Human Bone Localization in Ultrasound Image Using YOLOv3 CNN Architecture

R. Arif Firdaus Lazuardi*, Tita Karlita †, Eko Mulyanto Yuniarno ‡, I Ketut Eddy Purnama §, and Mauridhi Hery Purnomo ¶
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo - Surabaya 60111, Indonesia
Email: *lazuardi16@mhs.its.ac.id, †tita16@mhs.ee.its.ac.id, ‡ekomulyanto@ee.its.ac.id, §ketut@ee.its.ac.id, ¶hery@ee.its.ac.id

***Abstract*—Localization of human long bones in ultrasound images has quite complex challenges. This is due to a representation of the reflection of a sound wave emitted by a B-scan sensor. The ultrasound scan does not only display bone specimens, but also contains muscles, soft tissue, and other parts under the skin tissue Therefore we need a system that can automatically recognize bone specimens in ultrasound images. This study implements deep learning based learning systems using the convolutional neural network (CNN) method with YOLOv3. The training results from the network detector with IoU threshold 0.5 can recognize bone objects in $mAP_{@50}$, $mAP_{@75}$ and $mAP_{@50:95}$ with values of 99.98, 97.68 and 85.67 respectively. And for the results of training the network detector with IoU threshold 0.75 can recognize bone objects in $mAP_{@50}$, $mAP_{@75}$ and $mAP_{@50:95}$ with values of 99.96, 97.46 and 86.35 respectively.**

***Index Terms*—bone USG, CNN, YOLOv3**

I. Introduction

Health technology helps health practitioners such as doctors and nurses in the process of diagnosing, treating, and treating a patient's illness. Some of the health technologies use medical imaging techniques such as Computed Tomography (CT scan), Magnetic resonance imaging (MRI), and Ultrasonography (USG). MRI and CT scans are the most popular medical imaging techniques in the world of health. This is because the image representation of the two technological devices produces excellent image quality.

CT scans produce a lot of ions and radiation to patients, and MRI is an expensive technology. Compared to these two technologies, ultrasound is a medical device that has the potential to be used to represent hard structures such as bone. That is because ultrasound technology is safe to use, not invasive, based on sound waves, relatively cheaper, presents data in real-time, and portable.

However, medical image representation using ultrasound sensors for bone is not as good as other medical images. Because there are also catches of objects other than bones, such as; muscle, flesh, blood vessels, and soft tissue. So the main problem in this study is to detect the location of bone areas in ultrasound images.

Research related to medical imaging has been carried out, some of which use digital image processing algorithms such as those carried out by R. Jia to automatically segment bone in USG images using Local Phase Features [6]. The study was also conducted by J. Kowal to carry out automatic bone segmentation to minimize the risk of computer-based medical measures [5].

While the use of deep learning based algorithms has also been carried out by several researchers. Among them are to detect cervical cancer [4], segment cervical smear images [10], detect liver tumor disease [11], and classify breast cancer thermal images [12].

In recent years, the CNN method has undergone many developments, the architecture that is often used is the F-R CNN proposed by S. Ren et al. [7]. This architecture uses an approach by submitting

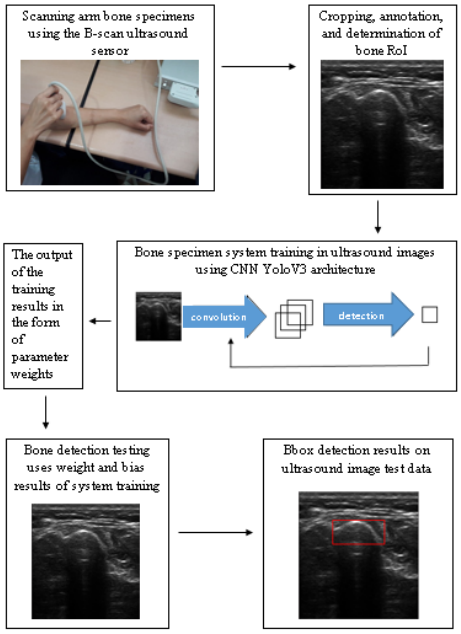

Fig. 1. Flow Diagram of Bone Localization in Ultrasound Images

proposals for several regions that are assumed to be objects, which are often termed as region proposal networks (RPN). Then from several proposals, the region of the object is classified by the network classifier to determine the score class predicition in each region of interest detected.

Faster R-CNN requires two stages to be able to detect and classify objects in an image. The first step is to determine a number of RoIs which are assumed to have an object. And the second stage is to classify the probability level of an object. This makes Joseph Redmon et al. Propose a new approach in creating object detection network architectures that only experience one stage of a network detector. The so-called YOLO architecture (You Only Look Once) [1] in 2016. This YOLO architecture uses the basic algorithm of the Regression Problem to create a network detector which at one stage can detect objects while recognizing them.

In 2016, K. Tita et al. began the study to be able to reconstruct 3D images of human long-bone specimens on ultrasound images. The study began by conducting automatic detection of bone objects in ultrasound images using an algorithm based on digital image processing [2]. K. Tita's research was carried out to extract bone features in ultrasound images. Then in 2018, K. Tita developed a

978-1-7281-2965-5/19/$31.00 ©2019 IEEE

RoI based deep learning automatic detection using CNN Faster R-CNN architecture [3].

Therefore in this study, a system based on the same algorithm was developed by K. Tita [3], but uses a different CNN architecture. In this study the CNN architecture used was YOLOv3 [9]. In Fig. 1 a block flow diagram of the system determines the location of the bone in the ultrasound image.

II. Methodology

The research is divided into four stages, the first stage is data acquisition, the second stage is labeling (annotation) in the image, the third stage is training data to the system using CNN with YOLO network architecture, and the fourth stage is testing the algorithm.

A. Dataset Preparation

The data taken is a series of ultrasound B-scan images on different subjects. The tool used to retrieve human long bones ultrasound image data is the TELEMED Linear Transducer L15-7L40H-5 B-Scan sensor with a frequency range of 7.0-15.0 MHz and 39 mm field of view.

USG image retrieval technique uses freehand scanning. Where the operator takes ultrasound image data by moving the ultrasound probe by hand freely. Free hand scanning is done in two ways. The first way is to do a gradual sweeping on the arm and save the representation of the image with the image extension *.jpg. And the second way, sweeping is done by moving the ultrasound sensor probe regularly along the arm, both the right hand and the left hand. Scanning results are stored in video format with *.avi extension, for the next video to be parsed so that data can be retrieved in each frame and stored with image extension *.jpg.

B. Data Annotation

The results of human long bones representation on ultrasound images not only display the representation of the bone area, but also display some ultrasound sensor configuration information. Because we don't use this information as a training parameter, we have to crop or cut images in certain areas, so what we get is a representation of bone, soft tissue, muscle, flesh, skin, etc.

The research team members have consulted a Radiologist to determine bone RoI. Information from someone who is an expert in the field of radiological imagery is very necessary and important. This is because, the information is the basis of the parameters that we practice with the system. So if the information has been tested properly.

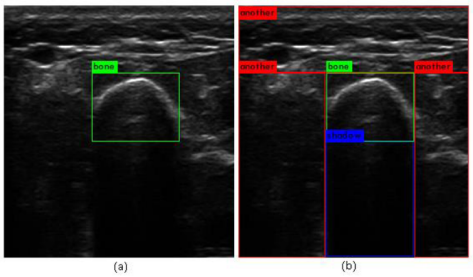

Fig. 2. (a) RoI manual object bone. (b) RoI another object and shadow based on RoI bone in figure (a)

Each ultrasound image of a human arm bone specimen is annotated as 3 objects; bone, another, and shadow. Bone is a bone surface whose pixel intensity is relatively high (white) and has a shape like a parabolic curve. Shadow is a shadow of the surface of the bone that has a relatively low pixel intensity (dark) compared to pixels on the surface of the bone, this is because ultrasound waves cannot penetrate the bone, so that waves returning to the receiver probe B-scan ultrasound can only represent the surface of the bone.

While another is an area outside of bone and shadow. The area is above and on both sides of the bone. Another can represent many things from the human body tissue, such as; muscles, blood vessels, etc. Because this research is focused on recognizing bone objects, the parts are not trained in detail to the system.

As seen in Fig. 2, in some ultrasound image datasets, another area has a form of texture and pixel intensity that almost resembles a bone object. In plain view, it can be distinguished that the object is not bone, because it does not have a shadow below it. Therefore at the training stage another and shadow object systems are also trained, so that the system can recognize and distinguish bone objects and those that resemble it in another area.

C. Algorithm Training

Implementation of the algorithm dividing the existing dataset is divided into 2 main data, namely training data and test data. While the training data is divided into 2 data, namely training data and evaluation data. CNN training consists of two stages. The first stage is feed-forward and the second stage is back-pass. When feed-forward, images are used through several convolution processes. For the YoloV3 architecture, there are a total of 53 convolution layers, this stage outputs predictive values on 3 scales; large, medium and small scale. In each prediction the value issued is a tensor box containing 4 bbox offset values, objectness, and score class. In each tensor there are 3 boxes that are predicted, then only 1 box is chosen which represents the grid to detect the presence of human long bones.

At the end of the feed-forward stage, the predicted RoI bone bbox value will be compared with the ground truth bbox in that image. The difference between the bbox ground truth and bbox prediction is measured using a loss function, so that the loss value will be studied by the system, and at the back-pass stage the algorithm will correct the parameters of weights and biases (weights).

In Fig. 3 there is an illustration of the flow of system training data. The training is divided into two scenarios, the first scenario is training the system using the IoU threshold > 0.5, and the second scenario is training the system using the IoU threshold > 0.75. The training uses these two scenarios to measure the system training response to the average loss value in each training iteration and mean average precision (mAP) calculated from the evaluation data when the training system takes place every 4 epochs.

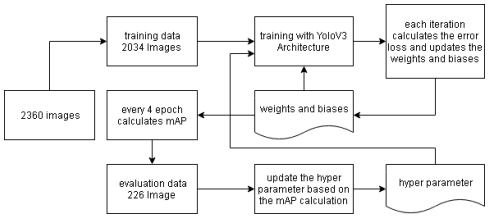

Fig. 3. System Training Data Flow

The research dataset taken from USG images of human long bones was taken from 7 people and divided into 9 subdata. Each is represented by the letters A, B, C, D, E, F, and G. The division of the dataset used for the training and testing stages is presented in Table I. The computational resources used in this study use the

facilities available at Google Colab. Where Google Colab provides computing resources with the Tesla K80 GPU capabilities and 12GB of GPU RAM capacity [8].

TABLE I
DATASET

Subdata	Training	Evaluation	Testing
A	540	60	120
B	450	50	100
C	360	40	80
D	450	50	100
E	54	6	-
F	180	20	100
G	-	-	120
Total	2034	226	620

D. System Testing

Metric evaluation used to calculate mean average precision (mAP) values for each class in each image. The mAP value is calculated by taking the average precision (AP) value. AP is obtained by calculating the area under the cartesian coordinate curve between precison (x axis) and recall (y axis).

Precision values measure how accurate the network detector is in predicting objects in an image. Precision is the percentage of predictions of successful objects, obtained from the comparison between true positive predictive value (TP) with the number of positive object detection, TP added by false positive (FP).

While the recall value measures how well the network detector finds all the correct objects (both objects and objects are found correctly). Recall is obtained by comparing the TP value with the number of all object detections obtained in that image, TP added with false negative (FN).

Mathematical equations of precision, recall, average precision, and mAP calculations are,

$$precision(p) = \frac{TP}{TP + FP}$$

$$recall(r) = \frac{TP}{TP + FN}$$

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

So the parameters to be analyzed in each training scenario are as follows; true positive (TP), false positive (FP), false negative (FN), precision, recall, mean average precision (mAP), and average loss. The average loss value shows how much error is obtained during system training in each iteration.

III. Results and Discussion

A. Training with IoU Threshold > 0.5 ($weights_{50}$)

At the beginning of the training in the 100th iteration, the average loss value is very high with a value of¿ 1000.0, so the mAP value in the evaluation image is also very small at 0.02. When the training period enters the 300th iteration the value of mAP is better at the value 0.41, even though the average loss value is in the 5.0.

Fig. 4 is a comparison of average loss and mAP values with IoU threshold 0.5. The 400th iteration of mAP is very good at 0.92 with an average loss value of 1.5. The more the iteration increases, the average loss value has a decreasing graph trend, until the 600 iteration has reached 1.0 and the mAP calculation reaches a maximum value of 1.0. This means that in the iteration, the value of weights @, 5 is good enough because it is able to correctly detect the test image correctly. The mAP value stays at a maximum value of 1.0 until the training iteration ends in the 6000th batch.

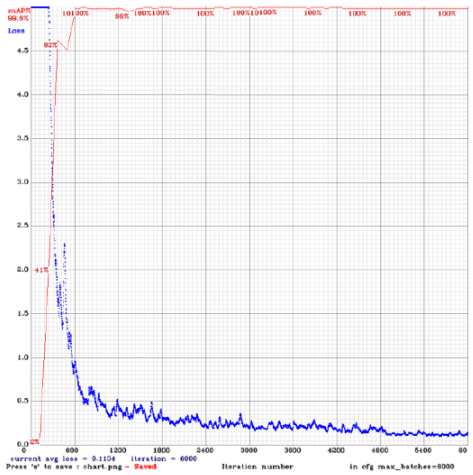

Fig. 4. Average loss and mAP with IoU Threshold > 0.5

And the average loss value continues to have a declining trend, starting stable when entering the 1200th iteration (<0.5), continuing to shrink until the end of the iteration reaches 0.05. This shows that the system continues to study the error value assigned in each iteration to update the weight and bias values used in the YoloV3 detector network.

B. Training with IoU Threshold > 0.75 ($weights_{75}$)

Fig. 5 is a comparison of average loss and mAP values with IoU threshold 0.5. As for system training with IoU threshold > 0.75, the average loss parameter enters a stable value since the 3000 batch, with a value below 0.5 and decreasing until the last batch at position 0.005. This shows that the system requires a longer iteration time to be able to adapt to the IoU threshold > .75. This happens because the results of detection of objects that are considered correct when having IoU are high above > 0.75.

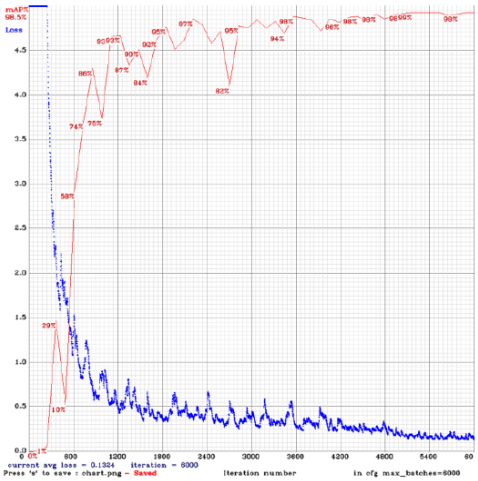

Fig. 5. Average loss and mAP with IoU Threshold > 0.5

Because the IoU threshold > 0.75 used is high enough to evaluate system parameters. Then the mAP value continues to oscillate in calculating detection accuracy in evaluation images per 4 epochs. Because, the system always tries to correct a large error value in each iteration so that the time needed is relatively longer. However, the trend of mAP is always getting better in more itersions until at the 1800th iteration, the mAP value is stable above 0.8.

C. Testing

For each of the training results in all scenarios (P1 and P2), the test data is tested using 2 scenarios. Testing the first scenario to sub data groups A, B, C, D, and F (U1 test scenario). And the second scenario testing was tested on sub group G data (U2 test scenario). The test data are presented in Table II and for the block scenario diagram testing is shown in Fig. 6.

TABLE II
TEST SCENARIO

Weights	Scenario	Subdata	Total
$weights_{50}$	U1	A,B,C,D,E,F	500
$weights_{50}$	U2	G	120
$weights_{75}$	U1	A,B,C,D,E,F	500
$weights_{75}$	U2	G	120

The division of these two scenarios is carried out to analyze the behavior of the system when the test is carried out on the sub-data which part of the data has been used as training data (scenario U1) and sub-data for which none of the data is used as training data (scenario U2).

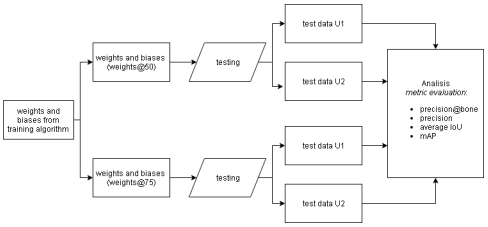

Fig. 6. Block Diagram of The Test Scenario

D. Analysis of Metric Evaluation

The value of the metric evaluation is taken using different IoU thresholds. Starting from the threshold value of 0.5 to 0.95 with a step of 0.05. Precision value is the percentage of system accuracy in detecting objects in an image. In this study, more detailed analysis was directed at bone objects, so the value of precision for the bone class was also presented.

1) $Precision_{bone}$*:* .

For U1 test data the use of $weights_{50}$ and$weights_{75}$ have the same precision values starting from the IoU threshold of 0.5 to 0.6. For $weights_{50}$ the value of high precision only exists when the IoU threshold > 0.65, the remaining $weights_{75}$ in all remaining IoU thresholds gives higher precision values. This shows that $weights_{75}$ has better accuracy than $weights_{50}$ in detecting bone RoI in ultrasound images.

As for the U2 test data, the use of $weights_{50}$ and $weights_{75}$ each has a superior value in some IoU random threshold sequences. This has meaning, that the weight and bias parameters of the training results of the P1 and P2 scenarios, when tested on ultrasound images that have never been used as training system materials, have the same opportunity to produce good bone object detection. In detail the data are presented in Table III.

TABLE III
$Precision_{bone}$

IoU	$U1_{50}$	$U1_{75}$	$U2_{50}$	$U2_{75}$
0.50	100.00	100.00	99.17	99.17
0.55	100.00	100.00	98.35	98.33
0.60	99.80	99.80	95.87	97.50
0.65	99.60	99.40	92.56	89.17
0.70	98.00	98.80	86.78	84.17
0.75	95.80	96.00	68.60	69.17
0.80	88.00	91.00	48.76	47.50
0.85	67.60	76.40	27.27	25.83
0.90	33.80	42.40	4.96	10.83
0.95	5.40	7.00	0.00	1.67

2) precision: .

For U1 test data the use of $weights_{50}$ and $weights_{75}$ each has a superior value in some IoU random threshold sequences. But $weights_{75}$ has relatively more precision values at the IoU threshold > 0.75. This might happen because precision calculates all the accuracy of the network detector in detecting all objects (bone, shadow, and another).

As for the U2 test data, the use of $weights_{75}$ has a relatively better value than $weights_{50}$. This means that the weight and bias parameters of P2 scenario training have a better tendency to detect the whole object being trained in the system. In detail the data are presented in Table IV.

TABLE IV
PRECISION

IoU	$U1_{50}$	$U1_{75}$	$U2_{50}$	$U2_{75}$
0.50	100.00	99.96	99.83	99.67
0.55	100.00	99.96	99.67	99.50
0.60	99.96	99.92	99.17	99.33
0.65	99.76	99.80	98.50	97.50
0.70	99.32	98.48	97.17	96.34
0.75	98.72	98.88	92.85	93.18
0.80	97.08	97.72	87.85	87.85
0.85	92.08	93.80	79.70	80.70
0.90	80.92	82.12	63.23	65.89
0.95	52.00	49.32	27.29	33.44

3) average IoU*:* .

For U1 test data using $weights_{75}$, almost all IoU thresholds have a higher value than $weights_{50}$. This means that network detector $weights_{75}$ produces the bbox predictive value that is closest to the bbox ground truth value. Likewise with U2 test data, network detector $weights_{75}$ has a better value than $weights_{50}$ in almost all IoU thresholds. Which means $weights_{75}$ can adapt well in detecting objects in a group of ultrasound images that previously have never been part of a system training dataset. In detail the data are presented in Table V.

4) mAP*:* .

In the U1 test data, the use of $weights_{50}$ and $weights_{75}$ has a higher value with the same proportion in the IoU random threshold

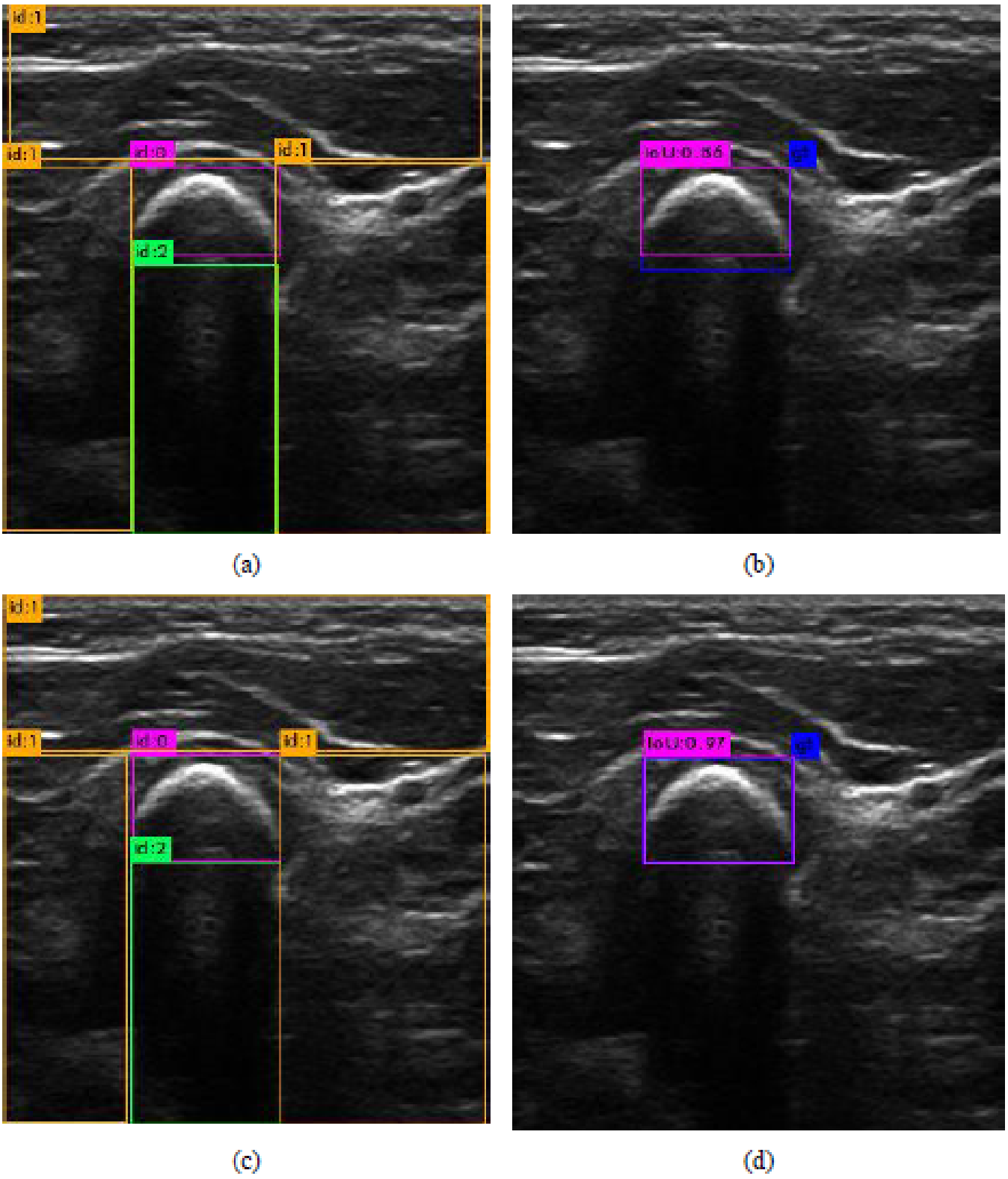

Fig. 7. (a,c) Results of Detection of All Objects. (b,d) Bone Objects and Bbox Ground Truth Are Displayed

value. However, network detector $weights_{75}$ has a tendency for relatively higher mAP values at the IoU threshold > 0.8. This means that network detector $weights_{75}$ with a relatively large threshold parameter is able to detect the existence of objects in an image.

And for U2 test data, use of $weights_{75}$ has a relatively similar pattern with U1 test data. Network detector $weights_{75}$ has a higher

TABLE V
AVERAGE IoU

IoU	$U1_{50}$	$U1_{75}$	$U2_{50}$	$U2_{75}$
0.50	93.34	93.46	89.66	89.91
0.55	93.34	93.46	89.57	98.82
0.60	93.32	93.43	89.27	89.73
0.65	93.19	93.36	88.85	88.58
0.70	92.89	93.15	87.96	87.79
0.75	92.45	92.71	84.80	85.49
0.80	91.18	91.80	80.91	81.35
0.85	87.03	88.56	74.18	75.42
0.90	77.24	78.31	59.76	62.41
0.95	50.33	47.75	25.34	32.26

value in a relatively large threshold parameter. Even though it was tested on a test image that had not previously been entered as training material. This also provides a tendency for behavior that network detector $weights_{75}$ are able to adapt to ultrasound images of human bone specimens. In detail the data are presented in Table VI.

TABLE VI
AVERAGE IoU

IoU	$U1_{50}$	$U1_{75}$	$U2_{50}$	$U2_{75}$
0.50	99.98	99.96	99.91	99.73
0.55	99.98	99.96	99.86	99.37
0.60	99.97	99.89	98.91	98.90
0.65	99.59	99.73	96.98	94.50
0.70	98.78	99.09	93.46	91.65
0.75	97.68	97.46	81.98	83.89
0.80	92.96	94.16	72.62	72.86
0.85	81.09	85.28	58.16	61.17
0.90	60.92	63.61	33.93	39.79
0.95	25.75	24.37	6.54	10.79

E. Visual Detection Results

Fig. 7 is the U1 test image with IoU acquisition which is highest in network detector $weights_{75}$ which has a value of 0.97 compared to the same test image the result of network detector $weights_{50}$ has a value of 0.86. Fig. 7 (a,b) is results of detector using weights $weights_{50}$. And Fig. 7 (c,d) is results of detector using $weights_{75}$. It appears in the figure that the $weights_{75}$ detector results have a very high IoU value. Almost equal to the value of the bbox ground truth. This shows that the network detector with $weights_{75}$ is relatively better compared to network detector $weights_{50}$.

IV. CONCLUSION

Weight and bias parameters of the training system P1 ($weights_{50}$) and P2 ($weights_{75}$) are used to detect groups of test images. The final results of each detection for each weighted and biased value in all test scenarios can well find the location of the bone object. This can be seen from the measurement of the value of $mAP_{@50}$, $mAP_{@75}$ and $mAP_{@50:95}$ for $weights_{50}$ in the U1 group test image (A, B, C, D, F) having their respective values of 99.98, 97.68, and 85.67. While testing for groups U2 (G) has values of 99.91, 81.98, and 74.23 respectively. And for $weights_{75}$ in the U1 group test image (A, B, C, D, F) have values of 99.96, 97.46, and 86.35 respectively. While testing for groups U2 (G) has values of 99.73, 83.39, and 75.27 respectively.

V. ACKNOWLEDGMENT

We thank to Kemenristekdikti (Kementerian Riset, Teknologi dan Pendidikan Tinggi) for financing the research under the PTUPT (Penelitian Terapan Unggulan Perguruan Tinggi) Scheme.

REFERENCES

[1] R.J Joseph, D. Santosh, G. Ross, F. Ali, You Only Look Once: Unified, Real Time Object Detection," *IEEE Conference on Computer Vision and Pattern Recoginition*, 2016.

[2] K. Tita, Y. Eko Mulyanto, P. I Ketut Eddy, P. Mauridhi Hery, Automatic Bone Outer Contour Extraction from B-Modes Ultrasound Images Based on Local Phase Symmetry and Quadratic Polynomial Fitting, *SPIE, Second International Workshop on Pattern Recognition*, 2017.

[3] K. Tita, Y. Eko Mulyanto, P. I Ketut Eddy, P. Mauridhi Hery, Deteksi Region of Interest Tulang pada Citra B-mode secara Otomatis menggunakan Region Proposal Neteworks, JNTETI, Vol. 8, No. 1, 2019.

[4] D.A. Dharmawan, Deteksi Kanker Serviks Otomatis Berbasis Jaringan Syaraf Tiruan LVQ dan DCT, *J. Nas. Tek. Elektro dan Teknol. Inf. Vol. 3, No. 4*, p 269-272, 2014.

[5] J. Kowal, C. Amstutz, F. Langlotz, H. Talib, dan M.G. Ballester, "Automated Bone Contour Detection in Ultrasound B-Mode Images For Minimally Invasive Registration in Computer-Assisted Surgery An In Vitro Evaluation," *Int. J. Med. Robot. Comput. Assist. Surg. MRCAS, Vol. 3, No. 4, pp. 341348*, 2007.

[6] R. Jia, S.J. Mellon, S. Hansjee, A.P. Monk, D.W. Murray, dan J.A. Noble, "Automatic Bone Segmentation in Ultrasound Images Using Local Phase Features and Dynamic Programming," *IEEE 13th Int. Symp. Biomed. Imaging, pp. 10051008*, 2016.

[7] S. Ren, K. He, R. Girshick, dan J. Sun, "Faster R-CNN: Towards Real- Time Object Detection with Region Proposal Networks," *ArXiv Prepr.ArXiv1506.01497., Vol. 74, pp. 114*, 2015..

[8] T. Carneiro, R. Victor, M. Da, T. Nepomuceno, G. Bian, dan V.H.C.D.E. Albuquerque, "Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications," *IEEE Access Trends, Perspect. Prospect. Mach. Learn. Appl. to Biomed. Syst. Internet Med. Things., Vol. 6, pp. 6167761685*, 2018.

[9] R.J Joseph, D. Santosh, G. Ross, F. Ali, YOLOv3: An Incremental Improvement," *arXiv: 1804.02767 [cs.CV]*, 2018.

[10] N.P. Husain dan C. Fatichah, "Segmentasi Citra Sel Tunggal Smear Serviks Menggunakan Radiating Component Normalized Generalized GVFS," *J. Nas. Tek. Elektro dan Teknol. Inf., Vol. 6, No. 1, pp. 107 114*, 2017.

[11] N. Syakrani, Y. Widhiyasana, dan A.A. Efendi, "Deteksi Tumor Hati dengan Graph Cut dan Taksiran Volume Tumornya," *J. Nas. Tek. Elektro dan Teknol. Inf., Vol. 7, No. 1, pp. 35-43*, 2018.

[12] O. Herliana, T.S. Widodo, dan I. Soesanti, "Klasifikasi Nomsupervised Citra Thermal Kanker Payudara Berbasis Fuzzy C-MEANS," *J. Nas. Tek. Elektro dan Teknol. Inf., Vol. 1, No. 3, pp. 55-59*, 2012.