Detection of objects dangerous for the operation of mining machines

Jakub Szymkowiak[1 [0000−0001−8110−3594]], Marek Bazan[1 [0000−0002−2455−6463]], Krzysztof Halawa[1 [0000−0001−6508−0468]], and Tomasz Janiczek[1 [0000−0002−8901−1368]]

Wrocław University of Science and Technology, 27 Wybrzeże Wyspiańskiego st. 50-370 Wrocław, Poland
252868@student.pwr.edu.pl, marek.bazan@pwr.edu.pl, krzysztof.halawa@pwr.edu.pl, tomasz.janiczek@pwr.edu.pl, https://wit.pwr.edu.pl/en/

Abstract. Deep learning was used to detect boulders that can damage excavators in opencast mines. Different network architectures were applied, i.e., modern YOLOv5, RetinaNet and Mask-RCNN. Studies were carried out in which the results obtained using a few networks were compared. The abovementioned neural networks were exploited in a framework for detection of oversized boulders on a conveyor belt operating in an opencast coal mine. The method is based on the analysis of a certain number of consecutive frames of the film from an industrial camera. The novelty relies on checking the detection of a boulder within subsequent frames and allowing the skipping of a prescribed small number of neighboring frames with false negative detections. This allows one to make a decision about stopping a conveyor belt after detecting a boulder in consecutive frames even when they are interleaved with frames that contained a boulder missed by a detector due to misleading environmental conditions such as shadows or sand. The method was tested on recordings from an opencast mine in Poland. The proposed method can help prevent the failure of expensive equipment.

Keywords: deep learning · YOLOv5 · object detection · digital image processing · opencast mining.

1 Introduction

Currently, there are many opencast mines around the world, where various raw materials are extracted. The vast majority of mineral resources is developed by opencast mining [22]. Often in such mines there are giant excavators [6], the height of which is similar to buildings that are several storeys. There is a high risk of damage if there is a big rock on the conveyor belt. Failures caused by unnoticed large stones are very costly, complicated to repair and cause mining to stop. Therefore, it is very important to develop methods that will reduce the risk of the excavator operator not noticing such dangerous cases. For the

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

automatic detection of dangerous boulders, this article proposes a method using deep neural networks. Among other things, modern and fast YOLOv5 networks were used.

This work was created in connection with a practical problem that occurred in an open-pit mine and can occur in other situations concerning stones recognition against a difficult background [20] [8] [25] [23].

During the transfer of spoil in the form of small gravel or sand, a large stone appeared on the conveyor belt from time to time, which caused significant danger to mining machines. To obtain models operating in different lighting conditions, both video streams taken in good daylight and movies shot under artificial lighting in the middle of the night were used. A camera with image stabilization hardware was mounted on the excavator above the conveyor belt. An example view from the camera is shown in Fig. 1. A stone is visible on the conveyor belt. Even such a relatively small object should stop the transport of the excavated material. In many cases, much larger stones have been mined, which, if overlooked, can cause severe damage.

Fig. 1. Camera view with a visible stone on the conveyor belt

2 Preparing a dataset

The 27 recordings over several days were used for a training process. Additional 10 recordings were excluded from training and put aside for testing the final framework after the training of the models is finished. After dividing 27 videos into individual frames, over 28,000 images were obtained. Then, appropriate labels were assigned to images in which at least one large rock was observed. A Python script was created to make labeling faster and easier. At least one stone was visible in 2399 images. Sometimes not only one stone may appear on a frame. However, such situations occur very rarely. Large stones were marked in

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

Fig. 2. Camera view with a visible stone on the conveyor belt (artificial light)

the images with bounding boxes. For this purpose, the makesense.ai platform was used, which enables the preparation of datasets with selected objects. Then, the dataset with ready-made annotations was processed on the roboflow.ai platform [18]. This platform allows users to download annotations in various formats. It was also used to divide the data into training, validation, and test sets. To inspect the stability of the training process the models were trained in two configurations of a data split:

1. the training set – 80% of the frames and the test and validation sets – 10% of the frames each,
2. the training set – 70% of the frames, and the test and validation sets – 15% of the frames each.

The sizes of images to which frames have to be scaled in datasets used for different object detection models are depicted in Section 3 when describing the models.

One has to note that all frames in each set contained a boulder. It is a requirement for object detection models training that images without positive detection are not provided. It means that not all frames from videos are included in the datasets, but only those containing a positive detection. In the datasets, frames are shuffled. One has to note that validation and test sets in the context of detection model building are different from validation and test sets that are formed from ten videos not used in the training. The validation set of the later datasets is used for fine-tuning of hyper-parameters and the test set for testing the whole framework.

3 Applied and tested models of neural networks

Experiments were carried out in which the results of the use of the following neural networks were checked: RetinaNet, Mask RCNN and YOLOv5. A brief comparison of these architectures is e.g., in [16], [5].

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

3.1 RetinaNet

RetinaNet is a one-shot object detection model that utilizes anchor boxes and focal loss to handle the class imbalance problem in object detection [12]. RetinaNet has achieved state-of-the-art results on several benchmark datasets and is widely used for object detection tasks in real-world applications. Because of these advantages, it was also decided to test its effectiveness in detecting stones. This network is built of one backbone and two other subnets that are responsible for different tasks. The backbone creates a convolutional feature map by analyzing the entire input image. One subnet performs convolutional object classification on the output of the backbone subnet. The latter subnet performs convolutional regression, creating bounding boxes [12]. The images processed by this architecture were scaled to a rectangular image size between 800x800 and 1333x1333 pixels. The architecture of RetinaNet is presented in Fig. 3. Fizyr's implementation of RetinaNet was used in the conducted experiments [7].

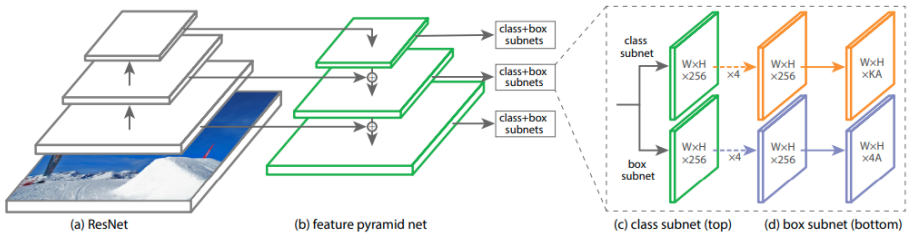

Fig. 3. The RetinaNet architecture [12].

3.2 Mask RCNN

The Mask R-CNN (Mask Region-based Convolutional Neural Network) extends the popular Faster R-CNN and, besides detecting objects, it also performs instance segmentation [9],[4]. An interesting case of use in the context of our problems was presented in [8] [15]. The Mask R-CNN first generates object proposals and then classifies the proposals and generates masks in the second stage by applying a fully convolutional network. The network which is based on Matterport's implementation but fully compatible with TensorFlow 2 [19] were applied. This implementation is based on the Feature Pyramid Network (FPN) and ResNet101 [11][1]. Combining the FPN and the ResNet allowed us to increase both the quality and the speed of data processing. The comparison between the application of FPN and ResNet is shown in Fig. 4, where the numbers signify the spatial resolution and channels. The arrows denote convolutional, deconvolutional, and fully-connected layers.

In Fig. 4 it can be seen that the head of Mask RCNN is built of two branches - one of them is responsible for generating masks and the other one for classifying

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

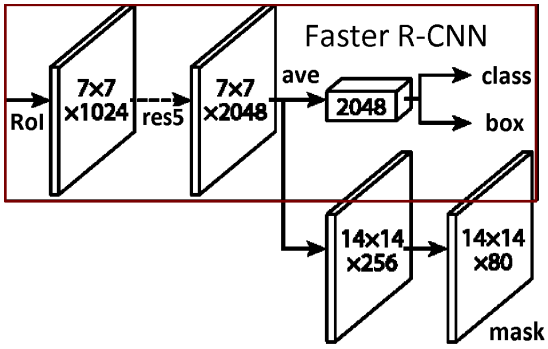

Fig. 4. The head architecture of the Mask RCNN [11].

and calculating bounding boxes. The latter consists of two parts: one that is for classification and the other is responsible for detection (they are marked as *class* and *box* respectively) [9]. The images processed by Mask RCNN architecture were scaled to a rectangular image size between 800x800 and 1024x1024 pixels.

3.3 YOLOv5

The prototype of YOLOv5 was YOLO (*You Only Look Once*), which was introduced in 2015 as a uniquely fast neural network for object detection [17]. YOLOv5 was created in 2020 as a significantly improved version of the first YOLO architecture [21]. YOLOv5 was implemented using PyTorch framework. The first YOLO network is built of 24 convolutional layers and two fully-connected layers [17]. The architecture of this network is shown in Fig. 5. YOLO is a one-shot object detection model with CSP (Cross-Stage Partial connections), which allows for better feature reuse and reduces computation time [24].

YOLOv5 is built of three parts:

- New CSP-Darknet53 (*backbone*) — contains 29 convolutional layers.
- SPPF, New CSP-PAN (Cross-Stage Partial-connection & Path Aggregation Network) (*neck*) [10],[2] — SPPF is built of three `MaxPooling2D` layers, two `ConvBNSiLU` layers, and one `Concat` layer. `ConvBNSiLU` is built of one convolutional layer followed by a batch-normalization layer and Sigmoid Linear Unit [16]. `Concat` is a layer that combines outputs of the preceding layers. The application of a `CSP` (*Cross-Stage-Partial connection*) allows us to improve the quality of object detection. It is a technique derived from CSPNet [14].
- YOLOv3 Head (*head*) — Three subnets built of convolutional layers that were used since YOLOv3 [13].

The images processed by the YOLOv5 architecture were scaled to the image size of 416x416 pixels. This is a standard image size for output from roboflow.ai [18]. Scaling images in other models to that size did not improve the results.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

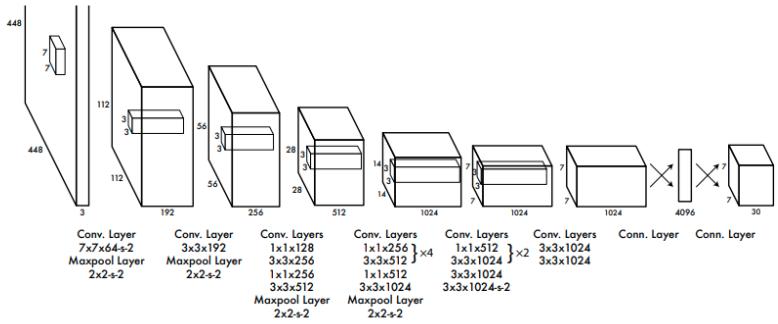

Fig. 5. The architecture of the first YOLO network [17]

4 Detection system

The presented detection system is designed as follows. On its input, the frames received from the camera are delivered in real-time. After detecting a stone, the system can send a signal to stop the conveyor belt automatically. Subsequent frames are analyzed on an ongoing basis. In order to minimize false alarms, the presence of a boulder is signaled when the neural network detects a boulder in in most frames out of 30 consecutive frames. Although there is a chance that the one stone is observable, it may be difficult to detect even by a well-trained model. Such situations have occurred, such, as dark images, blurry images, boulders covered in sand, etc. Such cases were taken it into account, and it was decided that there should be a possible break in the detection process. This break lasts no more than a small number of frames. For example, if a stone was detected in 10 frames in a row, then undetected in 4 frames in a row, and subsequently again detected in 20 frames in a row, it would be considered that a large stone had really appeared on the line. The optimal number of consecutive frames that may constitute a break in the sequence of frames where true detection occurred, is established using a validation set by maximizing the number of proper signals to stop a conveyor belt. For such a method a high-quality stone detector is required to minimize false negatives due to the model but not due to the environmental conditions. To our knowledge, such a method has not been published before.

Additionally, the created program prints information about a number of frames in which a stone appeared (not counting the sets of frames that fulfilled the conditions of sending a signal). The total time of analysis and analyzed frames per second is also shown. It should be noted that the system works in the same way for each network but the implementations differ in detail.

5 Results

This section first presents the results of the network operation on single images and then performance of the created detectors based on 30 consecutive frames from recordings. There was one stone in each recording.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

5.1 Evaluation of networks

After proper training of all networks, their performance was examined. Among other things, the F1-score was determined using precision and recall, which are represented by the following formulas:

$$p = \frac{t_p}{t_p + f_p} \tag{1}$$

$$r = \frac{t_p}{t_p + f_n} \tag{2}$$

where t_p is true positives, f_p denotes false positives, f_n is false negatives.

The formula for F1-score is the harmonic mean of precision and recall:

$$F_1 = \frac{2pr}{p + r} \tag{3}$$

where p is precision, r denotes recall.

Moreover, mean average precision (mAP) was calculated. This metric is commonly used to assess the quality of object detection models. mAP takes into account the trade-off between precision and recall, and accounts for both false positives and false negatives. This property makes mAP such a relevant measure. The method of calculating mAP is described, among others, in [3] Each network was evaluated on the validation data set. The results are shown in Table 1.

Table 1. Comparison of the networks for the following data split: training set 80%, validation set 10%, test set 10%

model	F1	precision	recall	mAP
YOLOv5 based	0.991	0.992	0.991	0.992
Mask RCNN based	0.611	0.588	0.637	0.884
RetinaNet based	0.328	0.199	0.950	0.994

Based on the analysis of the data in Table 1, it can be seen that the best results were obtained using the YOLO network. In order to ensure that the obtained results do not significantly depend on the choice of data for the test and validation sets, the experiments were repeated for the following data division: training set 70%, validation set 15%, test set 15%. The obtained results are presented in Table 2.

5.2 Evaluation of created detectors

The detection system were tested only on recordings that were not used during the training. The results are shown in the tables **??**-4. For clarification, here is the explanation what each column means:

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

Table 2. The YOLOv5-based model quality measures for a data split: training set 70%, validation set 15%, test set 15%.

model	F1	precision	recall	mAP
YOLOv5 based	0.989	0.988	0.992	0.995

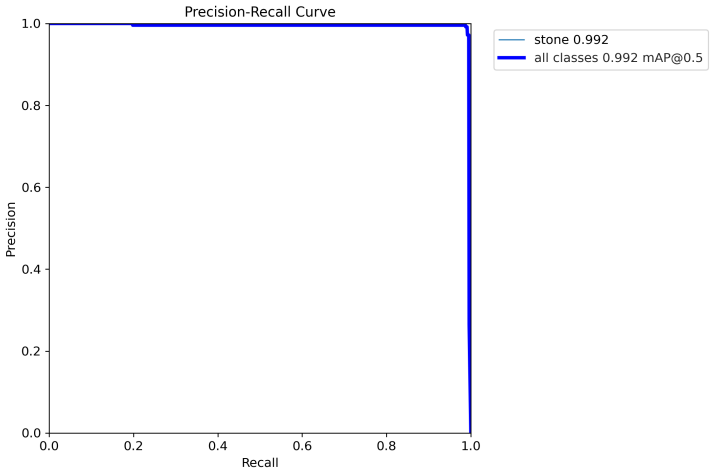

Fig. 6. Precision-recall curve for YOLOv5 for data split: training set 80%, validation set 10%, test set 10%

- video — input/output AVI file
- signals — a number of such series of detection that lasted at least 30 frames with possible breaks of 5 frames maximum,
- other detections — a number of other frames on which at least one stone was detected,
- time — total time of analyzing the recording,
- FPS — frames per second.

The following three detectors, which use 30 consecutive frames, were created: Retinanet based model, Mask RCNN based model and YOLOv5-based model.

Detector with Mask R-CNN For two video the model could not detect stones in a sufficient number of frames in a row. It should be noticed that one of these recordings is the same one where the stone was not well detected also by the YOLOv5-based model. It can also be noticed that, on the one hand, this model detects stones more frequently than the YOLOv5-based model. On the other hand, after analyzing the output files, one can see that it detects more *true positives* and also more *false positives*. In addition, the confidence is set to 0.9, so

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

Table 3. The results for Mask RCNN-based detector. Results for 5 skip frames.

video	signals	other detections	time [s]	FPS
test video 1	2	36	2796.0	0.37
test video 2	1	295	3103.1	0.34
test video 3	4	233	2741.4	0.36
test video 4	2	49	2789.9	0.37
test video 5	5	200	2679.9	0.36
test video 6	0	135	2643.9	0.38
test video 7	0	104	2891.5	0.38
test video 8	2	134	2806,8	0.38
test video 9	3	26	2456.0	0.38
test video 10	7	147	2736.2	0.38

the possible amount of *false positives* had already been decreased. Although there is a threat that this model may generate a false signal - it sometimes happens that it detects some parts of a machine or background as stones. Therefore there is a risk that such *false positive* may cause sending the signal to stop the line. Anything like this was not observed for the YOLOv5-based model. However, this model is one order of magnitude slower than the YOLOv5-based model. That is why it could be hard to use it in a real mine, even if it is quite accurate.

Detector with RetinaNet The RetinaNet-based model was able to detect a sufficient number of stones only for four recordings. It can be said it is less accurate than the other described models. In this case, there is a similar risk to the case of Mask RCNN-based model — it sometimes detects parts of a machine or background as stones, which can lead to sending false signals. This model is about twice as fast as the Mask RCNN-based model. It could not be used in a real mine because it is not accurate enough and still much slower than the YOLOv5-based model.

Detector with YOLOv5 For the YOLOv5-based detector we performed a fine-tuning of a number of skip frames. The result of this process in shown in Table 5. For this purpose, from the test set five videos are treated as validation videos on which the optimal number of skip frames is obtained by maximising the number of positive signals and retaining as low a number as possible of false positives or true positives that did not form a 30 successive frames sequence

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

Table 4. The results for RetinaNet-based detector. Results were obtained for 5 skip frames.

video	signals	other detections	time [s]	FPS
test video 1	2	68	1339.7	0.77
test video 2	0	85	1363.9	0.77
test video 3	0	69	1299.3	0.77
test video 4	0	88	1337.6	0.77
test video 5	1	225	1263.0	0.77
test video 6	0	58	1298.0	0.77
test video 7	0	108	1414.2	0.77
test video 8	0	55	1373.5	0.77
test video 9	2	73	1210.7	0.77
test video 10	3	130	1340.0	0.77

Table 5. The results for the YOLOv5-based detector for a varying number of skip frames on a validation set consisting of 5 videos. Fine-tuning a number of skip frames relies on maximisation of positive signals, as well as keeping the number of other detections frames as low as possible on videos from the validation set. The best results are bolded.

video	1 skip frame		3 skip frames		5 skip frames		7 skip frames		9 skip frames	
	signals	other detections	signals	other detections	signals	other detections	signals	other detections	signals	other detections
val. video 1	2	30	2	30	2	30	**2**	**30**	2	30
val. video 2	0	42	0	42	0	42	**0**	**42**	0	42
val. video 3	0	69	0	69	1	39	**1**	**39**	0	69
val. video 4	1	53	1	53	1	53	**2**	**23**	2	23
val. video 5	5	105	6	75	7	56	**7**	**56**	7	56

described by a column "other detections". The optimum on the validation set was attained for 7 skip frames. The performance on the remaining 5 test videos (not used in any stage of the training) is shown in Table 6.

One can see that for two videos (one in the validation set and one in the test set), the model was unable to detect a stone for a sufficient number of frames in

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

Table 6. The results for the YOLOv5-based detector for seven skip frames on a test set of five videos not seen by the model at any stage of the training.

video	signals	other detections	time [s]	FPS
test video 1	2	14	76.0	13.1
test video 2	1	136	86.0	12.7
test video 3	6	67	80.4	13.1
test video 4	0	22	80.0	13.1
test video 5	3	9	78.9	13.1

a row in order to generate a signal. One of these recordings is really dark, and in the other one, the stone is barely visible, which is caused by sand covering it. It can be said that for most cases, the YOLOv5-based model detects big stones well. It should also be noted that this model is very fast in comparison to the others. It processes 12-14 frames per second, which makes it unmatched if it comes to operation time. In conclusion, the YOLOv5-based detector would be right to use in real mines because the results would be delivered with low delay.

6 Conclusion

The use of deep models can be very desirable for detecting dangerous objects on a conveyor belt. Due to the monotony and lack of focus, there is a high risk of overlooking the boulders by the operator. The use of 30 consecutive frames allows to significantly improve the correctness of an operation.

The YOLOv5-based detector is the most suitable to use in a real mine because it is fast and accurate enough to detect dangerous objects in a video stream quite effectively. The results on test recordings are satisfying in terms of accuracy and processing speed. The Mask RCNN-based detector is slower but also accurate enough. Although it cannot match the YOLOv5-based detector in terms of processing speed, the RetinaNet-based model can detect objects only in some cases. It is twice as fast as the Mask RCNN-based detector but still one order of magnitude slower than the YOLOv5-based model. It is possible to try optimizing each of these detector models for a specific mine by manipulating the training hyperparameters. If the detection system is not equipped with a powerful GPU and detection time is crucial, it is recommended to use the YOLOv5-based model.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

Acknowledgments

The authors would like to thank Produs S.A. and The Bełchatów coal mine for their cooperation and research data and to members of the CyberTech circle.

References

1. Abdulla, W.: Mask RCNN for Object Detection and Segmentation. https://github.com/matterport/Mask_RCNN, accessed: 2022-11-14
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
3. Cao, Y., Chen, K., Loy, C.C., Lin, D.: Prime sample attention in object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11583–11591 (2020)
4. Cheng, R.: A survey: Comparison between convolutional neural network and yolo in image identification. In: Journal of Physics: Conference Series. vol. 1453, p. 012139. IOP Publishing (2020)
5. Córdova, M., Pinto, A., Hellevik, C.C., Alaliyat, S.A.A., Hameed, I.A., Pedrini, H., Torres, R.d.S.: Litter detection with deep learning: A comparative study. Sensors **22**(2), 548 (2022)
6. Dhillon, B.S.: Mining equipment reliability. Springer (2008)
7. Fizyr: Keras-RetinaNet. https://github.com/fizyr/keras-retinanet, accessed: 2022-11-10
8. Fujita, H., Itagaki, M., Ichikawa, K., Hooi, Y.K., Kawano, K., Yamamoto, R.: Fine-tuned pre-trained mask r-cnn models for surface object detection. arXiv preprint arXiv:2010.11464 (2020)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence **37**(9), 1904–1916 (2015)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
13. Nepal, U., Eslamiat, H.: Comparing yolov3, yolov4 and yolov5 for autonomous landing spot detection in faulty uavs. Sensors **22**(2), 464 (2022)
14. Parico, A.I.B., Ahamed, T.: Real time pear fruit detection and counting using yolov4 models and deep sort. Sensors **21**(14), 4803 (2021)
15. Qu, X., Wang, J., Wang, X., Hu, Y., Zeng, T., Tan, T.: Gravelly soil uniformity identification based on the optimized mask r-cnn model. Expert Systems with Applications **212**, 118837 (2023)
16. Rain Juhl, E.F.: Real-time Object Detection and Classification for ASL alphabet. http://cs231n.stanford.edu/reports/2022/pdfs/147.pdf, accessed: 2022-12-11
17. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10

18. Roboflow.ai: https://roboflow.com/ (2023), [Online; accessed 20-April-2023]
19. Sciancalepore, M.: Mask R-CNN for Object Detection and Segmentation (Working with tf 2.4.1). https://github.com/masc-it/Mask-RCNN, accessed: 2022-11-14
20. Suresh, M., Abhishek, M.: Kidney stone detection using digital image processing techniques. In: 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). pp. 556–561. IEEE (2021)
21. Thuan, D.: Evolution of yolo algorithm and yolov5: The state-of-the-art object detention algorithm (2021)
22. Velikanov, V., Kozyr, A., Dyorina, N.: Engineering implementation of view objectives in mine excavator design. Procedia Engineering **206**, 1592–1596 (2017)
23. Vishmitha, D., Yoshika, K., Sivalakshmi, P., Chowdary, V., Shanthi, K., Yamini, M., et al.: Kidney stone detection using deep learning and transfer learning. In: 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA). pp. 987–992. IEEE (2022)
24. Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H.: Cspnet: A new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 390–391 (2020)
25. Zhang, H.: Image processing for the oil sands mining industry [in the spotlight]. IEEE Signal processing magazine **25**(6), 200–198 (2008)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_10