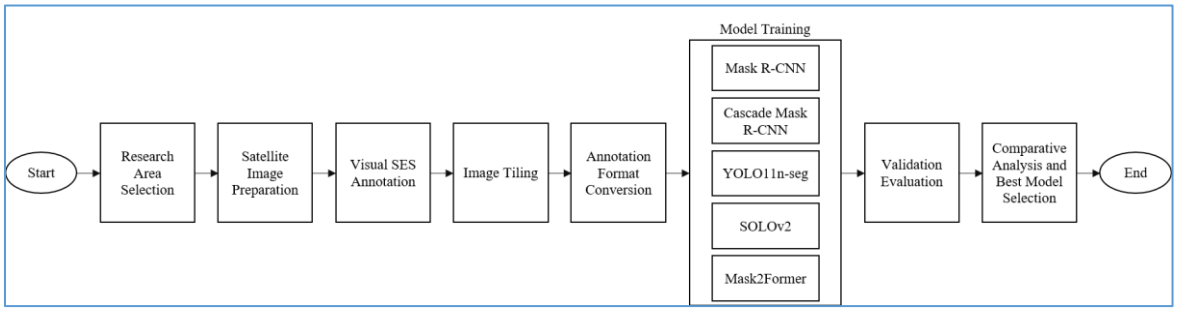

Figure 1 Research workflow

3.2 Data and Research Area

The data used in this study consist of high-resolution satellite images of residential areas in Banten and DKI Jakarta. The images were obtained from Google Earth, with the satellite imagery date recorded as 16 February 2026. The imagery has an approximate spatial resolution of 0.15 meters per pixel. These areas were selected because they contain varied residential patterns, including dense lower-income settlements, planned middle-class housing, and larger upper-class residential areas. This variation is important because the model needs to learn different visual characteristics of residential socioeconomic classes.

The dataset was divided into training and validation sets. The training data were used to train the models, while the validation data were used to evaluate and compare model performance. The dataset summary is shown in Table 2.

The dataset contains three visual socioeconomic classes. These classes do not represent direct measurements of income or expenditure. Instead, they are visual socioeconomic categories interpreted from physical characteristics visible in satellite imagery.

Table 2 Research dataset summary

Dataset	Images	Annotations
Training Data	1,209	7,052
Validation Data	213	1,353
Total	1,422	8,405

3.3 Socioeconomic Class Annotation Guidelines

This study uses three socioeconomic classes: lower, middle, and upper. These classes were determined through expert visual interpretation of house characteristics and the surrounding environment. The use of visual housing characteristics as socioeconomic indicators is supported by previous remote sensing studies, where physical welfare indicators such as roofing quality, building characteristics, and infrastructure are considered more observable in high-resolution imagery than direct income or consumption measures [3]. Expert-defined visual features such as building type, roof material, building size, settlement structure, building density, road quality, and vehicle presence have also been used to support wealth estimation from satellite imagery [7].

Table 3 Initial socioeconomic class guidelines

Class	Meaning	Initial visual guidelines
Class 1	Lower SES	Irregular environment; narrow alley access; not passable by cars; simple roofs such as metal sheets or light steel; housing-unit density >100 units/ha; distance between buildings <1.5 m.
Class 2	Middle SES	Relatively organized buildings; car access available; roofs generally made of clay tiles or similar materials; density of 80-100 units/ha; distance between buildings 1.5-3 m.
Class 3	Upper SES	Organized buildings; good car access; better building quality; more premium roof materials; density <80 units/ha; distance between buildings >3 m; close to main roads or commercial areas.

 ISSN:2302-8149
e-ISSN:2540-9719

In this study, the three classes are treated as expert-based visual socioeconomic classes, not direct measurements of household income or expenditure. The GIS expert team used the initial guidelines based on area regularity, road access, house size, roof condition, building density, distance between buildings, and proximity to main roads or commercial areas. The guidelines were used as an initial reference, while the final class decision was based on the overall visual-context interpretation of the experts. When visual interpretation from satellite imagery was unclear, additional checking was conducted using Google Street View to observe road condition, building façade, neighborhood quality, and surrounding infrastructure.

To assess inter-rater reliability, 100 house instances from imagery outside the training and validation datasets were assigned randomized identification numbers. Two members of the GIS expert team classified the same house instances into lower, middle, or upper visual socioeconomic classes using the annotation guidelines presented in Table 3. The resulting paired classifications were used to calculate unweighted, linear-weighted, and quadratic-weighted Cohen's kappa. Because the reliability samples were separate from the model-development dataset, the assessment was used only to measure consistency between annotators and did not affect the training or validation labels.[21].

3.4 Annotation and Pre-processing

House objects were manually annotated as polygons using ArcGIS Pro. Each visible house was delineated as an individual polygon and assigned to one of three visual socioeconomic classes. Polygon annotation was used because instance segmentation requires object-shape information, not only bounding box information. This allows the models to learn both the location and approximate shape of each house.

The annotated images were then prepared through image tiling and annotation adjustment. Image tiling was necessary because the original residential imagery covered relatively large areas, while deep learning models require smaller image inputs for training. All input tiles were prepared at 512 × 512 pixels. During tiling, house polygons were adjusted or clipped according to tile boundaries to maintain alignment between image tiles and their corresponding annotations.

After preprocessing, the annotations were converted into the formats required by each model. COCO format was used for Detectron2 and OpenMMLab-based models, while YOLO segmentation format was prepared for YOLO11n-seg. The conversion process preserved the same image split and class mapping so that the model comparison remained consistent.

3.5 Compared Instance Segmentation Models

This study compares five instance segmentation model configurations representing different segmentation approaches, as summarized in Table 4. Mask R-CNN was included as a two-stage instance segmentation model that produces bounding boxes, class labels, and segmentation masks for each detected object [14]. Cascade Mask R-CNN was included as a cascade-based model that progressively refines object proposals through multiple detection stages [15], [16]. YOLO11n-seg was included as a one-stage instance segmentation model designed for efficient object detection and mask prediction [18]. SOLOv2 was included as a box-free instance segmentation model that directly predicts instance masks using dynamic mask generation [17]. Mask2Former was included as a transformer-based segmentation model that applies masked attention to instance, semantic, and panoptic segmentation tasks [19].

All tested models were trained using the same residential socioeconomic dataset, three-class structure, input-tile size, training–validation split, and frozen validation set. However, their backbones, optimization methods, batch sizes, and training schedules were adjusted according to their respective frameworks and available computational resources. Therefore, the comparison should be interpreted as an evaluation of the tested model configurations rather than as a fully controlled comparison of model architectures alone. This setup allows the study to identify which tested configuration achieved the highest observed validation performance for the house-level visual socioeconomic classification task.

Table 4 Compared instance segmentation models

Model	General Approach	Main Output
Mask R-CNN	Two-stage detection-based	Bounding box, class label, segmentation

 ISSN:2302-8149
e-ISSN:2540-9719

	instance segmentation	mask
Cascade Mask R-CNN	Cascade-based instance segmentation	Refined bounding box, class label, segmentation mask
YOLO11n-seg	One-stage YOLO-based instance segmentation	Bounding box, class label, segmentation mask
SOLOv2	Box-free instance segmentation	Class label and segmentation mask
Mask2Former	Transformer-based segmentation	Class label and segmentation mask

3.6 Experimental Environment and Training Configuration

Manual annotation and spatial data preparation were conducted using ArcGIS Pro. Model training and validation were conducted using Google Colab with GPU acceleration. The runtime environment used an NVIDIA Tesla T4 GPU with approximately 15 GB of GPU memory, 12.7 GB of system RAM, around 78–100 GB of temporary storage, and 2 virtual CPU cores.

All models used the same input tile size of 512 × 512 pixels, the same three-class structure, and the same frozen validation set. A confidence threshold of 0.5 was used only for qualitative visualization and was not applied as a filtering threshold during COCO AP calculation, which used the evaluator's ranked prediction scores. The main training configuration of each model is summarized in Table 5.

The training schedule, batch size, learning rate, and optimizer were adjusted according to the implementation requirements and memory demands of each framework. Although the training configurations were not identical across frameworks, all models used the same input tile size, class structure, training-validation split, and evaluation dataset. Therefore, the comparison focuses on validation performance under the same house-level visual socioeconomic classification task.

Table 5 Model training configuration

Model	Framework	Backbone / Initialization	Training Setting	Optimization
Mask R-CNN	Detectron2	ResNet-50-FPN, COCO-pretrained	8,000 iterations, batch 4	SGD, LR 0.00025, steps 6,000 and 7,400
Cascade Mask R-CNN	Detectron2	ResNet-50-FPN, COCO-pretrained	8,000 iterations, batch 4	SGD, LR 0.00025, steps 6,000 and 7,400
YOLO11n-seg	Ultralytics	YOLO11n-seg, pretrained weights	30 epochs, batch 4	Auto optimizer, LR 0.01, default scheduler
SOLOv2	MMDetection / OpenMMLab	ResNet-50-FPN, pretrained backbone	30 epochs, batch 2	SGD, LR 0.001, milestones 20 and 27
Mask2Former	MMDetection / OpenMMLab	ResNet-50, COCO-pretrained configuration	8,000 iterations, batch 1	AdamW, LR 0.000025, steps 6,000 and 7,400

To make training exposure as comparable as possible, the training schedules were selected to provide a similar number of effective passes through the 1,209-image training dataset. For iteration-based models, the approximate number of effective epochs was calculated as the number of iterations multiplied by the effective batch size and divided by 1,209. Where exact equivalence was not possible because of framework-specific implementation and GPU-memory constraints, the closest feasible configuration was used, and the remaining differences are reported in Table 5.

3.7 Evaluation Method

 ISSN:2302-8149
e-ISSN:2540-9719

Model performance was evaluated using COCO evaluation metrics for both bounding boxes and segmentation masks [22]. Bounding box evaluation measures how well the model localizes house objects, while segmentation evaluation measures how well the model forms object masks. The evaluation metrics used in this study include segmentation AP, segmentation AP50, segmentation AP75, bounding box AP, bounding box AP50, and bounding box AP75.

All models were evaluated using the same frozen validation dataset, the same COCO-format ground-truth annotation file, the same category definitions, and the same IoU-based COCO metric definitions. The class mapping was kept consistent across all models. The original categories 1, 2, and 3 were mapped internally to zero-based model indices when required by each framework, and then converted back to the original category IDs for COCO-format evaluation. This procedure was used to ensure that the validation results remained comparable across models.

Bounding box metrics were included to provide a complete COCO-style comparison alongside segmentation metrics. For Mask R-CNN, Cascade Mask R-CNN, and YOLO11n-seg, bounding boxes were produced directly by the model outputs. Mask R-CNN and Cascade Mask R-CNN generate bounding boxes as part of the region-based detection pipeline, while YOLO11n-seg produces both bounding boxes and segmentation masks during inference. For SOLOv2, bounding boxes were not produced by a native bounding-box detection head because SOLOv2 is primarily a mask-based instance segmentation model. Therefore, the bounding boxes for SOLOv2 were derived from the predicted instance masks by calculating the minimum rectangular bounding box enclosing each predicted mask. These derived boxes were then converted into COCO bounding-box format [x,y,width,height] and evaluated as bounding-box predictions.

For Mask2Former, instance predictions were obtained through the MMDetection inference pipeline. The predicted bounding boxes were taken directly from the MMDetection instance-prediction output, converted into COCO bounding-box format [x,y,width,height], and evaluated using the same frozen validation dataset, COCO ground-truth annotation file, category definitions, and IoU-based evaluation settings used for the other models. The corresponding predicted instance masks were evaluated separately using COCO segmentation metrics. Therefore, no bounding boxes were derived from the Mask2Former masks. Although bounding-box metrics were included to support a complete COCO-style comparison, segmentation AP remained the primary metric for comparing instance-segmentation performance.

Intersection over Union was used to measure the overlap between predicted objects and reference annotations. IoU compares the intersection area and union area between the predicted object and the ground truth object. The IoU formula is shown in Equation (1).

$$IoU = \frac{\text{Area}(\text{Prediction} \cap \text{Ground Truth})}{\text{Area}(\text{Prediction} \cup \text{Ground Truth})} \quad (1)$$

Average Precision measures model performance based on the precision-recall relationship. Precision refers to the proportion of correct predictions among all predicted objects, while recall refers to the proportion of detected objects among all ground truth objects. AP can be expressed as the area under the precision-recall curve, as shown in Equation (2).

$$AP = \int_0^1 p\,(r)\,dr \quad (2)$$

In Equation (2), p(r) represents precision as a function of recall. In COCO evaluation, AP is calculated by averaging AP values across multiple IoU thresholds from 0.50 to 0.95 with a step of 0.05. The COCO AP formula is shown in Equation (3).

$$AP_{\text{COCO}} = \frac{1}{10} \sum_{t \in \{0.50, 0.55, \ldots, 0.95\}} A\,P_t \quad (3)$$

AP50 represents average precision at an IoU threshold of 0.50, while AP75 represents average precision at an IoU threshold of 0.75. AP50 indicates performance under a more moderate overlap requirement, while AP75 indicates stricter localization or segmentation quality. These metrics are

 ISSN:2302-8149
e-ISSN:2540-9719

suitable for this study because house-level visual socioeconomic classification requires both correct house localization and accurate house mask generation.

The final analysis compares the five models based on their validation metrics. The best-performing model is determined by considering both segmentation and bounding box results, with greater emphasis on segmentation performance because this study focuses on instance segmentation of individual houses. The comparison does not use an independent geographic test area; therefore, the results represent comparative validation performance under the same dataset configuration. Generalization to other cities or regions is treated as a limitation and future research direction.

4 Results and Analysis

4.1 Overall Validation Performance

The validation results of the five instance segmentation models are shown in Table 6. The comparison used the same frozen validation dataset, the same three visual socioeconomic classes, and the same input tile size of 512 × 512 pixels. The evaluation was conducted using COCO metrics for both segmentation masks and bounding boxes.

Table 6 Model Validation Performance

Model	Segm AP	Segm AP50	Segm AP75	BBox AP	BBox AP50	BBox AP75
Mask R-CNN	17.8126	43.239	10.49	21.4895	46.0369	16.3809
Cascade Mask R-CNN	18.646	44.1331	13.2865	22.7357	46.4578	20.0264
YOLO11n-seg	16.4577	39.8523	9.3802	19.5967	41.7148	15.8649
SOLOv2	11.0655	32.9338	4.3288	9.4967	25.8839	4.9865
Mask2Former	17.4	43.6	9.8	18.9	44.1	12.3

Table 6 shows that Cascade Mask R-CNN achieved the highest score across all reported validation metrics. It obtained the highest segmentation AP, segmentation AP50, segmentation AP75, bounding box AP, bounding box AP50, and bounding box AP75. This indicates that Cascade Mask R-CNN provided the most balanced validation performance for both house localization and mask segmentation in this dataset.

The difference between Cascade Mask R-CNN and Mask R-CNN was not very large at the AP50 level, but the improvement became clearer at the stricter AP75 level. Cascade Mask R-CNN achieved a segmentation AP75 of 13.2865, compared with 10.4900 for Mask R-CNN. It also achieved a bounding box AP75 of 20.0264, compared with 16.3809 for Mask R-CNN. This suggests that the cascade-based model was more effective when more precise overlap between prediction and reference annotation was required. This finding is consistent with the cascade detection concept, where object hypotheses are progressively refined through multiple stages, and with aerial-image research showing that Cascade Mask R-CNN can improve object detection performance compared with baseline Mask R-CNN [15], [16].

However, the performance difference between Cascade Mask R-CNN and Mask R-CNN should be interpreted cautiously. Cascade Mask R-CNN exceeded Mask R-CNN by 0.8334 segmentation AP and 1.2462 bounding box AP. These differences indicate the highest observed validation performance in this experiment, but this study did not conduct repeated training with different random seeds, bootstrap confidence intervals, or paired uncertainty analysis. Therefore, the results should not be interpreted as statistically significant differences, but as comparative validation results under the current experimental setting.

 ISSN:2302-8149
e-ISSN:2540-9719

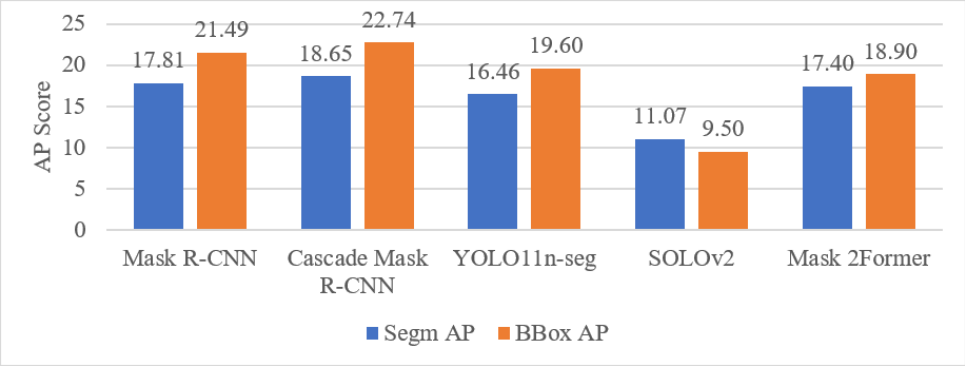

Figure 2 Model comparison of segmentation AP and bounding box AP

Figure 2 shows the comparison between segmentation AP and bounding box AP for each model. The figure highlights that bounding box AP was generally higher than segmentation AP for most models, although SOLOv2 showed the opposite pattern because its bounding boxes were derived from predicted masks rather than produced by a native bounding-box detection head. Overall, the results indicate that locating the general position of houses was generally easier than producing accurate segmentation masks that follow the shape of each house.

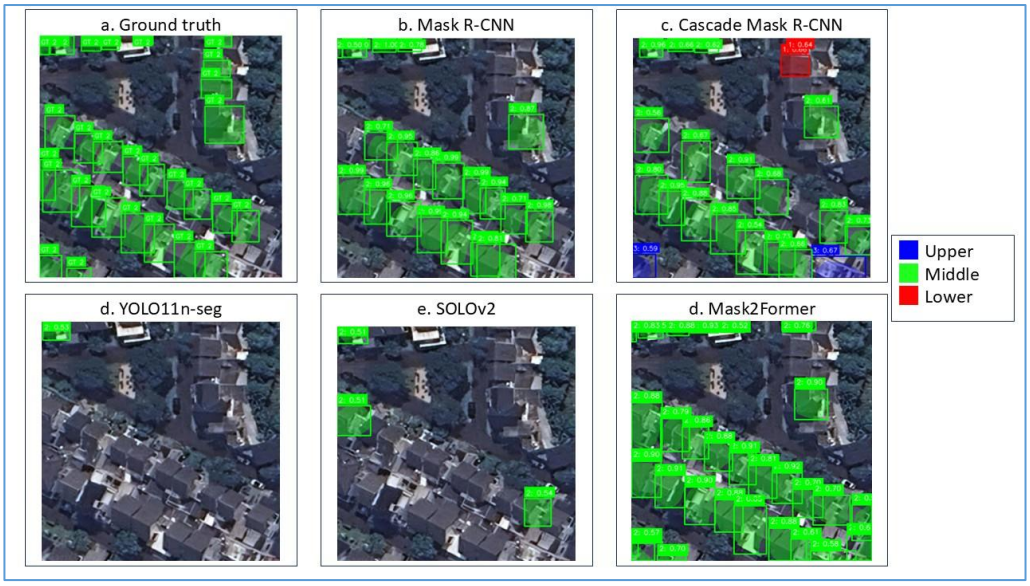

Figure 3 Qualitative comparison of model prediction results

In addition to quantitative evaluation, visual comparison was conducted to observe how each model detected, segmented, and classified house objects. Figure 3 shows an example validation image and the prediction outputs generated by the five evaluated models. This qualitative comparison helps illustrate differences in object detection, mask completeness, and the separation of adjacent houses.

4.2 Segmentation Performance Analysis

Segmentation performance is the main focus of this study because house-level visual socioeconomic classification requires object masks, not only object locations. Based on segmentation AP, the model ranking was Cascade Mask R-CNN, Mask R-CNN, Mask2Former, YOLO11n-seg, and SOLOv2. Cascade Mask R-CNN achieved the highest segmentation AP of 18.6460, followed by Mask R-CNN with 17.8126 and Mask2Former with 17.4000.

The segmentation AP values show that mask generation remains challenging in this dataset. One reason is the complexity of residential satellite imagery. Houses can be small, dense, irregular, visually similar, and located close to other houses. In dense settlements, the boundary between one roof and another may be unclear. Shadows, vegetation, roof color variation, and image tile boundaries can also make segmentation more difficult. As a result, the model may detect the general location of a house but fail to produce a mask that closely matches the reference polygon. These difficulties are

 ISSN:2302-8149
e-ISSN:2540-9719

consistent with remote sensing object detection studies, which show that scale variation, complex backgrounds, small object size, and spatial resolution can affect detection and segmentation performance [11], [12]. Large-scale building extraction studies also show that building size, density, and environmental variation can influence extraction quality [13].

The difference between segmentation AP50 and segmentation AP75 also supports this interpretation. All models achieved much higher AP50 than AP75. For example, Cascade Mask R-CNN achieved 44.1331 for segmentation AP50 but only 13.2865 for segmentation AP75. This means that the model could often detect houses at a moderate overlap threshold, but its performance decreased when the evaluation required more precise mask overlap. Therefore, the main difficulty was not only finding house objects, but also accurately separating and outlining them.

Mask2Former achieved a relatively strong segmentation AP50 of 43.6000, which was slightly higher than Mask R-CNN at 43.2390. However, Mask2Former had lower segmentation AP and AP75 than Mask R-CNN. This suggests that Mask2Former was able to identify house objects at a moderate overlap level, but it was less consistent when stricter segmentation quality was required. This may be related to the limited dataset size, training configuration, or the need for more extensive optimization when using transformer-based segmentation models [19].

4.3 Bounding Box Localization Performance Analysis

Bounding box evaluation measures how well each model localizes house objects. In this evaluation, Cascade Mask R-CNN achieved the best bounding box performance, with a bounding box AP of 22.7357 and bounding box AP50 of 46.4578. Mask R-CNN ranked second with a bounding box AP of 21.4895, followed by YOLO11n-seg with 19.5967, Mask2Former with 18.9000, and SOLOv2 with 9.4967.

It should be noted that bounding-box predictions were not generated in the same manner for all architectures. Mask R-CNN, Cascade Mask R-CNN, YOLO11n-seg, and Mask2Former provided bounding boxes directly through their respective inference outputs. In contrast, SOLOv2 does not use a native bounding-box detection head; therefore, its bounding boxes were derived from the predicted masks using tight axis-aligned enclosing rectangles. Consequently, the bounding-box metrics provide a useful COCO-style comparison of object localization, but segmentation AP remains the primary metric for evaluating and comparing instance-segmentation quality.

The bounding box AP values were generally higher than the segmentation AP values. This result is expected because bounding box localization is less detailed than segmentation. A bounding box only needs to cover the general object area, while a segmentation mask must follow the object shape more closely. In residential satellite imagery, this difference is important because many houses have irregular shapes, attached roofs, or unclear boundaries.

Cascade Mask R-CNN showed the strongest advantage in bounding box AP75. The model achieved 20.0264, which was higher than the other models. This indicates that Cascade Mask R-CNN was better at precise object localization. The cascade structure progressively refines object predictions, which may help the model adjust bounding boxes more accurately before producing the final prediction. This advantage is relevant for house-level segmentation because inaccurate localization can affect the quality of the final mask.

4.4 Model Architecture Comparison

The tested two-stage and cascade-based configurations produced the two highest segmentation AP values in this dataset. Cascade Mask R-CNN achieved the highest segmentation AP, followed by Mask R-CNN. These configurations use region proposals and region-based feature extraction, which may be beneficial for distinguishing small, dense, and visually similar house objects. Mask R-CNN generates bounding boxes, class labels, and instance masks through region-based detection and segmentation [14]. Cascade Mask R-CNN extends this process through multiple refinement stages, which can improve the quality of object localization before the final mask is produced [15].

The tested YOLO11n-seg configuration achieved moderate validation performance, with a segmentation AP of 16.4577 and a bounding box AP of 19.5967. These results were lower than those of Mask R-CNN and Cascade Mask R-CNN but higher than those of SOLOv2. YOLO-based segmentation models are designed for efficient one-stage detection and segmentation and may offer practical advantages for inference and deployment [18]. However, under the training configuration used in this study, YOLO11n-seg did not achieve the highest validation performance. This result

 ISSN:2302-8149
e-ISSN:2540-9719

suggests that the tested configuration was less effective at separating and delineating dense residential objects than the two-stage configurations.

The tested SOLOv2 configuration produced the lowest validation performance, with a segmentation AP of 11.0655 and a bounding box AP of 9.4967. SOLOv2 directly predicts instance masks without relying on a native bounding-box detection head [17]. In this study, the large number of small and adjacent houses may have made direct instance-mask prediction more difficult. Nevertheless, the result should not be interpreted as evidence that the SOLOv2 architecture is generally unsuitable for residential imagery, because its performance may be improved through additional configuration adjustment, longer training, or more extensive hyperparameter optimization.

The tested Mask2Former configuration achieved competitive segmentation performance, particularly at the AP50 threshold. Its segmentation AP50 of 43.6000 was slightly higher than that of Mask R-CNN, although its overall segmentation AP and AP75 were lower. Mask2Former uses transformer-based masked attention and can support multiple segmentation tasks [19]. ts performance in this study may have been influenced by the available dataset size, training exposure, batch size, and computational limitations. Therefore, the results indicate that Cascade Mask R-CNN achieved the highest observed performance among the tested configurations, but they do not establish that one architecture is universally superior to the others.

4.5 Annotation Reliability and Object Complexity

Annotation reliability was evaluated to measure the consistency of visual socioeconomic class interpretation between annotators. The reliability check used 100 annotation pairs from the three socioeconomic classes. The results are shown in Table 7.

Table 7 Inter-rater reliability results

Metric	Result
Number of annotation pairs	100
Exact agreement	74.00%
Unweighted Cohen's kappa	0.6139
Linear-weighted Cohen's kappa	0.711
Quadratic-weighted Cohen's kappa	0.8077

Table 7 shows that the weighted coefficients were higher than the unweighted coefficient because adjacent-class disagreements received smaller penalties than disagreements between more distant classes. Together with the exact agreement of 74.00%, these results indicate reasonably consistent expert labeling. However, the reliability results measure consistency between annotators and do not establish that the visual socioeconomic classes are equivalent to measured household income or expenditure [23].

However, visual socioeconomic class is not always clear from satellite imagery alone. Some houses may have similar roof materials but different surrounding conditions. Other houses may be located at the boundary between lower and middle, or middle and upper classes. Therefore, the kappa results should be interpreted as evidence of annotation consistency, not as proof that the visual classes directly represent actual household income or expenditure.

This issue is consistent with previous socioeconomic mapping studies, which explain that visual indicators from imagery can support socioeconomic estimation but are not direct measurements of income or consumption [3]. It is also related to the problem of human interpretation and model explainability in satellite-based socioeconomic prediction, where model outputs and expert interpretation may both be affected by visual ambiguity [7].

Annotation consistency can affect model performance because the model learns from the visual patterns provided in the annotations. When class boundaries are visually clear, the model has a better chance of learning consistent patterns. However, when class differences are subtle or ambiguous, classification becomes more difficult. Therefore, moderate AP values do not only reflect model limitations, but also the difficulty of the visual socioeconomic labeling task itself.

Object shape and density also affect segmentation quality. Houses in planned residential areas may have more regular shapes and clearer spacing, making them easier to segment. In contrast, dense settlements often contain small houses, narrow gaps, irregular roof shapes, and overlapping visual

 ISSN:2302-8149
e-ISSN:2540-9719

patterns. These conditions can cause under-segmentation, over-segmentation, missed detections, or masks that merge adjacent houses. This explains why stricter metrics such as AP75 were much lower than AP50. This interpretation is also consistent with remote sensing studies showing that small objects, dense object distribution, and complex backgrounds can reduce detection and segmentation accuracy [11], [12].

4.6 Implications for House-Level Visual Socioeconomic Classification

The comparison indicates that Cascade Mask R-CNN was the most suitable model among the tested methods for this dataset. Its stronger performance in both segmentation and bounding box metrics suggests that cascade-based refinement is useful for detecting and segmenting individual houses from satellite imagery. This finding is important because the task requires the model to identify each house as a separate object and assign it to a visual socioeconomic class.

The results also show that model comparison is necessary for this type of task. If the study only used one segmentation model, it would be difficult to determine whether the result was caused by the dataset, the annotation process, or the model architecture. By comparing five models under the same dataset condition, this study provides a clearer basis for selecting an appropriate method for house-level visual socioeconomic classification.

The output of the model can be used as a house-level visual socioeconomic layer. This layer does not represent direct income, expenditure, or household survey data. Instead, it represents visual socioeconomic interpretation based on residential physical characteristics visible from satellite imagery. Therefore, the output is most appropriate as supporting spatial information that can complement other datasets such as demographic data, accessibility data, road networks, administrative boundaries, and commercial facility data.

4.7 Result Limitations

Several limitations must be considered when interpreting the results. First, the evaluation was conducted using a frozen validation set, not an independent geographic test area. Therefore, the results show comparative validation performance under the same dataset configuration, but they do not yet confirm generalization to other cities or regions.

Second, the dataset was built from residential areas in Banten and DKI Jakarta. Residential patterns in other cities may differ in roof material, building density, road structure, vegetation, settlement arrangement, and image quality. As a result, the model may require additional training data before being applied to other geographic contexts.

Third, the socioeconomic labels are based on visual interpretation. Although the annotation used visible housing characteristics and additional checks when necessary, the labels are not direct measurements of household income or expenditure. Future research should validate the predicted classes with survey data, income data, expenditure data, or other socioeconomic indicators.

Fourth, the training configurations were adjusted according to each framework and computational limitation. This means the comparison reflects practical implementation under the same dataset and available resources, but not necessarily the maximum possible performance of each architecture. Future research should test additional hyperparameter settings, larger backbones, longer training schedules, and independent geographic validation areas.

5 Conclusion

This study compared five instance-segmentation configurations for house-level visual socioeconomic classification using high-resolution satellite imagery. Cascade Mask R-CNN achieved the highest observed validation performance, producing the strongest segmentation and bounding-box metrics under the current experimental setting. The results also showed that precise mask delineation remained more difficult than object localization, particularly for small, dense, and visually ambiguous houses. The study contributes a comparative evaluation of multiple instance-segmentation approaches for classifying individual houses into expert-defined lower, middle, and upper visual socioeconomic categories. Inter-rater analysis indicated reasonable annotation consistency, although the classes remain visual proxies rather than direct measurements of income or expenditure. The findings are limited to the frozen validation data from Banten and DKI Jakarta and should not be interpreted as evidence of geographic generalization or statistically significant model superiority. Future research

 ISSN:2302-8149
e-ISSN:2540-9719

should evaluate independent geographic regions, validate the visual classes against external socioeconomic data, and conduct broader architecture-specific hyperparameter testing.

Acknowledgement

The author would like to express sincere gratitude to the academic supervisors and reviewers for their guidance, feedback, and constructive suggestions throughout the development of this research. The author also acknowledges PT. Asia Riset Institusi for providing support during the preparation of the dataset and research process. Special appreciation is given to the GIS expert team for their assistance in residential area interpretation, polygon annotation, and visual socioeconomic classification. Their contribution was important in preparing the annotated dataset used in this study.

References

[1] A. Arshad, J. Zulfiqar, M. H. Zaib, A. Khan, and M. J. Khan, "*Mapping Socioeconomic Conditions using Satellite Imagery: A Computer Vision Approach for Developing Countries*," *Journal of Economy and Technology*, Vol. 1, pp. 144–163, Nov. 2023, DOI: 10.1016/j.ject.2023.11.001.

[2] F. S. K. Agyemang, R. Memon, L. J. Wolf, and S. Fox, "*High-Resolution Rural Poverty Mapping in Pakistan with Ensemble Deep Learning*," *PLoS One*, Vol. 18, No. 4 April, Apr. 2023, DOI: 10.1371/journal.pone.0283938.

[3] O. Hall, F. Dompae, I. Wahab, and F. M. Dzanku, "*A Review of Machine Learning and Satellite Imagery for Poverty Prediction: Implications for Development Research and Applications*," Oct. 01, 2023, *John Wiley and Sons Ltd*. DOI: 10.1002/jid.3751.

[4] F. Biljecki and K. Ito, "*Street view Imagery in Urban Analytics and GIS: A Review*," Nov. 01, 2021, *Elsevier B.V.* DOI: 10.1016/j.landurbplan.2021.104217.

[5] A. Singleton, D. Arribas-Bel, J. Murray, and M. Fleischmann, "*Estimating Generalized Measures of Local Neighbourhood Context from Multispectral Satellite Images using a Convolutional Neural Network*," *Comput. Environ. Urban Syst.*, Vol. 95, Jul. 2022, DOI: 10.1016/j.compenvurbsys.2022.101802.

[6] E. Suel, S. Bhatt, M. Brauer, S. Flaxman, and M. Ezzati, "*Multimodal Deep Learning from Satellite and Street-Level Imagery for Measuring Income, Overcrowding, and Environmental Deprivation in Urban Areas*," *Remote Sens. Environ.*, Vol. 257, May 2021, DOI: 10.1016/j.rse.2021.112339.

[7] H. Sarmadi, I. Wahab, O. Hall, T. Rögnvaldsson, and M. Ohlsson, "*Human Bias and CNNs' Superior Insights in Satellite based Poverty Mapping*," *SCI. Rep.*, Vol. 14, No. 1, Dec. 2024, DOI: 10.1038/s41598-024-74150-9.

[8] O. L. F. de Carvalho *et al.*, "*Instance Segmentation for Large, Multi-Channel Remote Sensing Imagery using Mask-RCNN and a Mosaicking Approach*," *Remote Sens. (Basel).*, Vol. 13, No. 1, pp. 1–24, Jan. 2021, DOI: 10.3390/rs13010039.

[9] T. Hou and J. Li, "*Application of Mask R-CNN for Building Detection in UAV Remote Sensing Images*," *Heliyon*, Vol. 10, No. 19, Oct. 2024, DOI: 10.1016/j.heliyon.2024.e38141.

[10] M. Amo-Boateng, N. Ekow Nkwa Sey, A. Ampah Amproche, and M. Kyereh Domfeh, "*Instance Segmentation Scheme for Roofs in Rural Areas based on Mask R-CNN*," *Egyptian Journal of Remote Sensing and Space Science*, Vol. 25, No. 2, pp. 569–577, Aug. 2022, DOI: 10.1016/j.ejrs.2022.03.017.

[11] S. Gui, S. Song, R. Qin, and Y. Tang, "*Remote Sensing Object Detection in the Deep Learning Era—A Review*," Jan. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. DOI: 10.3390/rs16020327.

[12] J. Butler and H. Leung, "*A Heatmap-Supplemented R-CNN Trained using an Inflated IoU for Small Object Detection*," *Remote Sens. (Basel).*, Vol. 16, No. 21, Nov. 2024, DOI: 10.3390/rs16214065.

[13] S. Chen, Y. Ogawa, C. Zhao, and Y. Sekimoto, "*Large-Scale Individual Building Extraction from Open-Source Satellite Imagery via Super-Resolution-based Instance Segmentation Approach*," *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 195, pp. 129–152, Jan. 2023, DOI: 10.1016/j.isprsjprs.2022.11.006.

[14] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "*Mask R-CNN*," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., Dec. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

[15] Z. Cai and N. Vasconcelos, "*Cascade R-CNN: Delving into High Quality Object Detection*," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. DOI: 10.1109/CVPR.2018.00644.

[16] H. Y. Oh, M. S. Khan, S. B. Jeon, and M. H. Jeong, "*Automated Detection of Greenhouse Structures using Cascade Mask R-CNN*," *Applied Sciences (Switzerland)*, Vol. 12, No. 11, Jun. 2022, DOI: 10.3390/app12115553.

[17] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "*SOLOv2: Dynamic and Fast Instance Segmentation*," in *Advances in Neural Information Processing Systems*, 2020.

[18] Z. Qiu, X. Huang, Z. Sun, S. Li, and J. Wang, "*GS-YOLO-Seg: A Lightweight Instance Segmentation Method for Low-Grade Graphite Ore Sorting based on Improved YOLO11-Seg*," *Sustainability (Switzerland)*, Vol. 17, No. 12, Jun. 2025, DOI: 10.3390/su17125663.

[19] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "*Masked-Attention Mask Transformer for Universal Image Segmentation*," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. DOI: 10.1109/CVPR52688.2022.00135.

[20] S. Mukodimah and C. Fauzi, "*Comparison of Tree Implementation, Regression Logistics, and Random Forest to Detect Iris Types*," *Technology Acceptance Model*, Vol. 12, No. 2, Oct. 2021.

[21] R. Bakeman, "*KappaAcc: A Program for Assessing the Adequacy of Kappa*," *Behav. Res. Methods*, Vol. 55, No. 2, 2023, DOI: 10.3758/s13428-022-01836-1.

[22] T.-Y. Lin *et al.*, "*Microsoft COCO: Common Objects in Context*," in *ECCV*, European Conference on Computer Vision, Sep. 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/

[23] J. Cohen, "*Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit*," *Psychol. Bull.*, Vol. 70, No. 4, pp. 213–220, 1968, DOI: 10.1037/h0026256.