Detection of one-horned rhino using multispectral images

Simantika Choudhury[1], Amlan Jyoti Das[2], Navajit Saikia[3], Subhash Chandra Rajbongshi[4]

[1]Department of Electronics and Communication Engineering, Gauhati University,
Guwahati, India,
csimantika@gauhati.ac.in

[2]Obaforta India Pvt. Ltd.,
Guwahati, India,
amlan@spinaanalytica.com

[3] Department of Electronics and Telecommunication Engineering, Assam Engineering College,
Guwahati, India,
navajit.ete@aec.ac.in

[4] Department of Electronics and Communication Engineering, Gauhati University,,
Guwahati, India,
subhash73@gauhati.ac.in

Abstract*: *Animal detection and surveillance is an important field of research to address the needs for protection of endangered species among others. The challenges in animal detection include low-contrast and poor image quality which is commonly observed during night time. Researchers have mostly worked on day-light, low-contrast and thermal images. To handle the challenges of detection during night time, multispectral images in combination with deep architectures may be used for better detection performance. In the present work, one-horned rhino is considered for detection because they are getting endangered for reasons like poaching, natural calamities and diseases. A novel multispectral one-horned rhino dataset is introduced and the multispectral data is obtained by combining the channels of color images and the corresponding thermal images. Instance segmentation based techniques YOLACT and YOLACT++ are used here to detect rhinos with the above multispectral dataset. The performances of the detectors are studied in terms of mAP and FPS.

***Keywords*: animal detection, multispectral, instance segmentation, YOLACT, one-horned rhino.**

(Article history: Received: 7th March2023 and accepted 17th August 2023)

I. INTRODUCTION

Detection is an important topic of research in computer vision and the requirement of animal detection is necessary due to various reasons like protecting endangered animals, preventing collision of animals on roads, animal security, etc. One-horned rhino is one of those animals which are getting endangered because of brutal human activity such as poaching and it can be prevented by monitoring wildlife. Researchers have already worked on animal detection using computer vision based techniques. As time approached, deep learning based networks became popular and gave promising performance in animal detection [1]. It is found that most of the literatures used animal datasets which contains daylight [2], low-contrast images [3], etc. The challenge in animal detection is mostly observed during night time due to poor image quality as well as darkness. Also, identifying the uniqueness of the detected animals present in a particular scene is quite difficult. In order to mitigate these challenges, deep learning architectures based on instance segmentation may be used to get richer and meaningful output in combination with multispectral dataset to give better performance in different illumination conditions.

This work focuses on a deep learning approach for the purpose of segmenting and detecting one-horned rhino mostly during night time which may help in protection and surveillance of the rhinos. The methodology is based on instance segmentation based deep networks using multispectral rhino images of different variations. Instance segmentation is a combination of object detection and semantic segmentation. It gives both bounding-box and mask output. The bounding box helps in localizing the object of interest and the mask helps in providing pixel-level information for that particular object. The advantage of instance segmentation is that it gives an additional feature of classifying every individual object in an image which leads to finer inference output and meaningful information. A labeled multispectral rhino dataset is introduced in this work which contains two channels: color and thermal. Multispectral data is used because it contains information of both the channels. The color image will provide the texture and surrounding information, whereas, the thermal image will provide attention to the object of interest. This multispectral data is then used as input to the single-stage instance segmentation techniques to detect and segment out rhinos. The single-stage techniques take speed into consideration so that it may be used in real time applications.

II. RELATED WORKS

Object detection provides localization of image objects which are classified and semantic segmentation provides finer inference by pixel-level prediction of each object class. These two computer vision tasks are evolved and combined

together to a deep learning based technique called instance segmentation. It creates different labels for separate instances of objects belonging to same class which solves the problem of object detection and semantic segmentation simultaneously. Some instance segmentation based algorithms are introduced by researches which gives better performance in accuracy as well as speed. Mask R-CNN [4] is a two-stage approach to generate instance masks. It adds an additional layer to the Faster R-CNN [5] object detection network which generates binary masks. BlendMask [6] is a fully convolutional instance segmentation method which is more efficient and faster compared to other two-stage techniques because they are simpler. It uses low-level granularity information from semantic segmentation along with instance level information. YOLACT [7] generates instance masks by linearly combining prototype masks with the mask coefficients. YOLACT++ [8] introduces deformable convolutional networks to the backbone architecture of YOLACT to provide better quality prototype masks.

Detection in nature scene is quite a challenging task. Computer vision tasks mostly use color and gray channels, but, in order to handle the challenges, only using these channels might not be always helpful. Hence, there is a need to explore and move outside the visible spectrum such as thermal and near-infrared images. In [9], multispectral data is used for detecting multiple objects using a YOLOv3 detector. Researchers also worked with multispectral satellite data for object recognition using FCN [10]. In [11], power equipment fault detection is done using multispectral data and instance segmentation technique SOLOv2.

Marine animal detection is done in [12] using multispectral data and traditional detection methods. In [13], dog skin diseases are classified using multispectral images and deep learning approaches. In [2], one-horned rhino is detected by a YOLOv3 object detector for daylight color images. In [3], low-contrast color images are used to detect rhinos. A GAN based enhancement technique known as EnlightenGAN is used as a preprocessing step before giving the images as input to the object detector. In our work, one-horned rhino is considered for automatic detection using multispectral images in combination with deep learning based instance segmentation algorithms.

The main contributions of this work are summarized in the following:

i. Two popular instance segmentation based techniques, namely, YOLACT [7] and YOLACT++ [8], are used for detection and segmentation of one-horned rhino. These network architectures are trained and tested by tuning the hyper-parameters in order to improve detection performance.
ii. A multispectral dataset of one-horned rhinos from North-East India is proposed which consists of RGB and thermal images. To the best of our knowledge, this is the first multispectral dataset that contains greater one- horned rhino for automatic detection.
iii. The dataset is annotated (mask/ pixel-level annotations) in COCO format (.json) with the help of a labeling tool known as Labelme [14].

The rest of the paper is organized as follows. The overview of the theories related to the system model is presented in Section III. Section IV describes the proposed multispectral one-horned rhino dataset and its characteristics. In Section V, the proposed methodology of the animal detection system and its implementation details is explained. Section VI presents the experimental analysis. Finally the paper is concluded in Section VII.

III. Related Theories

The theories related to this work are briefly discussed here in this section.

A. *YOLACT*

It is an instance segmentation based framework which forgoes an explicit localization step and can be used in real time. It breaks up instance segmentation into two parallel tasks: (a) some sets of prototype masks are generated over an entire image, (b) predicts mask coefficients for instances. These two tasks are then linearly combined to get the final instance segmented output which includes detection of bounding box as well as mask. The network architecture of YOLACT is shown in Fig. 1.

B. *YOLACT++*

The network architecture of YOLACT++ is similar to YOLACT, only a few additional modifications are done. The differences between these two architectures are: (a) deformable convolutions are added to the backbone network to produce good quality and precise prototype masks, (b) fast mask re-scoring network is added which re-ranks the masks based on their quality, and (c) better choices of anchors are made to increase recall value.

IV. Proposed Multispectral One-Horned Rhino Dataset

Previously proposed one-horned rhino datasets were based on daylight [2] and low-contrast images [3]. Here, a multispectral one-horned rhino imagery dataset is introduced which contains color and thermal images. The images are captured during low-light and night conditions using thermal imagers, namely, TESTO868 (320x240 resolution) and TESTO872 (640x480 resolution). These images are resized to a fixed dimension of 640x360.

The dataset is categorized into two classes: ‘rhino’ and ‘group of rhinos’. The dataset contains a total of 594 color images and their corresponding thermal images. These images are then divided into training set, validation set and testing set. The training set contains 500 images, the validation set contains 62 images and the testing set contains 32 images. These images are annotated manually using Labelme tool. Polygon masks are generated for each annotation and the output is saved in COCO (.json) format.

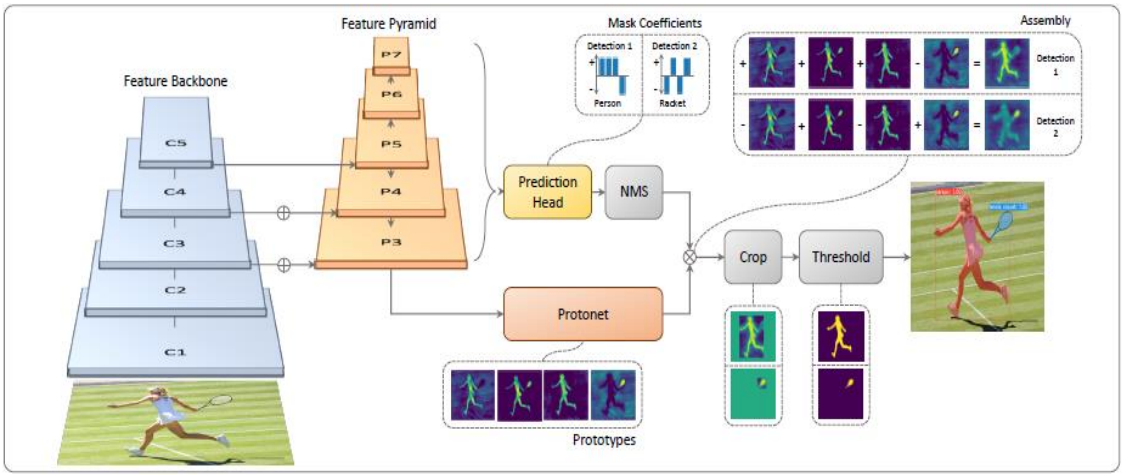

Fig. 1. Network Architecture of YOLACT[7]

Some characteristics of this multispectral one-horned rhino dataset are as follows:

i. It contains images of rhinos present in cluttered and occluded environment.
ii. It contains images of rhinos from various perspectives like lateral, frontal and rear views.
iii. It contains images of rhinos with different activities like drinking, sitting, sleeping, walking, etc.
iv. It contains color and thermal images in different illumination conditions.

Some examples of the dataset are shown in Fig. 2. Here, color image and its corresponding thermal image are shown in different illumination conditions. It may be seen that in night time, the color image is completely dark and the presence of rhino and human is not visible through naked eye, whereas, in the corresponding thermal image, the objects are clearly visible.

V. PROPOSED METHODOLOGY

This section discusses the proposed methodology of this work. The multispectral dataset is used for training an instance segmentation model to detect one-horned rhinos. Figure 3 shows the block diagram of the methodology.

Fig. 2. Dataset samples in different illumination conditions (upper row: RGB/color image, bottom row: corresponding thermal image)

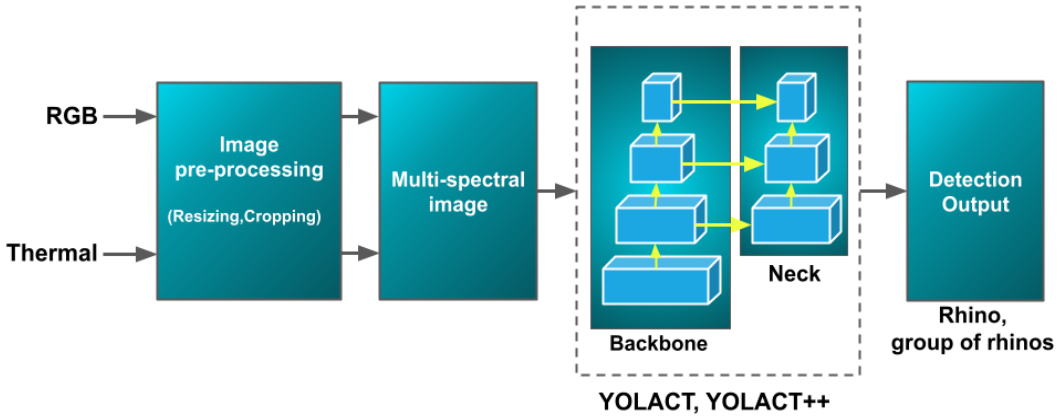

Fig. 3. Block diagram of the proposed methodology

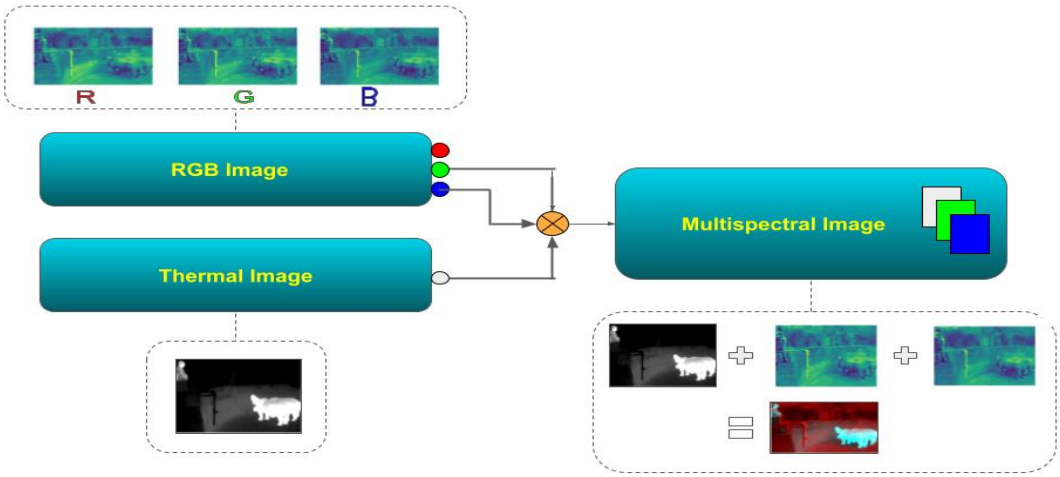

Fig. 4. Multispectral data generation

The steps involved are summarized below:

i. The raw color and thermal images from the thermal imagers are resized to a fixed dimension of 640x360.

ii. The resized images are then cropped to remove parallax effect of the color image and its corresponding thermal image.

iii. Multispectral data is generated using these pre-processed color and thermal images by channel combination method. The red channel from the color image is replaced with the thermal image to get the multispectral output. Fig. 4. gives the pictorial representation of generation of the multispectral data.

iv. The multispectral data is then used for training YOLACT and YOLACT++ instance segmentation models to finally detect and segment 'rhinos' and 'group of rhinos'.

A. *Implementation Details*

The models are trained using a ResNet based pre-trained weights for transfer learning. The hyper-parameters used for training are selected based on trial-based optimization. While training, the initial learning rate used is 0.001 and later on, it is reduced to $5e^{-4}$. The optimizer used is SGD (Stochastic Gradient Descent) with a momentum of 0.9. The loss functions used are: classification loss, box regression loss and binary cross entropy loss. The networks are trained for 2500 epochs with a batch size of 8.While testing, the confidence threshold is set to 0.001, the IoU (Intersection over Union) thresholds of 0.5 and 0.85 are selected for both bounding box-level and mask-level. The rest of the parameters during testing are similar to those considered during training. All the networks are trained and tested using

a single NVIDIA GeForce RTX 2080 GPU of 11 GB memory size. The implementations are done using Python (version 3.7) and Pytorch (version 1.6) environments. The following section will provide the experimental analysis of the results achieved.

VI. EXPERIMENTAL RESULTS

To evaluate the instance segmentation based model architectures, two evaluation metrics are taken into consideration: mAP (mean average precision) for bounding box and masks, and FPS (Frames per second). The mAP will give the average precision for both the classes and FPS will measure the detection speed. Higher values of mAP and FPS indicate better detection performance and speed. Higher FPS also means that the model may be considered to be used in real time applications.

Table 1 shows the performance of the instance segmentation models tested on our multispectral dataset for bounding boxes as well as masks. The two models, namely, YOLACT and YOLACT++, are also compared with Mask R-CNN which is also an instance segmentation algorithm but it is two-staged. It seems from the table that YOLACT is performing the best in terms of mAP at 0.5 and 0.85 IoU for both bounding box and mask. It achieves 88.47% mAP at 0.5 IoU and 86.78% mAP at 0.5 IoU for bounding box. Again, for mask-level, it achieves 88.60% mAP at 0.5 IoU and 80.85% mAP at 0.85 IoU. But, it is also seen that YOLACT++ is better in terms of speed. It achieves FPS of 19.42 whereas YOLACT achieves FPS of 18.71.

TABLE I. DETECTION PERFORMANCE OF NETWORK ARCHITECTURES

Models		**mAP at 0.5 IoU**	**mAP at 0.85 IoU**	**FPS**
Mask R-CNN [4]	Bounding box-level	70.11%	56.51%	5.39
	Mask-level	68.74%	47.66%	
YOLACT [7]	Bounding box-level	**88.47%**	**86.78%**	18.71
	Mask-level	**88.60%**	**80.85%**	
YOLACT++ [8]	Bounding box-level	77.09%	60.22%	**19.42**
	Mask-level	77.31%	51.73%	

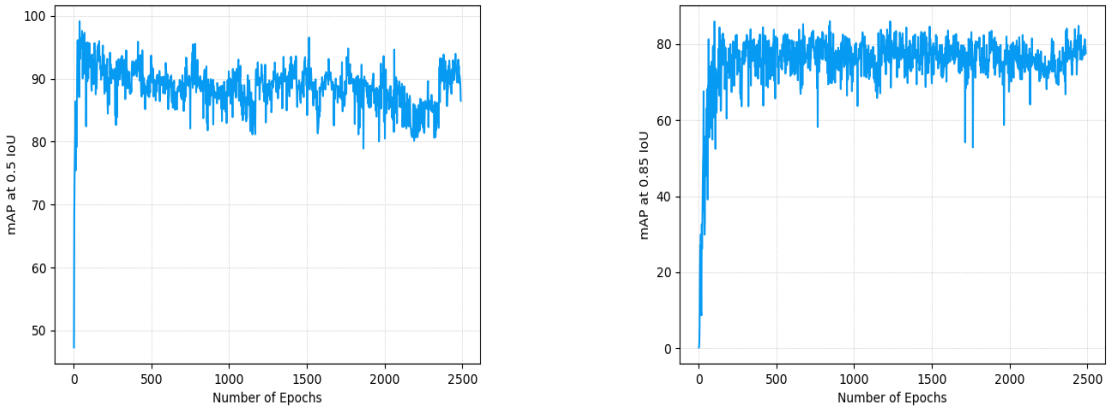

Fig.5. Epoch vs mAP plot at 0.5 (left) and 0.85 (right) IoU for bounding box

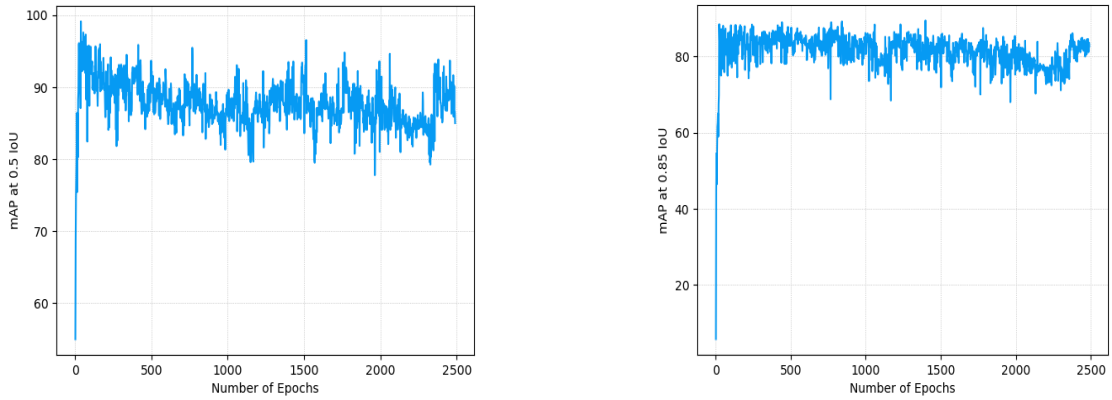

Fig.6. Epoch vs mAP plot at 0.5 (left) and 0.85 (right) IoU for mask

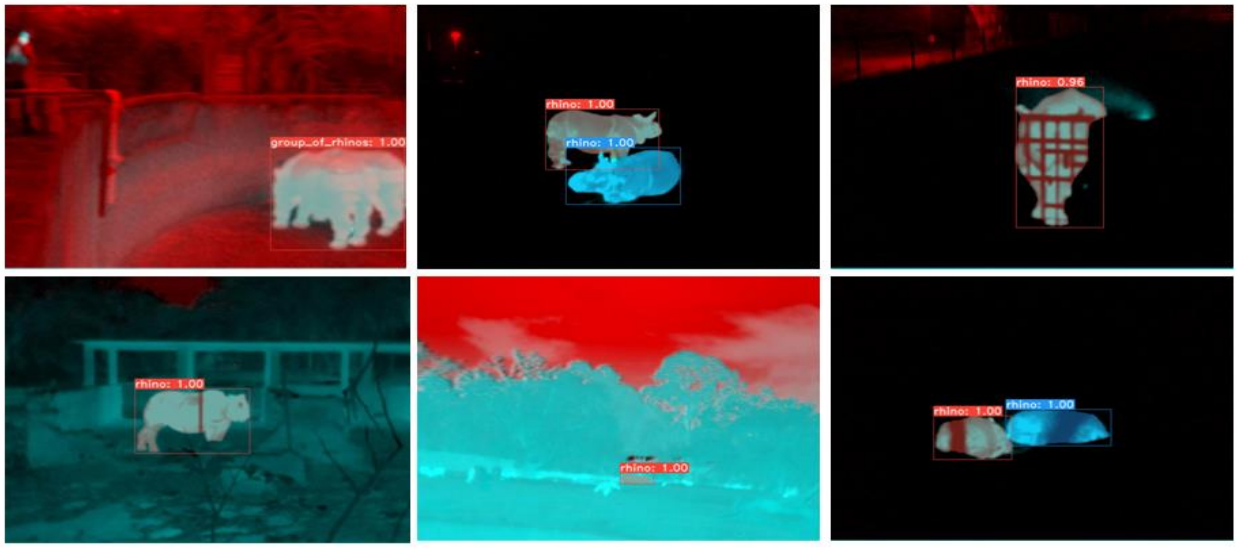

Fig.7. Detection outputs of YOLACT architecture

The mAP values for YOLACT++ are a little lower than YOLACT because of the selection of backbone architecture. In YOLACT, ResNet backbone network is used which seems to perform well for smaller size dataset. The addition of deformable convolutions and mask re-scoring network in YOLACT++ gives improvement in terms of speed for our dataset. These two models are compared with Mask-RCNN and it is seen that both YOLACT and YOLACT++ are quite faster and also achieves better mAP values. The epoch vs mAP plots (at 0.5 and 0.85 IoU) for bounding box and mask are shown in Figure 5 and Figure 6. Some detection outputs of the YOLACT architecture is shown in Fig. 7.

VII. CONCLUSION

This paper used instance segmentation based deep architectures YOLACT and YOLACT++ for detection and segmentation of one-horned rhino in different illumination conditions. A multispectral one-horned rhino dataset was proposed here which consists of color and thermal images. The dataset was manually labeled with pixel-level annotations using a labeling tool. The deep networks were hyper-tuned empirically to achieve better detection performance. The performances of the models were evaluated based on two metrics: mAP and FPS. YOLACT performed best in terms of mAP and YOLACT++ performed best in terms of speed which may be used in real world applications. It can be said that the overall performance of YOLACT is best compared to the other networks because it is maintaining a tradeoff between detection performance and speed. It can also be concluded that single-staged instance segmentation techniques perform better than two-stage network like Mask R-CNN in terms of speed as well as detection. In future, there is a need for increasing the size of the multispectral dataset for efficient use in deep learning models. Addition of more animal classes will also enhance the performance of the model. Some recent instance segmentation and transformer based algorithms may be used to further improve animal detection systems in real-time.

REFERENCES

[1] S. Choudhury, N. Saikia and A. J. Das, “Recent Trends in Learning Based Techniques for Human and Animal Detection”, in Joint National Conference on Emerging Computing Technologies & its Applications (NCECTA 2019), April, 2019, PSG College of Technology, Coimbatore, Tamil Nadu, India.

[2] S. Choudhury, N. Bharti, N. Saikia and S. Rajbongshi, “Detection of One-horned Rhino from Green Environment Background using Deep Learning”, Journal of Green Engineering, vol. 10, pages 4657-4678, September, 2020.

[3] S. Choudhury, N. Saikia, S. Rajbongshi and A. Das, “Employing generative adversarial network in low light animal detection”, In: Kumar, S., Hiranwal, S., Purohit, S.D., Prasad, M. (eds) Proceedings of International Conference on Communication and Computational Technologies . Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-19-3951-8_75.

[4] K. He, G. Gkioxari, P. Dollár and R. Girshick, “Mask R-CNN”, *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks”, In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15). MIT Press, Cambridge, MA, USA, 91–99, 2015.

[6] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang and Y. Yan, "BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 8570-8578, doi: 10.1109/CVPR42600.2020.00860.

[7] D. Bolya, C. Zhou, F. Xiao and Y. J. Lee, "YOLACT: Real-Time Instance Segmentation," 2019 IEEE/CVF International Conference on

Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 9156-9165, doi: 10.1109/ICCV.2019.00925.

[8] D. Bolya, C. Zhou, F. Xiao and Y. J. Lee, "YOLACT++ Better Real-Time Instance Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 2, pp. 1108-1121, 1 Feb. 2022, doi: 10.1109/TPAMI.2020.3014297.

[9] M.O. Gani, S. Kuiry, A. Das, M. Nasipuri and N. Das, "Multispectral Object Detection with Deep Learning", arXiv e-prints, 2021. doi:10.48550/arXiv.2102.03115.

[10] P. Gudžiu, O. Kurasova, V. Darulis, "Deep learning-based object recognition in multispectral satellite imagery for real-time applications", Machine Vision and Applications 32, 98 (2021). https://doi.org/10.1007/s00138-021-01209-2.

[11] J. Shu, J. He, L. Li, and T. Reddy G, "MSIS: Multispectral Instance Segmentation Method for Power Equipment", In Intell. Neuroscience 2022, https://doi.org/10.1155/2022/2864717.

[12] J. Lopez, J. Schoonmaker and S. Saggese, "Automated detection of marine animals using multispectral imaging," 2014 Oceans - St. John's, St. John's, NL, Canada, 2014, pp. 1-6, doi: 10.1109/OCEANS.2014.7003132.

[13] S. Hwang, H.K. Shin, J.M. Park, "Classification of dog skin diseases using deep learning with images captured from multispectral imaging device", Mol. Cell. Toxicol. 18, 299–309 (2022). https://doi.org/10.1007/s13273-022-00249-7

[14] B. C. Russell, A. Torralba, K. P. Murphy, "LabelMe: A Database and Web-Based Tool for Image Annotation", In International Journal Computer Vision, 77, 157–173 (2008). https://doi.org/10.1007/s11263-007-0090-8.

AUTHOR PROFILE

Simantika Choudhury has completed her Bachelor of Engineering in Electronics and Telecommunication Engineering from Gauhati University. She has also completed her Master of Technology in Signal Processing and Communication from Gauhati University. She is currently pursuing her Ph.D. in Electronics and Communication Engineering, Gauhati University. Her research areas include image processing, computer vision, machine learning and deep learning.

Amlan Jyoti Das has completed his Ph.D. from Electronics and Communication Engineering, Gauhati University. He is currently working as a Data Scientist in Obaforta India Pvt. Ltd. His areas of interest are computer vision, machine learning and deep learning.

Navajit Saikia is presently working as an Associate Professor in the Department of Electronics and Telecommunication Engineering of Assam Engineering College. Signal Processing and Communications are two areas of his teaching interests. His research interests include image processing, speech processing, information security and reversible logic. He has published several research papers in journals and conference.

Subhash Chandra Rajbongshi completed his M.Tech. and Ph.D. degree from Gauhati University, Assam, India. He is currently working as Scientific Ocer at Gauhati University. His area of interest includes computer vision, image processing, signal processing etc. He has also published research papers in journals and conference proceedings.