IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 18, 2021, accepted April 7, 2021, date of publication April 12, 2021, date of current version April 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3072636

Contour Extraction of Individual Cattle From an Image Using Enhanced Mask R-CNN Instance Segmentation Method

ROTIMI-WILLIAMS BELLO, AHMAD SUFRIL AZLAN MOHAMED, AND ABDULLAH ZAWAWI TALIB

School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang 11800, Malaysia

Corresponding author: Ahmad Sufril Azlan Mohamed (sufril@usm.my)

This work was supported by the Division of Research and Innovation (RCMO), Research University under Grant 1001 / PKOMP / 8014001 and School of Computer Sciences, Universiti Sains Malaysia.

ABSTRACT In animal husbandry, the traceability of individual cattle, their health information, and performance records greatly depend on computer vision and image processing-based approaches. However, some of these approaches perform below expectations in obtaining real-time information about individual cattle. No doubt, inaccurate segmentation and incomplete extraction of each cattle object from an image are notable contributory factors. As accurate segmentation is a prerequisite for obtaining real-time information about individual cattle, and since the algorithm of Mask R-CNN relies on the algorithm of simultaneous localization and mapping (SLAM), for the construction of the semantic map, which sometimes exchanges image background for the foreground, there is a need to enhance the available approaches towards achieving precision animal husbandry. To achieve this, an enhanced Mask R-CNN instance segmentation method is proposed to support indistinct boundaries and irregular shapes of cattle bodies. The methods employed in the research are in multiple folds: (1) Pre-enhancement of the image using generalized color Fourier descriptors (GCFD); (2) Provision of optimal filter size that was smaller than ResNet101 (the backbone of Mask R-CNN) for the extraction of smaller and composite features; (3) Utilization of multiscale semantic features using region proposals; (4) A fully connected layer of existing Mask R-CNN integrated with a sub-network for enhanced segmentation and (5) Post-enhancement of the image using Grabcut. Experiments on the datasets of cattle images produced better results when compared to other state-of-the-art methods with 0.93 mAP.

INDEX TERMS Animal husbandry, cattle, Grabcut, instance segmentation, Mask R-CNN.

I. INTRODUCTION

In many countries all over the world, the agricultural sector contributes to the economy more than any other sector. As meat and dairy are the two most widely demanded products of cattle for human sustainability with their quality production depending on the good welfare of the producing cattle, there is a need to raise the welfare and management standard of livestock farming including that of the breeders [1]. An identification problem is one of the major problems confronting cattle breeders. Cattle that are not properly marked or labeled for identification will be very difficult to claim their ownership if missed or swapped. From the classical methods to the modern-day methods, different methods of identification have been proposed in the literature. In the past, lots of conventional constructs such as tattoo, tags, photographs, drawings, descriptions, branding (hot and freeze), and ear notching were identification methods put in place by cattle breeders for cattle identification to prevent identification problems should there be any incidence of missing, swapping, ownership disputes and false insurance claims [2], [3].

However, these methods of cattle identification are with flaws and less satisfactory. Therefore, some improved methods that are distinguishable have been proposed [4] for accurate and reliable identification as artificial marks, no matter the permanence, give room for duplication of different animals with which swapping can be practiced.

Before the advent of the muzzle print method [5], cattle identification has been by sketching the color markings on them on paper for registration and identification purposes.

This classical method of identification causes trouble among the breeders when their cattle are sold or are on an official test due to a lack of artistic ability on the part of the

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

breeders which makes the matching of the sketches and the markings on the animal disagree. It was with this permanent identification problem in the mind of every breeder that the practicable suggestion of using muzzle print as means of permanent identification was made by O. H. Baker of the American Jersey Cattle Club in Petersen's paper entitled ''The identification of the bovine by means of nose-prints'' [5]. Petersen's paper was the first published paper to suggest a permanent cattle identification method based on muzzle print principles widely accepted today.

However, the muzzle print method involves a substantial amount of images and high computational time for individual animal identification correlation.

In achieving huge and robust animal husbandry, the accessibility to behavioral and wellbeing information of individual cattle cannot be overlooked as this plays a great role in supporting the management in making the decision that relates to livestock matters [6]. As earlier iterated, cattle are being monitored using different conventional techniques such as radio frequency identification (RFID) and sensor-based machines [7] against what is obtainable in the state-of-the-art vision-based settings where segmentation of the image is a precondition for a robust and efficient cattle monitoring. Different studies have been performed on the segmented images for the extraction of visual features to carry out the evaluation and behavior analysis of animal welfare such as length and width, curvature, and posture of the animal body [8]–[12]. The accuracy and efficiency of image segmentation are very apparent in furthering the analysis of the image in vision-based individual cattle monitoring and performance recording. However, considering the conventional algorithm of SLAM which mask region-based convolutional neural network (Mask R-CNN) relies upon for instance segmentation, the position of the map point information is the only geometric point that is either densely or sparsely located in the space.

Judging the position of these spatial points to be feasible avails us with comparatively accurate information about cattle object location, but that does not avail us with a higher semantic information level. Current progress recorded in deep learning enriches us with a direction for overcoming this challenge. The potency of deep neural networks in feature learning [13], [14] has enabled noteworthy progress in the field of computer vision, object detection, and segmentation. In the aspect of object detection, the Faster RCNN method [15] from which the Mask R-CNN [16] method was coined has greatly contributed to the robust detection of objects [17]. In the aspect of object segmentation, the MASK R-CNN has great image object detection and segmentation tactics [18].

Nevertheless, the algorithm of MASK R-CNN instance segmentation cannot completely differentiate between the image foreground and background during segmentation.

Motivated by these limitations, the algorithm for cattle image segmentation in the semantic map is enhanced in this paper by combining it with the algorithm of Grabcut. The implication of the enhanced algorithm is the increment in cattle segmentation accuracy in the course of dynamically constructing the semantic map [19]. By employing the enhanced algorithm, the cattle object will be identified more accurately in the image and the idea of localizing will be accomplished better. The work in this paper involves detecting and extracting key images that contain the cattle object, get the images inputted into the convolutional network and enhanced the process by subjecting the images to features descriptor for the cattle instance segmentation, then, apply the Grabcut for the contour extraction. In this paper, the contour extraction of individual cattle from an image using an enhanced Mask R-CNN instance segmentation method is proposed. The work in this paper is an attempt to achieve real-time cattle traceability, health information, and performance recording in animal husbandry applications [20], [21].

The followings are the research contributions:

- Pre-enhancement of the image using generalized color Fourier descriptors (GCFD);
- Provision of optimal filter size that was smaller than ResNet101 (the backbone of Mask R-CNN) for the extraction of smaller and composite features, thereby, the number of parameters required for the training was decreased;
- Utilization of multiscale semantic features using region proposals;
- A fully connected layer of existing Mask R-CNN integrated with a sub-network for enhanced segmentation;
- Post-enhancement of the image using Grabcut.

The arrangement of the paper takes the following order: Section II relates the work in this paper to some related works of literature in object detection and segmentation, Section III introduces the materials and methods employed in achieving the proposed approach, Section IV illustrates the implementation, Section V presents the results and discussion, and Section VI concludes the work and suggests future research direction.

II. RELATED WORKS

Recently, different kinds of semantic neural networks that can segment targets have been proposed in the literature.

Below are some illustrations to support this proposition, they are as follows:

The first appearance of fully convolutional networks was in 2015 [22] in which a new chapter was opened in the computer vision community for semantic segmentation.

Just recently, AlexNet and GoogLeNet classification networks are being modified by different researchers as full convolutional networks to produce robust and accurate segmentation by merging the semantic information of the deep and rough network layers with the trivial information of the shallow and fine network layers. The novelty of SegNet [23] as a segmentation network is apparent in the way in which the lower resolution input feature map is up-sampled by the decoder. The feature map of the U-Net [24] encoder is succinctly tied to the decoder's up-sampling feature map at each stage resulting in the formation of a trapezoidal structure.

Dilated convolution was proposed by DeepLab V1 [25], in the proposal, there's no reduction in the resolution of the feature graph in the last two operations of the maximum pooling, and empty convolution is employed in the convolution after the next to the last maximum pool. The conditional stochastic field is employed after the operation as post-processing to re-establish the boundary details to get a precise positioning effect. A multi-scale sturdy segmentation method was proposed by DeepLab V2 [26] for hollow spatial pyramid pooling, and by merging the deep convolutional neural network (DCNN) method and probabilistic graph model, the target boundary's location is improved. A multi-grid method employed for introducing different cavitations in the residual block was proposed by DeepLab V3 [27] where the features of image-level are joined to the module of the hollow space pyramid pooling, and thereafter, employing batch normalization techniques.

A semi-supervised technique that includes a generator network was proposed in generating against networks (GANs); this is to enable the provision of additional training samples for multi-object classifiers as discriminators in the GAN framework. DeepLab [28] is an extension of the three versions of DeepLab V1, DeepLab V2, and DeepLab V3.

Image segmentation as an extension of object detection strives to accurately define objects' classes at a pixel-wise level. Two different categories of image segmentation are in existence; they are (1) semantic segmentation and (2) instance segmentation. Semantic segmentation is a segmentation task that involves pixel-labeling of each image's pixel for a particular object's classes; however, there is no differentiation among the objects which belong to the same object class [29]. A usually employed model of semantic segmentation is fully convolutional networks (FCN) which is a variant of CNN for transforming pixels of an image into categories of the pixel. By employing masks to represent the object in an image, each object instance can be identified by instance segmentation which also simultaneously identifies object class prediction and mask extraction [30]. Three steps are usually involved in instance segmentation, namely identification of regions of the proposal using region proposal network (RPN), object class prediction, and object mask extraction. The convolution operation assists mask extraction in encoding the spatial layout of the input object. Just recently, in the study [18], the authors proposed some instance segmentation methods, this is in line with the work of Bai and Urtasun [31], where deep learning and transform of watershed were combined for the production of an energy map, and segmentation of object instances was realized by cutting a single energy level. Also, in the study [32], the authors proposed the extension of a fully convolutional network in livestock practice to achieve beef cattle segmentation.

An instance segmentation that is iterative in operation was proposed in the study [29]; learning of implicit shapes ahead of improving the labeling quality of the predicted pixel-wise was the major novelty work in their proposed instance segmentation method. Moreover, in the study [33], the authors proposed a DeepMask that produces the proposals of segmentation object unswervingly from the pixels of the image before classifying them into categories that are different from one another. This was extended in [34] where little changes were made to the DeepMask to produce SharpMask. In the study [16], the authors proposed a novelty Mask R-CNN framework that could perform detection of object key points in an image and instance segmentation. Put together, this entire breakthrough in computer vision research reflects the practicality of convolutional neural networks (CNNs) based methods in the segmentation of cattle under controlled environments.

Many algorithms that are based on deep learning have opened ways for dramatic research advancement in computer vision such as object detection [4], [35], semantic segmentation [36], and instance segmentation [33], [16].

According to [15], a good number of object detection algorithms produce a bounding box for each target detected followed by classification of the bounding boxed objects. In [37], the region-based convolutional neural network (R-CNN) method employs selective search to produce region proposals and employs deep CNN for classifying the object proposals. Nevertheless, the R-CNN method for extracting features from the proposal region is not so cheap. Region proposals are produced in Faster R-CNN [37], [15] via a secluded first branch called region proposal network (RPN) which images are passed through to generate a set of anchors otherwise known as rectangular object bounding boxes, and the second branch called Fast R-CNN detector is employed for the feature extraction from each candidate box before performing classification and bounding box regression.

To summarize the above-related works, the image's objects should not only be correctly located by instance segmentation method, but they should be able to perfectly get segmented by the method. So, instance segmentation [38] known with this quality can be taken as a combination of object detection and semantic segmentation.

III. MATERIALS AND METHODS

This section presents the materials and methods which are employed in carrying out the tasks that are involved in this study. The section is divided into the following subsections: data acquisition and pre-processing; overview of the proposed approach framework; extraction of the keyframe from image rows; image enhancement using generalized color Fourier descriptors (GCFD); Mask R-CNN instance segmentation network; region proposal network and loss function; enhanced Mask R-CNN for cow characterization; model development; an abridged model of ResNet; the enhanced Mask R-CNN structure and Grabcut image segmentation.

A. DATA ACQUISITION AND PRE-PROCESSING

Our research targets in this work are the trypanotolerant Muturu and Keteku cattle commonly found in Nigeria, West Africa. The cattle, 10 in number were kept back in a ranch with little or no conspicuous form of identification.

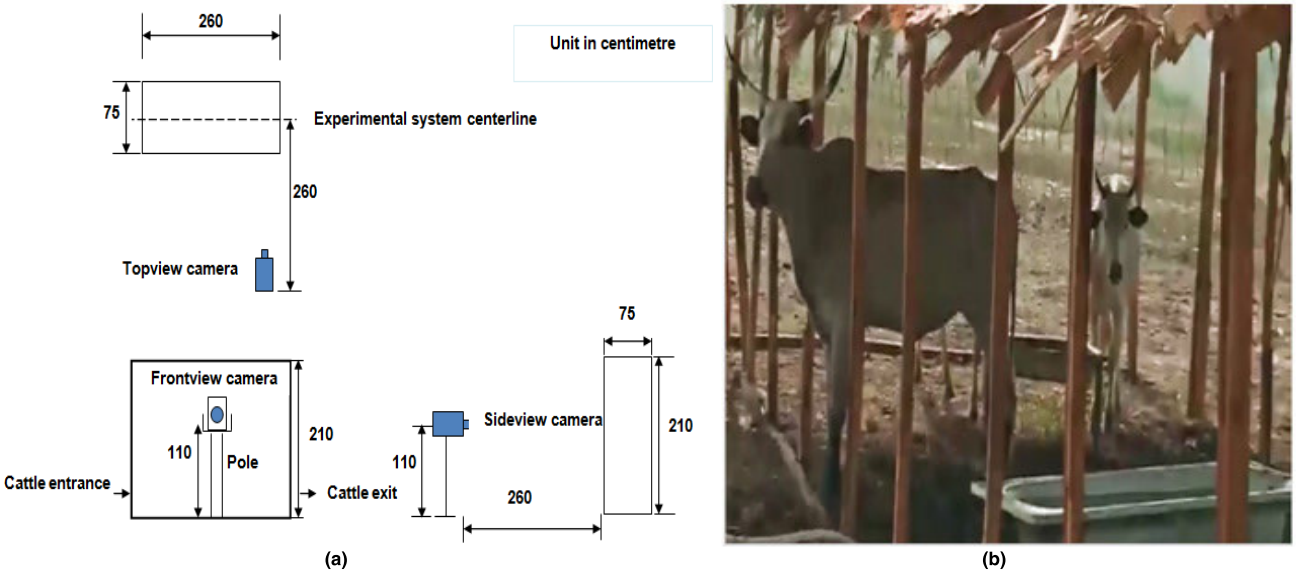

FIGURE 1. (a) Dimensional sketch of the individual cattle recognizing system (b) Instance of cattle in the ranch.

They are the breeds that are mostly reared for their meat, and sometimes as farm equipment. The body length and the body height of each cow are 86.6 cm and 95.0 cm respectively. The dimensional sketch of the individual cattle recognizing system, and the example of cattle in the ranch, as shown in Fig. 1 (a) and (b) respectively depict the platform employed for the data acquisition. All the cattle species were studied to recognize individual cattle characteristics, each cow with 100 images resulting in 1000 images in entirety. 400 images, that is, 4 cows' subject $\times$ 100 images of each subject were employed for the training of the network in the training phase. 600 images, that is, 6 cows' subject $\times$ 100 images of each subject were employed for testing the network in the testing phase.

By the middle of September 2019, a practicality test was carried out to mine the image data, and analysis of the mined image data was performed consequently by image processing.

The augmentation technique widely used in populating photographs was employed in this work to augment the acquired images as this is the only means by which high accuracy can be achieved by the model of deep learning which is mostly dependent on the amount and diversity of the training data. Therefore, by using augmentation techniques such as shearing, rotation, and translations of image height and width the issue of diversity in the experimental data was solved with the addition of 4000 datasets from which 600 images were added to the training dataset and 3400 images were added to the testing dataset. Moreover, the proposed model is trained on the pre-trained Microsoft Common Objects in COntext (MS COCO) weights of Mask R-CNN making it fit for the acquired experimental data (own cow dataset).

To get the case study in order, a charged coupled device (CCD) camera was made use of for capturing each cow. To acquire images of requisite size, the camera was sited on a pole very high and fairly of limited distance from the centerline of the experimental system. The system for processing each cow image was tactically sited in a location that cattle usually pass through almost every day with reduced illumination variation which aids in producing noise-free and blur-free images. The system employed for detecting and segmenting the cow can operate on any Windows-based computer. A more rapidly operational computer system is preferred to any others for the image processing that involves so many calculations and processing on the go. The specifications of the computer employed for the cattle segmentation and contour extraction task are 16 Gigabytes of RAM, Intel Core i5 processor @2.4 GHz, 2 terabytes of hard disk space, a graphics card, a computer monitor for monitoring the processing of multiple images, and a CCD digital camera. Open computer vision (OpenCV) and its library were used as the specification for the image-processing and computer vision elements execution.

The dataset of the cattle is a herculean task for the segmentation network to segment when the following aspects are considered:

- The frequent change in position of the cattle object and assumption of different positions especially when moving. To address this, the network is expected to possess a much stronger capacity for generalization;
- The coat patterns resemblance that exists among the cattle with no conspicuous form of identification makes it difficult if not impossible to differentiate between two specific cattle, this, nearly makes the differentiation of one cow from another impossible for human eyes in the event of partial occlusion;
- The impact of illumination variation on the algorithm of machine learning poses a challenge to the entire segmentation process, as machine learning can mistakenly assume patches for cattle's features;

TABLE 1. Network hyperparameters for the model.

Specifications	Amount in number
Learning rate	0.001
Weight decay	0.0001
Momentum of learning	0.9
Dimension of image (minimum)	512
Dimension of image (maximum)	512
Detection confidence (minimum)	0.5
Number of batches	5
Size of batch	200
Epochs	5
Iterations per epoch	5
Steps per epoch	1000
Mask shape	28×28
Number of anchor classes (cow & background)	2

FIGURE 2. Flowchart of cow image instance segmentation and patches-free contour extraction using the enhanced Mask R-CNN and Grabcut.

- The impossibility to distinguish between the segmented cow instance and the image background is also a great challenge in the cow detection task.

B. OVERVIEW OF THE PROPOSED APPROACH FRAMEWORK

To monitor the traceability, health information, and performance of individual cattle in computer vision-based animal husbandry; the pre-requisite for such task is an effectual cattle segmentation which in no small measure helps in furthering the analysis of the image [39]. Proposed in this work is an enhanced Mask R-CNN instance segmentation method with an adapted generalized color Fourier descriptors (GCFD) and Grabcut algorithm for achieving cow instance segmentation in the typical cattle ranch setting. As shown in Fig. 2, for given cows in a row, it is important to determine the keyframes from the non-key frames, and this goes by the step of keyframe extraction.

The selected keyframes which are affected by the variation in illumination are subjected to image enhancement through the application of GCFD. After which the cow body areas are detected and segmented by the enhanced Mask R-CNN model which is constructed and trained mainly for that purpose. The results produced by the segmentation process help in the structural mapping of the spatial locations of the cow features such as head, trunk, and legs which collectively, in the end, enable precise extraction of cow contour lines. The ResNet101 layers [40], [41] which form the extraction mechanism of the convolutional neural network are employed for the feature extraction from the inputted image layers, the extracted features are subjected to the color descriptor which is used to obtain useful colorimetric information in the process of handling color images. The obtained feature map is passed to the region proposal network (RPN) to produce regions of interest (ROI) which are afterward selected by the ROIAlign layer in correspondence to the feature map (serving as ground truth to generate intersection over union) based on the RPN's output. The sizes of the feature map are fixed before they are sent to the fully connected layer (FCN) for the object class, bounding box, and mask predictions. For every cow in the training datasets, ground truth was manually annotated followed by the training of the network after labeling for optimization of parameters.

Table 1 shows the network hyperparameters for the model.

C. EXTRACTION OF KEY FRAME FROM IMAGE ROWS

In the image rows, cattle were tracked to know the ones that are not actively in form, that is, the ones that are not healthily in form due to disease infection or another. This alone, resulted in collecting datasets of repeatable images.

This is in addition to the welfare-based evaluation where the breeders were demanding frames with cattle movement and position changing. This is to guard against the low efficiency of the whole system by not applying the same method to each frame. So, to minimize using the images that are collected in multiple places, and concentrating on the information of key motion for the overall efficiency of the system, each frame in the image rows is categorized as a keyframe and non-key frame.

Keyframes are the frames with motion images of cattle by and large containing relevant health information for behavioral evaluation.

D. IMAGE ENHANCEMENT USING GCFD

GCFD is applied in the image enhancement to get the edges of the cow based on the color images produced from every frame in the captured video which is later used in describing the features of the cow for robust cow characterization after

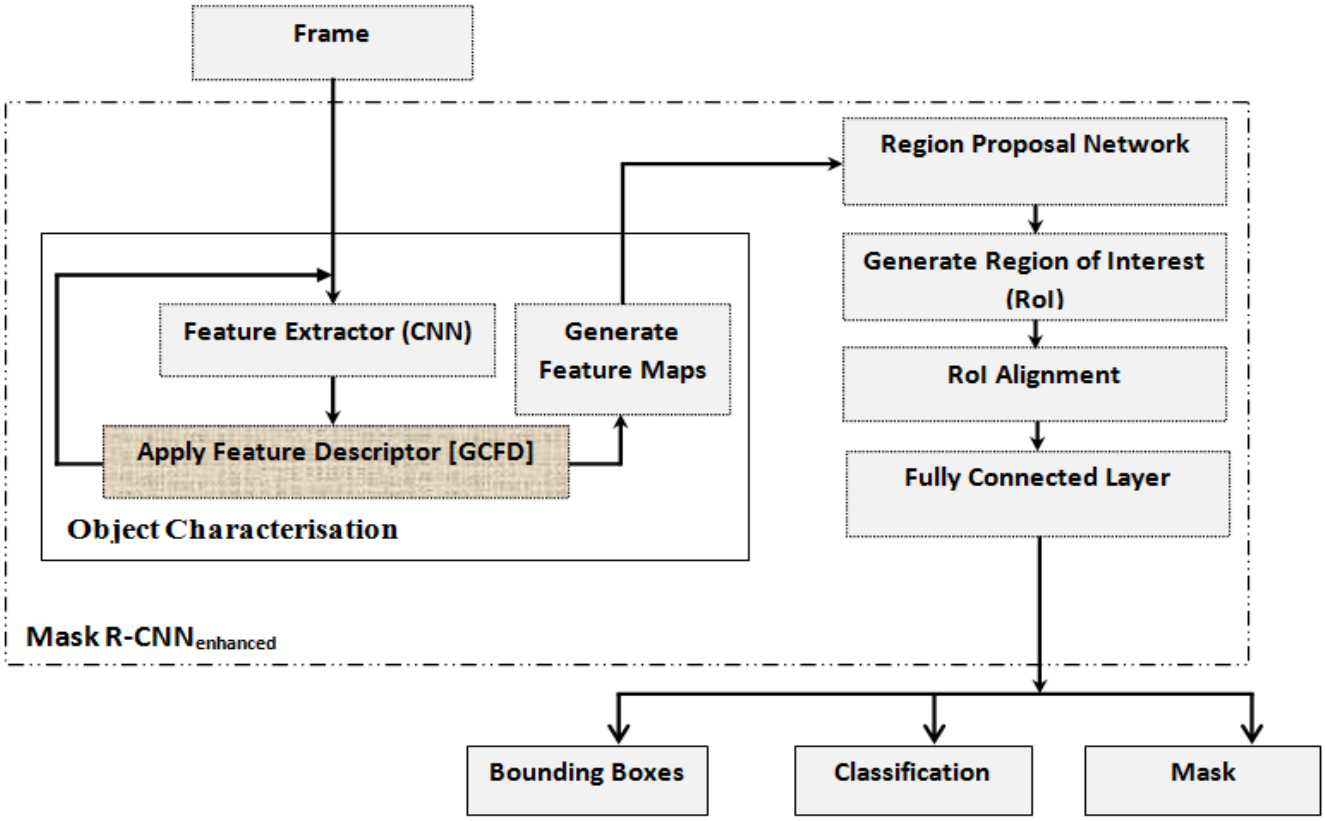

FIGURE 3. Image of the proposed model.

segmentation. The descriptor helps in improving the detected curves of the image patches without modifying the color (contrast, brightness, and saturation) of the patches. To produce GCFD, the ideal method is to divide the image color into each particular channel of color which are red (r), green (g), and blue (b). Each of these channels is computed and their results are three sets of descriptors. These descriptors are produced by employing the combination of two descriptors which are computed from both parallel and orthogonal projections in a 2 dimension-fast Fourier transform (FFT). The formula which defines how these two projections are combined is as follows [42]:

$$\mathrm{GCFD_B}(f) = \{\mathrm{GCFD}_{\|\mathrm{B}}(f) + \mathrm{GCFD}_{\perp_\mathrm{B}}(f)\} \quad (1)$$

where $\mathrm{GCFD_B}(f)$ = Computation of GCFD, $\mathrm{GCFD}_{\|B}(f)$ = GCFD in parallel part, and $\mathrm{GCFD}_{\perp_\mathrm{B}}(f)$ = GCFD in orthogonal part.

GCFD corresponds to the computation of classical generalized Fourier descriptors (CGFD) on the Clifford Fourier transform's parallel and orthogonal part. About the parallel part, GCFD computes on the red channel while the chromatic plane of green and blue is computed on the orthogonal part. When a segmented cow image is processed using the GCFD feature vector, a vector of 16 doubles is produced. The first and the ninth values are very high because they represent the first descriptor of GCFD in parallel GCFD$\|$B (f) and GCFD in orthogonal GCFD$^{\perp}$B (f) in that order. Using a fast Fourier transform, parallel values and orthogonal values are obtained respectively, and the values are combined to generate a set of feature maps representing the image itself.

E. MASK R-CNN INSTANCE SEGMENTATION NETWORK

Mask R-CNN is known for its flexibility as an instance segmentation method and is built on the success of Faster R-CNN by incorporating an additional branch called an image segmentation mask. Fig. 3 illustrates the reconstructed architecture of Mask R-CNN [16] for instance segmentation. The implementation of Mask R-CNN follows the adoption of the procedure with two stages where the first stage is identical to the Faster R-CNN's RPN. The second stage involves Mask R-CNN outputting a binary mask for each region of interest in parallel to predicting the offset of the class and the box. This whole process is different from what is obtainable in most of the recent systems where mask prediction dictates classification [38].

The approach in the model follows the success of Fast R-CNN which applies in parallel the classification of bounding-box and regression.

1) REGION PROPOSAL NETWORK AND LOSS FUNCTION

The network for region proposal carries out convolution operation on the pixel sliding window of feature maps generated from the convolutional neural layers. There is a selection of k anchors with different aspect ratios and scales for each center

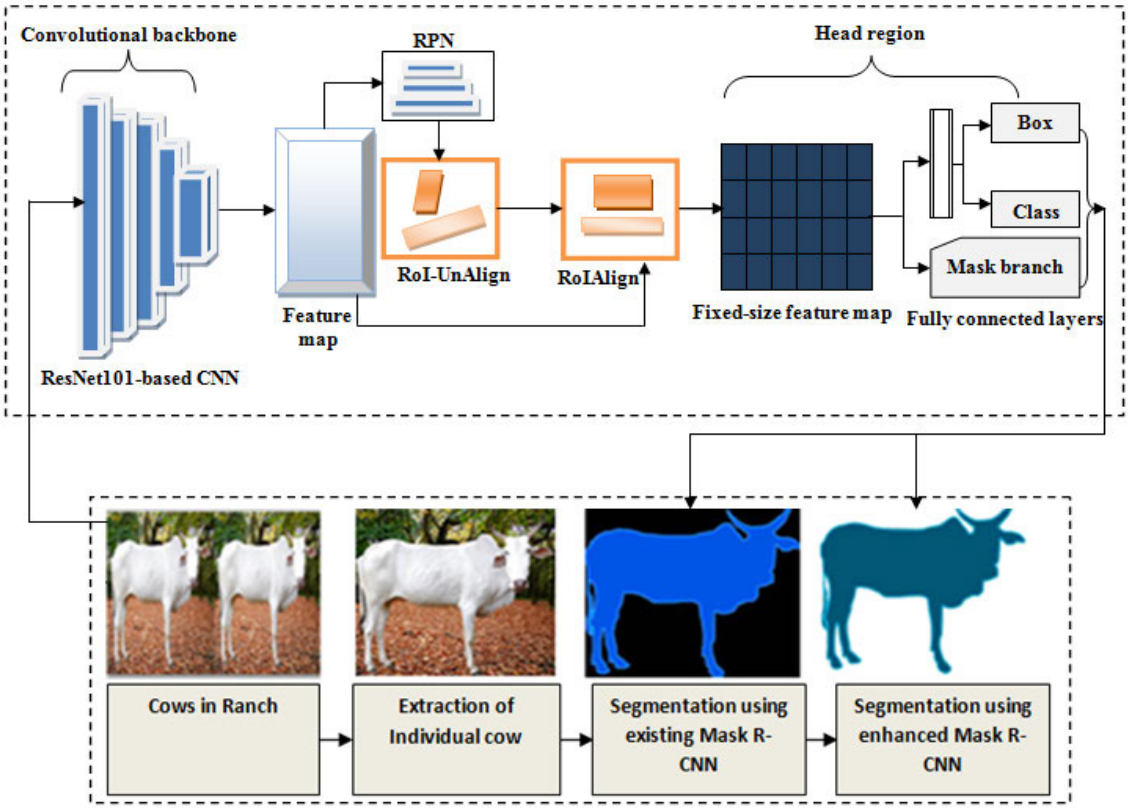

FIGURE 4. The structural framework of the enhanced Mask R-CNN for cattle image instance segmentation.

point in the feature map, after which the selected k anchors are mapped to the original feature maps resulting in a huge amount of region proposals. The individual points found in the feature maps are for the production of feature codes for the corresponding window regions which are corresponded to the low-dimensional feature codes in Mask R-CNN which are performed by a convolution operation. The difference between the value predicted and the ground truth is indicated by the loss function in the training of the network. Moreover, the roles played by loss function in the training of model for cattle segmentation are very important. In our proposed enhanced Mask R-CNN for cattle instance segmentation, a combined loss function is employed in training the regression of the bounding box, and object class prediction including the mask prediction branch. The loss function employed for this task is according to the following equation:

$$L = L_{ce} + L_{be} + L_{me} \tag{2}$$

where L is the loss function, L_{ce} is the classification error, L_{be} indicates the regression error of the bounding box, and L_{me} is the mask error.

2) ENHANCED MASK R-CNN FOR COW CHARACTERIZATION

The work in this study focuses on the segmentation of individual cattle object from an image without any patches or background objects using the instance segmentation method. The proposed enhanced Mask R-CNN for cattle instance segmentation is carried out by combining the algorithm of Grabcut with the Mask R-CNN for patches-free instance segmentation and contour extraction of the detected individual cattle from an image. The algorithm of instance segmentation finds a representation mask for every image's object in contrast to the algorithm of semantic (class) segmentation which precisely represents objects' classes at a pixel-wise level without distinguishing between two objects of the same class.

Enhanced Mask R-CNN is an improvement on Mask R-CNN, known for its flexibility as an instance segmentation method and built on the success of Faster R-CNN by incorporating an additional branch called segmentation mask. Fig. 4 illustrates the structural framework of the enhanced Mask R-CNN cattle image instance segmentation. The framework comprises two components: (1) The backbone component that is responsible for convolutional operation such as extraction of features over the complete image; (2) the head component that is responsible for performing object class, bounding-box, and mask predictions. The network of RPN is responsible for computing the region proposals passed to it by GCFD-described feature map and the computed features are matched with the feature map by the ROIAlign (Region of Interest Alignment) layer before sending to the fully connected network through the fixed-size feature map for simultaneous operations of predicting the cattle object class, refining bounding box and generating robust segmentation mask.

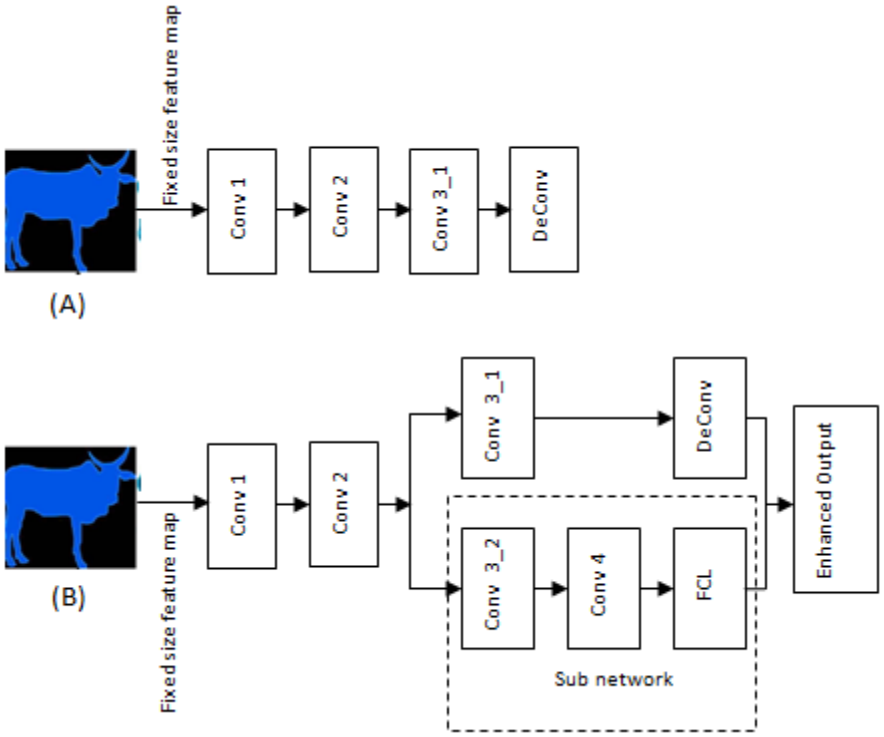

FIGURE 5. (A) Existing Mask R-CNN (B) Enhanced Mask R-CNN for cattle image instance segmentation.

Since feature pyramid network (FPN) utilizes both semantic information and resolution feature maps at a very high level for perfect localization, an FPN that is based on the ResNet101 network is employed in this study as the model backbone to realize profitable accomplishment in both the speed and accuracy [43], whereas Faster R-CNN with ResNet101 is responsible for the head structural design. CNN is employed first in the task of segmenting and extracting features of cattle from an image using ResNet101 which is later passed on to the descriptors (GCFD) for gainful colorimetric information generation, the generated feature maps are passed on to the RPN which employs the approach of sliding window on the generated feature maps to process and generate regions of interest in the form of bounding box proposals. The arbitrarily size spatial interest regions in the features are mapped to a fixed size spatial resolution by ROIAlign using interpolation that is bilinear in form. As a final point, the class object, the bounding box, and the mask predictions are simultaneously performed by the Mask R-CNN head component.

3) MODEL DEVELOPMENT

Fig. 5 and Fig. 8 show the proposed method as an extension of the existing Mask R-CNN. In the proposed method, we have enhanced the existing Mask R-CNN by (1) providing an optimal filter size that was smaller than ResNet (the backbone of Mask R-CNN) for the extraction of smaller and composite features, thereby, the number of parameters required for the training was decreased, and this led to increased in the computation efficiency; (2) utilizing multiscale semantic features using region proposals and (3) integrating a fully connected layer of existing Mask R-CNN with a sub-network for enhanced segmentation. Fig. 5 (A) shows the existing Mask R-CNN and Fig. 5 (B) shows the enhanced Mask R-CNN for cattle image instance segmentation.

4) AN ABRIDGED MODEL OF RESNET

The activation layers in the backpropagation process that are skipped in ResNet are the major reasons for most of its issues. This is due to the unavailability of a formula to describe the changing process that takes place in the ResNet parameters which in turn leads to inaccuracy of the gradient formula. Moreover, there is no clarification of the layer that deserves more training over another in the training process.

An abridge model of ResNet was proposed to solve these issues using the algorithm of backpropagation. Based on the formula of the new gradient, new rules were obtained that are made of different parameters for ResNet. The obtained rules provide optimal filter size that was smaller than ResNet for the extraction of smaller and composite features, thereby, the number of parameters required for the training was decreased, and this led to an increase in the computation efficiency. Fig. 6, Fig. 7, and Fig. 8 show the architecture of ResNet, the block of ResNet, and the enhanced architecture of ResNet respectively.

The enhanced performance of ResNet was attained by employing a deep network that has a set of blocks to handle the issues of gradient vanishing [16]. Fig. 8 shows the ResNet

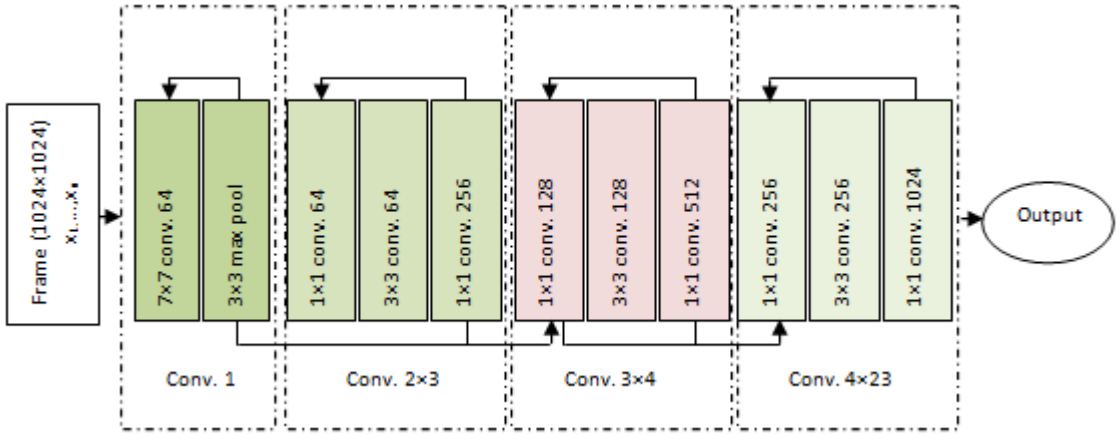

FIGURE 6. Architecture of ResNet.

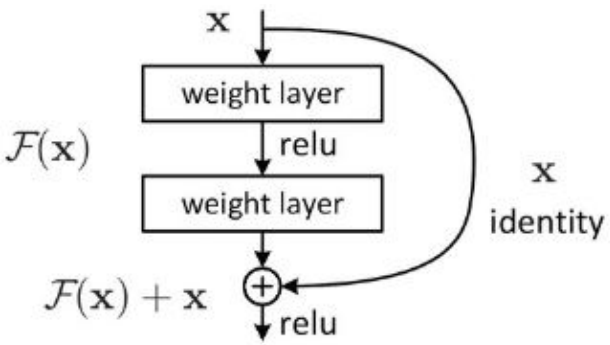

FIGURE 7. Block of ResNet [16].

block, and the formula for establishing the two-layer block is presented as follows:

$$H(\mathrm{x}) = F(\mathrm{x}, \{W_i\}) + \mathrm{x} \tag{3}$$

where,

x = Building block input.

H(x) = Building block output vectors.

$F(\mathrm{x}, \{W_i\})$ = The learned residual mapping in the training process.

Based on Fig. 8, training for convolutional layers with the best block was carried out as enumerated below:

(a) 1 repetition for the 1st convolution block.

(b) 4 repetitions for the 2nd convolution block.

(c) 4 repetitions for the 3rd convolution block.

(d) 14 repetitions for the 4th convolution block.

5) THE ENHANCED MASK R-CNN STRUCTURE

The structure of the enhanced Mask R-CNN model is divided into three separate branches. The first branch called network backbone is used for feature extraction and generation of ROI. This branch consists of ResNet101+RPN+Feature Pyramid Network (FPN). Multi-scale feature maps were generated in this branch before mapping each of its points to the input image for the acquisition of matching ROI. The second branch called ROIalign (region of interest alignment) is used for pooling the generated ROIs from the first branch to fixed-size feature maps in order to overcome any form of quantization error. The third branch is used for generating a mask. All the fixed-size feature maps from the second branch pass through the fully connected layer (FCL) to generate a cow object mask in addition to bounding box regression and cow object classification. The above three modules are illustrated in Fig. 4.

6) GRABCUT IMAGE SEGMENTATION

Grabcut provides the avenue for cutting edge segmentation method which helps in removing heterogeneous objects from an image while the homogenous objects are retained.

That is, there is a clear distinction between the foreground object and background object during and after segmentation. The interactive method of Grabcut facilitates the involvement of the graphical segmentation approach and maximum flow technology [44]. The cow object model and the graph background are substituted by the Grabcut with the fusion fast Fourier transform (FFT) of the three sets of descriptors, namely red-green-blue (RGB) three-channel, and there is a division of the image by incessantly separating the estimation of the segmentation and the interactive iteration of the model parameter learning.

Algorithm 1 shows the algorithm of Grabcut that produces Fig. 4.

IV. IMPLEMENTATION

Keras Python library with a graphic processing unit is employed for the construction, optimization, and evaluation of the model designed for the cattle instance segmentation.

Python is chosen as the programming language because the efficiency of its code and all-inclusive support of its algorithms allow the importing model of ResNet101 and utilization of data flow graphs to carrying out the calculation in deep learning. The hardware and software information employed for the implementation are shown in Table 2. All the cow species constitute the entire dataset with careful study to recognize individual cow characteristics, each cow with

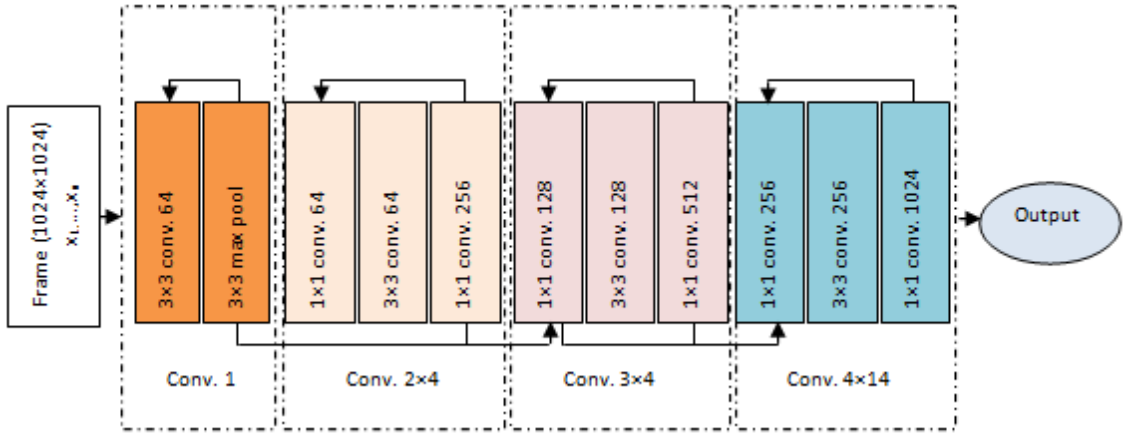

FIGURE 8. The enhanced architecture of ResNet.

Algorithm 1 Algorithm of Grabcut

Grabcut$_{adapt}$: Incorporating structural mapping with Grabcut

Step 1: Initialization
Load Mask R-CNN generated image
Function Grabcut
Apply direct box selection for the target
Select the box with the target
Select all the pixels outside the box as the background pixel
Select all the pixels inside the box as the potential target
for each pixel 'n' on the outside **do**
  Initialize the label of the pixel 'N' for the background pixel
  for each pixel 'n' on the inside **do**
    Initialize the label of the pixel 'N' for potential target
    if $\alpha_n = 1$
      Get some pixels that belong to the target
    else if $\alpha_n = 0$
      Get some pixels that belong to the background
    end If
Step 2: Minimization of iteration
Function structuremapping
  Assignthe FFT component to each pixel
  for a given image data **do**
    Optimize the parameters of the FFT
    Split estimation
    Repeat the steps of minimization of iteration
    Smooth the segmented boundary and other post-processing using border mat
  end for

100 images resulting in 1000 images in entirety. 400 images, that is, 4 cows' subject × 100 images of each subject were employed for the training of the network in the training phase. 600 images, that is, 6 cows' subject × 100 images of each subject were employed for testing the network in the testing phase. By the middle of September 2019, a practicality test was carried out to mine the image data, and analysis of the mined image data was performed consequently by image processing.

The format of the raw images extracted from the videos was in Joint Photography Expert Graphics (JPEG) format at 5120 by 3840 pixels which was later reduced in pixels using MATLAB to increase the processing speed and overcome unnecessary overfitting during the network training. Different keyframes are chosen for both the training datasets and the testing datasets with a ratio of 2 to 1. The area of the cattle was manually labeled using the LabelMe tool [45] to generate ground truth to meet the requirement of model training which lasted a whole day with a 0.001 learning rate. Also, to not destroy the convolutional layers extraction ability, all the layers of the backbone component were stationary leaving only the network head for independent training by employing the training dataset in each case. Each ROI loss comprised of the following losses: loss of classification, loss of bounding box, and loss of mask. All these losses occurred during the training of the network, however, the loss of mask only exists in positive regions of interest. Therefore, if the intersection over union (IOU) between the ROI and the ground truth yielded at least a convincing threshold, the outcome is assigned positive, or else it is assigned negative.

V. RESULTS AND DISCUSSION

Fig. 11 shows the comparison of the enhanced Mask R-CNN instance segmentation and the contour extraction experiment (Fig. 4) performed in this study to state-of-the-art methods. Fig. 4 clearly illustrates the transformation of the cattle image having passed through the process of instance segmentation and contour extraction. The observation in Fig. 4 is the complete removal of heterogeneous objects and the patches caused by illumination variation, and the spatial contour layout of the cattle's body is completely extracted from the image, unlike what is obtained in some other instance segmentation tasks.

The essence of instance segmentation is to detect and get the object classification using a bounding box; this technique

TABLE 2. Software and hardware requirements.

Software	Type/Version	Hardware	Type/Version
Operating system	64-bit Windows 10	CPU	Intel Core i5 processor@2.4GHz
IDE	Visual studio 2019	RAM	16 Gigabytes
Python library	Keras	Graphics card	GeForce GTX 1080 Ti
MATLAB	R2019b	Hard-disk	2 Terabytes
		Camera module	Vision Datum LEO 640H-200gc High-Speed 200fps Sharp RJ33 CCD Gigabit Ethernet 3d
		Monitor	10.1 inch IPS HD Portable LCD Gaming Monitor PC display VGA HDMI interface for PS3/PS4/XBOx360/CCTV/ Camera

TABLE 3. Results of instance segmentation for cow images (best outputs are in bold).

Operation on cow image	Method	Data type	mAP	Time (s)
Instance segmentation	Mask R-CNN	Raw	0.90	0.73
		Enhanced	0.92	0.73
	Enhanced Mask R-CNN	Raw	0.92	0.75
		Enhanced	**0.93**	**0.75**

TABLE 4. Results of contour extraction for cow images (best outputs are in bold).

Operation on cow image	Minimum	Maximum	ADE	Time (s)
Contour extraction using Mask R-CNN	0.035	64.17	33.56	0.77
Contour extraction using enhanced Mask R-CNN	0.029	61.19	**30.46**	**0.71**

is extended in Mask R-CNN with the addition of mask segmentation branch for object parallel prediction.

The employment of the Grabcut algorithm complements the SLAM algorithm employed in Mask R-CNN to produce patches-free cow image segmentation and contour extraction. The beauty in the enhanced Mask R-CNN is that the GCFD that was used in enhancing the image reduces the time-wasting effects of the ResNet101 algorithm [46] during the conversion of the image from $1024 \times 1024 \times 3$ (RGB) to $32 \times 32 \times 2048$ feature maps, and mitigates the negative effects of variation in illumination during the cattle image capture exercise thereby reducing the misjudgment of pixels between the actual cattle body and the shadows. The relationship graph between the cluster's number and the sum of squared within-cluster (SSWC) is shown in Fig. 9, and this helps in selecting the optimal number of clusters ($k = 30$) where the change in SSWC starts to level off.

The enhanced Mask R-CNN cattle instance segmentation method is invariant to the external influence of any objects that may possess similarity in color with the coat patterns of the experimental cattle. Both the accuracy recorded for the segmentation and the time it takes to process the whole extraction are in Table 3 and Table 4 respectively. A union function of the cattle segmented without Grabcut and the cattle segmented using the enhanced Mask R-CNN instance segmentation method is employed for measuring the accuracy of the image segmentation process. The accuracy of the segmentation process is measured by:

$$\text{Accuracy} = \frac{\text{A} \cap \text{B}}{\text{A} \cup \text{B}} \times 100 \tag{4}$$

where A refers to the bounding box of the target object and B refers to the bounding box of the ground truth.

To evaluate the contour line that is extracted, the average distance error is calculated as follows:

$$\text{ADE} = \frac{\text{A}_{\text{union}} - \text{A}_{\text{overlap}}}{\text{T}_{\text{contour}}} \tag{5}$$

where A_{union} is the region embraced by both ground truth and predicted mask, $\text{A}_{\text{overlap}}$ is the region overlapped amongst the ground truth and predicted mask, and $\text{T}_{\text{contour}}$ is the ground-truth contour line's pixel number.

The accuracy of the enhanced Mask R-CNN instance segmentation method on the enhanced image datasets is approximately 93% with 0.75 seconds processing time, and the enhanced image datasets generate roughly 1% accuracy more than the image datasets that are not enhanced. This simply implies that there is a great improvement in the quality of image during the cattle segmentation task with the involvement of enhanced Mask R-CNN and the result is an improvement over the Mask R-CNN and MaskSplitter [32] cattle instance segmentation methods. The contour of individual cattle can be easily extracted considering the segmentation results of the enhanced Mask R-CNN-based method. The enhanced image datasets are employed for the extraction experiment of the cattle's contour since the segmentation results of the enhanced image datasets are superior to the dataset of the raw images.

As shown in Fig. 4, the result of contour extraction using the enhanced Mask R-CNN segmentation method is similar to the actual contour of the experimental cattle which is preferred to what is obtained using Mask R-CNN [47]

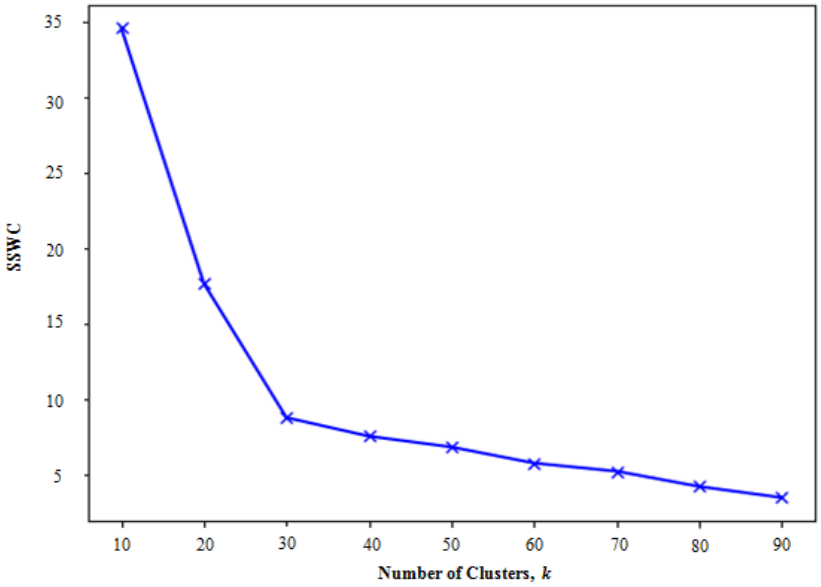

FIGURE 9. The elbow method for determining the optimal k value based on SSWC.

TABLE 5. Metrical results of the segmentation of Muturu cow images using Mask R-CNN$_{enhanced}$-model 1.

	AP@ 0.50	AP@ 0.55	AP@ 0.60	AP@ 0.65	AP@ 0.70	AP@ 0.75	AP@ 0.80	AP@ 0.85	AP@ 0.90	AP@ 0.95
1st Cow	0.94	0.90	0.86	0.80	0.59	0.49	0.45	0.31	0	0
2nd Cow	0.91	0.89	0.89	0.83	0.60	0.57	0.52	0.40	0	0
3rd Cow	0.88	0.87	0.87	0.72	0.69	0.58	0.53	0.41	0	0
4th Cow	0.87	0.92	0.90	0.79	0.66	0.57	0.51	0.40	0	0
5th Cow	0.92	0.88	0.84	0.75	0.67	0.59	0.53	0.45	0	0
mAP	0.90	0.89	0.87	0.78	0.64	0.56	0.51	0.39	0	0

where the extracted contour is having background color (binary mask). The more the dataset, the more the improvement in segmentation and contour extraction performance, so, the underperformance experienced in some parts of this study is a result of less access to training datasets. The difference in pixel length (center errors) [48], which is the measurement of the difference between the predicted object and the ground-truth object, calculated as average distance error (ADE) of the extracted contour (Equation 5) is shown in Table 4 where 30.46 ADE of the extracted contour was gained by the enhanced Mask R-CNN instance segmentation method making it appreciably better than what is obtained in [32], [47].

ImageNet [49], MS COCO dataset [50], and Pascal VOC [51] are the most commonly employed datasets in object detection research, and competition involving object detection. However, the proposed model, being an extension of the Mask R-CNN model with pre-trained Microsoft Common Objects in COntext (MS COCO) weights, employed MS COCO datasets. The datasets contain more than 80 different classes with over 250 thousand data of different settings made available as datasets for training and validation. Cow dataset from the MS COCO datasets has over 2071 images whereby 1986 images were apportioned as a training dataset and 87 images were apportioned as validation and testing dataset.

The own cow dataset was employed for the model implementation and validation, while both the MS COCO cow dataset and the own cow dataset were used for the testing with their results presented in Tables 3-12. Two models, namely model 1 and model 2 were developed in addition to the existing Mask R-CNN to test and compare the accuracy of the models. Table 5 shows the metrical results of the segmentation of Muturu cow images using the first developed model (model 1), Table 6 shows the metrical results of the segmentation of Muturu cow images using the second developed model (model 2), Table 7 shows the metrical results of the segmentation of Muturu cow images using the existing model (Mask R-CNN), Table 8 shows the metrical results of the segmentation of Keteku cow images using the first developed model (model 1), Table 9 shows the metrical results of the segmentation of Keteku cow images using the second developed model (model 2), Table 10 shows the metrical results of the segmentation of Keteku cow images using the existing model (Mask R-CNN), Table 11 shows the metrical results of the segmentation of MS COCO cow images using the developed model. Table 12 shows the precision-recall (mAP) of the developed model 1 and model 2, and the existing model at different IOU threshold values. The implications of these results are analyzed as follows: at 0.50 IOU value, both developed model 1 and model 2 produced mAP of 0.90 each less than 0.92 that what was obtained by the existing model

TABLE 6. Metrical results of the segmentation of Muturu cow images using Mask R-CNN$_{enhanced}$-model 2.

	AP@ 0.50	AP@ 0.55	AP@ 0.60	AP@ 0.65	AP@ 0.70	AP@ 0.75	AP@ 0.80	AP@ 0.85	AP@ 0.90	AP@ 0.95
1st Cow	0.94	0.89	0.86	0.80	0.59	0.49	0.45	0	0	0
2nd Cow	0.91	0.88	0.85	0.83	0.60	0.57	0.52	0	0	0
3rd Cow	0.88	0.87	0.87	0.72	0.69	0.58	0.53	0	0	0
4th Cow	0.87	0.91	0.90	0.79	0.66	0.57	0.51	0	0	0
5th Cow	0.92	0.86	0.82	0.75	0.67	0.59	0.53	0	0	0
mAP	0.90	0.88	0.86	0.78	0.64	0.56	0.51	0	0	0

TABLE 7. Metrical results of the segmentation of Muturu cow images using Mask R-CNN.

	AP@ 0.50	AP@ 0.55	AP@ 0.60	AP@ 0.65	AP@ 0.70	AP@ 0.75	AP@ 0.80	AP@ 0.85	AP@ 0.90	AP@ 0.95
1st Cow	0.90	0.89	0.86	0.82	0.59	0.49	0.47	0	0	0
2nd Cow	0.92	0.88	0.85	0.83	0.60	0.57	0.53	0	0	0
3rd Cow	0.91	0.87	0.87	0.85	0.69	0.58	0.53	0	0	0
4th Cow	0.94	0.93	0.90	0.84	0.66	0.57	0.52	0	0	0
5th Cow	0.93	0.93	0.82	0.86	0.67	0.59	0.53	0	0	0
mAP	0.92	0.90	0.86	0.84	0.64	0.56	0.52	0	0	0

TABLE 8. Metrical results of the segmentation of Keteku cow images using Mask R-CNN$_{enhanced}$-model 1.

	AP@ 0.50	AP@ 0.55	AP@ 0.60	AP@ 0.65	AP@ 0.70	AP@ 0.75	AP@ 0.80	AP@ 0.85	AP@ 0.90	AP@ 0.95
1st Cow	0.97	0.68	0.76	0.76	0.64	0.49	0.46	0	0	0
2nd Cow	0.95	0.73	0.80	0.81	0.60	0.57	0.49	0	0	0
3rd Cow	0.95	0.72	0.83	0.82	0.74	0.55	0.52	0	0	0
4th Cow	0.94	0.69	0.82	0.80	0.65	0.55	0.50	0	0	0
5th Cow	0.98	0.73	0.83	0.81	0.72	0.54	0.53	0	0	0
mAP	0.96	0.71	0.81	0.80	0.67	0.54	0.50	0	0	0

TABLE 9. Metrical results of the segmentation of Keteku cow images using Mask R-CNN$_{enhanced}$-model 2.

	AP@ 0.50	AP@ 0.55	AP@ 0.60	AP@ 0.65	AP@ 0.70	AP@ 0.75	AP@ 0.80	AP@ 0.85	AP@ 0.90	AP@ 0.95
1st Cow	0.97	0.79	0.76	0.68	0.64	0.49	0.46	0.31	0	0
2nd Cow	0.95	0.80	0.81	0.73	0.60	0.57	0.49	0.39	0	0
3rd Cow	0.95	0.84	0.82	0.72	0.74	0.55	0.52	0.31	0	0
4th Cow	0.94	0.82	0.80	0.69	0.65	0.55	0.50	0.39	0	0
5th Cow	0.98	0.85	0.81	0.73	0.72	0.54	0.53	0.44	0	0
mAP	0.96	0.82	0.80	0.71	0.67	0.54	0.50	0.37	0	0

TABLE 10. Metrical results of the segmentation of Keteku cow images using Mask R-CNN.

	AP@ 0.50	AP@ 0.55	AP@ 0.60	AP@ 0.65	AP@ 0.70	AP@ 0.75	AP@ 0.80	AP@ 0.85	AP@ 0.90	AP@ 0.95
1st Cow	0.93	0.92	0.90	0.80	0.59	0.49	0.45	0	0	0
2nd Cow	0.91	0.88	0.88	0.83	0.60	0.57	0.52	0	0	0
3rd Cow	0.89	0.89	0.87	0.72	0.69	0.58	0.53	0	0	0
4th Cow	0.88	0.93	0.90	0.79	0.66	0.57	0.51	0	0	0
5th Cow	0.92	0.90	0.90	0.75	0.67	0.59	0.53	0	0	0
mAP	0.91	0.90	0.89	0.78	0.64	0.56	0.51	0	0	0

TABLE 11. Metrical results of the segmentation of MS COCO cow images using Mask R-CNN$_{enhanced}$.

	AP@ 0.50	AP@ 0.55	AP@ 0.60	AP@ 0.65	AP@ 0.70	AP@ 0.75	AP@ 0.80	AP@ 0.85	AP@ 0.90	AP@ 0.95
1st Cow	0.93	0.94	0.91	0.82	0.90	0.57	0.53	0	0	0
2nd Cow	0.87	0.92	0.90	0.83	0.89	0.59	0.51	0	0	0
3rd Cow	0.88	0.93	0.88	0.85	0.88	0.58	0.53	0	0	0
4th Cow	0.93	0.90	0.89	0.84	0.92	0.49	0.52	0	0	0
5th Cow	0.89	0.91	0.92	0.86	0.87	0.57	0.47	0	0	0
mAP	0.90	0.92	0.90	0.84	0.89	0.56	0.51	0	0	0

when tested on the Muturu cow images. At 0.50 IOU value, when both developed model 1 and model 2 were tested on the Keteku cow images, each model achieved mAP of 0.96. However, the existing model produced mAP of 0.91 from the same dataset at the same 0.50 IOU value. When the developed model was tested on the MS COCO cow images, mAP of 0.90 was achieved with 0.02 less than what was obtained at 0.55 IOU value. These analyses and the summary in Table 12 show that the own datasets (Muturu and Keteku) performed better than the MS COCO dataset, this is not unconnected to the MS COCO too many iconic cow images difficult to train our model with. Also, the proposed model performs better than the existing model in terms of accuracy and speed as shown in Table 3, Table 4, and Table 12.

TABLE 12. Precision-recall (mAP) of Mask R-CNN$_{enhanced}$ and Mask R-CNN at different IOU threshold values.

Model	Cow object	AP @									
		0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
Mask R-CNN$_{enhanced}$ (Model 1)	Muturu	0.90	0.89	0.87	0.78	0.64	0.56	0.51	0.39	0	0
	Keteku	0.96	0.71	0.81	0.80	0.67	0.54	0.50	0	0	0
mAP		**0.93**	0.80	0.84	0.79	0.66	0.55	0.51	0.20	0	0
Mask R-CNN$_{enhanced}$ (Model 2)	Muturu	0.90	0.88	0.86	0.78	0.64	0.56	0.51	0	0	0
	Keteku	0.96	0.82	0.80	0.71	0.67	0.54	0.50	0.37	0	0
mAP		**0.93**	0.85	0.83	0.75	0.66	0.55	0.51	0.19	0	0
Mask R-CNN	Keteku	0.91	0.90	0.89	0.78	0.64	0.56	0.51	0	0	0
	Muturu	0.92	0.90	0.86	0.84	0.64	0.56	0.52	0	0	0
mAP		0.92	0.90	0.88	0.81	0.64	0.56	0.52	0	0	0

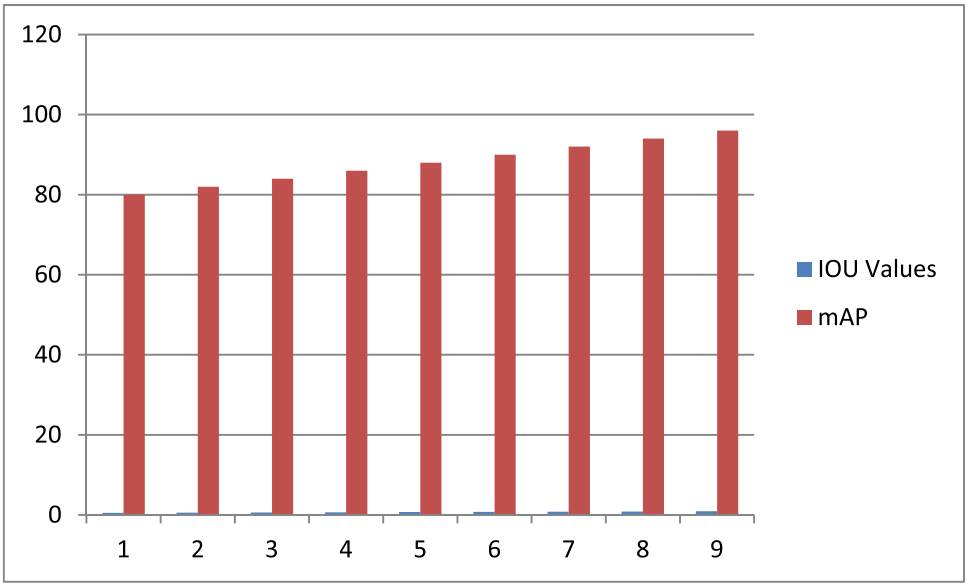

FIGURE 10. Precision-recall (mAP) graph of Mask R-CNN$_{enhanced}$ and Mask R-CNN at different IOU threshold values.

TABLE 13. Comparison of enhanced Mask R-CNN with state-of-the-art methods (best mAP output is in bold).

Segmentation method	Backbone network	mAP
Mask R-CNN [16]	ResNet101	0.92
MaskSplitter [32]	VGG16	0.71
Fully Convolutional Instance-aware Semantic Segmentation (FCIS) [38]	ResNer101-C5-dilated	0.56
Faster R-CNN [43]	ResNer101-FPN	0.90
YOLO v2 [52]	DarkNet19	0.91
Mask Single Shot Detector (Mask SSD) [53]	ResNet101-FPN-B6	0.82
DeepMask [33]	VGGNet [55]	0.53
SharpMask [34]	VGGNet [55]	0.82
Multi-task Network Cascades (MNC) [54]	ResNet101-C4	0.42
Enhanced Mask R-CNN (Proposed)	ResNet101	**0.93**

Fig. 10 shows the precision-recall (mAP) graph of Mask R-CNN$_{enhanced}$ and Mask R-CNN at different IOU threshold values. Fig. 11 shows the comparison of the output of the enhanced Mask R-CNN with the state-of-the-art methods.

A. PERFORMANCE EVALUATION

The performance of the proposed enhanced Mask R-CNN for cattle instance segmentation and extraction of the contour is evaluated using mean average precision (mAP) of intersection over union (IOU) which is defined as the area of intersection by the area of the union of a predicted object bounding box and a ground-truth object bounding box (Equation 4), and average distance error (Equation 5) which is used for measuring the distance between the mapped cow contour lines and the actual cow contour lines in terms of pixel length (center errors), and also for the measurement of the difference between the predicted object and the ground-truth object. The mAP is a very common tool for measuring average precision in image segmentation evaluation. Together with average distance error, mAP is employed in this study as the metrics

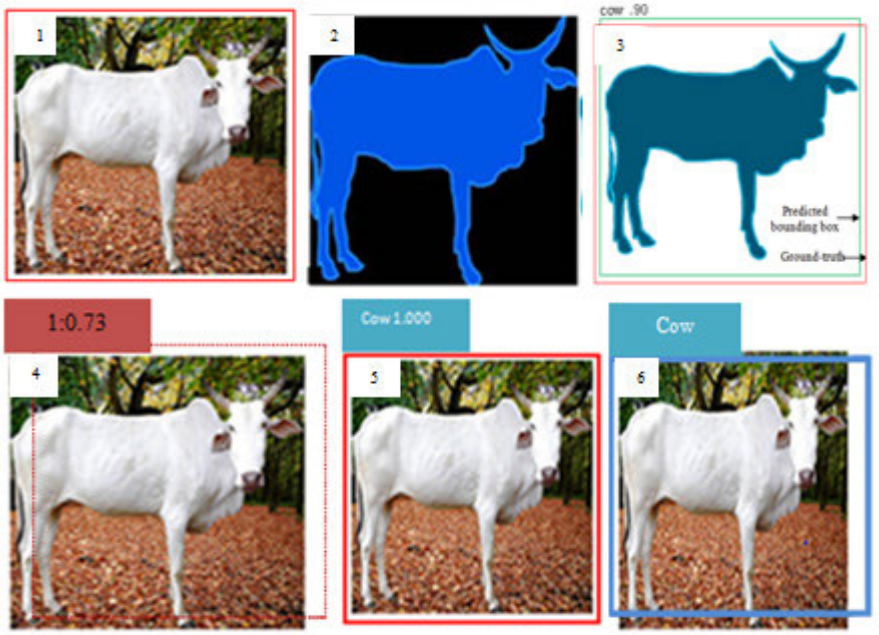

FIGURE 11. (1) Cow image with the ground truth (2) Cow image segmentation using Mask R-CNN (3) Cow image segmentation using enhanced Mask R-CNN (4) Cow image segmentation using SSD (5) Cow image segmentation using Faster R-CNN (6) Cow image segmentation using YOLOv2.

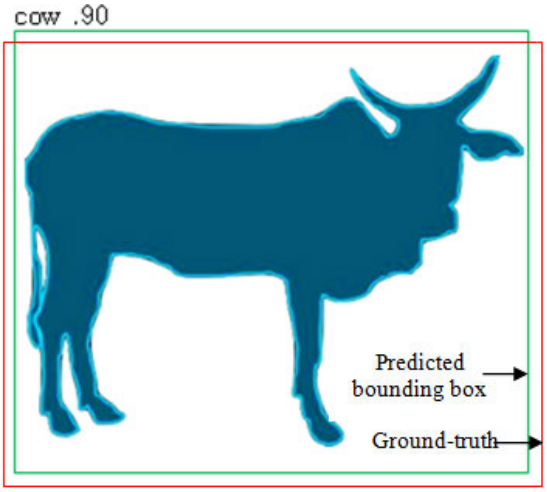

FIGURE 12. Enhanced Mask R-CNN showing the bounding boxes (ground truth & predicted cow object), the class label, the prediction confidence, and the mask of the detected cow object.

for evaluation as presented in Tables 3-12. Fig. 12 shows the bounding boxes (ground truth & predicted cow object), the class label, the prediction confidence, and the mask of the detected cow object.

B. COMPARISON OF ENHANCED MASK R-CNN WITH STATE-OF-THE-ART METHODS

In furtherance of evaluating the performance of the proposed enhanced Mask R-CNN instance segmentation method, a comparison was made between the proposed approach and state-of-the-art instance segmentation methods namely MaskSplitter [32], Fully Convolutional Instance-aware Semantic Segmentation (FCIS) [38], Faster R-CNN [43], YOLO v2 [52], Mask Single Shot Detector (Mask SSD) [53], DeepMask [33], SharpMask [34], Multi-task Network Cascades (MNC) [54], and Mask R-CNN [16] as shown in Table 13. MaskSplitter and Mask SSD comprise the fully convolutional network, and the mask of ground truth for refining the cattle mask representation in images without the prediction of a bounding box.

MaskSplitter and Mask SSD make use of neither bounding boxes nor RPNs. The MaskSplitter framework is trained to learn how to output mask representations of three different types, namely two bad and one good. The framework is comprised of the algorithm that dictates the mask representations type and the number of true cow objects predicted; loss functions of Euclidean and pixel-wise sigmoid; and a set of fully-connected layers and convolutional neural networks, one for every prediction's type. On the other hand, using Mask SSD, instance segmentation is added on single-stage detectors, which helps in detecting image objects while at the same time producing for each instance, a segmentation mask.

However, the mAP results for MaskSplitter and Mask SSD instance segmentation methods show less accuracy than the Mask R-CNN from which the enhanced Mask R-CNN inherits all its merits.

The enhanced Mask R-CNN is capable of generating gainful colorimetric information in the course of features extraction while at the same time generating a superiority segmentation mask for every instance. It also ensures structural mapping and extraction of cow contour from an image without any patches or background. Moreover, the time and mean

average precision of instance segmentation were significantly improved using the enhanced model.

VI. CONCLUSION

A practically and efficiently enhanced Mask R-CNN instance segmentation framework for contour extraction of individual cattle from an image has been proposed in this study. There are three main enhancements on the proposed enhanced Mask R-CNN: (1) provision of optimal filter size that was smaller than ResNet101 (the backbone of Mask R-CNN) for the extraction of smaller and composite features, thereby, the number of parameters required for the training was decreased; (2) utilization of multiscale semantic features using region proposals and (3) a fully connected layer of existing Mask R-CNN integrated with a sub-network for enhanced segmentation. Moreover, the enhanced Mask R-CNN achieved 0.93 mAP when evaluated, and the method demonstrated accurate simultaneous localization and mapping.

While the domain of instance segmentation of cow image is well studied, we acknowledge only two previous studies that have made use of Mask R-CNN for this similar idea [39], [47]. Those studies did not utilize our proposed methods in their work. We intend to improve on training the network to learn how to predict individual masks separately such that the segmentation of overlapping regions and explicit differentiation of body parts of cattle objects will be possible.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENT

The authors received funding from the Division of Research and Innovation (RCMO), Research University Grant (1001 / PKOMP / 8014001) and School of Computer Sciences, Universiti Sains Malaysia for this publication.

REFERENCES

[1] T. Van Hertem, L. Rooijakkers, D. Berckmans, A. P. Fernández, T. Norton, D. Berckmans, and E. Vranken, "Appropriate data visualisation is key to precision livestock farming acceptance," *Comput. Electron. Agricult.*, vol. 138, pp. 1–10, Jun. 2017.

[2] S. Kumar, A. Pandey, K. S. R. Satwik, S. Kumar, S. K. Singh, A. K. Singh, and A. Mohan, "Deep learning framework for recognition of cattle using muzzle point image pattern," *Measurement*, vol. 116, pp. 1–17, Feb. 2018.

[3] R. Bello, A. Z. Talib, and A. S. A. Mohamed, "Deep learning-based architectures for recognition of cow using cow nose image pattern," *Gazi Univ. J. Sci.*, vol. 33, no. 3, pp. 831–844, 2020.

[4] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 25, pp. E5716–E5725, Jun. 2018.

[5] W. E. Petersen, "The identification of the bovine by means of nose-prints," *J. Dairy Sci.*, vol. 5, no. 3, pp. 249–258, May 1922.

[6] J. M. Bos, B. Bovenkerk, P. H. Feindt, and Y. K. van Dam, "The quantified animal: Precision livestock farming and the ethical implications of objectification," *Food Ethics*, vol. 2, no. 1, pp. 77–92, Dec. 2018.

[7] T. T. Zin, I. Kobayashi, P. Tin, and H. Hama, "A general video surveillance framework for animal behavior analysis," in *Proc. 3rd Int. Conf. Comput. Meas. Control Sensor Netw. (CMCSN)*, May 2016, pp. 130–133.

[8] N. C. Lynn, T. T. Zin, and I. Kobayashi, "Automatic assessing body condition score from digital images by active shape model and multiple regression technique," in *Proc. Int. Conf. Artif. Life Robot., Miyazaki*, 2017, pp. 311–314.

[9] A. Nasirahmadi, S. A. Edwards, and B. Sturm, "Implementation of machine vision for detecting behaviour of cattle and pigs," *Livestock Sci.*, vol. 202, pp. 25–38, Aug. 2017.

[10] M. F. Hansen, M. L. Smith, L. N. Smith, K. A. Jabbar, and D. Forbes, "Automated monitoring of dairy cow body condition, mobility and weight using a single 3D video capture device," *Comput. Ind.*, vol. 98, pp. 14–22, Jun. 2018.

[11] S. F. Tebug, A. Missohou, S. S. Sabi, J. Juga, E. J. Poole, M. Tapio, and K. Marshall, "Using body measurements to estimate live weight of dairy cattle in low-input systems in senegal," *J. Appl. Animal Res.*, vol. 46, no. 1, pp. 87–93, Jan. 2018.

[12] A. L. Zhang, B. P. Wu, C. T. Wuyun, D. X. Jiang, E. C. Xuan, and F. Y. Ma, "Algorithm of sheep body dimension measurement and its applications based on image analysis," *Comput. Electron. Agricult.*, vol. 153, pp. 33–45, Oct. 2018.

[13] P. Tang, C. Wang, X. Wang, W. Liu, W. Zeng, and J. Wang, "Object detection in videos by high quality object linking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1272–1278, May 2020.

[14] R. W. Bello, A. Z. Talib, A. S. A. Mohamed, D. A. Olubummo, and F. N. Otobo, "Image-based individual cow recognition using body patterns," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 92–98, 2020.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[17] L. Neumann, A. Zisserman, and A. Vedaldi, "Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection," in *Proc. NIPS Workshop Mach. Learn. Intell. Transp. Syst.*, 2018, pp. 1–8.

[18] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "MaskLab: Instance segmentation by refining object detection with semantic and direction features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4013–4022.

[19] A. Fathi, Z. Wojna, V. Rathod, P. Wang, H. O. Song, S. Guadarrama, and K. P. Murphy, "Semantic instance segmentation via deep metric learning," 2017, *arXiv:1703.10277*. [Online]. Available: http://arxiv.org/abs/1703.10277

[20] J. Gardenier, J. Underwood, and C. Clark, "Object detection for cattle gait tracking," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2206–2213.

[21] X. Song, E. A. M. Bokkers, P. P. J. van der Tol, P. W. G. G. Koerkamp, and S. van Mourik, "Automated body weight prediction of dairy cows using 3-dimensional vision," *J. Dairy Sci.*, vol. 101, no. 5, pp. 4448–4459, May 2018.

[22] A. Arnab and P. H. S. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 441–450.

[23] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[24] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[25] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 437–446.

[26] Z. Hayder, X. He, and M. Salzmann, "Shape-aware instance segmentation," 2016, *arXiv:1612.03129*. [Online]. Available: https://arxiv.org/abs/1612.03129

[27] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother, "InstanceCut: From edges to instances with MultiCut," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5008–5017.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[29] K. Li, B. Hariharan, and J. Malik, "Iterative instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3659–3667.

[30] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 247–251.

[31] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5221–5229.
[32] A. Ter-Sarkisov, R. Ross, J. Kelleher, B. Earley, and M. Keane, "Beef cattle instance segmentation using fully convolutional neural network," 2018, *arXiv:1807.01972*. [Online]. Available: http://arxiv.org/abs/1807.01972
[33] P. O. Pinheiro, R. Collobert, and P. Dollár, "Learning to segment object candidates," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1990–1998.
[34] P. O. Pinheiro, T. Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 75–91.
[35] A. Fuentes, S. Yoon, J. Park, and D. S. Park, "Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information," *Comput. Electron. Agricult.*, vol. 177, pp. 1–11, Oct. 2020.
[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[37] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.
[38] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2359–2367.
[39] B. Xu, W. Wang, G. Falzon, P. Kwan, L. Guo, G. Chen, A. Tait, and D. Schneider, "Automated cattle counting using mask R-CNN in quadcopter vision system," *Comput. Electron. Agricult.*, vol. 171, pp. 1–12, Apr. 2020.
[40] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 297–312.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[42] J. Mennesson, C. Saint-Jean, and L. Mascarilla, "Color object recognition based on a Clifford Fourier transform," in *Guide to Geometric Algebra in Practice*. London, U.K.: Springer, 2011, pp. 175–191.
[43] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[44] X. Wu, S. Wen, and Y. A. Xie, "Improvement of mask-RCNN object segmentation algorithm," in *Proc. Int. Conf. Intell. Robot. Appl.*, 2019, pp. 582–591.
[45] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, May 2008.
[46] N. F. F. Alshdaifat, A. Z. Talib, and M. A. Osman, "Improved deep learning framework for fish segmentation in underwater videos," *Ecol. Informat.*, vol. 59, pp. 1–11, Sep. 2020.
[47] Y. Qiao, M. Truman, and S. Sukkarieh, "Cattle segmentation and contour extraction based on Mask R-CNN for precision livestock farming," *Comput. Electron. Agricult.*, vol. 165, pp. 1–9, 2019.
[48] M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2137–2155, Nov. 2016.
[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.
[51] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[52] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[53] H. Zhang, Y. Tian, K. Wang, W. Zhang, and F.-Y. Wang, "Mask SSD: An effective single-stage approach to object instance segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 2078–2093, 2020.
[54] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3150–3158.
[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

ROTIMI-WILLIAMS BELLO received the B.Tech. degree in mathematics and computer science from FUTMINNA, Nigeria, the M.Tech. degree in computer science from FUTA, Nigeria. He is currently pursuing the Ph.D. degree with the School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia. His research interests include vision and image processing, computer security and cryptography, machine learning, and data mining.

AHMAD SUFRIL AZLAN MOHAMED received the BIT degree (Hons.) from Multimedia University, Malaysia, the M.Sc. degree from the University of Manchester, U.K., and the Ph.D. degree from the University of Salford, U.K. He is currently with the School of Computer Sciences Universiti Sains Malaysia, Pulau Pinang, Malaysia. His research interests include image processing, video tracking, facial recognition, and medical imaging.

ABDULLAH ZAWAWI TALIB received the B.Sc. degree from Bradford, U.K., the M.Sc. degree from Newcastle Upon Tyne, U.K., and the Ph.D. degree from Wales, U.K. He is currently with the School of Computer Sciences Universiti Sains Malaysia, Pulau Pinang, Malaysia. His research interests include computer graphics and visualization, geometric computing, and scientific computing.