BALINESE SHADOW PUPPET CHARACTERS DETECTION IN THE WAYANG PETENG PERFORMANCE USING THE YOLOv5 ALGORITHM

I Gusti Ngurah Bagus Putra Asmara[1], Made Windu Antara Kesiman[2], Gede Indrawan[3]

[1,2,3] Program Studi Magister Ilmu Komputer, Universitas Pendidikan Ganesha

email: agungputra6716@gmail.com[1], antara.kesiman@undiksha.ac.id[2], gindrawan@undiksha.ac.id[3]

Abstract

To generate greater public interest in Balinese shadow puppet performances, it is crucial to explore novel ways of educating viewers about the characters showcased in the plays, as many individuals may need to become more familiar with them. In Object Detection, an algorithm is called You Only Look Once (YOLO). This research utilizes the YOLOv5 algorithm to detect Balinese shadow puppet characters in the "wayang peteng" performances. The dataset consists of 5040 images, divided into training, validation, and test data, with a ratio of 7:2:1 (This ratio helps in effectively training and evaluating the YOLOv5 model on a diverse set of data). Four YOLO models are trained, each with a different number of epochs (a single iteration of training when the entire dataset has been passed forward and backward through the neural network), resulting in 12 models. All models are tested using the test data images to obtain precision, recall, and mean Average Precision (mAP) metrics. Additionally, three videos measure the average frames processed per second. The research findings reveal that the YOLOv5n model with 200 epochs achieves the best results, with a precision value of 1, recall of 1, mAP@0.5 of 0.995, mAP@0.5-0.95 of 0.985, and 128.20 frames per second.

Keywords: Computer Vision, Object detection, Wayang, YOLO

INTRODUCTION

Bali is one of the provinces in Indonesia with culture, which is the main element in the daily life of the Balinese people. The culture in Bali is still preserved today because this culture cannot be separated from the daily life activities of the Balinese people. One aspect of Balinese art is wayang kulit. Balinese shadow puppet is one of the Balinese cultures that are thick with noble values, morals, art, entertainment, education, knowledge, philosophy, and religion [1][2][3]. In a Balinese shadow puppet performance, sometimes the audience barely recognizes the characters in the wayang. Not knowing the characters in a wayang performance will make it difficult to understand the storyline and the message conveyed. This phenomenon has decreased the number of shadow puppet shows in general and in the number of spectators [4]. If people are not interested in the performance of Balinese shadow puppets, then the performance of Balinese shadow puppets will be threatened with extinction.

Balinese shadow puppets as a culture can be preserved in two forms: cultural experience

and knowledge [5]. Cultural knowledge means preserving culture through information media to educate the public about Balinese shadow puppet performances [5]. To preserve and increase the interest of the Balinese people in the performance of Balinese shadow puppets, an innovative method is needed to educate the public about the characters in the wayang performance, considering that only some know the characters played in the performance.

In Computer Vision, there is a method called Object detection. Object detection is a method for identifying objects in an image. Previous researchers have used the Object Detection method with various algorithms to detect objects in images with varying speeds and accuracy. In 2019, Prince Kumar and colleagues compared the R-CNN, Fast R-CNN, Faster R-CNN, and YOLO models, concluding that YOLO produced the best results with real-time detection capabilities [6]. Other studies have compared the performance of YOLO, SSD, R-CNN, Fast R-CNN, and R-FCN with the results that YOLO has the advantage of speed in processing images which reaches 45 images per second [7]. In addition, research from Min Li and

friends and research from Srivastava comparing R-CNN, YOLO, and SSD concluded the same thing that YOLO produces the best results with the best speed and accuracy as well as the ability to detect in real time [8][9]. From the results of the four studies, it was concluded that the YOLO algorithm could perform object detection with good performance and is supported by the ability to detect objects in real time.

Ramya conducted research by comparing the performance of YOLOv3, YOLOv4, and YOLOv5 in detecting blood cells. This research concluded that YOLOv5 detects blood cell objects with the best precision, recall, and accuracy values of YOLOv3 and YOLOv4 [10]. By paying attention to the advantages of the YOLOv5 algorithm, this research uses the YOLOv5 algorithm to detect objects of Balinese shadow puppet characters in wayang peteng performances.

Research on the detection of character objects in wayang performances was carried out by Susanto and Mulyo in 2019, classifying digital images into four classes of wayang characters, including Arjuna, Gatotkaca, Srikandi, and Hanoman. The flow in this research starts with preprocessing by changing the image of training data and test data into binary images. Then apply edge detection process to the binary image and extracts area, perimeter, and eccentricity features, which will later become input to the neurons of the artificial neural network in the training process. After the model was generated from the training process, testing was carried out on the 30 test data provided. This research resulted in an artificial neural network to classify the four classes of wayang characters with an accuracy of 96% [11]. This research has not been able to determine the position of the puppet character as a result of classification, has not been able to detect more than one object in one image, and has not been able to detect real-time puppet character objects in wayang performances.

LITERATURE STUDY
A. Wayang Peteng

Wayang peteng is a Balinese shadow puppet performance held at night and has a broader theme than wayang lemah, namely in the form of entertainment or spirituality . Wayang peteng is performed using lights or torches to reflect the wayang shadows that will be captured on the screen. According to the theme of the story, the wayang peteng is divided into nine types, namely: Wayang Parwa, Wayang Ramayana, Wayang Gambuh, Wayang Calonarang, Wayang Cupak, Wayang Sasak, Wayang Arja, Wayang Tantri, and Wayang Babad.

B. Object Detection

Computer vision and image processing are closely related to object detection; a technology computers utilize that detects certain classes of semantic objects (such as people, buildings, or cars) in digital images and videos[12]. Computer vision tasks commonly utilize object detection techniques, including image annotation, vehicle counting[13], and activity recognition [14]. Object detection methods are generally grouped into approaches based on artificial neural networks (ANN) and non-ANNs. The non-ANN approach first needs to define features using one of the methods; then, to carry out classification tasks, one can utilize techniques like Support Vector Machine (SVM). At the same time, ANN techniques can end-to-end object detection without defining features precisely and are usually based on Convolutional Neural Networks (CNN).

Examples of non-ANN approaches are Viola–Jones object detection framework with Haar features, Scale-Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HOG). An example of the ANN approach is as follows: R-CNN, Fast R-CNN, Faster R-CNN, cascade R-CNN, SSD, YOLO, RefineDet, Retina-Net, Deformable Convolutional Networks.

C. YOLO Algorithm

YOLO stands for "You Only Look Once" which is a high-speed object detection algorithm using a single convolution network [15]. YOLO breaks the image into grids, then each part of the grid is classified and localized. Then YOLO, predicting where to place the bounding box. Predicting the bounding box is done with a regression-based algorithm by simultaneously predicting the bounding box for the entire image, making it faster. YOLO uses an architecture that resembles the CNN architecture, where YOLO only uses convolution and pooling layers. The last convolution layer is adjusted to the number of classes and the desired number of prediction boxes.

YOLOv5 is the newest series from YOLO. YOLOv5 is a development of YOLOv4 with an increased speed of up to 140 frames per second, increased accuracy, and increased ability to detect small objects. In addition, the size of YOLOv5 is relatively small, almost 90% smaller than YOLOv4, which makes it possible to use YOLOv5 in embedded devices.

D. YOLOv5 Concept

YOLOv5 illustration can be seen on Figure 1. YOLOv5 will first separate the input image into cells. Each cell is tasked with predicting the bounding box if the center of the bounding box is in the cell concerned. Each cell predicts a bounding box involving the x, y, width (w), and height (h) coordinates of the bounding box and a confident score. Class prediction is also done in each cell.

To perform object detection, each grid cell predicts several B bounding boxes with parameters and their confidence scores for these bounding boxes[16]. The confidence score reflects the presence or absence of an object in the bounding box. The value of the confidence score is defined as Equation 2

$$IOU_{box2}^{box1} = \frac{intersection(box1, box2)}{union(box1, box2)}$$

$$confidence\_score = p(Object) \; x \; IOU_{pred}^{truth} \qquad (2)$$

Where *p(Objek)* is the probability of the presence or absence of an object in a cell, and IOU_{pred}^{truth} is the Intersection Over Union (IOU), which is the value of the intersection of the area of the predicted result box and the ground truth box as shown in equation 1. Value *p(Objek)* has a range from zero to one, so the confidence score will be close to zero if no object is detected in a cell. On the other hand, the value of the trust score will be close to the value of IOU_{pred}^{truth}. In addition, each bounding box has four other parameters (x,y,w,h) with center coordinates (x,y), width (w), and height (h), as shown in Figure *2*. By adding the confidence score parameter, each bounding box has five parameters.

Finally, YOLO applies Non-Maximum Suppression (NMS) to remove all bounding boxes that do not contain objects or bounding boxes that contain the same object as other bounding boxes. Examples before and after applying the NMS can be seen in Figure 3. The steps for implementing the NMS are as follows: (1) Determine the "X" bounding box that has the highest class prediction probability; (2) Define a new bounding box "Y" that intersects with a bounding box "X" and predict the same class; (3) Calculate the IOU value (by equation 2.1) of the "X" and "Y" bounding boxes; (4) If the IOU value is more than 50%, then remove the "Y" bounding box; (5) Repeat step 2 to step 4 until no bounding box intersects with the "X" bounding box; (6) Repeat step 1 until none of the bounding boxes that predict the same class intersect.
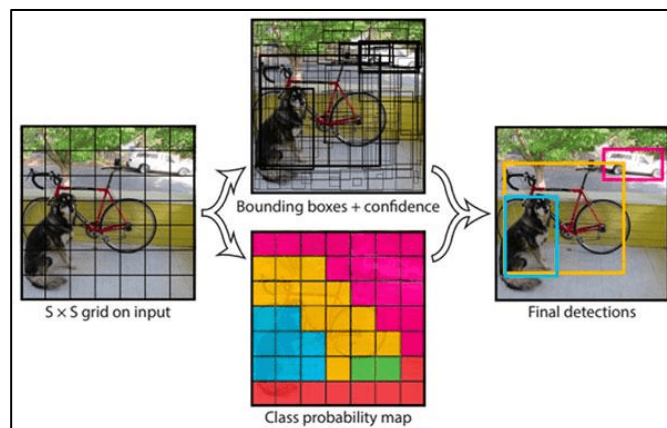

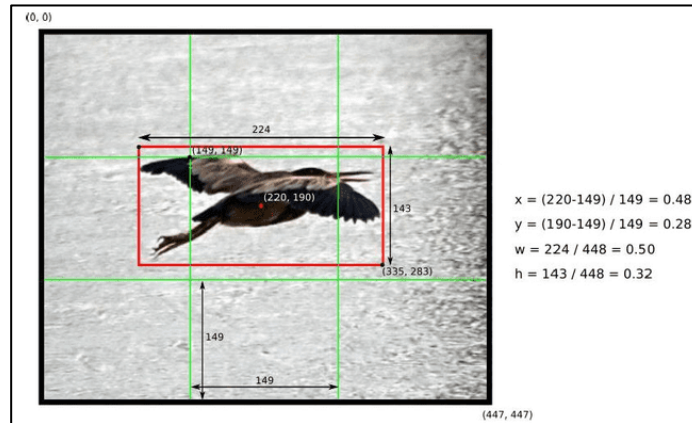Figure 1. The YOLO model applies a 7x7 cell grid to the input image

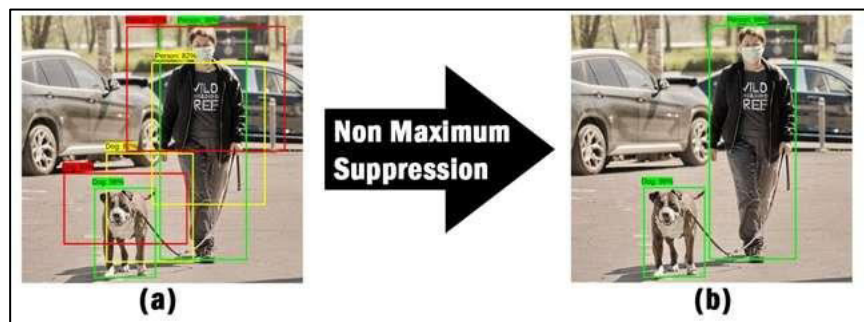Figure 2. An example of a bounding box parameter in a 3x3 grid cell

Figure 3. Examples before being subject to NMS (a) and after being subject to NMS (b)

E. YOLOv5 Architecture

YOLOv5 architecture can be seen in Figure 4. There are three parts in the YOLOv5 network, namely Backbone, Neck, and Head. The YOLOv5 architecture has an input size of 640x640 pixels, according to the image size used. Then, the input image is passed through the backbone, the CSPDarknet network. CSPDarknet consists of 53 layers of 3x3 convolutions, CSP blocks, and residue blocks. This architecture is designed to extract features from low to high-resolution images and can capture object context. After passing through the backbone, we get an output feature map with a size of 80x80. The output from the backbone is then passed through additional convolution layers called the neck.

YOLOv5 applies a Path Aggregation Network (PANet) to improve information flow. PANet adopts a Feature Pyramid Network (FPN) structure by adding a bottom-up path that enhances the deployment of low-level features. The neck produces an output feature map with a size of 80x80 which is then forwarded to the convolution layers called heads.

In YOLO, several head layers produce an output feature map with a size of 80x80. After the head, another convolution layer is performed for object detection. This convolution layer produces an output feature map the same size as before, namely 80x80. Then, a convolution layer performs upsampling or increases the resolution. Upsampling is done to double the size of the feature map so that the output size of the feature map becomes 40x40. The exact process is carried out on the neck, head, detection layers, and upsampling layers to produce an output feature map with a size of 40x40 and 20x20. After the process on all grid cells is complete, the feature map output from each grid cell is combined into one tensor, which contains all object detection predictions.
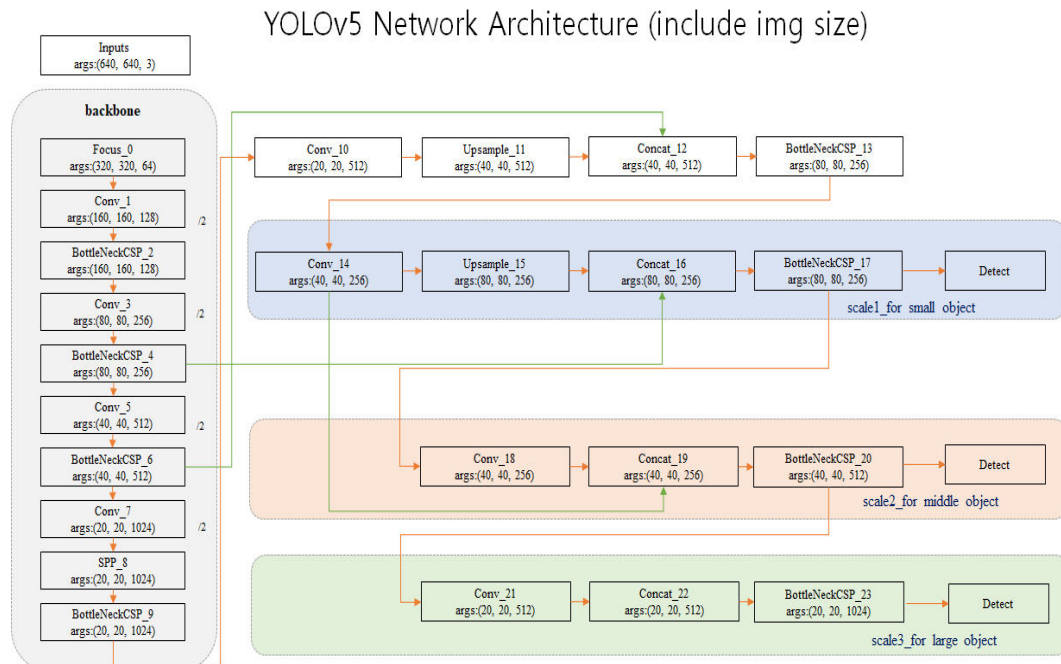
Figure 4. YOLOv5 Architecture

METHOD

This methodology section describes the steps in detecting Balinese shadow puppet characters in a wayang peteng performance using the YOLOv5 algorithm. The steps that were passed in this research were data collection, preprocessing, training, and testing.

A. Data Collection

The data collected is a video file with a resolution of 1280x720 pixels at 30 frames per second and RGB color mode. Videos are downloaded directly from Cenk Blonk's official Youtube channel. The number of videos downloaded is 12, the details of which can be seen in Table 1. All videos that have been downloaded are then entered into the preprocessing stage to obtain a dataset.

Table 1. Video Source of Wayang Peteng Performance

No	Video Title	Hyperlink	Original Duration
1	Masihkah kita penting sekolah?	https://youtu.be/uCUIEH2tTaw	12.11
2	Sebuah Renungan (Setan Dalam Diri)	https://youtu.be/6iG6hVzxHsU	12.48
3	Tidak Cukup Hanya Cinta	https://youtu.be/11RoWgRG3OA	12.40
4	Jangan Suka Berprasangka Buruk	https://youtu.be/KpxUX0GKOJ0	11.27
5	Apem 2M	https://youtu.be/Giby2jGuvKU	14.23
6	Program Studi Ilmu Liak	https://youtu.be/KbsW2AO5AlM	13.39
7	Kepala Dosa	https://youtu.be/KlVLni9_N6k	14.04
8	Reklamasi Teluk Benoa	https://youtu.be/XD-Z3_8x8co	10.28
9	Intinnya Tetap Bersyukur	https://youtu.be/rz_RODvf1ls	9.49
10	Harga Barang Naik, Gaji Tetap?	https://youtu.be/zgq_deRNfLc	13.53
11	Filsafat Cinta	https://youtu.be/tQEfoAtXyGM	10.45
12	Perubahan dan Kematian yang pasti di Dunia ini?	https://youtu.be/6ecYThkPTkw	13.01

B. Preprocessing

The preprocessing step is done to prepare data for research. The preprocessing step involves cutting video data, changing video resolution, extracting images from videos, and annotating ground truth objects in the image.

Cutting the video is done to remove parts of the video that are not needed, such as the intro, ad appearances, and greetings for Galungan and Kuningan holidays. The video-cutting process is carried out using the Adobe Premiere Pro 2020 application. All videos produced from this process have a duration of 7

minutes. All parts of the video contain shadow puppet characters.

Video resolution is also changed using the Adobe Premiere Pro 2020 application. The resolution of all videos is changed to 640x640 pixels.

The process of extracting images from videos using the Roboflow service Roboflow provides a tool to extract images from video input based on the number of frames required to capture per second. The extraction process takes one image per second from each video. The total number of images obtained from this process is 420x12 = 5040 images, with the number of objects of each class contained in all images being 420. An example of extracting images from a video using Roboflow can be seen in Figure 5.

The ground truth annotation process is also carried out using Roboflow. The object classes annotated in all images can be seen in Table 2. All images and annotation results are divided into three groups, namely training data, validation data, and test data, with a ratio of 7:2:1.
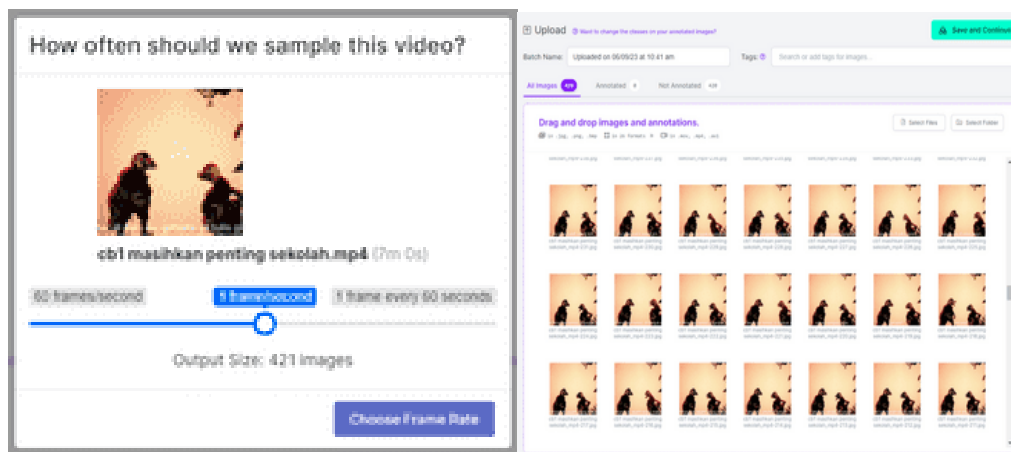

Figure 5. An example of extracting images from a video using Roboflow

Table 2. Class, ID and Image Example of Wayang Character Objects

Class (Id)	Blonk (0)	Cenk (1)	Delem (2)	Merdah (3)	Sangut (4)	Tualen (5)
Example						

C. Training

The training was carried out on 4 YOLOv5 models, each of which was trained using three epochs, namely 100, 200, and 300, so that 12 YOLOv5 models were obtained. The training process uses the Google Colab environment by utilizing the Tesla T4 GPU to speed up the training process. The training process uses hyperparameters with details in Table 3. The details of the four YOLOv5 models used can be seen in Table 4. The data used are training data and validation data.

Table 3. YOLOv5 Hyperparameters

Parameter	Value
Weight Decay	0.0005
Optimizer	SGD (Stochastic Gradient Descent) / Adam
Learning Rate	0.01

Momentum	0.937
Batch Size	16
Subdivisions	8
Input Size	640x640
Number of Classes	6
Grid Size	32
IoU Threshold	0.45
Confidence Threshold	0.25
NMS Threshold	0.45
Model Size	YOLOv5n , YOLOv5s, YOLOv5m, YOLOv5l
Head Type	CSP (Convolutional Spatial Pyramid)
Activation Function	Mish

Table 4. Comparison of YOLOv5N, YOLOv5S, YOLOv5M, and YOLOv5L

	YOLOv5n	YOLOv5s	YOLOv5m	YOLOv5l
Depth multiple	0.33	0.33	0.67	1
Width multiple	0.25	0.5	0.75	1

D. Testing

The testing process is carried out in two stages. The first stage is to detect the test data to get metric precision and recall, mAP@0.5 and mAP@0.5-0.95. The second stage is to detect in real-time against three videos to get the detection speed metric in frames per second. The videos used are videos number 3, 5, and 10 in Table 1. Footage of the detection results of the test data images can be seen in Figure 6. The author also conducted an experiment to perform detection with input from the camera. This can be seen in Figure 7.

Detection using a webcam is only to prove YOLOv5's ability to detect in real time using input from a webcam which is indicated by the number of frames per second that can be processed. Detection using a webcam cannot be done in the Google Collab environment, so it is done using a personal computer. The personal computer used has a Core i7 9700KF processor specification with 16GB 3200MHz RAM. YOLOv5 detection performance results using a webcam can be seen in Table 6.
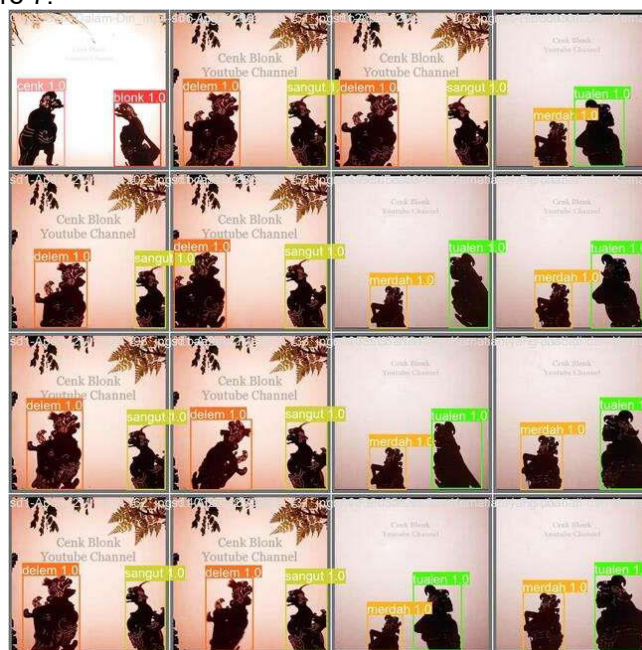

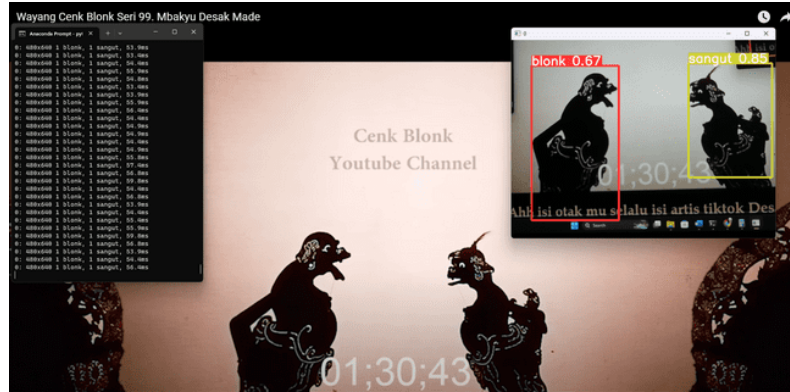Figure 6. Example of detection results on image test data

Figure 7. Example of detection in real time using the camera

RESULT AND DISCUSSION

The testing phase obtains a metric table seen in Table *5*. The resulting metric values are the average of the metrics for all detected object classes.

Table 5. YOLOv5 Models Test Result

Models	Epoch	Frame Per Second	Precision	Recall	mAP_0.5	mAP_0.5: 0.95
YOLOv5n.pb	100	127,71	0,999	1	0,995	0,979
YOLOv5n.pb	200	128,20	1	1	0,995	0,985
YOLOv5n.pb	300	135,14	0,999	1	0,995	0,986
YOLOv5s.pb	100	104,17	0,999	1	0,995	0,983
YOLOv5s.pb	200	106,38	1	1	0,995	0,987
YOLOv5s.pb	300	107,53	1	1	0,995	0,988
YOLOv5m.pb	100	57,14	1	1	0,995	0,984
YOLOv5m.pb	200	57,05	1	1	0,995	0,989
YOLOv5m.pb	300	56,82	1	1	0,995	0,988
YOLOv5l.pb	100	30,33	1	1	0,995	0,986
YOLOv5l.pb	200	30,27	0,999	1	0,995	0,988
YOLOv5l.pb	300	30,39	0,999	1	0,995	0,989

Table 6. YOLOv5 Models Detection Speed using Webcam

Models	Epoch	Frame Per Second
YOLOv5n.pb	100	21.3
YOLOv5n.pb	200	25
YOLOv5n.pb	300	20.87
YOLOv5s.pb	100	8.24
YOLOv5s.pb	200	8.68
YOLOv5s.pb	300	8.83
YOLOv5m.pb	100	4.57
YOLOv5m.pb	200	4.4
YOLOv5m.pb	300	4.31
YOLOv5l.pb	100	2.32
YOLOv5l.pb	200	2.37
YOLOv5l.pb	300	2.37

From Table 5 and Table 6, it can be concluded that the YOLOv5n.pb model at epoch 200 performs well with maximum precision and recall values (1) and high mAP_0.5 values. This model also has a good detection speed of around 128 frames per second and 25 frame per second when detecting using webcam as input source. The YOLOv5s.pb model also shows solid performance with high precision, recall, and mAP_0.5 values but has a slightly slower detection speed than YOLOv5n.pb. The YOLOv5m.pb and YOLOv5l.pb models perform well but have a lower detection speed. In choosing the best model, it is necessary to consider the balance between detection speed and object detection performance. One of the goals of this research is to produce an optimal model that can be used to perform object detection in real-time. Hardware type affects detection performance in real time. Detection using a better processor or using a GPU device can improve detection performance using YOLOv5.

While doing the research, the writer experienced two problems. First, several videos on Cenk Blonk's YouTube channel feature other types of characters listed in Table 2, as shown in Cenk Blonk's video series 119 entitled "Eblonk Melajah Kebatinan." The video shows the characters Cenk and Blonk wearing traditional Balinese accessories called headbands. The video clip of "Eblonk Melajah Kebatinan" can be seen in Figure 8. Therefore, further research can add video data in the form of other types of original appearances of the characters to be detected so that the resulting model is more general.

In the testing process by detecting in real-time with input in the form of video or webcam, detections are sometimes found that is False Positive, where the background of the wayang peteng performance, such as kayon, house objects, plants, or others, is detected as one of the six specified objects. An example of this incident can be seen in Figure 9. This was due to the absence of background image data used in the training process in this research. The author does not find an empty image that only contains the background of the wayang peteng performance. One of the ways to get the background image in a video of the wayang peteng performance is to manipulate the original image and remove the wayang character objects using an image processing application, such as Adobe Photoshop.

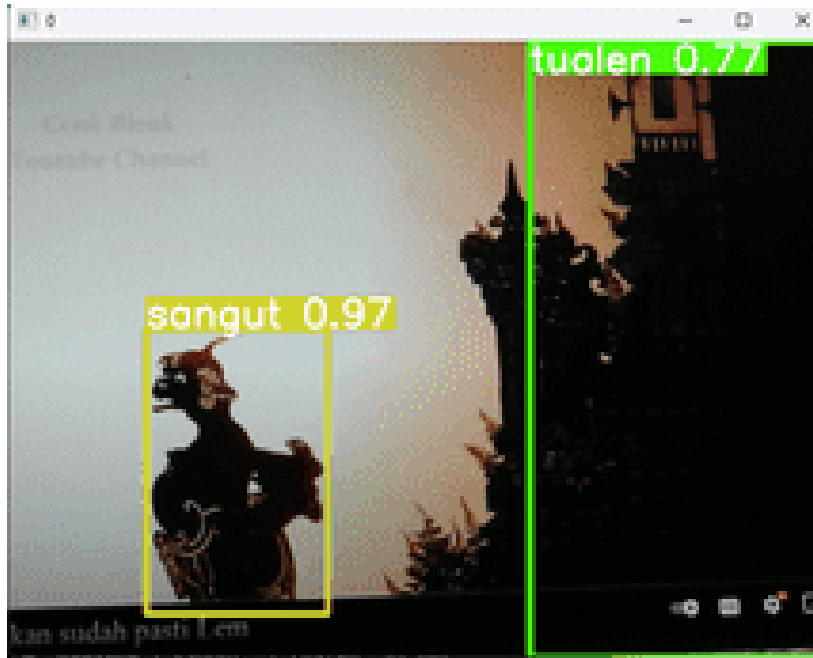Figure 8. Another type of Cenk and Blonk figures

Figure 9. Example of False Positive detection on background

CONCLUSION

In selecting the best model, it is necessary to consider the balance between detection speed and object detection performance. This research determined that the YOLOv5n model with an epoch of 200 produced the best results with perfect metric precision and recall (1), the value of mAP0.5 was 0.995, and mAP0.5-0.95 was 0.985.

Other researchers interested in developing a model for object detection of Balinese shadow puppet characters can add video data in the form of other types of the original appearance of the characters they want to detect so that the resulting model is more general. In addition, background image data is needed from Balinese shadow puppet performances in the training process so that false positive detection does not occur. One way to get the background image in a video of the wayang peteng performance is to manipulate the original image and remove the wayang character object using an image processing application, such as Adobe Photoshop.

REFERENCES

[1] M. Covarrubias, *Pulau Bali: Temuan yang menakjubkan*. Udayana University Press, 2014

[2] I. M. Marajaya, "Pertunjukan Wayang Kulit Bali Dari Ritual Ke Komersialisasi," *Kalangwan: Jurnal Seni Pertunjukan*, vol. 5, pp. 21–28, 2019

[3] I. M. Marajaya, "Estetika Pertunjukan Wayang Kulit Bali," *Wayang (Jurnal Ilmiah Seni Pewayangan)*, 2006

[4] I. M. (Universitas H. I. D. Yudabakti, "Marginalisasi dan Revitalisasi Wayang Kulit Parwa di Kabupaten Gianyar pada Era Globalisasi," *Journal of Bali Studies*, vol. 6, no. 1, pp. 223–252–223–252, 2016.

[5] M. Kusbiyanto, "Upaya Mencegah Hilangnya Wayang Kulit Sebagai Ekspresi Budaya Warisan Budaya Bangsa," *Jurnal Hukum & Pembangunan*, vol. 45, 2015

[6] P. Kumar, V. Garg, P. Somvanshi, and P. C, "A Comparative Study of Object Detection Algorithms in A Scene," *International Journal of Engineering Research & Technology (IJERT)*, vol. 8, no. 05, 2019.

[7] H. Zota and M. Dhande, "A Comparative Study of Widely Used Image Detection Algorithms," no. July, pp. 5183–5187, 2020.

[8] M. Li, Z. Zhang, L. Lei, X. Wang, and X. Guo, "Agricultural greenhouses detection in high-resolution satellite images based on convolutional neural networks: Comparison of faster R-CNN, YOLO v3 and SSD," *Sensors (Switzerland)*, vol. 20, no. 17, pp. 1–14, 2020

[9] S. Srivastava, A. V. Divekar, C. Anilkumar, I. Naik, V. Kulkarni, and V. Pattabiraman, "Comparative analysis of deep learning image detection

algorithms," *J Big Data*, vol. 8, no. 1, 2021

[10] V. Gupta Pola, A. Bhavya Vaishnavi, and S. Suraj Karra, "Comparison of YOLOv3, YOLOv4 and YOLOv5 Performance for Detection of Blood Cells," *International Research Journal of Engineering and Technology*, pp. 4225–4229, 2021

[11] A. Susanto and I. Mulyono, "Rekognisi Wayang Kulit Menggunakan Jaringan Syaraf Tiruan," in *SENDI_U*, Oct. 2019.

[12] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. Papastathis, and M. G. Strintzis, "Knowledge-Assisted Semantic Video Object Detection," *IEEE Trans Circuits Syst*, vol. 15, no. 10, pp. 1210–1224, 2005.

[13] A. A. Alsanabani, M. A. Ahmed, and A. M. Al Smadi, "Vehicle Counting Using Detecting-Tracking Combinations: A Comparative Analysis," *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, vol. PartF16834, pp. 48–54, 2020

[14] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," *Proceedings of the IEEE International Conference on Computer Vision*, 2007

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem. pp. 779–788, 2016

[16] S. Mody, H. Mehta, P. Mantri, B. Ali, and A. Khivasara, "Safety Gear Equipment Detection for Warehouse and Construction Sites Using YOLOv5," *International Research Journal of Engineering and Technology*, vol. 9, no. June, 2021