Mid-Upper Arm Circumference Measurement Using Digital Images: A Top-Down Approach with Panoptic Segmentation Using Mask R-CNN

Maya Silvi Lydia[1]*, Pauzi Ibrahim Nainggolan[2], Desilia Selvida[3],
Doli Aulia Hamdalah[4], Dhani Syahputra Bukit[5], Amalia[6], Rahmita Wirza Binti O. K. Rahmat[7]
Department of Computer Science, Universitas Sumatera Utara, Medan, 20155, Indonesia[1, 4, 6]
Computer Vision and Multimedia Laboratory, Universitas Sumatera Utara, Medan, 20155, Indonesia[2, 3]
Department of Public Health, Universitas Sumatera Utara, Medan, 20155, Indonesia[5]
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang 43400, Malaysia[7]

***Abstract*—Assessing nutritional status, particularly among children and pregnant women, necessitates accurate measurement of Mid-Upper Arm Circumference (MUAC). This research introduces a novel system for MUAC estimation from digital images using the Mask R-CNN algorithm, employing a top-down panoptic segmentation strategy. The proposed model was designed to identify the upper arm region within human body images and compute MUAC values autonomously. Mask R-CNN was selected due to its capacity to perform precise segmentation of objects within visually complex scenes, especially in the mid-upper arm area. Model training was conducted using a dataset of annotated images, with subsequent evaluation confirming its ability to reliably detect and measure MUAC. The system was validated using 72 image samples, yielding a mean absolute error (MAE) of 2.31 cm when compared to manual measurements. Among these samples, 29.2% (21 individuals) exhibited a measurement discrepancy of 0 to 1 cm, 27.8% (20 individuals) showed a 1 to 2 cm difference, and 43.1% (31 individuals) demonstrated deviations exceeding 2 cm. Despite some variations in measurement accuracy, the system presents a promising tool for enhancing the automation and efficiency of nutritional assessments.**

***Keywords*—Mask R-CNN; mid-upper arm circumference; digital images; segmentation; mean absolute error**

I. Introduction

Recent advancements in artificial intelligence and image processing have significantly impacted the medical and healthcare sectors. A prominent area of application is anthropometry, where these technologies have improved the accuracy and efficiency of assessing nutritional and health status through body measurements [1]. Amongst the various anthropometric indicators, Mid-Upper Arm Circumference (MUAC) has gained recognition as a critical metric for evaluating individual health and physical condition [2] [3] [4].

MUAC is particularly valued for its effectiveness in detecting nutritional deficiencies, especially in vulnerable populations such as pregnant women [5]. As a simple, non-invasive, and rapid screening tool, MUAC is instrumental in identifying individuals at risk of Chronic Energy Deficiency (CED). CED during pregnancy poses a major public health concern, as it is linked to elevated risks of infection, anemia, hemorrhage, and adverse neonatal outcomes, including low birth weight (LBW) [6]. MUAC measurement below the widely accepted threshold of 23.5 cm is generally considered indicative of increased CED risk, necessitating further clinical evaluation and targeted nutritional support [7] [8]. Owing to its low cost, ease of use, and minimal equipment requirements, MUAC is particularly suitable for deployment in limited resource settings [9].

Traditionally, MUAC is measured by wrapping a flexible measuring tape around the midpoint of the upper arm. This process involves identifying anatomical landmarks, specifically the shoulder and the elbow—and determining the midpoint between them. The tape is then positioned horizontally and applied with sufficient tension to avoid measurement errors due to slackness or excessive compression [7]. Despite its practicality, this manual method is susceptible to variability, primarily due to inconsistencies in tape dimensions and user technique, which may compromise the reliability and accuracy of the measurements [10].

The advancement of technology, digital approaches are emerging as potentially more accurate, efficient, and scalable solutions [11][12][13][14][15]. Liu, Sowmya, and Khamis [16] showed that the photogrammetric method can produce precise MUAC measurements, with a Technical Error of Measurement (TEM) value of 3.76 mm and its reliability coefficient (R) reaching 0.99. However, this research has yet to tap the potential of the Mask R-CNN method, which has proven to be highly effective and can be generalized to various tasks, including complex analyses in medical imaging [17][18][19] and human pose estimation [20], making it highly suitable for the detailed segmentation required for accurate MUAC calculation from digital images.

Top-Down Panoptic Segmentation approach utilizing Mask R-CNN is among the pertinent innovations in developing digital image-based MUAC measurement systems. In [20], Mask R-CNN framework represents a robust methodology capable of concurrently performing object detection and generating precise segmentation masks for each detected object instance.

Panoptic segmentation as introduced in [21]. It supports comprehensive scene interpretation by unifying instance and

*Corresponding Author.

semantic segmentation in single framework. This technique assigns each pixel to either distinct objects ("things", such as the upper arm) or homogeneous background regions ("stuff"), thereby generating semantically rich image representations, a critical requirement for anthropometric analysis based on digital images.

The Top-Down approach is particularly advantageous for digital anthropometry, especially when implemented using models such as Mask R-CNN. This method operates in a hierarchical manner, beginning with the detection of regions of interest such as the upper arm through bounding boxes, followed by fine-grained segmentation within these localized areas. Mask R-CNN exemplifies this approach by coupling coarse object detection with subsequent mask refinement. Central to this process is the Region Proposal Network (RPN), which identifies candidate object regions [22] prior to the mask generation stage. This sequential pipeline enhances computational efficiency by concentrating processing resources on the most relevant portions of the image.

Integrating panoptic segmentation with a Top-Down architecture, the accuracy of MUAC measurement is substantially improved. This combination streamlines image analysis workflows, making it a promising innovation for applications in anthropometry and broader health-related fields.

Despite recent progress in digital anthropometry, existing image-based MUAC methods still leave three practical gaps. First, most pipelines do not employ a top-down instance/panoptic segmentation strategy (e.g., Mask R-CNN with RPN/RoIAlign/FPN) to robustly isolate the upper-arm region under real-world variation in pose, clothing, and background, limiting measurement reliability. Second, pixel-to-centimeter scaling is commonly tied to specialized hardware or fixed capture setups, offering no simple, low-cost calibration standardizable across sites in resource-constrained settings. Third, there is limited end-to-end validation against manual MUAC using interpretable metrics that jointly assess region localization (e.g., IoU) and circumference agreement (e.g., MAE in cm) on real subjects. In response, we: i) develop a top-down panoptic-segmentation pipeline based on Mask R-CNN to detect and segment the upper arm from complementary front and side views, ii) formalize an ellipse-based MUAC estimation that fuses both views, and iii) introduce a lightweight, field-feasible pixel-to-centimeter scaling procedure using a ubiquitous wall power outlet as reference. We then evaluate detection/segmentation and MUAC agreement against manual measurements on 72 individuals.

Accordingly, this study develops a top-down panoptic-segmentation pipeline based on Mask R-CNN (with RPN, RoIAlign, and FPN) to detect and segment the upper-arm region from two complementary views (front and side), and formalizes MUAC estimation using an ellipse-based circumference approximation that combines the front- and side-view measurements. To realize field feasibility, the pipeline includes a lightweight pixel-to-centimeter scaling procedure that uses a standardized wall power-plug as a ubiquitous reference object. The system is implemented in Detectron2 [28] and evaluated using Intersection over Union (IoU) [25] for detection/segmentation and mean absolute error (MAE, cm) for MUAC agreement with manual measurements, alongside distributional error analysis and training-configuration comparisons.

The significance of this work is twofold. Practically, it reduces dependence on operator skill and measurement hardware by shifting MUAC assessment to a camera-based, low-cost workflow that can be standardized across sites and is well-suited to limited-resource environments. Technically, it demonstrates how top-down panoptic segmentation can strengthen digital anthropometry by delivering reliable region isolation prior to metric estimation, laying the groundwork for near-real-time and mobile implementations that can support large-scale nutritional screening and follow-up [21].

To guide the reader, the remainder of this study is organized as follows: Section II presents the related works. Section III (Materials and Methods) details the dataset and annotation protocol, image-capture standards, reference-object scaling, the MUAC estimation formulation for front/side views (ellipse-based), the MAE metric, and the Mask R-CNN–based top-down panoptic segmentation pipeline. Section IV (Implementation and Results) first describes the dataset preparation and training environment, then details the system architecture. It also presents a comparative evaluation across training iterations and model variants (loss and accuracy), and reports the test results—detection IoU; MUAC–vs–manual agreement summarized by MAE (in cm) with its error distribution; and qualitative top-down panoptic-segmentation visualizations. Section V (Discussion) analyzes the findings, limitations, and avenues for improvement. Section VI (Conclusion) summarizes the contributions and key findings. Finally, acknowledgments are provided.

II. Related Work

Prior digital anthropometry solutions span: i) manual tape measurement, ii) monocular photogrammetry and regression, iii) 3-D imaging systems, and iv) learning-based detection/segmentation pipelines.

1) Manual tape. MUAC is commonly measured by wrapping a dedicated MUAC tape around the upper arm; however, design differences (e.g., material thickness and scale placement) introduce systematic measurement discrepancies, and there is no single global tape specification—prompting calls for common design standards and standardized reporting [10].

2) Single-camera photogrammetry/regression. Liu et al. showed that MUAC (and height) can be estimated from images using linear regression under controlled capture with high reliability (TEM$\approx$3.76 mm, R$\approx$0.99). These systems typically rely on strict imaging geometry and explicit calibration/landmarking, and do not perform instance-level region isolation of the upper arm [16].

3) 3-D imaging. Automated 3-D anthropometry has demonstrated precise body measurement extraction, including in low-resource settings, but requires dedicated hardware and controlled environments that can increase deployment cost and operational complexity [11], [29], [14].

4) Learning-based segmentation. Mask R-CNN and related top-down components (RPN, RoIAlign, FPN) [20], [24] deliver strong instance segmentation across diverse medical-imaging tasks, including organ and lesion analysis [17]–[19]. To our knowledge, integrating this paradigm into an end-to-end MUAC pipeline—with explicit centimeter scaling and paired-view fusion—remains underexplored.

Our study advances category (iv) by: 1) using a top-down Mask R-CNN pipeline to robustly isolate the upper-arm region from both front and side views, 2) fusing the views via an ellipse-based circumference formulation, and 3) introducing a lightweight, field-feasible pixel-to-centimeter calibration using a ubiquitous wall power outlet as the reference object. We evaluate both localization (IoU) and agreement with manual MUAC (MAE), on 72 subjects, thereby reporting interpretable, end-to-end accuracy rather than proxy metrics alone. Compared to (ii), our approach avoids hand-crafted landmarking and yields explainable, mask-level measurements; compared to (iii), it requires only commodity cameras and a common reference object, improving deployability in resource-constrained settings [16], [11], [29].

Across categories, open challenges remain: a) robust segmentation under clothing–background similarity and sub-optimal pose; b) calibration methods that are reference-free or self-calibrating; c) generalization beyond standardized capture; and d) near-real-time, mobile deployment. These gaps motivate methods that combine reliable instance-level isolation with simple, field-feasible calibration and interpretable circumference estimation.

III. Materials and Methods

This section establishes the methodological foundation of the proposed MUAC-from-image system. We first outline the Mask R-CNN–based top-down panoptic segmentation pipeline—covering the backbone-FPN architecture, RPN/ROIAlign heads, and instance masks—then formalize MUAC estimation from front/side views via an ellipse-based circumference formulation. We also define the MAE metric and detail the dataset acquisition and image standards, COCO-formatted annotations (person, hand, upper arm) prepared in Roboflow, and the pixel-to-centimeter scaling derived from a wall power-plug reference.

A. Mask R-CNN

In the realm of instance segmentation, Mask R-CNN introduces a substantial enhancement, enhancing the Faster R-CNN framework by integrating both object detection and detailed pixel-wise delineation of individual objects [20]. Its primary innovation lies in the addition of a separate, parallel pathway dedicated to producing binary segmentation masks for each proposed region, known as a Region of Interest (RoI). This structure allows the model to simultaneously perform object classification, precise localization, and segmentation at the instance level.

The model employs a two-stage processing pipeline, as illustrated in Fig. 1. Initially, the Region Proposal Network (RPN) [22] identifies candidate object regions within the input image, generating a set of preliminary bounding boxes. These proposals are subsequently refined in the second stage, where features corresponding to each RoI are extracted using the RoIAlign operation. This operation is critical for maintaining spatial alignment between the input feature map and the sampled features, which directly contributes to the precision of the segmentation output. Utilizing these aligned features, the network proceeds to classify each object, adjust the bounding box coordinates, and produce a binary segmentation mask delineating the object's spatial boundaries.

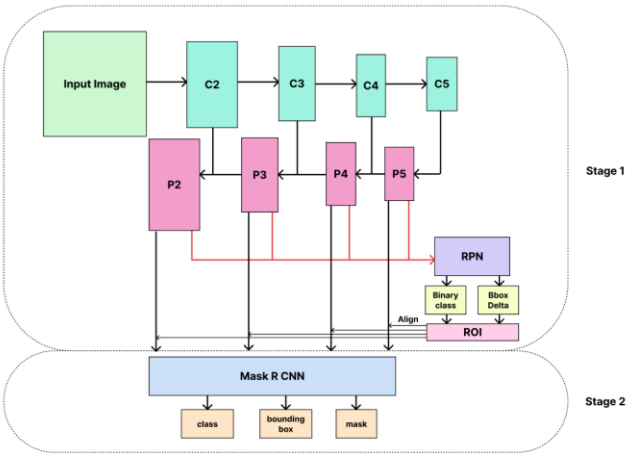

Fig. 1. Architecture of Mask R-CNN.

A key advancement in Mask R-CNN is the replacement of the RoIPool layer with RoIAlign. RoIPool's quantization step often introduces spatial misalignments that compromise mask accuracy. RoIAlign mitigates this issue by avoiding quantization, preserving the exact spatial information during feature extraction. This refinement substantially improves both detection precision and the quality of the generated masks.

Since its introduction, Mask R-CNN has consistently demonstrated cutting-edge performance across prominent visual recognition benchmarks, particularly the COCO (Common Objects in Context) dataset [23]. Its exceptional accuracy has contributed to its broad adoption in diverse computer vision applications, extending well beyond its original roles in instance segmentation and object detection to encompass more advanced tasks, including human pose estimation.

The operational workflow of the Mask R-CNN architecture is organized into two main phases. Initially, a deep convolutional neural network—typically a ResNet variant—serves as the backbone to derive hierarchical feature maps from the input image. The extracted features are then passed to the RPN, which identifies potential object-containing regions. In the subsequent phase, the RoIAlign mechanism is employed to extract accurately aligned features corresponding to each proposed region. These refined features are subsequently utilized to simultaneously carry out object categorization, refine bounding box coordinates, and generate segmentation masks at the pixel level for individual instances.

In first stage, backbone network typically ResNet produces multi-scale feature maps denoted as C2 through C5, corresponding to different levels of convolutional hierarchy. These maps capture progressively abstract representations, with

deeper layers encoding semantic information and shallower layers retaining fine-grained spatial details.

To address scale variation among objects, Mask R-CNN integrates a Feature Pyramid Network (FPN) [24]. FPN constructs a multi–level feature hierarchy (P2–P5) by combining high-level semantic features from deeper layers with spatially detailed information from earlier layers through a top–down pathway and lateral connections. This multi–scale representation enhances the model`s ability to accurately detect and segment objects of various sizes within the same image.

Once the feature map is formed, RPN generates proposals for areas potentially containing objects. Each area proposal is assessed based on two main outputs:

- Binary Class: Determines whether the area contains objects or not.
- Bounding Box Delta: Correction of the bounding box on the proposal area to improve the accuracy of object position prediction.

Next, the area proposals generated by RPN are processed using ROI Align, a technique that aims to extract features with a fixed size, even though the size of the area proposals varies. ROI Align improves the shortcomings of the ROI Pooling method, which tends to produce quantization errors in the detection process.

During the second phase, the region proposals produced by the Region Proposal Network (RPN) are further processed by the Mask R-CNN head, which simultaneously carries out object classification, refines bounding box coordinates, and generates segmentation masks. This stage produces three principal outputs:

- Object Classification: The model assigns a category label to each detected object based on its learned features.
- Bounding Box Regression: The system adjusts the coordinates and dimensions of the bounding boxes to more accurately enclose the detected objects.
- Mask Prediction: In addition to detecting objects, Mask R–CNN produces a binary mask for each identified instance. This mask delineates the specific pixels associated with the object, enabling precise instance-level segmentation within the image.

This stage enables Mask R–CNN to provide outputs like bounding boxes, class labels, and binary masks that map object areas more accurately. Thus, Mask R–CNN is not only able to detect the position of objects but also provide detailed segmentation for each detected object. Final result of Mask R–CNN visualized by displaying the bounding box, class label, and mask on the input image.

B. Measurement Upper Arm Circumference (MUAC)

MUAC serves as a simple yet effective anthropometric measure for evaluating the nutritional status of pregnant women, particularly in screening for CED. Measure value below 23.5 cm is recognized as indicative of elevated CED risk, which is associated with adverse maternal outcomes and an increased likelihood of delivering infants with LBW.

Traditionally way assessed using manual measuring tapes, MUAC in this study is estimated through a digital method that leverages upper arm image segmentation. The approach begins with Mask R-CNN, which detects and accurately segments the upper arm within the image. From the segmentation mask's bounding box, specific geometric features—namely diagonal and horizontal lengths—are extracted to estimate MUAC.

For images captured from a frontal view, MUAC is inferred from the length of a diagonal line extending from the top-right to the bottom-left corner of the bounding box. In lateral view images, estimation is based on the horizontal distance from the midpoint of the left vertical edge of the bounding box to its geometric center. This view-dependent measurement technique is adapted from established methodologies [16], which also employed diagonal and horizontal representations to approximate upper arm dimensions. The proposed automated method aims to enhance the precision and consistency of MUAC estimation in digital anthropometric analysis.

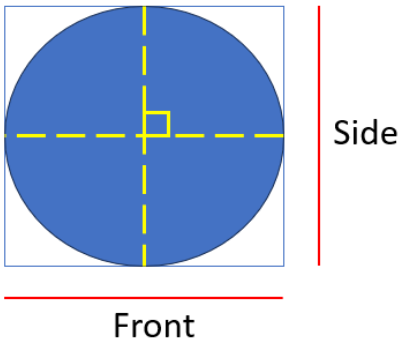

Fig. 2. A method for measuring MUAC.

The calculation method for MUAC is shown in Fig. 2. The front is the result of MUAC measurement with the picture taken from the front, and the side is the result of MUAC measurement with the picture taken from the side. Because the shape of the MUAC is closer to an ellipse than a circle, we will use the elliptical equation [see Eq. (1)]:

$$\text{Circumference} = 1/2\pi\ (\text{Front} + \text{Side}) \quad (1)$$

C. Mean Absolute Error (MAE)

MAE is one of the most frequently used evaluation metrics to measure the error rate in a regression or forecasting model. MAE quantifies the average magnitude of the absolute differences between observed values and corresponding model predictions, providing a straightforward indication of the model's predictive accuracy [26].

MAE is a commonly used performance metric that quantifies the average magnitude of errors between predicted values ($\hat{y}_i$) and actual ground truth values (y_i). It calculates the absolute differences without taking into account the direction of the errors, thereby providing a clear and intuitive interpretation in the same units as the original data. Due to this property, MAE is particularly well-suited for image-based measurement applications, such as those investigated in this study. The mathematical formulation of MAE is presented as follows [see Eq. (2)]:

$$MAE = \frac{1}{n} i + \sum_{i=1}^{n} (yi - \hat{y}i) \quad (2)$$

where,

- yi is the manual measurement value (reference)
- $\hat{y}i$ is the system measurement value
- n is the amount of data

D. Dataset

This study collected the dataset in person at the Faculty of Public Health, University of North Sumatra. Pictures were taken while the individual stood upright, with the left hand unobstructed by clothing and raised slightly, while the right hand was pointed straight down, as shown in Fig. 3. There was a total of 392 images, as shown in Fig. 4, but only 144 images were suitable for MUAC measurement. The reason is three-fold: i) Four pictures were taken from different angles during data collection. Only two images per individual were used in this study, so the remaining two were discarded. ii) The profile data of the individual in the image was not found, making the image unusable. iii) The captured image does not conform to the image standard.

Fig. 3. A dual-angle picture in-pose for MUAC measurement.

Fig. 4. The annotated set of images.

Images that do not meet the image standard cannot be included in this study because testing results will differ significantly from those that meet the image standard. Also, the measurement results will not match because the images that do not meet the image standards have different camera-to-object distances. The image standard detail such as Dimension: 2340 × 4160 pixel; Resolution:72 DPI; File Type: JPG; Camera to object distance: 180 cm. The collected data set was manually annotated on the Roboflow platform [27]. Annotation is done by marking bounding areas and bounding boxes to the objects to be detected in the image. The images in this dataset are put into three classes: person, hand, and upper arm. One hundred forty-four images were manually annotated. Each person has two images: a front-view image and a side-view image. The data labeling is done on a total of 72 individuals.

After annotating the images, a standard dataset suitable for model training was built. The dataset creation process was performed using the same annotation platform. This dataset is generated in JSON format, which is adapted to the COCO (Common Objects in Context) format so that it can be used directly in the Mask R-CNN architecture.

We used an in-house, task-specific dataset—rather than public corpora—because generic scene datasets lack MUAC-specific labels, paired front/side views, and the controlled capture conditions described above that are required for reliable pixel-to-centimeter scaling. These properties align with our method's needs (upper-arm detection/segmentation and dual-view geometry) and enable a faithful end-to-end evaluation with Mask R-CNN. We acknowledge the modest dataset size and controlled setting as limitations and plan to expand to more diverse populations and capture conditions in future work.

E. Reference Object Method

To establish a scale for measurements within the image, we employ a reference object method by using a known object—a standard power plug- in this study—as a reference for scaling. This approach relies on the known dimensions of the reference object to infer measurements within the image. The power plug is a standard object because its dimensions are typically standardized and can be reliably measured.

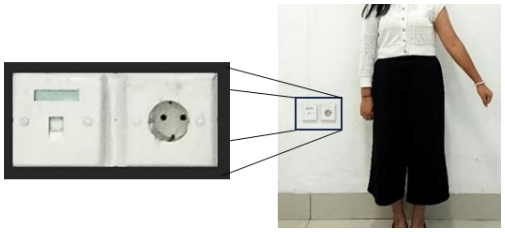

Fig. 5. (Left) Locating the power plug on the left side of the wall, and (Right) cropping the power plug with an image editor.

In Fig. 5, two images are presented to illustrate the process. The Right image shows an MUAC individual standing in front of a wall with a power plug visibly mounted on it. This initial image captures the scene as it appears naturally, with the power plug as the calibration reference object. The Left image in Fig. 5 shows the result after the power plug has been isolated by cropping the original image using Adobe Photoshop. This cropping step focuses on the power plug, removing extraneous elements from the image to allow for precise measurement.

Fig. 6. Determining the power plug's height using a measuring tape.

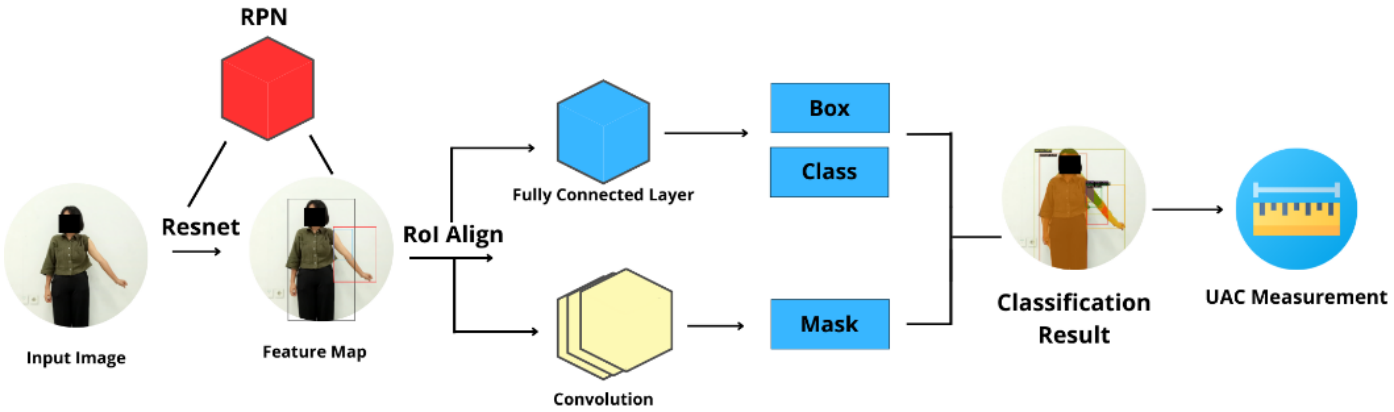

Fig. 7. General architecture of proposed system.

Fig. 6 shows the physical measurement of the actual power plug using a measuring tape, which determines its height to be 8.5 cm. Together, the equivalent height in pixels after measuring the power plug with digital imaging software is the width 297 pixels and height 136 pixels. These measurements are critical for calculating the scale factor that relates pixel dimensions in the image to real world dimensions.

The scale is calculated using the following equation:

$$Scale = y : \hat{y} \tag{3}$$

The variable y refers to the physical height of the power plug (8.5 cm, and the variable $\hat{y}$ refers to the height of the power plug in pixels (136 pixels). After plugging the values measured beforehand into the equation, the scale is computed: Scale = 8.5 cm: 136 px.

Simplifying this ratio, we divide both sides by 8.5, yielding 1 : 16. This means that 16 pixels in the image correspond to 1 cm in the real world. In other words, the pixel-to-centimeter conversion factor is 16 pixels/cm [see Eq. (3)]. This scale factor is then applied to all object detection and segmentation results obtained from the image. For example, if an object detected in the image has a height of 48 pixels, its real-world height would be 48 / 16 = 3 cm.

IV. Implementation and Results

This section presents the implementation and empirical findings: dataset preparation and training on Google Colab using Detectron2 with Mask R-CNN-FPN, the end-to-end system architecture, and a performance study across training iterations (1000 versus 2000). We then report testing and evaluation results—including IoU-based detection accuracy, MUAC measurement comparisons with the resulting MAE, and the accuracy distribution—and illustrate the top-down panoptic segmentation outputs that underpin the measurements.

A. Dataset and Training

In this study, the dataset was prepared and processed using the Roboflow platform, a tool specifically developed to optimize workflows in computer vision applications. During the data collection phase, relevant images were gathered and subsequently organized within Roboflow to ensure compatibility with model training requirements. This preparation process likely encompassed several key steps, including image annotation (e.g., labeling objects within images), data augmentation (e.g., applying transformations such as rotation or scaling to enhance dataset variability), and partitioning the dataset into training, validation, and test subsets. Roboflow's integrated functionalities—such as automated preprocessing and annotation tools—facilitate the creation of well-structured and machine learning-ready datasets.

Following the completion of dataset preparation, model training was conducted using Google Colab (Google LLC). Cloud-based platform offers access to significant computational resources, including Graphics Processing Units (GPUs), at little to no cost. Google Colab was using as the training environment primarily due to its robust GPU support, which is essential for efficiently accelerating the computationally intensive process of training deep learning models. The training efforts within this environment centered explicitly on the Mask R-CNN model, a novel architecture renowned for its efficacy in object detection and instance segmentation.

This research employs Detectron2 [28], a comprehensive library developed by Facebook AI Research (FAIR). Detectron2 facilitates the execution of complex, high-precision inference tasks, a capability leveraged in this study for model evaluation on the prepared dataset. For the specific architectural implementation herein, the Detectron2 framework was configured to utilize the Mask R-CNN model in conjunction with FPN.

B. System Architecture

The diagram depicts the data flow through the system, starting from the input images, followed by feature extraction using a backbone network. The system then branches into two main tasks: 1) a box classification and regression head, which predicts the object class and refines the bounding box coordinates, and 2) a mask prediction head, which generates a segmentation mask for each detected object. The final output is designed to produce measurements for the MUAC using the aforementioned algorithms. Fig. 7 illustrates the general architecture of the proposed system, providing a visual overview of the workflow.

C. Model Performance Evaluation with Varying Iterations

To evaluate the impact of training iterations on model performance, we compare the total loss and classification loss of different models trained for 1000 and 2000 iterations, as shown

in Table I. Total loss is a comprehensive metric that measures the overall error of the model, encompassing components such as classification, bounding box regression, segmentation, and object proposal. Classification loss quantifies the model's error in assigning objects to their correct classes. Lower values for both metrics indicate improved model performance.

TABLE I. PERFORMANCE COMPARISON FOR ALL TESTED MODELS

Model	Accuracy (P)	Total Loss	Class Loss
Faster R–CNN R50-FPN-3x (1000 iterations)	0.96	0.33	0.08
Faster R–CNN R50-FPN-3x (2000 iterations)	0.98	0.19	0.05
Mask R–CNN R50-FPN-3x (1000 iterations)	0.96	0.41	0.08
Mask R–CNN R50-FPN-3x (2000 iterations)	0.98	0.26	0.04
Mask R–CNN R101-C4 (1000 iterations)	0.98	0.40	0.04
Mask R–CNN R101-C4 (2000 iterations)	0.99	0.26	0.02

Table I provides a comparative summary of performance results for multiple models, including Faster R–CNN and Mask R–CNN variants. For instance, as illustrated by data also visualized in Fig. 8, extending the training of the Mask R–CNN R50-FPN-3x model from 1000 to 2000 iterations led to a reduction in total loss from 0.41 to 0.26 and in classification loss from 0.08 to 0.04. Similarly, for the Faster R-CNN R50-FPN-3x model, an identical increase in training iterations (from 1000 to 2000) yielded a decrease in total loss from 0.33 to 0.19 and classification loss from 0.08 to 0.05. This trend of improved loss metrics with more extensive training was consistent across other evaluated models. For example, the Mask R-CNN R101-C4 model exhibited a drop in total loss from 0.40 to 0.26 and in classification.

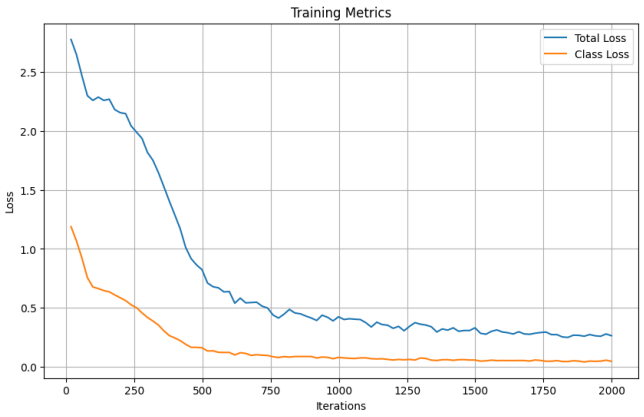

Fig. 8. Training loss curves for Mask R-CNN R50-FPN-3x. The plot illustrates the decrease in total loss and class loss over 2000 training iterations. The continued reduction in both loss metrics suggests that training for 2000 iterations yields a more refined model compared to training for only 1000 iterations.

Consistent with the observed improvement in loss metrics, the accuracy (P)—the proportion of correct predictions—also increased with more training iterations. For example, the accuracy of the Mask R-CNN R50-FPN-3x model improved from 0.96 to 0.98 when its training was extended from 1000 to 2000 iterations. Similarly, the Mask R-CNN R101-C4 model's accuracy rose from 0.98 to 0.99 under the same increase in training duration.

The presented results demonstrate that increasing training iterations from 1000 to 2000 consistently reduces total and classification losses across all tested models. This reduction is indicative of enhanced model convergence and lower overall error. The concurrent improvements observed in accuracy further corroborate this finding. Therefore, predicated on these empirical observations, this study adopts models trained for 2000 iterations, as they exhibit demonstrably superior performance characterized by minimized error and heightened accuracy.

D. Testing and Evaluation

Following the completion of model training, evaluation and testing procedures are conducted within the Colab notebook environment, where inference results are visualized directly on the input images. This workflow is seamlessly integrated with cloud storage services, which are utilized to manage both the datasets and the trained model files. To assess the accuracy of object detection, the IoU metric is employed as a key performance indicator. The IoU results for all evaluated images are presented in Table II.

TABLE II. IOU CALCULATION RESULTS

Individual	IoU for front-view image	IoU for side-view image
1	0,78	0,75
2	0,80	0,82
3	0,79	0,84
4	0,90	0,73
5	0,87	0,83
6	0,77	0,61
7	0,90	0,78
8	0,87	0,79
9	0,82	0,87
10	0,83	0,79
11	0,81	0,78
12	0,81	0,83
13	0,0	0,63
14	0,81	0,76
15	0,82	0,70
16	0,86	0,92
17	0,81	0,74
18	0,85	0,90
19	0,88	0,87
20	0,83	0,71
21	0,89	0,69
22	0,80	0,62
23	0,83	0,82
24	0,81	0,70
25	0,85	0,85
26	0,78	0,78

27	0,88	0,63
28	0,83	0,69
29	0,85	0,88
30	0,76	0,79
31	0,81	0,78
32	0,87	0,67
33	0,89	0,79
34	0,73	0,64
35	0,78	0,85
36	0,89	0,68
37	0,76	0,81
38	0,89	0,82
39	0,89	0,88
40	0,79	0,88
41	0,85	0,70
42	0,72	0,59
43	0,76	0,79
44	0,81	0,88
45	0,86	0,70
46	0,76	0,69
47	0,87	0,76
48	0,85	0,69
49	0,76	0,79
50	0,89	0,77
51	0,79	0,65
52	0,83	0,79
53	0,89	0,68
54	0,80	0,68
55	0,82	0,88
56	0,79	0,78
57	0,86	0,88
58	0,89	0,85
59	0,87	0,74
60	0,85	0,84
61	0,81	0,79
62	0,86	0,88
63	0,91	0,73
64	0,88	0,81
65	0,82	0,71
66	0,74	0,76
67	0,87	0,69
68	0,87	0,86
69	0,90	0,82
70	0,81	0,77
71	0,85	0,78
72	0,83	0,80

After inference, the output is extracted to the CPU for easier processing. From here, the bounding boxes, masks, class labels, and prediction scores are extracted from the results. The class indices corresponding to the objects to be detected are then identified, namely "person", "hand", and "upper arm".

Next, iterations are performed on each class to find the best detection based on the highest score value of each detected object. If the processed class is "upper arm", then the mask's diagonal and horizontal lengths are calculated using the CalculateDiagonalLength and HorizontalCutLength algorithms. The best mask for each class is then saved for further visualization. The saved information includes the bounding box, mask, score, and class label of the detection result. Subsequently, after measuring the MUAC of 72 individuals, the results obtained can be seen in Table III.

TABLE III. COMPARISON OF MUAC MEASUREMENTS

Individual	Manual measurement (cm)	System measurement (cm)
1	22,5	25,61
2	26	25,71
3	29	31,69
4	27,5	32,58
5	24	24,92
6	32,5	29,73
7	29	31,60
8	25,5	26,20
9	23,5	24,73
10	27	31,20
11	32	33,17
12	29	29,63
13	28,5	54,46
14	24	25,32
15	26	24,33
16	26,5	26,98
17	24	26,39
18	26	29,04
19	21,5	26,98
20	24	26,69
21	23	26,49
22	27	25,12
23	28	32,09
24	29	28,65
25	33	33,26
26	24	24,53
27	28	30,81
28	27	29,93
29	22	23,84
30	27	29,53
31	28	31,50
32	21	23,94
33	30,5	31,30
34	26	26,59
35	26	27,18

36	35	40,52
37	35	32,09
38	26,5	29,14
39	27,5	28,95
40	25,5	26,39
41	23	23,25
42	27	25,32
43	27	30,52
44	22	23,16
45	29	33,36
46	27	25,81
47	26,5	28,36
48	24,5	23,35
49	29	32,58
50	26	30,12
51	30	35,62
52	20	20,02
53	23	23,16
54	23	24,92
55	25,5	26,30
56	23	23,65
57	29	32,97
58	29	26,79
59	22	20,80
60	30	29,53
61	24	25,90
62	29,5	29,63
63	24	25,51
64	31	31,4
65	30	32,18
66	32	34,05
67	29	29,14
68	31,5	33,85
69	28	29,24
70	26	26,79
71	20,5	21,88
72	26	27,08
73	22,5	25,61
74	26	25,71

The overall system accuracy can be seen in Table IV, which provides a breakdown of the system's accuracy when tested on a dataset of 72 individuals using Google Colab as the training and inference platform.

TABLE IV. ACCURACY DISTRIBUTION

Total Data	72 Individuals
0 - 1 cm Results	21 Individuals
1 - 2 cm Results	20 Individuals
Results > 2 cm	31 Individuals

The distribution of results presented in Table IV elucidates the system's overall accuracy profile and identifies key limitations. While a substantial proportion of MUAC predictions achieve high accuracy (29.2% exhibiting a 0–1 cm difference from actual values) and moderate precision (27.8% with a 1–2 cm difference), the considerable percentage (43.1%) of predictions deviating by more than 2 cm suggests a lack of uniform precision across the study population. This inconsistency in performance could plausibly arise from various factors. These may include disparities in lighting conditions during image acquisition, instances of occlusion affecting the arm region, or natural variations in the anatomical appearance of the MUAC region across individuals, all of which can adversely affect the performance of the underlying Mask R–CNN model responsible for object detection and segmentation.

An evaluation was conducted using the Mean Absolute Error (MAE) metric, as defined in Eq. (2), to assess system performance quantitatively. This metric computes the mean absolute deviations of the system-generated measurements and the manual (ground truth) measurements for all individuals in the dataset. The analysis yielded an MAE of 2.31 cm. This result, representing the average discrepancy between the system's estimates and the actual values, highlights the necessity for additional improvements to improve the system's measurement precision.

E. Top-Down Panoptic Segmentation Approach

The system's measurement process was visualized using Detectron2's Visualizer, a dedicated utility for illustrating the outcomes of object detection and segmentation procedures. This tool graphically depicts the boundaries and masks of detected objects (specifically the upper arm in this context), superimposing key measurement indicators, such as diagonal and horizontal lines, onto the segmented areas. Such visual outputs are invaluable for qualitatively evaluating the system's detection and segmentation performance. Furthermore, they play a critical role in validating quantitative measurements by enabling direct comparison with ground truth data, which ultimately supports the assessment of the system's output reliability.

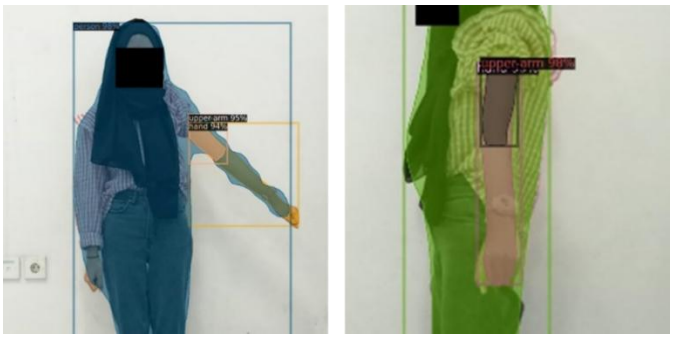

Fig. 9. Top-down detection and segmentation results for individual 52.

Fig. 9 is an example case with an estimated MUAC measurement that is the closest to the ground truth. The length of the diagonal line intersecting the detected mask area is 84 pixels. Meanwhile, the horizontal line bisecting the mask is 120 pixels long. The measurement results in pixel units are then

converted into centimeter (cm) units using a conversion factor, where 1 cm is equivalent to 16 pixels [derived from Eq. (3)].

After the conversion, it is found that the length of the diagonal line intersecting the mask (from the front image) is 5.25 cm, while the length of the horizontal line intersecting the mask (from the side image) is 7.5 cm. Next, these values are used in, as shown in Eq. (1), the elliptical formula to calculate the overall MUAC, which is 20.0175 cm.

For further examples, the following two instances are system MUAC measurements with a significant difference from the corresponding ground truth MUAC.

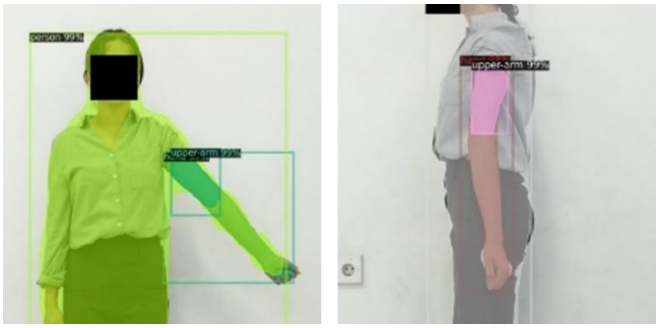

Fig. 10. Top-down detection and segmentation results for individual 48.

Fig. 10 illustrates an example of an image showing the result of a MUAC measurement that deviates by 1 to 2 cm from the ground truth value. In the image, the length of the diagonal line intersecting the detected mask area is 98 pixels (converted to 6.125 cm). Meanwhile, the length of the horizontal line bisecting the mask is 140 pixels (converts to 8.75 cm). Based on the data presented in Table III, the ground truth of individual 48 is 24.5 cm. Thus, the difference between the computed result and the ground truth is 1.15 cm.

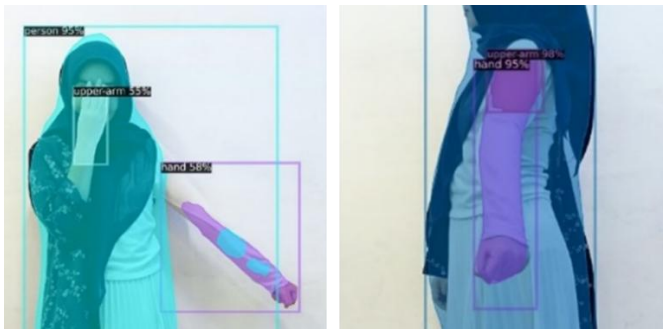

Fig. 11. Top-down detection and segmentation results for individual 13.

In another case, Fig. 11 is an example of images with the results of a MUAC measurement with a difference of > 2 cm compared to the actual MUAC measurement. The diagonal line intersecting the mask is 346 pixels (converts to 21.65 cm), while the horizontal line intersecting the mask is 209 pixels (converts to 13.06 cm). Plugging these values into the elliptical formula, we get the circumference measurement of 54.46 cm.

V. Discussion

The results of the IoU calculation in Table I shows that the model achieves reasonably good accuracy. The average for IoU calculation results for images taken from the front reaches 0.82, which is quite close to 1. The average IoU calculation results for images taken from the side reach 0.77, shows that the model is better at detecting pictures taken from the front than those taken from the side.

In Table III, the actual MUAC of individual 52 is 20 cm. So, the calculation result in this case has a difference of about 0.15 cm with the ground truth. The difference is not too high due to the person doing a moderately good pose where the arm is raised high enough so that the system can detect the upper arm well. The person also raises their sleeves, so that the upper arm is more straightforward to detect.

In some cases, as seen with individual 13, there is a notable variation between the estimated MUAC and the ground truth. In the case of individual 13, the size difference reaches 25.96 cm. This size difference is important because the system does not detect the upper arm correctly. As can be seen in Fig. 11, the system detects the upper arm on the person's left hand. This detection error is caused by the individual wearing a white shirt that resembles the background color of the picture. Furthermore, the participant's sleeve remained unraised, a circumstance that posed an additional challenge to the system's ability to perform accurate detection and segmentation of the target region.

Future Work: Building on our error analysis—evidenced by the front- versus side-view IoU gap (0.82 vs. 0.77) and failures such as individual 13, where clothing–background similarity and unraised sleeves led to mis-segmentation—we outline four concrete directions with measurable targets: 1) Robustness to clothing – background similarity and atypical poses: targeted augmentations (color/texture jitter, occlusion/cutout, pose perturbation), hard-negative mining, and boundary-refinement heads; target $\geqslant$ +0.03 absolute IoU on side view and $\geqslant$ 20% reduction in MUAC errors > 2 cm. 2) Calibration beyond site-specific objects: ablate the current outlet-based scaling against portable fiducials (checkerboard/ArUco) and reference-free/self-calibrating options; target $\leqslant$ 1.5$\times$ variance relative to the best method and < 5% catastrophic scale failures. 3) View fusion and measurement modeling: benchmark the current ellipse formulation against (a) robust statistics over multiple per-view cuts and (b) a small interpretable regressor from per-view mask features; target $\geqslant$ 0.3 cm MAE reduction without sacrificing interpretability. 4) Real-time, on-device deployment and prospective validation: compress the model (pruning/quantization), profile on mobile hardware, and run a head-to-head study against trained operators using standardized MUAC tapes; target $\leqslant$ 200 ms end-to-end latency per view with $\leqslant$ +0.2 cm MAE degradation versus desktop, intraclass correlation $\geqslant$ 0.9, and Bland – Altman limits of agreement within $\pm$3 cm [29].

VI. Conclusion

The Mask R-CNN model accurately detects the upper arm and other body parts. This shows good potential for object segmentation, especially in the case of MUAC measurement. Based on the IoU calculation, 0.82 was obtained for images taken from the front angle and 0.77 for pictures taken from the side. These high IoU values indicate that the applied Mask R-CNN model can produce segmentations close to the annotated ground truth mask, especially in images that meet the image

standards. This value also indicates that the model can detect upper-arm objects better when the image is taken from the front. This result supports the effectiveness of the system for MUAC measurement.

The MUAC measurement results obtained through the digital imaging system show variations in accuracy depending on the quality of the image. As seen in Fig. 9, when the photos are taken with closer compliance to the image standard, the system measurements can closely resemble the manual measurements (higher accuracy). However, significant differences between system measurements and manual measurements can occur, as shown in Fig. 10 and Fig. 11, indicating that the system's accuracy will decrease, if the individual does not show the upper arm clearly and the color of the sleeve matches or resembles the background color.

Capturing images with a frontal viewing angle, proper distance, and adequate lighting tends to produce more accurate measurement results. This emphasizes the importance of quality and consistency in the image capture process to ensure that the automated measurement system can provide accurate and reliable results.

Evaluation using the Mean Absolute Error (MAE) metric resulted in an average error of 2.31 cm between system measurements and manual measurements. Of the 72 image samples tested, 21 individuals (29.2%) had measurement differences of 0-1 cm, 20 individuals (27.8%) had differences of 1-2 cm, and 31 individuals (43.1%) had differences of more than 2 cm. Thus, further development is expected to overcome the problems found in this study and improve the performance and accuracy of the MUAC detection and measurement system.

This study has several limitations. The dataset is modest and was collected under a standardized imaging protocol, which may limit generalizability. Pixel-to-centimeter calibration depends on a wall power-outlet reference that may not be available or standardized across sites. MUAC estimation fuses paired front- and side-view measurements via an ellipse approximation, making the final circumference susceptible to view-specific errors. The pipeline is also sensitive to clothing/background similarity and pose. Future work will expand data diversity and size, explore reference-free calibration, and strengthen the robustness of the segmentation and measurement pipeline.

ACKNOWLEDGMENT

This work was supported by "The DRTPM Kemendikbudristek of Republic Indonesia", under Research Grant 2023, in accordance with the Letter of Assignment Agreement, Number: 52/UN5.2.3.1/PPM/KP-DRTPM/B/2023.

REFERENCES

[1] C. Kyle and K. Jhon, Anthropometric Measurement. StatPearls Publishing, Treasure Island (FL), 2025. http://europepmc.org/books/NBK537315.

[2] B. G. Sisay, D. Haile, H. Y. Hassen, and S. H. Gebreyesus, "Performance of mid-upper arm circumference as a screening tool for identifying adolescents with overweight and obesity," PLoS ONE, vol. 15, no. 6, pp. 1–13, 2020, https://doi.org/10.1371/journal.pone.0235063.

[3] L. Thorup et al., "Mid-upper arm circumference as an indicator of underweight in adults: a cross-sectional study from Nepal," BioMed Central Public Health, vol. 20, no. 1, pp. 1–7, 2020, https://doi.org/10.1186/s12889-020-09294-0.

[4] E. Van Tonder et al., "Mid-upper arm circumference (MUAC) as a feasible tool in detecting adult malnutrition," South African Journal of Clinical Nutrition, vol. 32, no. 4, pp. 93–98, 2019, https://doi.org/10.1080/16070658.2018.1484622.

[5] A. Fakier, G. Petro, and S. Fawcus, "Mid-upper Arm Circumference : A Surrogate for Body," South African Medical Journal, vol. 107, no. 7, pp. 606–610, 2017, https://doi.org/10.7196/SAMJ.2017.vi7.12255.

[6] Febri Adriati and Shanty Chloranyta, "Status Gizi Ibu Hamil Berdasarkan Pengukuran Lingkar Lengan Atas (LILA)," Jurnal Kesehatan Panca Bhakti Lampung, vol. 10, no. 2, p. 127, 2022, https://doi.org/10.47218/jkpbl.v10i2.194.

[7] K. G. Mishra, V. Bhatia, and R. Nayak, "Association between mid-upper arm circumference and body mass index in pregnant women to assess their nutritional status," Journal of Family Medicine and Primary Care, , 2020, https://doi.org/10.4103/jfmpc.jfmpc_57_20

[8] D. T. Hendrixson, P. N. Lasowski, A. S. Koroma, and M. J. Manary, "Newborn Mid–Upper Arm Circumference Identifies Low–Birth Weight and Vulnerable Infants: A Secondary Analysis," vol. 6, no. 10, p. nzac138, 2022, https://doi.org/10.1093/cdn/nzac138.

[9] G. R. Babu et al., "Mid-upper arm circumference in pregnant women and birth weight in newborns as substitute for skinfold thickness: findings from the MAASTHI cohort study, India," BioMed Central Pregnancy and Childbirth, vol. 21, no. 1, pp. 1–11, 2021, https://doi.org/10.1186/s12884-021-03915-1.

[10] R. Rana et al., "Mid-Upper Arm Circumference Tapes and Measurement Discrepancies: Time to Standardize Product Specifications and Reporting," Global Health Science and Practice, vol. 9, no. 4, pp. 1011–1014, 2021, https://doi.org/10.9745/GHSP-D-21-00273.

[11] N. N. Kaashki et al, "Deep Learning-Based Automated Extraction of Anthropometric Measurements from a Single 3-D Scan," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–14, 2021, https://doi.org/10.1109/TIM.2021.3106126.

[12] S. L. Cichosz et al., "Precise Prediction of Total Body Lean and Fat Mass From Anthropometric and Demographic Data: Development and Validation of Neural Network Models," Journal of Diabetes Science and Technology, vol. 15, no. 6, pp. 1337–1343, 2020, https://doi.org/10.1177/1932296820971348.

[13] Š. Dana, R. Adam, and M. Martin, "Automatic Estimation of Anthropometric Human Body Measurements," VISAPP, 2022, vol. 4, no. 1, pp. 537–544, https://doi.org/10.5220/0010878100003124.

[14] M. Edoardo et al., "Digital Anthropometry: A Systematic Review on Precision, Reliability and Accuracy of Most Popular Existing Technologies," Nutrients, vol. 15, no. 2, pp. 1–39, 2023, https://doi.org/10.3390/nu15020302.

[15] D. S. Bukit, M. S. Lydia, P. I. Nainggolan, H. I. Mahmud, R. M. Sembiring, R. M. Hasibuan, D. Athirah, F. M. Mufida, and R. W. B. O. K. Rahmat, "Leveraging Machine Learning Techniques for Stunting Detection and Height Growth Prediction in Children Aged 0–5 Years," ELTICOM 2024. pp. 130–134, 2024, https://doi.org/10.1109/ELTICOM64085.2024.10864967.

[16] Yingying Liu, Arcot Sowmya, and Heba Khamis, "Single camera multi-view anthropometric measurement of human height and mid-upper arm circumference using linear regression," PLoS ONE, vol. 13, no. 4, pp. 1–22, 2018, https://doi.org/10.1371/journal.pone.0195600.

[17] H. Qinhua et al., "An effective approach for CT lung segmentation using mask region-based convolutional neural networks," Artificial Intelligence in Medicine, vol. 103, p. 101792, 2020, https://doi.org/10.1016/j.artmed.2020.101792.

[18] M. Fan et al., "Mass Detection and Segmentation in Digital Breast Tomosynthesis Using 3D-Mask Region-Based Convolutional Neural Network: A Comparative Analysis," Frontiers in Molecular Biosciences, vol. 7, pp. 1–15, 2020, https://doi:10.3389/fmolb.2020.599333.

[19] Linqin Cai, Tao Long, Yuhuan Dai, and Yuting Huang, "Mask R-CNN-Based Detection and Segmentation for Pulmonary Nodule 3D Visualization Diagnosis," IEEE Access, vol. 8, pp. 44400–44409, 2020, https://doi:10.1109/ACCESS.2020.2976432.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick, “Mask R-CNN” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no .2, pp. 386–397, 2020, https://doi.org/10.1109/TPAMI.2018.2844175.

[21] K. Alexander et al., “Panoptic Segmentation,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9396–9405. https://doi.org/10.1109/CVPR.2019.00963.

[22] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017. https://doi.org/10.1109/TPAMI.2016.2577031

[23] Lin, TY. et al. (2014). Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48

[24] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 936-944, https://doi.org/10.1109/CVPR.2017.106

[25] F. Hou et al., “Improved Mask R-CNN with distance guided intersection over union for GPR signature detection and segmentation,” Autom Constr, vol. 121, p. 103414, Jan. 2021, https://doi.org/10.1016/J.AUTCON.2020.103414.

[26] Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, 45–76. https://doi.org/10.28945/4184.

[27] B. Dwyer, J. Nelson, and T. Hansen, “Roboflow,” 2024, 1.0. Accessed: Apr. 20, 2025. [Online]. Available: https://roboflow.com

[28] Wu, Yuxin, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. "Detectron2." (2019): https://github.com/facebookresearch/detectron2.

[29] M. A. et al., “Accuracy of Fully Automated 3D Imaging System for Child Anthropometry in a Low-Resource Setting: Effectiveness Evaluation in Malakal, South Sudan,” JMIR Biomed Eng, vol. 7, no. 2, p. e40066, Oct. 2022, https://doi.org/10.2196/40066.