Indonesian Street Food Calorie Estimation using Mask R-CNN and Multiple Linear Regression

Nadya Aditama
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
nadya.aditama@gmail.com

Rinaldi Munir
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
rinaldi@informatika.org

2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T) | 978-1-6654-5858-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPC2T53885.2022.9776804

***Abstract*— Indonesian people need to know about the calorie information of the street food. One of effective ways to get that is using image-based calorie estimation technologies. However, there are two limitations. First, the problem of occluded food and second is the low R Squared value in measurement using linear regression model with area feature. This research proposed the Mask R-CNN model for amodal instance segmentation task to get the complete object shape and multiple linear regression model with area, perimeter, length, and width to predict the food weight. This research proposed Indonesian street food dataset that has six classes. There are 1646 images of the dataset and total instance of each food are 644 bakwan, 812 bolu, 918 cireng, 679 serabi, 711 tahu, and 766 tempe. The number of data point in multiple linear regression model is 230 bakwan, 200 bolu, 250 cireng, 240 serabi, 230 tahu, and 230 tempe. The proposed multiple linear regression model has the highest R Squared score in all classes with the average R Squared 0.80425. Mask R-CNN ResNeXt-101-FPN in amodal instance segmentation task reaches the best F1 Score. In occluded scenario this model gets F1 Score 0.821 in IoU threshold 0.85. In non-occluded scenario the model gets F1 Score 0.994 in IoU threshold 0.9. Even though the F1 Score is high, there are some false detections and the bad segmentation quality. In calorie prediction, the proposed model is not reducing MAE score in some classes due to the segmentation quality and food characteristic.**

Keywords—Mask R-CNN, Amodal Instance Segmentation, Multiple Linear Regression, Food Calorie Estimation, Indonesian Street Food.

I. INTRODUCTION

According to World Health Organization (WHO), obesity is the serious problem. It can increase the risk of heart disease and stroke that caused death [1]. Obesity can be controlled by knowing calorie information in food so the people can control what they eat. There is various street food in Indonesia, but some people do not know about the calorie information of the food. One of effective ways to get the information is using image-based calorie estimation technologies.

There is various image-based calorie estimation approach today. Research by [2] use Mask R-CNN, the instance segmentation method that has been developed by [3] to get the food shape on the top view and calculate the food volume, convert it to weight, and estimate the calories by calorie conversion from food weight. But this research only predicts the rectangular food shape. Research by [4], also use Mask R-CNN to get the food shape, but this research used linear regression to predict the food weight before converting it to calorie units. This method is more flexible because it can predict various food shape. But there are two limitations in [4]. First, there is a problem in occluded food. If the same object is occluding or too close, it will be counted as one object. Incomplete shape is one of the problems to get the full area features. Second, according to [5] in predicting sheep weight, linear regression model using area feature has lower R Squared value than multiple linear regression model using length and width feature which means multiple linear regression is better than linear regression with only one variable.

This research proposes a model which can detect calories from occluded food using Mask R-CNN for amodal instance segmentation task to get the complete shape of the object and multiple linear regression to predict the food weight in gram units and convert it to calorie units. Mask R-CNN for amodal instance segmentation has been evaluated by [6]. This research also proposed an Indonesian street food image dataset include the weight and calorie information.

The remainder of this paper is organized as follows. Section II explains the related works. Section III gives the explanation about the system flow and the details of the method. Section IV present the experimental result and some explanation about the dataset. At the end, the conclusion and future works is presented in Section V.

II. RELATED WORKS

Food recognition for Indonesian Food has been implement in research [7], [8], and [9]. Research in [7] proposed random forest classifier with combination of color and texture feature for the recognition model. The result reaches 93.5% of accuracy [7]. Research by [8] and [9] applied Convolutional Neural Network (CNN) architecture for the food recognition system. Both of research reaches the accuracy over 90% in predicting the Indonesian food. But those research is limited to the food class prediction, not the food calories.

There is various image-based food calorie estimation approach. Research by [10], proposed the multi-task CNN method to recognize the food class and estimate calorie estimation at the same time with two branch fully connected layer. Although this method can reduce the error measurement compared to single-task CNN, this research does not consider the food size for the prediction. The same authors [11], proposed the method to estimate food calories in multiple food dish. In [11], the food dish is detected by Faster R-CNN model and the calorie is predicted by regression-based CNN. Same as the previous research, this research does not consider the food size.

Research by [2], proposed the calorie estimation based on calculation of food volume and convert it to food weight before converting it to calorie units. This research focused on food rectangular shape. Research by [4] proposed the calorie estimation based on the number of pixels in the contour represent the segmentation area and predict the food weight by linear regression model.

Linear regression model for predicting the object weight has been done by [5]. Research by [5] proposed the method to

978-1-6654-5858-0/22/$31.00 ©2022 IEEE

predict the weight of the sheep using multiple linear regression with length and width feature. The result has been achieved accuracy measurement by 98.75% with R Squared of 0.993 [5]. This research also compared the model with linear regression with area feature. According to [5], linear regression with area feature has lower R Squared value than multiple linear regression with length and width feature.

Amodal instance segmentation is the segmentation task that segment the occluded part. Traditional instance segmentation is only concerned with visible part of each instance [6]. The idea of amodal instance segmentation task is the person ability to predict the shape of the object even if the object is partially occluded. Research by [6] has been evaluated some instance segmentation model to do amodal instance segmentation task using KINS dataset, KITTI dataset that has been annotated amodally. Mask R-CNN can do this task with reasonable result, even though PANet has better result than Mask R-CNN based on mAP score. But this research does not mention what backbone model is used in the Mask R-CNN model. This research will observe the backbone model in Mask R-CNN.

III. Methods

The authors will develop an image-based food calorie estimation using two main methods based on Fig. 1.

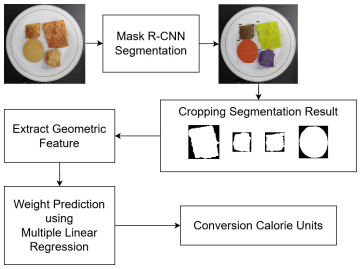


Fig. 1. Flow diagram of the proposed method.

First, Mask R-CNN will be trained using amodal annotations dataset. In amodal annotation, if the object is partially covered with the other object, the annotation will follow the real shape of the object. Fig. 2 shows the example of amodal annotation.

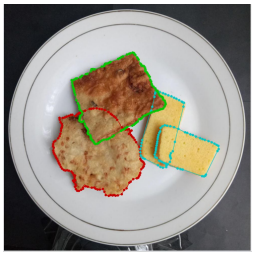


Fig. 2. The example of amodal annotation

Mask R-CNN model will segment the food object, so the model can get the shape of the object. After that, the system will extract geometry features. The authors propose four geometric features such as area, perimeter, length, and width. These features will be used to predict food weight using multiple linear regression. The food weight will be converted into calorie units.

A. Mask R-CNN

Mask R-CNN is one of the instance segmentation methods that has been developed by Facebook AI Research (FAIR). Mask R-CNN is the extension of Faster R-CNN by adding the branch for predicting segmentation masks on each Region of Interest (RoI), in parallel with existing branch for classification and bounding box regression [3]. Fig. 3 shows the architecture of Mask R-CNN.

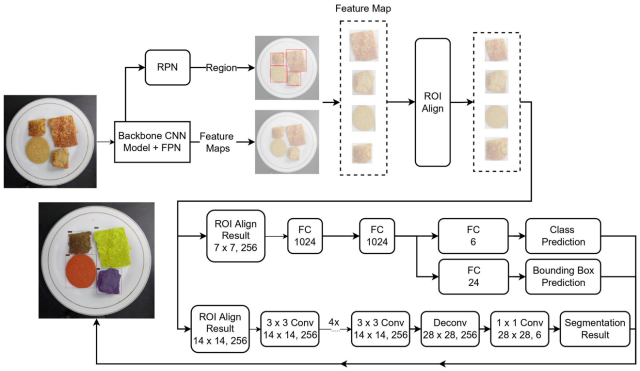


Fig. 3. Mask R-CNN architecture.

As shown in Fig. 3, four important parts in Mask R-CNN model is backbone model, Region Proposal Network (RPN), RoIAlign, and head architecture. Backbone model is the CNN Architecture to extract the feature map. Research by [3] evaluated ResNet and ResNeXt architecture as the backbone model of Mask R-CNN and added Feature Pyramid Network (FPN) architecture in the backbone model to get the better feature map in different scale. Region Proposal Network (RPN) proposes candidate object bounding boxes [3]. Those candidate object size will be adjusted by RoIAlign [4]. In head architecture, there are three architecture branches to predict the bounding box, object class, and segmentation result. Mask R-CNN use Fully Convolutional Network (FCN) to segment the object.

Mask R-CNN can do amodal instance segmentation task with the reasonable result. This has been evaluated by [6] in KINS dataset. But [6] do not mention which backbone model is better in this model. This study will observe the backbone pretrained model of Mask R-CNN in amodal instance segmentation task.

B. Multiple Linear Regression

If there are more than one independent variable in regression model it will be called Multiple Linear Regression. As mentioned before, the authors proposed four geometry feature such as area, perimeter, length, and width. Equation (1) shows the function of the proposed model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \qquad (1)$$

Where y is the food weight, x_1 is the area, x_2 is the perimeter, x_3 is the length, x_4 is the width, ε is the random error terms and $\beta_0, \beta_1, \ldots, \beta_4$ is the parameter of the regression model that will be estimated from the data. Length and width are calculated by generating the rotated bounding box and calculating the distance between the midpoint of the length and width bounding box. Perimeter is calculated by adding the connected pixel distance in the contour. Area is the number of pixels inside the contour.

IV. Experimental Result

A. Dataset

In this study, there are six foods in the proposed dataset such as bakwan, bolu, cireng, serabi, tahu, tempe. The image has been captured using OPPO F1s smartphone with 13 MP camera. The image is also captured from the top-view. The distance of the camera is fixed, ±30 cm. The resolution of the picture is 3120 x 4160 pixel. The image has been cropped and reduce to 1000 x 1000 pixel to reduce the computational resources for training process. The weight of each food is measured using digital scale 0.1 gram. Table I shows the food calorie information in each class that has been retrieved from fatsecret.co.id.

TABLE I. Calorie Information of Food

ID	Food	Calories per 100 grams (kcal)	Shape
0	Bakwan	228	Asymmetric
1	Bolu	297	Rectangular
2	Cireng	348	Asymmetric
3	Serabi	216	Circle-like
4	Tahu	271	Rectangular
5	Tempe	241	Rectangular

Indonesian food has various shapes. Fig. 4 shows the food image example of each class.

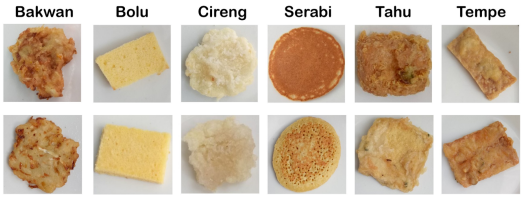


Fig. 4. Food image example of each class.

In this study, the dataset has been divided into training and validation for the segmentation model. This partition is used to get the best backbone model. Table II shows the data partition of segmentation model.

TABLE II. Data Partition for Segmentation Model

Data Partition	Position	Number of Images
Train	Occluded	652
	Non-Occluded	616
Val	Occluded	81
	Non-Occluded	77

In occluded images, not all the food in the plate is occluded. The occluded image has at least 1 object that partially occluded with cutlery and other food. In this dataset, the occluded level is estimated between 10% until 50% occluded, even though there are some images that partially occluded above 50%. The testing data will be used to test the final model, the combination of detection and estimation model. In testing data in occluded position, four unseen food items in each class are added to the test partition to see the ability of model to segment and predict the unseen food. Table III shows the amount of image in testing data.

TABLE III. Data Partition for Final Model

Data Partition	Position	Number of Images
Test	Occluded	145
	Non-Occluded	75

Table IV shows the total number of instances in each class and each partition.

TABLE IV. Total Number of Instance in Each Class

Food	Train	Val	Test	Total
Bakwan	467	59	118	644
Bolu	623	71	118	812
Cireng	706	98	114	918
Serabi	514	65	100	679
Tahu	544	65	102	711
Tempe	600	65	95	766

In estimation model, each item will be captured by 10 times in different position because if the position different, the amount of pixel will be different. Table V shows the data partition for the estimation model.

TABLE V. Data Partition for Estimation Model

Food Class	Data Partition	Total Item	Total Image
Bakwan	Train	23	230
	Test	5	50
Bolu	Train	20	200
	Test	5	50
Cireng	Train	25	250
	Test	5	50
Serabi	Train	24	240
	Test	5	50
Tahu	Train	23	230
	Test	5	50
Tempe	Train	23	230
	Test	5	50
Total		168	1680

B. Segmentation Model

The authors observed the best Mask R-CNN pretrained model from COCO dataset for amodal instance segmentation task based on the mean Average Precision (mAP) segmentation score. This research not only observed the Mask R-CNN FPN backbone model based on Fig. 3, but also observed Mask R-CNN ResNet-C4. ResNet-C4 has the different head structure that has been explained in [3]. To train the Mask R-CNN model, Stochastic Gradient Descent (SGD) will be used as the optimizer, with base learning rate 0.0001 and momentum 0.9. The number of iterations for the training process is 10000.

In this research, the training time of each backbone model will be observed as well. Table VI shows the result of the observation.

TABLE VI. MASK R-CNN PRETRAINED BACKBONE PERFORMANCE

Backbone Model	Training Time	mAP Train	mAP Validation
ResNet-50-C4	02:38:43	89.53 %	88.30 %
ResNet-101-C4	03:33:45	90.19 %	88.84 %
ResNet-50-FPN	02:19:38	93.49 %	90.31 %
ResNet-101-FPN	03:28:23	93.95 %	91.74%
ResNeXt-101-FPN	07:05:31	94.53%	91.47%

According to the result, Mask R-CNN with ResNeXt-101-FPN get the highest mAP Training score. But ResNeXt-101-FPN has the longest training time between five backbone models. The mAP validation between ResNet-101-FPN and ResNeXt-101-FPN has a little difference. The difference is only 0.27%. The authors will observe those two backbones model in final model.

C. Estimation Model

The authors observed three kinds of geometric feature. First, the area feature as done by [4]. Second, the length and width as done by [5]. Third, the proposed four feature. In estimation model, the regression model will predict the weight of food in gram units. Table VII shows R Squared score in each class.

TABLE VII. R SQUARED SCORE IN EACH CLASS

Feature	Class	R Squared (Weight)	Average R Squared
Area	Bakwan	0.6581	0.785216667
	Bolu	0.9524	
	Cireng	0.7464	
	Serabi	0.9588	
	Tahu	0.5537	
	Tempe	0.8419	
Length and Width	Bakwan	0.6687	0.79535
	Bolu	0.9470	
	Cireng	0.7857	
	Serabi	0.9404	
	Tahu	0.5783	
	Tempe	0.8396	
Length, Width, Perimeter, Area	Bakwan	0.6811	**0.80425**
	Bolu	0.9573	
	Cireng	0.7886	
	Serabi	0.9621	
	Tahu	0.5880	
	Tempe	0.8484	

According to the result, all classes get the highest R Squared value in proposed feature. That means, the data is more fit in multiple linear regression model. Even though the proposed feature has the highest R-Squared value, there is a difference at the measurement result based on the ground truth. The performance of estimation is measured with metric evaluation Mean Absolute Error (MAE). Table VIII shows the measurement result.

TABLE VIII. MAE SCORE IN ESTIMATION MODEL

Feature	Class	MAE (Weight)	Average MAE
Area	Bakwan	6.3240	5.291933
	Bolu	**3.0605**	
	Cireng	7.8405	
	Serabi	**4.2299**	
	Tahu	**4.9583**	
	Tempe	5.3384	
Length and Width	Bakwan	6.2522	5.31165
	Bolu	4.3188	
	Cireng	7.4662	
	Serabi	4.3082	
	Tahu	5.4221	
	Tempe	**4.1024**	
Length, Width, Perimeter, Area	Bakwan	**6.2035**	**5.254567**
	Bolu	3.1728	
	Cireng	**7.2903**	
	Serabi	4.2586	
	Tahu	5.7266	
	Tempe	4.8756	

As shown in Table VIII, multiple linear regression can reduce weight measurement error in three classes, tempe, bakwan, and cireng. But multiple linear regression model with length and width feature can reduce the error more than the proposed feature in tempe class because of the rectangular shape of tempe.

D. Final Model

The final model merges the trained Mask R-CNN model to segment the food object and the linear regression model get the food weight and convert the weight to food calories. There are two scenarios in this research. First, is the occluded scenario. The occluded scenario has at least one object that is occluded by other food or the cutlery. Second is the non-occluded scenario. In non-occluded scenario, there is no occlusion in the images.

In this research, the authors observed backbone model ResNet-101-FPN and ResNeXt-101-FPN. Also, the authors observed the Intersection of Union (IoU) threshold between segmentation result and ground truth. The authors observed IoU threshold 0.9 and 0.85. This value is chosen because this system needs the good segmentation result to predict the food calorie. For the result, average precision, recall, and F1 Score from all classes has been calculated. Table IX shows the F1 Score of the final model segmentation.

TABLE IX. F1 SCORE OF SEGMENTATION RESULT IN FINAL MODEL

IoU	Scenario	Backbone Model	Precision	Recall	F1 Score
0.9	Occluded	ResNet-101-FPN	0.697	0.718	0.706
		ResNeXt-101-FPN	0.716	0.722	0.719
	Non-Occluded	ResNet-101-FPN	0.967	0.971	0.966
		ResNeXt-101-FPN	0.992	0.996	0.994
0.85	Occluded	ResNet-101-FPN	0.799	0.823	0.81
		ResNeXt-101-FPN	0.818	0.824	**0.821**
	Non-Occluded	ResNet-101-FPN	0.967	0.971	0.966
		ResNeXt-101-FPN	0.992	0.996	**0.994**

As shown in Table IX, the best F1 Score in occluded scenario is ResNeXt-101-FPN in IoU threshold 0.85. That means this model can segmented occluded object better in IoU threshold 0.85. In non-occluded scenario, the the model can segment non occluded object better in IoU threshold 0.9. Even the IoU threshold is reduced to 0.85, the detection result is still the same.

The F1 Score between ResNet-101-FPN and ResNeXt-101-FPN has a little difference in all scenarios. For the calorie estimation, the ResNeXt-101-FPN is chosen based on the F1 Score value. The segmentation result is showed in Fig. 5.

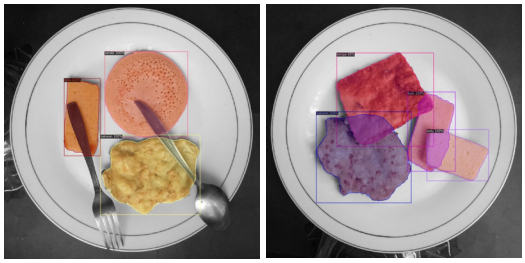
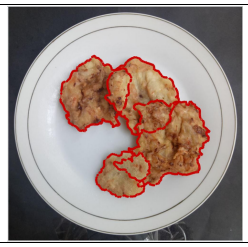
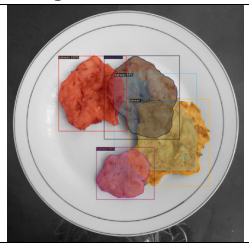


Fig. 5. The segmentation result of occluded object

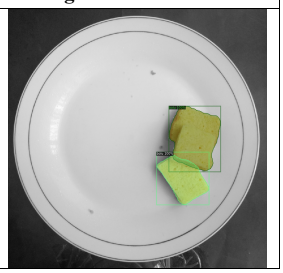
Some images achieve the good result of the segmentation. Even though ResNeXt-101-FPN has better F1 Score than ResNet-101-FPN, but there are some false segmentations in the model. First, there are some false positive segmentations between the occluded objects. Table X showed the false positive segmentation failure in some images.

TABLE X. FALSE POSITIVE SEGMENTATION EXAMPLE

Ground Truth	Segmentation Result
	

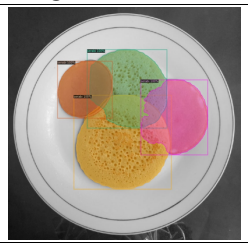
Second, in some images, if there are two or more food same class occluded it will be segmented as one object. Table XI shows the second segmentation failure.

TABLE XI. SEGMENTATION FAILURE IN CLOSED OBJECT

Ground Truth	Segmentation Result
	

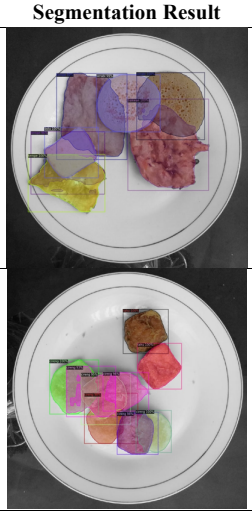
The third is the segmentation quality. Some image cannot produce the good segmentation quality in occluded object. Table XII showed the example of bad segmentation.

TABLE XII. THE BAD SEGMENTATION QUALITY IN SOME IMAGES

Ground Truth	Segmentation Result
	

When predicting the unseen object items, the segmentation result is good, even though there are some false positive detections between the occluded objects. That means, the models can generalize the new data, even though there are still some problems at false detection and quality segmentation. Table XIII shows the segmentation result.

TABLE XIII. THE SEGMENTATION RESULT IN UNSEEN OBJECT ITEMS

Ground Truth	Segmentation Result
	

In calorie estimation, the system only predicts the true positive segmentation result in IoU threshold 0.85 for the occluded scenario and IoU threshold 0.9 for the non-occluded scenario. Same as estimation model, the MAE score is used for model evaluation. Fig. 6 shows the estimation result.

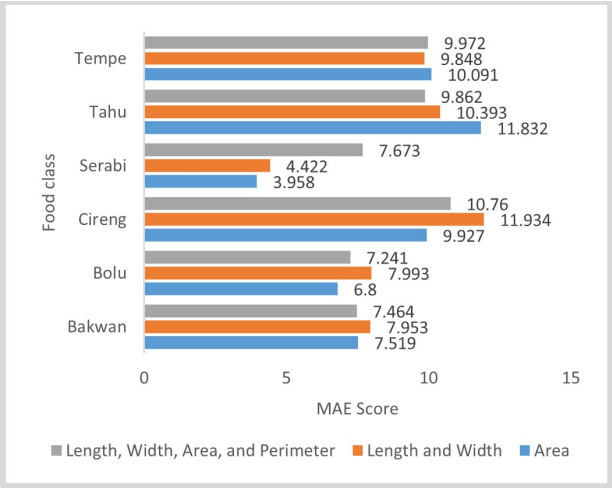


Fig. 6. MAE result in occluded object.

As shown in Fig. 6, the proposed model has been reducing MAE score in class bakwan and tahu in occluded scenario. Class cireng, bolu, and serabi has the lowest MAE score in linear regression with area feature. The result is caused by the quality of the segmentation so using the four features will not be good. In class tempe, the multiple linear regression with length and width feature has the lowest MAE score between all the regression models. Fig. 7 shows the MAE result in non-occluded object.

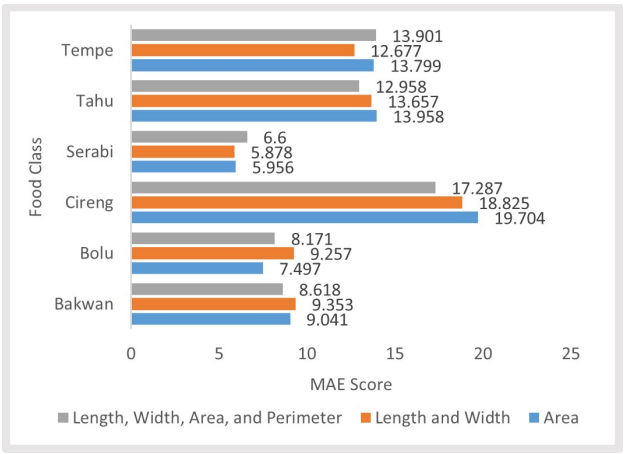


Fig. 7. MAE result in non-occluded object

In Fig. 7, the proposed feature has been reducing MAE score in class bakwan, tahu, and cireng in this scenario. In class tempe and serabi, the multiple linear regression with length and width feature has the lowest MAE score between all the regression models. This can be caused by the food symmetrical shape. Tempe has the rectangular shape and serabi has circle-like shape. Even though bolu and tahu has the rectangular shape, the MAE score is better in the proposed feature for tahu and MAE score is better in area feature for bolu.

There are other factors about why the proposed feature does not reduce MAE score in some class besides the quality of the segmentation. This can be due to the various food characteristic such as the density of the food, and how the food cooked. Even though the thickness of the food has been uniformed, the little difference between all the food thickness can affect the measurement.

V. Conclusion and Future Works

The proposed multiple linear regression model has the highest R Squared score in all class with average R Squared 0.80425. But when do the prediction in estimation model with ground truth, the proposed model is not reducing the MAE score in some class. In final model, some class get the good result in linear regression area in occluded scenario, which caused by the segmentation quality. For the non-occluded scenario, tempe and serabi get the good result in multiple linear regression with length and width and bakwan, tahu, and cireng get the good result in multiple linear regression with purposed model. That means more feature in the model is not necessarily reduce the measurement error.

Mask R-CNN ResNeXt-101-FPN has the best F1 Score in final model for two scenarios. For the occluded scenario the model reaches F1 Score 0.821 in IoU threshold 0.85. For the non-occluded scenario, the model reaches F1 Score 0.994 in IoU threshold 0.9. Even though the F1 Score is high, there are some false detections between the occluded object and some images produce not good segmentation quality.

Overall, amodal instance segmentation task is good enough to predict the occluded object calories. In the future, it is necessary to use another method to produce the good segmentation quality in amodal instance segmentation task and increase the variation of the data to get the better result.

References

[1] "Obesity and overweight," *World Health Organization*, 09-Jun-2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. [Accessed: 18-Mar-2021].

[2] R. D. Yogaswara, E. M. Yuniarno, and A. D. Wibawa, "Instance-Aware Semantic Segmentation for Food Calorie Estimation using Mask R-CNN," *Proc. - 2019 Int. Semin. Intell. Technol. Its Appl. ISITIA 2019*, pp. 416–421, 2019.

[3] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980-2988, 2017.

[4] M. L. Chiang, C. A. Wu, J. K. Feng, C. Y. Fang, and S. W. Chen, "Food Calorie and Nutrition Analysis System based on Mask R-CNN," *2019 IEEE 5th Int. Conf. Comput. Commun. ICCC 2019*, pp. 1721–1728, 2019.

[5] A. S. Abdelhady, A. E. Hassanien, Y. M. Awad, M. El-Gayar, and A. Fahmy, "Automatic Sheep Weight Estimation Based on K-Means Clustering and Multiple Linear Regression," *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018*, pp. 546-555, 2019.

[6] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia, "Amodal Instance Segmentation with KINS Dataset," *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 3009–3018, 2019.

[7] Y. A. Sari *et al.*, "Indonesian Traditional Food Image Identification using Random Forest Classifier based on Color and Texture Features," *2019 Int. Conf. Sustain. Inf. Eng. Technol.*, pp. 206–211, 2019.

[8] A. Wibisono, H. A. Wisesa, Z. P. Rahmadhani, P. K. Fahira, P. Mursanto, and W. Jatmiko, "Traditional food knowledge of Indonesia: a new high-quality food dataset and automatic recognition system," *J. Big Data*, vol. 7, no. 1, 2020.

[9] S. Giovany, A. Putra, A. S. Hariawan, L. A. Wulandhari, and E. Irwansyah, "Indonesian Food Image Recognition Using Convolutional Neural Network," in *Artificial Intelligence Methods in Intelligent Algorithms*, 2019, pp. 208–217.

[10] T. Ege and K. Yanai, "Simultaneous estimation of food categories and calories with multi-task CNN," *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. MVA 2017*, pp. 198–201, 2017.

[11] T. Ege and K. Yanai, "Estimating food calories for multiple-dish food photos," *Proc. - 4th Asian Conf. Pattern Recogniton, ACPR 2017*, pp. 646–651, 2017.