Expert Systems With Applications 194 (2022) 116537

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Detecting rocks in challenging mining environments using convolutional neural networks and ellipses as an alternative to bounding boxes

Patricio Loncomilla [a,*], Pavan Samtani [b], Javier Ruiz-del-Solar [a,b]

[a] Advanced Mining Technology Center, Universidad de Chile, Avenida Tupper 2007, Santiago, Chile
[b] Departamento de Ingeniería Eléctrica, Universidad de Chile, Avenida Tupper 2007, Santiago, Chile

ARTICLE INFO

Keywords:
Object detection
Rock detection
Convolutional neural networks
Deep learning

ABSTRACT

The automation of heavy-duty machinery and vehicles used in underground mines is a growing tendency which requires addressing several challenges, such as the robust detection of rocks in the production areas of mines. For instance, human assistance must be requested when using autonomous LHD (Load-Haul-Dump) loaders in case rocks are too big to be loaded into the bucket. Also, in the case of autonomous rock breaking hammers, oversized rocks need to be identified and located, to then be broken in smaller sections. In this work, a novel approach called Rocky-CenterNet is proposed for detecting rocks. Unlike other object detectors, Rocky-CenterNet uses ellipses to enclose a rock's bounds, enabling a better description of the shape of the rocks than the classical approach based on bounding boxes. The performance of Rocky-CenterNet is compared with the one of CenterNet and Mask R-CNN, which use bounding boxes and segmentation masks, respectively. The comparisons were performed on two datasets: the Hammer-Rocks dataset (introduced in this work) and the Scaled Front View dataset. The Hammer-Rocks dataset was captured in an underground ore pass, while a rock-breaking hammer was operating. This dataset includes challenging conditions such as the presence of dust in the air and occluded rocks. The metrics considered are related to the quality of the detections and the processing times involved. From the results, it is shown that ellipses provide a better approximation of the rocks shapes' than bounding boxes. Moreover, when rocks are annotated using ellipses, Rocky-CenterNet offers the best performance while requiring shorter processing times than Mask-RCNN (4x faster). Thus, using ellipses to describe rocks is a reliable alternative. Both the datasets and the code are available for research purposes.

1. Introduction

The automation of heavy-duty machinery and vehicles used in underground mining is an important requirement for increasing the safety, productivity, and operational continuity of mining operations. The most advanced underground mines already use teleoperated or autonomous heavy-duty machinery, such as LHD (Load-Haul-Dump) loaders, trucks, and rock breaking hammers (Salvador et al., 2020). However, the automation of this kind of equipment is a difficult task due to the challenging environments of underground mines having (i) limited visibility due to the presence of dust in the air and the poor illumination of specific areas, such as the *draw points* where material needs to be loaded, (ii) complex geometries (irregular and narrow tunnels, uneven ground, etc.) that complicate the movement of the heavy-duty equipment, and (iii) restrictions on the places where sensors can be installed, since in many cases there is no special infrastructure where sensors can be installed, and they can be installed only in places where they will not be affected by blasting.

In addition, given that the mining equipment and vehicles need to load, transport, and/or break rocks, the robust detection of rocks in production areas of mines is an important task to be addressed. For instance, in the autonomous loading with an LHD the rock pile must be characterized so that the LHD can collect the material (see Fig. 1 (a)). One problem in this task is that there may be oversized rocks or boulders, which could make the autonomous loading process impossible given that these rocks require secondary size reduction by explosives (see Fig. 1 (b)). Thus, for an autonomous loading system to be reliable, it must be able to detect the presence of oversized rocks, and to alert human assistance. In the case of autonomous rock breaking hammers (see Figs. 2 and 3), detecting individual rocks is needed in order to assess which rocks need to be broken by the hammer, and in which specific positions they must be struck to be broken and reduced in size. In this

* Corresponding author.
E-mail addresses: ploncomi@ing.uchile.cl (P. Loncomilla), pavan.samtani@ing.uchile.cl (P. Samtani), jruizd@ing.uchile.cl (J. Ruiz-del-Solar).

https://doi.org/10.1016/j.eswa.2022.116537
Received 1 February 2021; Received in revised form 7 January 2022; Accepted 8 January 2022
Available online 17 January 2022

(a) (b)

Fig. 1. (a) An LHD performing autonomous loading in a real mining tunnel. (b) An example of an oversized rock.

task, having a good approximation of both the shape and size of the rock is vital. The size allows deciding whether or not a rock requires breaking, while the shape and geometry of the rocks can be used to decide on the striking point. It must be noted that the minimum size threshold after which a rock requires processing, depends on the application. In the case of autonomous loading, large rocks are those large enough to not fit into the shovel, while in the rock breaking application large rocks are those large enough to not fall through the grizzly.

Hence, a relevant question is how to implement robust detection of rocks in productive underground mining areas, which at the same time fulfill the operational requirements of mines (e.g., LHDs cannot stop for analyzing the environment; they need to assess the rock pile while operating since material throughput is the highest priority of the mine). Given the astonishing performance of deep neural networks, specifically Convolutional Neural Networks (CNNs), in solving computer and robot vision problems, it is tempting to think that these might be applicable to the task. To be so, certain aspects need to be addressed, such as: (i) how to obtain enough data for training the CNN models, and (ii) how to represent rocks better by the models; is the standard bounding box representation of objects used by object detectors the best parameterization for representing rocks when both accurate, and at the same time, fast detections are required?

In this context, the main goal of this paper is to address these issues by providing labeled data that can be used to train CNN based rock detectors, and proposing a new neural architecture, named Rocky-CenterNet for the rock detection task. Important components of Rocky-CenterNet are the use of ellipses for representing rock geometry, and the way in which the orientation of these ellipses is estimated. Additionally, this paper provides a highly detailed analysis of how rock representation impacts the detection results.

In summary, the main contributions of this work are: (i) Rocky-CenterNet, a novel network architecture for detecting rocks in real time. The main characteristic of this architecture is the use of ellipses for describing rock's geometries. (ii) A new real-world dataset, named Hammer-Rocks, that contains images of rocks in an actual mining environment, captured while a rock breaking hammer was working in a production area of a copper underground mine, and (iii) an exhaustive comparison of CNN based models aimed at detecting rocks, which use different representations for the shape of the rocks -bounding boxes, ellipses, and segmentation masks-, considering several metrics to measure the quality of detections and the runtime of the process. Both the Hammer-Rocks dataset and the source code of Rocky-CenterNet are available for research purposes[1,2].

[1] https://github.com/amtc-rock-detectors

[2] https://datos.uchile.cl/dataset.xhtml?persistentId=doi%3A10.34691%2FFK2%2F1GQBHK

2. Related work

Several studies addressing the detection of rocks in images, aimed at different applications, have been published in recent years. In the work of Li et al. (2020), an unmanned aerial vehicle surveillance system for railway scenarios was introduced. An object detector based on SSD (Multi-block SSD) was implemented for detecting three classes of obstacles (persons, stones, and trains) in railways. The system achieved 98.9% precision and 92.1% recall on stones. While, in the work of Liu, et al., (2020), Faster R-CNN was used for detecting 8 classes of rocks, with the purpose of improving the evaluation of the stability of a rock mass, and the formulation of support schemes. The system achieved 96% recognition probability on images containing only one rock, and 80% on images containing multiple rocks.

Fanara et al., (2020) designed a system able to detect block falls in images from the north polar region of Mars is implemented and tested. Detection of block falls is challenging because they are small given the best available resolution of Mars' satellite imagery. For detecting block falls, two images captured at different times were used. The system uses HoG features and an SVM classifier for detecting block fall candidates. Then, an MSER region detector (Matas et al., 2004) and a threshold-based detector is applied over the image difference, for selecting accepted rock falls. The system is able to detect 75% of block falls with a false positive rate of 8.5%. On the other hand, Furlán et al. (2020) trained a network based on SSD for detecting rocks in environments similar to that on Mars. By using a ResNet-50 backbone on SSD, the system achieved a 0.253 mAP, running at 25 frames per second on a computer with 2 GPUs. A detector of lunar rock falls based on RetinaNet was implemented and reported by Bickel et al. (2018). The system achieved average precisions between 0.89 and 0.69, depending on the confidence threshold and intersection-over-union values used. An improved version of that system was proposed by Bickel et al., (2020), where six different architectures based on RetinaNet are trained either on Martian or lunar rock fall data, or a combination of both. The system is able to achieve a maximum overall recall of up to 0.78 and a maximum overall precision of up to 1.0, with a mean average precision of 0.71. Although the main topic of these studies is detection of rocks from images, the applications are different from ours, which is the automation of processes in mining under very challenging conditions. Also, the detectors based on deep learning are data-dependent, since the best performing architectures for an application depend on the dataset being used for training the networks. Then, these methods cannot be used out-of-the-box, in our application.

In the works of Niu et al., (2019) and Niu (2020), a Yolov3 detector (Redmon and Farhadi, 2018) was trained for detecting rocks, with the objective of developing an autonomous rock-breaking hammer. The detector is placed in a grate plate located at the surface, not in an actual underground mining environment, and, therefore, their environmental conditions are less challenging than ours were. Somua-Gyimah et al., (2019) presented an object detector based on SSD. Objects included belong to 9 terrain classes (including rocks), 8 mobile equipment classes,

or mine's personnel. The system achieved 80.9% precision and 91.3% recall. Although the objective of these works is similar to ours (detecting rocks for automating mining processes), the rocks are represented by means of bounding boxes, while in our work the rocks are represented as ellipses. The use of bounding boxes prevents applying these methods to tasks that require a more accurate description of the shape of the rocks, such as segmenting individual rocks for determining striking points in different rock positions, which is required for developing autonomous hammers. The determination of several striking points is required because the breaking can fail at a point, and then another one must be selected. Also, images shown in the works of Niu et al., (2019), Niu L. , (2020) and Somua-Gyimah et al., (2019) were captured in environments with good illumination, no dust, and without occlusions. However, the dataset used in our work contains images which were captured in a real underground mining environment, while a rock-breaking hammer was operating. Images acquired under these conditions are affected by several factors, like heavy dust, poor illumination, and occlusions, both between the rocks, and between the hammer and the material. The dust is generated in two situations, when LHDs discharge material over the grizzly, and when the hammer is breaking the rocks. Thus, images are noisier and harder to interpret.

A rock detection CNN developed in our laboratory, named Rocky-YOLO, is presented in the work of Lobos et al., (2019). It is based on a modification of the YOLOv2 algorithm for representing detections as ellipses instead of bounding boxes. Rocky-YOLO was a preliminary version of Rocky-CenterNet, the network architecture introduced in this work. In the original YOLOv2, the last layer from the network predicts bounding boxes by regressing its location (b_x, b_y), width b_w, height b_h and a measure of confidence b_o, which is constrained between 0 and 1. In Rocky-YOLO, the last layer is modified for predicting ellipses by regressing the major axis a, the minor axis b, and the orientation θ. Besides this modification, Rocky-YOLO and YOLOv2 are similar. The system detects oversized rocks in a custom dataset representing an extracting point at 1/5 scale, containing both stones and oversized rocks. The approach presented in Rocky-YOLO serves as the basis for this work and is used as a baseline in the experiments reported in Section 5.

3. Rocky-CenterNet – Detection of rocks using ellipses

In this paper, a novel network architecture named Rocky-CenterNet is proposed. It is based on CenterNet (Zhou et al., 2019), which is a one-stage object detector in which center points are estimated, and then the other properties of detection, like bounding box widths and heights, are regressed. CenterNet works by generating several feature maps. In the case of rock detection, five feature maps are generated: a heat map representing the probability of the presence of a rock at each location, two feature maps for correcting the position of the rocks, and two feature maps for estimating the width and height of the rocks.

Rocks are represented by ellipses instead of boxes in Rocky-CenterNet, since ellipses can better describe rocks' shapes because they can have different forms and can adopt different orientations in images. However, although horizontal and vertical rocks can be described by bounding-boxes, predicting horizontality or verticality requires estimating the orientation, which can be achieved by using ellipses. Elliptical detections are shown in Fig. 4.

Each ellipse is detected on a heat map, and then its major axis a, minor axis b, and orientation θ is estimated. Then we used the same parameterization as with CenterNet, but incorporated the following differences:

- The width feature map is replaced by a feature map to regress the major axis (a) of the ellipse.
- The height feature map is replaced by a feature map to regress the minor axis (b) of the ellipse.
- We added a feature-map with $c_{ang}+1$ channels, with c_{ang} the number of angle intervals, to predict the orientation of the ellipse.

The orientation of the ellipse corresponds to the angle formed between the minor axis and the major axis; this value is in the range [0°-180°). The prediction of the orientation of an ellipse is obtained through classification and regression ($c_{ang} \geq 2$). In this approach, the range [0°-180°) is divided into c_{ang} bins, each covering an interval with size $B = 180^{\circ}/c_{ang}$. Then, the angle is classified first as belonging to one of the intervals by using a Softmax function. Once the interval is determined, the difference between the angle value and the interval's lower bound is regressed. Therefore, the following loss term, L_{ang}, is added to the loss function, of the network, L_{det}:

$$L_{ang} = \frac{1}{N}\sum_{k=1}^{N}\left(\sum_{i=1}^{c_{ang}} p_{ik} \bullet \log(1/\hat{\mathrm{p}}_{\mathrm{ik}}) + \mathrm{c}_{\mathrm{ang}} \bullet (\mathrm{r}_{\mathrm{k}} - \hat{\mathrm{r}}_{\mathrm{k}})^2\right) \quad (1)$$

$$L_{det} = L_k + \lambda_{size} L_{size} + \lambda_{off} L_{off} + \lambda_{ang} L_{ang} \quad (2)$$

where c_{ang} corresponds to the number of angle intervals, p_{ik} is 1 if the kth angle is in the ith interval, and 0 otherwise; $\hat{p}_{ik}$ is the predicted

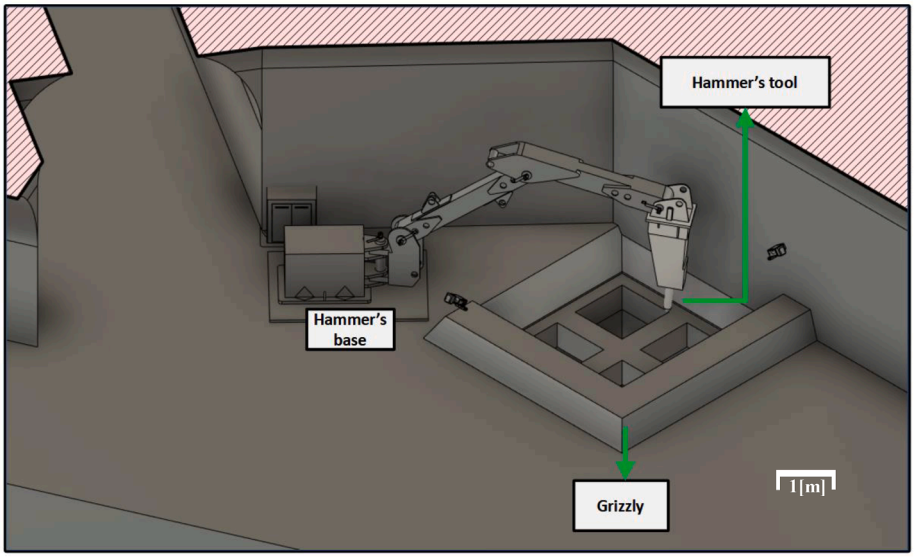


Fig. 2. A rock breaking hammer environment, which includes: a hammer and a grizzly. Oversized boulders need to be reduced to fall through the grizzly.

probability for the of the kth angle being in the ith class; r_k is the kth angle regression value; $\hat{r}_k$ is the predicted regression value; and N is the number of examples in the batch. The other terms of the loss function, L_{det}, are described in the original CenterNet work (Zhou et al., 2019).

The proposed classification plus regression approach is similar to the "anchor box approach" taken in the YOLO object detection network (Redmon and Farhadi, 2017). We find that using this formulation to define an ellipse's orientation, gives better results in the rock detection task.

As in YOLO (Redmon et al., 2016) and various other object detectors, Non-Maximum Suppression (NMS) can be used in Rocky-CenterNet too. We used Soft-NMS, as used in CenterNet (Zhou et al., 2019), but instead, use the bounding box that encloses the ellipse after the transformation. Using bounding boxes for NMS results in different detections than when using ellipses, as more detections are eliminated. However, this effect is only significant when ellipses are very elongated and diagonally oriented.

One difficulty that arises in the use of ellipses is the transformation of the annotated ground-truth ellipses, when applying an affine transformation M_t on the image. For solving this, we obtained the parameters of the ellipse's general equation in its standard (3) and matrix forms (4). Using the matrix representation M_e, and the affine-transformation matrix M_t, the new parameters of the ellipse are obtained from the resulting transformed matrix form of the ellipse (5).

$$A\cdot x^2 + B\cdot x\cdot y + C\cdot y^2 + D\cdot x + E\cdot y + F = 0 \tag{3}$$

$$\begin{pmatrix} x & y & 1 \end{pmatrix}\cdot M_e \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = 0, M_e = \begin{bmatrix} A & B/2 & D/2 \\ B/2 & C & E/2 \\ D/2 & E/2 & F \end{bmatrix} \tag{4}$$

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{pmatrix} = M_t \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \Rightarrow (\tilde{x}\,\tilde{y}1)\cdot \left(\mathrm{M_t^{-1}}\right)^{\mathrm{T}}\cdot \mathrm{M_e}\cdot \mathrm{M_t^{-1}}\cdot \begin{pmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{pmatrix} = 0 \tag{5}$$

This transformation is applied on ground-truth and network-predicted ellipses, since the original size of the image does not necessarily match the network's input size, so transforming the ellipse is required together with transforming/resizing the image. This transformation is also used when applying Data Augmentation, as random translations and scale changes can be represented as affine transformations.

4. Experiments

4.1. Methods to be compared

In this work, four methods for detecting rocks are compared: CenterNet, Rocky-CenterNet, Mask R-CNN, and Rocky-YOLO, which is used as a baseline. CenterNet (Zhou et al., 2019) is a one-stage object detector in which detections are represented as bounding boxes. It works by detecting the center of the rocks, and then regressing the properties of the bounding box. The reader can refer to the original paper for a detailed description of the method. Rocky-CenterNet, the method proposed in this work, is a modification of the CenterNet detector in which the detections are represented as ellipses. The reader can refer to Section 3 for a detailed description of the proposed method. Mask R-CNN (He et al., 2017) is a method that detects objects, and then predicts a segmentation mask for each, which represents the shape of the object with a high level of detail. A complete description of the method is provided in the original paper. Finally, Rocky-YOLO (Lobos et al., 2019), a detector based on YOLO that also uses ellipses.

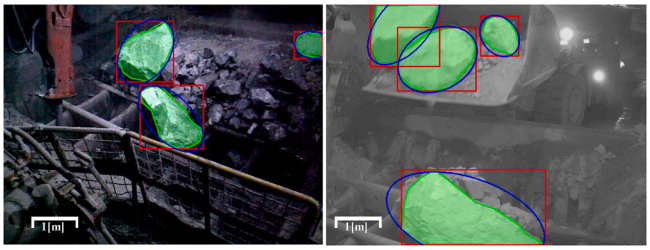


Fig. 4. Examples of rock annotations in various formats: polygons (green), ellipses (blue) and bounding boxes (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

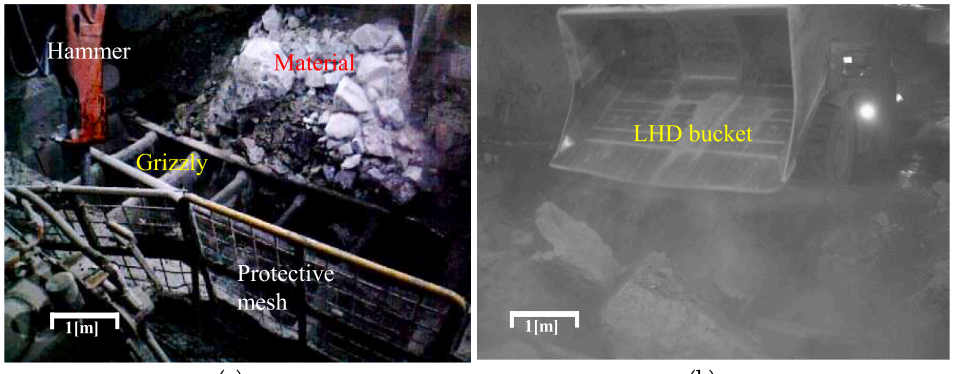
4.2. Datasets

Experiments using two datasets are reported: The *Hammer-Rocks* dataset, and the *Scaled Front View* dataset. In both datasets, mostly rocks that in the eyes of an operator would require reduction or repositioning to go through the further mining process stages, are considered. This approach of learning to detect just oversized rocks has two benefits. On the one hand, the labeling process is shorter and requires less effort, while on the other hand, procedures for classifying the kind of rock are not needed, as they are included into the detector.

4.2.1. Hammer-Rocks dataset

This dataset includes images that were captured in two different sessions in a Chilean underground mine, in an ore pass with a rock breaking hammer in operation (see Fig. 5).

In the dataset's images the oversized rocks to be detected were labeled manually as polygons, as these enable describing the shape of the rocks with a high level of accuracy. Dust accumulates in the air when LHDs dump material, making it difficult, or even impossible, to observe



(a) (b)

Fig. 3. (a) A Rock-breaking hammer in an ore pass (productive area of an underground mine). (b) An LHD dumping material in an ore pass. Part of this material (boulders) will be broken by the hammer.

the rocks in some images. Also, in some cases, it can be hard to decide which rocks are oversized. These ambiguities are inherent to the task (detecting rocks in underground environments), and cause the problem to be highly challenging. Labeled data was divided in batches, each containing images captured during 5 min. For each session, batches were assigned to either the training, validation or test set, such that 60% of the images were used for training the detectors, 20% for model validation, and 20% for test. The total number of images contained in each subset, and in the full dataset, is shown in Table 1. Some images of the Hammer-Rocks dataset are shown in Fig. 5.

Additionally, a custom criterion was used for evaluating the performance on the Hammer-Rocks dataset. For this dataset, rocks were divided in three categories: small rocks, which do not require further processing and can fall through the grizzly without intervention, medium sized rocks, which require some form of repositioning to fall through the grizzly, and large rocks, which require size reduction with the hammer to fall through the grizzly. The category labels for each the rocks were determined by expert developer engineers from our research center, which have practical knowledge about mining applications. Then, rocks belonging to the medium or large categories were considered the rocks which must be detected by the algorithms.

4.2.2. Scaled Front View dataset

Additionally, the Scaled Front View dataset from the work of Lobos et al., (2019) was also used in this work. This dataset consists of 556 images, captured under 6 different conditions, i.e., during 6 different sessions. To enable meaningful evaluation of the methods, images of each session were assigned either to the training, validation, or testing, resulting in a division of 345 / 100 / 110 for the training / validation / test sets. Unlike the previous datasets, this one is annotated with ellipses instead of polygons. As this dataset was captured in a laboratory setting, the size of the smallest rock is 10 cm, while the size of the larger one is 25 cm. An example image of this dataset and its annotations is shown in Fig. 6.

Images for this dataset were taken in six different sessions, each with different conditions and rock configurations. Images from each session was either assigned to train, validation and test sets, as show in Table 2.

4.3. Training procedure

The detectors being considered are trained by using the training sets from both the Hammer-Rocks dataset, and the Scaled Front View dataset:

- Mask R-CNN is trained over the two rock datasets, using their polygonal annotations. Ellipses in the Scaled Front View dataset are converted to polygons, using a 32-point approximation.
- CenterNet is trained over the two rock datasets, using the bounding boxes of the polygon that encloses the rock's shape.
- Rocky-CenterNet (the proposed method) is trained on the two rock datasets, using ellipses for describing the rock's shapes.
- Rocky-YOLO is trained using the same configuration as Rocky-CenterNet, i.e. ellipses are used for describing the rock's shapes.

The training configurations for each model are detailed in the following subsections:

4.3.1. Mask R-CNN

The model is first pretrained for the Instance Segmentation task on the MSCOCO Dataset (Lin, et al., 2014). Training on the rock datasets is then done in two phases, initializing the backbone with MSCOCO weights. In the first stage, only the head layers are trained for 500 epochs. In the second stage, all layers are trained for 700 epochs. The optimizer used for training the model is Stochastic Gradient Descent with Momentum ($\beta = 0.9$), and the gradient norms were clipped to 10. A batch size of 8 was used to estimate gradients during training. The learning rate was initially set at 10^{-4}, and then was reduced to the following values:

- $5{\cdot}10^{-5}$ at epoch 100
- 10^{-5} at epoch 200
- $2{\cdot}10^{-6}$ at epoch 500
- 10^{-6} at epoch 700
- $2{\cdot}10^{-7}$ at epoch 800

Data Augmentation is used to create synthetic data, and thus augment the training set and reduce overfitting. The techniques used are random crop & pad, random horizontal flips, and color augmentations (Krizhevsky et al., 2012). Additionally L2-weight regularization, with a factor of 0.0001 was used as an additional measure to avoid overfitting.

The number of epochs can be considered high in the second stage. However, in this stage both the backbone and the head's layers are jointly trained. Then, the number of epochs is in line with what would be expected, when considering that the original Mask R-CNN was trained for a similar number of epochs, but with 100x-1000x larger learning rates. Also, using a lower batch-size, lower learning rates, and a higher number of epochs, can offer a regularization effect, and stable learning (Goodfellow et al., 2016).

4.3.2. CenterNet and Rocky-CenterNet

Similar training approaches are used for both CenterNet and Rocky-CenterNet. Given that CenterNet's task is to detect objects and regress their bounding boxes, the model is pretrained on the detection task in the MSCOCO dataset. On the other hand, Rocky-CenterNet's weights are initialized using Xavier initialization (Glorot and Bengio, 2010), and trained from scratch. Both models are trained using Data Augmentation. The techniques used are random crop & pad, random horizontal flips, and color augmentations. The optimizer used was Adam (Kingma and Ba, 2015).

CenterNet showed the best results when trained using a batch size of 32. The learning rate was initially set at $4{\cdot}10^{-4}$, and was reduced to the following values:

- $4{\cdot}10^{-5}$ at epoch 20
- $4{\cdot}10^{-6}$ at epoch 50
- $4{\cdot}10^{-7}$ at epoch 80

On the other hand, Rocky-CenterNet showed the best results using a batch size of 8. The learning rate schedule used an initial learning rate of 10^{-4}, which was then reduced to:

- 10^{-5} at epoch 20
- 10^{-6} at epoch 50
- 10^{-7} at epoch 80

For Rocky-CenterNet, $c_{ang} = 5$, was used, i.e., the orientation bins had a size of $B = 36^{\circ}$. This value was selected because it provided the best performance for the task of predicting the orientation of the rocks.

Additionally, the use of rotation augmentation was tested in Rocky-CenterNet. Mild rotations (up to 5°) did not show improvement in terms of training and validation set performance, while severe rotations (up to 30°) dropped performance. This phenomenon is explained because random changes in aspect ratio (resulting from random crops & pad) also cause minor changes in the ground-truth ellipse orientations. Additionally, more severe rotation augmentations probably caused disruption in learning the environment's structure (grizzly, hammer, protective mesh, etc.)

4.3.3. Rocky-YOLO

The same training procedure used for Rocky-CenterNet was used for Rocky-YOLO. In this case, as with Rocky-CenterNet, the model does not

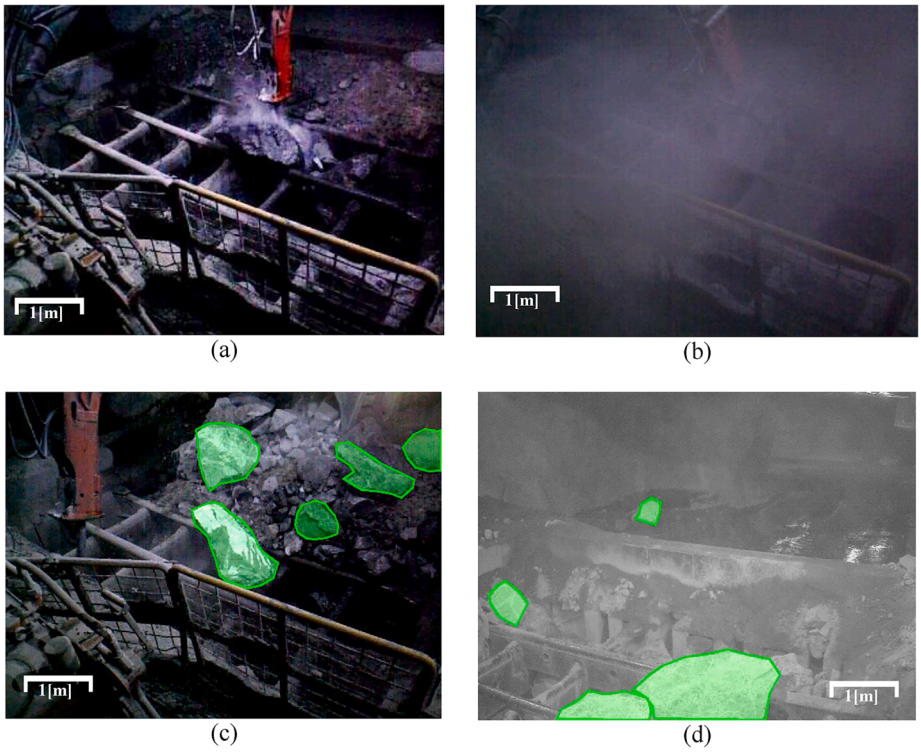


Fig. 5. Images from the Hammer-Rocks dataset: (a) rock breaking hammer, (b) image with dust, and (c)-(d) images with annotated rocks.

Table 1
Statistics for datasets captured in rock breaking hammer environments.

Name of dataset	Total time	Total images	Images used for training	Images used for validation	Images used for testing
Hammer-Rocks session-1	3 h	707	435	139	133
Hammer-Rocks session-2	1.6 h	546	333	106	107
Hammer-Rocks full	4.6 h	1253	768	245	240

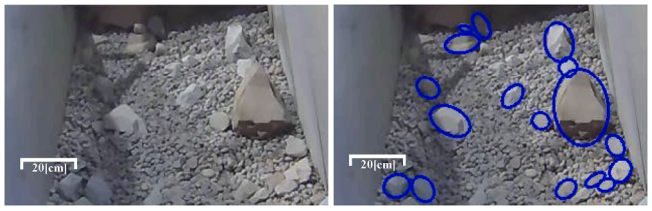


Fig. 6. Image and ground-truth annotated ellipses from the Scaled Front View dataset.

Table 2
Scaled Front View dataset division.

Session ID	Total Images	Training	Validation	Test
1	86			X
2	99	X		
3	205	X		
4	42	X		
5	100		X	
6	23			X

benefit by using weights trained for detection on the MSCOCO dataset, and thus, uses randomly initialized weights. In terms of Data Augmentation, the same techniques used for Mask-RCNN and Rocky-CenterNet / CenterNet were used. The optimizer used was RMSProp with $\beta = 0.9$.

Best performance was obtained using a batch size of 32. The learning rate schedule was defined with an initial learning rate of 10^{-2}, which was later reduced to:

- 10^{-3} at step 50
- 10^{-4} at step 100
- $5{\cdot}10^{-5}$ at step 150
- 10^{-6} at step 250

4.4. Analysis of representations for the ground-truth annotations

In the Hammer-Rocks dataset, rocks are annotated as polygons, and in the Scaled Front View dataset, rocks are annotated as ellipses. On the other hand, the different detectors tested in this work represent rocks as segmentation masks, ellipses, or boxes. Therefore, evaluating the quality of each representation type, and defining the procedures for transforming annotations between different representations are needed. It must be noted that runtime of the detectors must also be considered, and that the best representation to use depends on the application.

To obtain an ellipse from a polygonal annotation, we propose the use of the Minimum Volume Enclosing Ellipsoid Algorithm (Khachiyan, 1996), which permits obtaining the minimum area ellipse, given a set of points such that all points are enclosed. Obtaining a bounding box from a polygonal annotation, or from an ellipse, is straightforward, and can be done using the horizontal and vertical bounds of the polygon or ellipse.

In this work we propose the use of ellipses to approximate the shapes of rocks. The reasoning behind this is that, although rocks come in various shapes and sizes, ellipses might be more flexible and precise than bounding boxes for representing their shapes. In addition, ellipses are a more suitable representation for several applications. As an example, in the application of autonomous loading for an LHD, determining the sizes of the rocks from images is needed for evaluating whether these can fit

into the shovel, or not. However, areas from bounding boxes usually overestimate the rock's actual dimensions. Thus, using ellipses instead of bounding boxes helps to improve the estimation of the areas of the rocks.

A second application is the automation of rock breaking hammers. In this case, a striking point on the rock must be selected. As bounding boxes are a bad fit for the shape of the rocks, points inside bounding boxes do not necessarily belong to the rock. This problem can be solved by using ellipses for representing the rocks, since the chance of this happening is significantly reduced. Also, for breaking large rocks, the striking point must be near its border, as these rocks are reduced in several steps. Selecting points near the border of the rock is not possible when using boxes, but possible when using ellipses. Finally, as shown in Section 5.5, detecting rocks as ellipses is as fast as detecting rocks as bounding boxes, while detecting rocks as segmentation masks is significantly slower than using boxes. For instance, CenterNet-DLA is shown to be 2.5x faster than Mask R-CNN in the work of Zhou et al., (2019).

Annotated polygons have a larger overlap with the ellipses than with the bounding boxes in the Hammer-Rocks dataset. For measuring this effect, the distributions for the overlap between the original polygons, and the resulting ellipses and bounding boxes were computed. The histograms showing the distributions are presented in Fig. 7.

From the computed distributions, we observe that approximating rocks with ellipses yields intersection over union (IoU) values over 0.49 in all rocks, and that over half the annotated polygons overlap>75% with the resulting ellipses. Then, although rocks are far from being perfectly elliptical, the use of ellipses offers a better fit for approximating rock shapes, compared with bounding boxes. Hence, it can be concluded that the representations of rocks, ordered from best to worst ability to represent the true shapes, are: (i) polygons, (ii) ellipses, and (iii) bounding boxes.

5. Experimental results

5.1. Methodology

As mentioned above, the three rock detectors were tested on two datasets: the Hammer-Rocks dataset, and the Scaled Front View dataset. Average precision and recall were computed for several different sizes of rocks, following the COCO protocols[3]. Average precisions correspond to the area under the precision-recall curve for a given IOU threshold. Average recalls correspond to the fraction of the ground truth rocks detected when considering at most 100 detections per image. If no IOU is specified in a metric, then it is defined as the mean of metrics computed at several IOU values. The metrics used are defined in Table 3.

For each dataset, the average precision and average recall metrics were computed with respect to various ground-truth representations (polygons, bounding boxes, and ellipses). Then, since the representation to be used depends on the application, the performances of the methods for each representation were compared. Also, the influence on these measurements of both rock sizes, and the intersection-over-union allowed were analyzed.

5.2. Results using the Hammer-Rocks dataset

Average precision and recall were computed for several different sizes of rocks, following the COCO protocols. The results are presented on Tables 4 and 5. These are formatted as follows: in the vs *polygon* measurements, the detections outputted (segmentation masks in Mask R-CNN, rectangles in CenterNet, and ellipses in Rocky-CenterNet) are compared with the ground-truth polygons. In the vs *bounding box* metrics, the rectangle that encloses the Mask R-CNN segmentation mask and the rectangle predicted by CenterNet are evaluated against the ground-truth polygon's bounding box. And in the vs *ellipse* metrics, the ellipse predicted by Rocky-CenterNet and Rocky-YOLO are compared against the ground-truth ellipse. Metrics of the type vs *bounding box* were not calculated for Rocky-CenterNet because, in general, bounding ellipses for polygons are larger than the polygons itself. Thus, the boxes computed from Rocky-CenterNet's detected ellipses are not directly comparable with the ground-truth boxes.

From Tables 4 and 5, Mask R-CNN achieves the best average precision and recall, followed by Rocky-CenterNet, and CenterNet, when comparing each model's detections with the ground-truth polygons. CenterNet achieves the best results when comparing them with ground-truth bounding boxes. These results are seen throughout all rock sizes. Also, the metrics on this dataset are similar for various rock sizes, although Mask R-CNN has a slighter lower performance on small rocks.

It must be noted that the metrics are higher for the bounding box representation than for the ellipse and polygon representations, in most cases. This effect is explained by the area of intersection for a given configuration being dependent on the shapes involved, and does not necessarily imply that boxes are the most accurate representation. This effect is illustrated in Fig. 8.

It can also be noted that the ellipses predicted by Rocky-CenterNet obtain high AR and AP values when being compared against ground-truth ellipse annotations. Moreover, Rocky-CenterNet also improves over Rocky-YOLO by a large margin for all rock sizes. This shows that Rocky-CenterNet can learn elliptical representations effectively and better than Rocky-YOLO, and that this task can be accomplished with similar performance in terms of AP and AR, as the task of detecting polygons from the segmentation masks of Mask R-CNN, or bounding boxes with CenterNet.

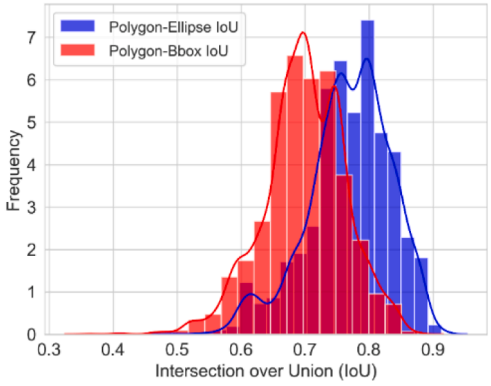
In addition, average precision and recall were computed for different levels of overlap, measured by its intersection over union, following the COCO protocols. The objective of this experiment was to analyze the performance of the detectors in representing the shapes for each kind of representation accurately. The results were obtained on test images from the Hammer-Rocks dataset. Results are presented on Tables 6 and 7.

One factor to be considered in the metrics obtained by comparing the ellipse predicted by Rocky-CenterNet with the ground-truth polygon, is that the error can be divided in two categories. The first one corresponds to a non-avoidable error that stems from the approximation of polygons with ellipses, caused by the fact that the annotated shapes do not correspond perfectly to ellipses. The second category is a fitting error, which corresponds to the error between the predicted and ground-truth ellipses. Clearly, we wish to decrease the latter as much as possible.

This previously mentioned phenomenon can be observed in the drops of Average Recall (AR) and Average Precision (AP), while increasing the IoU threshold. When representing ground-truth annotations as polygons, Mask R-CNN achieves the best performance on all AP and AR metrics, followed by Rocky-CenterNet, and then by CenterNet. For both CenterNet and Rocky-CenterNet, the metrics $AP^{IOU=.50}$ and $AR^{IOU=.50}$ are comparable to that of Mask R-CNN. However, for the stricter metrics $AP^{IOU=.75}$ and $AR^{IOU=.75}$, the performance of both CenterNet and Rocky-CenterNet are much poorer than that of Mask R-CNN. Then, as the metrics at IOU = 0.50 are affected less by the shape of the object than the metrics at IOU = 0.75, it can be concluded that all three detectors are able to detect objects annotated as polygons. However, the ability of Mask R-CNN to learn polygonal shapes is greater than that of Rocky-CenterNet, and even greater than that of CenterNet.

When representing the shapes of the rocks as bounding boxes or ellipses, the drop in the metrics between IOU = 0.50 and IOU = 0.75 is much smaller than in the case of polygonal annotations. This fact shows that the boxes predicted by both Mask R-CNN and CenterNet fit to ground-truth bounding boxes, and the ellipses predicted by Rocky-CenterNet fit to ground-truth ellipses. Additionally, it can be noted that Rocky-CenterNet is able to improve largely over Rocky-YOLO for all IoU levels, then, for detecting rocks as ellipses, Rocky-CenterNet is recommended.

[3] https://cocodataset.org/#detection-eval



(a)

	Ellipse	BBox
Minimum	0.4929	0.3640
Maximum	0.9135	0.8768
Mean	0.7692	0.6994
Q1	0.7307	0.6639
Median	0.7725	0.7010
Q3	0.8186	0.7476

(b)

Fig. 7. (a) Histograms showing the overlaps between annotated polygons and their bounding boxes (BBox), and between annotated polygons and ellipses in the Hammer-Rocks dataset. (b) Summary table with key overlap measurements between polygons and their bounding boxes, and between polygons and obtained ellipses.

Table 3
Average precision and recall as defined in the COCO protocol.

$AP^{IOU=0.50}$	Average precision at IOU = 0.50 (Pascal VOC metric)
$AP^{IOU=0.75}$	Average precision at IOU = 0.75 (strict metric)
AP	Average precision = mean of ($AP^{IOU=0.50}$, $AP^{IOU=0.55}$, …, $AP^{IOU=0.95}$)
AP^{small}	Average precision for objects with area < 32x32
AP^{medium}	Average precision for objects with area between 32x32 and 96x96
AP^{large}	Average precision for objects with area > 96x96
$AR^{IOU=0.50}$	Average recall at IOU = 0.50
$AR^{IOU=0.75}$	Average recall at IOU = 0.75
AR	Average recall = mean of ($AR^{IOU=0.50}$, $AR^{IOU=0.55}$, …, $AR^{IOU=0.95}$)
AR^{small}	Average recall for objects with area < 32x32
AR^{medium}	Average recall for objects with area between 32x32 and 96x96
AR^{large}	Average recall for objects with area > 96x96

Thus, the best methods on this dataset are: Mask R-CNN for representing polygonal annotations, CenterNet for representing bounding boxes, and Rocky-CenterNet for representing ellipses. Although polygonal annotations provide the best descriptions of the rocks, and bounding boxes the worst ones, the kind of representation best suited to a task depends on the application. In the case of autonomous loading with LHDs and rock breaking hammers, ellipses are the most suitable representations, when considering both approximation quality (as shown in Section 4.4), and inference times (as shown in Section 5.5). Also, for this application, ellipses are easier to use for selecting breaking points, as segmentation masks provide an excess of non-useful information.

5.3. Results using the Scaled Front View dataset

Average precision and recall were computed for several different sizes of rocks, following the COCO protocols. The results were obtained on test images from the Scaled Front View dataset. Since this dataset is annotated by using ellipses, ground-truth polygons are not available. Results are presented in Tables 8 and 9.

From these tables, it can be concluded that when ground-truth annotations are bounding boxes, CenterNet achieves the best results in most cases, followed by Rocky-CenterNet, and then by Mask R-CNN. When ground-truth annotations are represented as ellipses, Rocky-CenterNet achieves the best results, with the second-best performance achieved by CenterNet on small rocks, and Mask R-CNN on large rocks. The performances of all the detectors, but Rocky-YOLO, increase with the size of the rocks, which can be explained by the fact that this dataset contains a large amount of rocks per image (see Fig. 6), but only rocks considered to be oversized are annotated. In addition, Rocky-CenterNet improves over Rocky-YOLO in all the metrics evaluated in Tables 8 and 9.

Thus, when the representation to be used in an application is a

Table 4
Average precision (AP) for different detection areas, following the COCO criterion, for the Hammer-Rocks datasets. Results for several annotation representations (polygons, bounding boxes, and ellipses) of different rock sizes (all, small, medium, large) are considered.

Measurements/size	vs polygon			vs bounding box		vs ellipse	
	Mask R-CNN	CenterNet (w/flip[1])	Rocky CenterNet	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Rocky-YOLO
AP (all)	0.714	0.511	0.595	0.721	0.762	0.732	0.496
AP (small)	0.649	0.501	0.614	0.674	0.736	0.731	0.499
AP (medium)	0.737	0.522	0.601	0.746	0.762	0.755	0.509
AP (large)	0.735	0.518	0.603	0.722	0.786	0.738	0.502

[1]w/flip indicates that both normal and horizontally flipped versions of the images are used for performing detections.

Table 5
Average recall (AR) for different detection areas, following the COCO criterion, for the Hammer-Rocks datasets. Results for several annotation representations (polygons, bounding boxes, and ellipses) for different rock sizes (all, small, medium, large) are considered.

Measurements/size	vs polygon			vs bounding box		vs ellipse	
	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Rocky-YOLO
AR (all)	0.770	0.572	0.66	0.777	0.819	0.793	0.546
AR (small)	0.726	0.563	0.644	0.754	0.825	0.793	0.573
AR (medium)	0.778	0.574	0.666	0.794	0.810	0.808	0.523
AR (large)	0.783	0.576	0.665	0.769	0.827	0.775	0.556

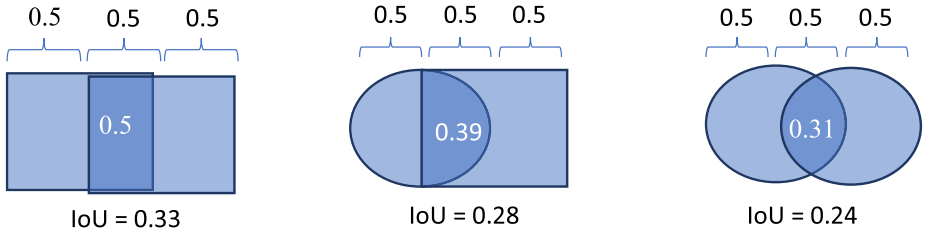


Fig. 8. For similar configurations, the intersection over the union between two squares is higher than that between a circle and a square, or between two circles. Example values are shown on the figure.

Table 6
Average precision (AP) for different overlaps, following the COCO criterion on the Hammer-Rocks dataset. Results for several annotation representations (polygons, bounding boxes, and ellipses) were considered.

Metric	vs polygon			vs bounding box		vs ellipse	
	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Rocky-YOLO
AP	0.714	0.511	0.595	0.721	0.762	0.732	0.496
$AP^{IOU=.50}$	0.934	0.909	0.938	0.933	0.955	0.947	0.768
$AP^{IOU=.75}$	0.750	0.045	0.151	0.746	0.813	0.758	0.171

Table 7
Average recall (AR) for different overlaps for the Hammer-Rocks dataset, following the COCO criterion. Results for several annotation representations (polygons, bounding boxes, and ellipses) were considered.

Metric	vs polygon			vs bounding box		vs ellipse	
	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Rocky-YOLO
AR	0.770	0.572	0.660	0.777	0.819	0.793	0.546
$AR^{IOU=.50}$	0.962	0.957	0.974	0.962	0.991	0.982	0.810
$AR^{IOU=.75}$	0.821	0.130	0.348	0.813	0.870	0.827	0.311

Table 8
Average precision (AP) for different detection areas, following the COCO criterion, for the Scaled Front View dataset. Results for several annotation representations (bounding boxes and ellipses) for different rock sizes (all, small, medium, large) are considered.

Measurement/size	vs bounding box			vs ellipse			
	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Rocky-YOLO
AP (all)	0.591	**0.660**	0.598	0.537	0.549	**0.588**	0.361
AP (small)	0.572	**0.622**	0.580	0.510	0.520	**0.569**	0.369
AP (medium)	0.659	**0.739**	0.677	0.640	0.618	**0.668**	0.344
AP (large)	0.767	0.880	**0.883**	0.834	0.628	**0.883**	0.000

Table 9
Average recall (AR) for different detection areas, following the COCO criterion, for the Scaled Front View dataset. Results for several annotation representations (bounding boxes and ellipses) for various rock sizes (all, small, medium, large) are considered.

Measurement/size	vs bounding box			vs ellipse			
	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Rocky-YOLO
AR (all)	0.657	**0.731**	0.671	0.598	0.618	**0.653**	0.429
AR (small)	0.649	**0.711**	0.660	0.582	0.602	**0.641**	0.439
AR (medium)	0.694	**0.782**	0.723	0.670	0.666	**0.708**	0.388
AR (large)	0.767	0.889	**0.900**	0.833	0.652	**0.900**	0.000

bounding box, CenterNet provides the best results. However, when the representation to be used is an ellipse, the best method is Rocky-CenterNet. It must be noted that ellipses provide a better fit on annotated rocks than bounding boxes, as is discussed in Section 4.4. So, Rocky-CenterNet provides the best performance in representing the shapes of the rocks in this dataset.

Moreover, average precision and recall were also computed for different levels of overlap, measured by its intersection over union, following the COCO protocols. Since this dataset is annotated by using ellipses, ground-truth polygons are not available. The objective of this experiment was to analyze the performance of the detectors in representing the shapes for each kind of representation accurately. The results were obtained on test images from the Scaled Front View dataset. Results are presented in Tables 10 and 11.

When representing ground-truth annotations as ellipses, CenterNet provides the best metrics at IOU = 0.50, but Rocky-CenterNet is even better at IOU = 0.75. Then, since the metrics at IOU = 0.50 are less affected by the shape of the object than the metrics at IOU = 0.75, it can be concluded that the ability of detecting objects of both methods is similar, but the ability of Rocky-CenterNet at learning elliptical shapes is

Table 10
Average precision (AP) for different overlaps, following the COCO criterion, for the Scaled Front View dataset. Results for several annotation representations (bounding boxes and ellipses) were considered.

Metric	vs bounding box			vs ellipse			
	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Rocky-YOLO
AP	0.591	**0.660**	0.598	0.537	0.549	**0.588**	0.361
$AP^{IOU=.50}$	0.847	**0.908**	0.873	0.824	**0.880**	0.865	0.585
$AP^{IOU=.75}$	0.432	**0.552**	0.390	0.247	0.137	**0.374**	0.045

greater than those of both CenterNet and Mask R-CNN. Also, when ground-truth annotations are represented as bounding boxes, CenterNet performs better than the other methods in all the metrics. It can be noted also that Rocky-CenterNet improves by a large margin over Rocky-YOLO for all IoU levels. As for all the metrics shown in this work, Rocky-CenterNet improves over Rocky-YOLO by close to 50 %, thus, use of the latter method is not recommended.

Once again, CenterNet is the best performing method when using bounding boxes, except when considering large rocks, in which case it is beaten by Rocky-CenterNet. Also, Rocky-CenterNet is again the best performing method when using ellipses. This is important to be considered along with the fact that ellipses allow a better approximation of a rock's shape compared to bounding boxes, as shown in Fig. 7. Also, ellipses provide a better representation than bounding boxes for two tasks, as described in Section 4.4: autonomous loading with LHDs, and autonomous rock breaking hammers. Moreover, in both tasks, ellipses are also more suitable than segmentation masks, as the latter provide excessive information which is not useful for solving the tasks of interest. Thus, as ellipses are the most useful representation for these two tasks, and Rocky-CenterNet is the detector best suited to detect ellipses, it is a viable option for detecting rocks in these two applications.

5.4. Visual results

Examples of detections using the different models are shown in Figs. 9 and 10. Ground-truth annotations are shown in green, Mask R-CNN detections in red, while Rocky-CenterNet's detected ellipses and CenterNet's detected bounding boxes are in blue. It can be observed that Mask R-CNN shows a higher tendency to generate false positives, compared to Rocky-CenterNet and CenterNet. Although these smaller-rock detections can be considered as valid rock detections, they do not correspond to rocks in the medium to large size category, and therefore were not considered as rocks of interest while labeling, causing these detections to be incorrect. Then, while the scale invariance property from Mask R-CNN is desirable for most applications, in the case of rock detection it causes the detector not learning to differentiate oversized rocks from small, irrelevant rocks. This flaw is compensated by the slightly higher quality of an approximation of the shapes of the rocks. On the other hand, Rocky-CenterNet and CenterNet usually detect the same rocks, with less of a tendency for false positives, when compared to Mask R-CNN. But once again, the advantage of Rocky-CenterNet, due to the use of ellipses, gives the proposed method an advantage.

5.5. Inference times

The inference times of the three best performing methods under analysis are measured using three different computer platforms, two corresponding to standard platforms used for evaluating objects detectors - Tesla V100 GPU and Titan Xp GPU-, and one corresponding to an industrial PC which can be used in mining environments considered in this work. The results are shown in Table 12.

First, inference times were measured using a Tesla V100 GPU contained in the DGX-1 cluster[4], From Table 12, Mask R-CNN requires about 2.5–3.5 times the processing time of CenterNet or Rocky-CenterNet for processing an image, when using this GPU. Also, the processing times required by CenterNet and Rocky-CenterNet are similar, with Rocky-CenterNet having a slight edge. Secondly, inference times were also measured on a Titan Xp GPU for the three models.[5] As shown in Table 12, in this platform Rocky-CenterNet processes an image 1.7 times faster than CenterNet, and 4 times faster than Mask R-CNN.

It is important to consider than in real mining application as the ones considered in this work –rock pile characterization from an LHD vehicle and control of a rock breaking hammer-, the two mentioned computer architectures cannot be used because of the robustness requirements. Instead, an industrial computer equipped with a GPU is the only alternative to be used to run this kind of CNN in these harsh environments. Taking these requirements into consideration, the third comparison was carried out using a top of the line industrial PC Abox-5210G[6], equipped with an NVIDIA GeForce GTX 1650 GPU[5], an Intel Core i7-10700TE CPU, and 32 GB ram. It can be observed that in this case, Rocky-CenterNet runs at around 20 frames per second, CenterNet at around 10 frames per second, and Mask R-CNN at 5 frames per second; Rocky-CenterNet processes images 1.9 times faster than CenterNet, and 4 times faster than Mask R-CNN.

When analyzing if these inference times allows the use of these architectures in real time, it must be considered that in the mining applications considered in this work, the industrial PC will run other processes along with the rock detector, reducing the available computer resources for rock detection. In the case of using the rock detector for characterizing the rock pile from an LHD vehicle, the internal LHD's PC will need to control the LHD navigation (self-localization, path planning, path following. etc.), acquire and analyze the images, and acquire and process the 2D LIDAR data, all of these in real-time. In the case of the rock hammer application, the industrial PC will need to control the hammer, acquire and analyze the images coming from two different cameras, and acquire and process the data coming from two 3D LIDAR sensors. Considering these requirements, Rocky-CenterNet seems to be the only detector which is fast enough for these applications.

5.6. Analysis of the results

Experiments using the Hammer-Rocks dataset show that Rocky-CenterNet provides better performance than CenterNet in all cases when the ground-truth annotations are represented as polygons. These results are shown on Tables 4 - 7. It is evident that Rocky-CenterNet provides a better approximation of the shapes of the rocks than CenterNet. Also, Rocky-CenterNet shows better performance than Rocky-YOLO in all the tests performed. Experiments using this dataset also show that Mask R-CNN provides the most accurate prediction for the shapes of the rocks. However, as shown on Table 12 , Mask R-CNN provides a frame rate that is much lower than that of the other detectors, as Mask R-CNN runs at only 5 fps in an industrial computer equipped with a NVIDIA GeForce GTX 1650 GPU[5], being 2x slower than CenterNet and 4x slower than Rocky-CenterNet. Also, the speed tests were performed with no other processes being run in the industrial computer, but

[4] https://www.amax.com/products/nvidia-products/nvidia-dgx-1/

[5] https://www.nvidia.com/en-us/geforce/graphics-cards/gtx-1650/

[6] https://www.sintrones.com/ABOX-5210G.html

Table 11
Average recall (AR) for different overlaps, following the COCO criterion, for the Scaled Front View dataset. Results for several annotation representations (bounding boxes and ellipses) were considered.

Metric	vs bounding box			vs ellipse			
	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Mask R-CNN	CenterNet (w/flip)	Rocky CenterNet	Rocky-YOLO
AR	0.657	**0.731**	0.671	0.598	0.618	**0.653**	0.429
$AR^{IOU=.50}$	0.892	**0.967**	0.936	0.873	**0.941**	0.929	0.664
$AR^{IOU=.75}$	0.550	**0.655**	0.514	0.377	0.307	**0.477**	0.156

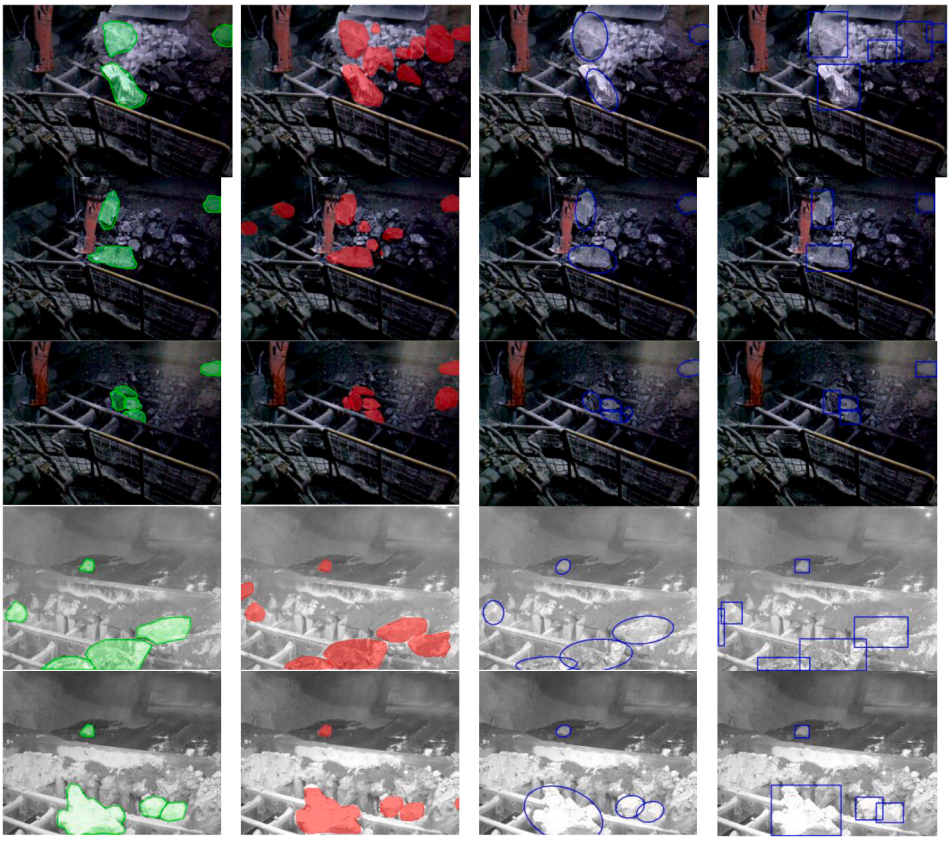


Fig. 9. Image examples of the Hammer-Rocks Dataset. Each row corresponds to a different image. From left to right: Ground-truth annotations in green, Mask R-CNN detections in red, Rocky-CenterNet detected ellipses in blue, and CenterNet detected bounding boxes in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

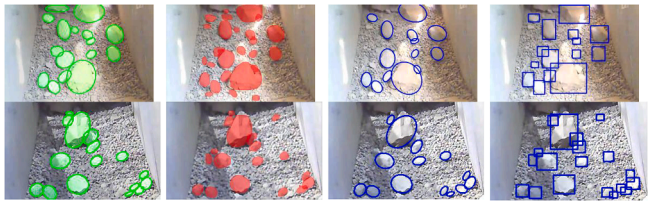


Fig. 10. Image examples of the Scaled Front View Dataset. Each row corresponds to a different image. From left to right: Ground-truth annotations in green, Mask R-CNN detections in red, Rocky-CenterNet detected ellipses in blue, and CenterNet detected bounding boxes in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in real robotic applications like autonomous loading of LHDs and autonomous rock breaking hammers, a large number of processes must run in parallel to rock detection. Then, the 5 fps frame rate achieved by Mask R-CNN is an optimistic estimation of its runtime in real applications. Also, the frame rate provided by Mask R-CNN is not high enough for being used in applications which require real time processing. Hence, when real time processing is required, Rocky-CenterNet emerges as a feasible alternative, as Mask R-CNN is far from running in real time.

Table 12
Inference times for the different methods.

Model	Inference time on Tesla V100	Inference time on Titan Xp	Inference time on GTX 1650[6] (industrial PC)
Mask R-CNN	0.056 sec/image	0.072 sec/image	0.205 sec/image
CenterNet (w/flip[1])	0.020 sec/image	0.031 sec/image	0.098 sec/image
Rocky-CenterNet	0.016 sec/image	0.018 sec/image	0.052 sec/image

[1] w/flip indicates that both normal and horizontally flipped versions of the images are used for performing detections.

Although in this paper CNN architectures used in previous works about rock detection such as SSD, Faster R-CNN, RetinaNet and YOLOv3, were not tested, the literature (Redmon & Farhadi, 2017; Zhou

et al., 2019) shows that CenterNet outperforms those methods in object detection tasks, using bounding boxes.

In the task of autonomous rock breaking (images which are captured in the Hammer-Rocks dataset), the use of ellipses shows some benefit over the use of segmentation masks for selecting breaking points. This benefit stems from ellipses being a simpler representation of a rock's shape, compared to an arbitrary polygon, which provides an unnecessary level of detail. On the other hand, Rocky-CenterNet achieves the highest performance when rocks are represented as ellipses, Thus, for autonomous rock breaking hammers, Rocky-CenterNet is a suitable rock detector since it not only provides a good trade-off between processing speed and performance in representing the shapes of the rocks, but also provides a highly convenient representation for the detected rocks.

The Scaled Front View dataset contains images captured in an experimental setup corresponding to a 1/5 scaled front with a pile, for testing autonomous loading for an LHD (Lobos et al., 2019). In this dataset, in which rocks are labeled as ellipses, Rocky-CenterNet achieved better results than CenterNet in 12 of the 14 metrics available when compared to ground-truth ellipses, while Mask R-CNN achieved the worst results. These results are shown on Tables 8 – 11. Then, Rocky-CenterNet is the best performing detector when rocks are represented as ellipses. On the other hand, from the discussion in Section 4.4, ellipses are better suited than both bounding boxes and segmentation masks for representing the shapes of rocks in the task of autonomous loading with LHDs. Therefore, since both Rocky-CenterNet and CenterNet provide similar processing speeds, which are much faster than that of Mask R-CNN, Rocky-CenterNet is a suitable alternative for detecting rocks in the task of autonomous loading with an LHD.

6. Conclusions

In this work, we present a novel approach to rock detection, named Rocky-CenterNet. Unlike other approaches, Rocky-CenterNet models rock detections as ellipses. Additionally, a new dataset named Hammer-Rocks is released, containing images captured in an underground ore pass while a rock-breaking hammer was working. This dataset is particularly challenging because of heavy dust in the environment, and ambiguity with respect to which rocks can be considered oversized by humans. Both difficulties are inherent to the problem to be solved. Rocks in this dataset were manually labeled and annotated as polygons. We found that in this dataset, ellipses provide a better approximation of the shapes of the rocks, than bounding boxes.

Four rock detection algorithms, which model detections as either bounding boxes (CenterNet), ellipses (Rocky-CenterNet and Rocky-YOLO), or segmentation masks (Mask R-CNN) were compared exhaustively. The comparisons consider both the performance of the detectors and the processing times required by each. In the Hammer-Rocks dataset, Mask R-CNN is the algorithm that provides the most accurate predictions of the shapes of the rocks. However, its frame rate is much lower than those of the other detectors. Also, on both the Hammer-Rocks dataset and on the Scaled Front View dataset, Rocky-CenterNet provides a better approximation for the shapes of the rocks than CenterNet, when rocks are represented as ellipses. Moreover, Rocky-CenterNet requires similar or lower processing times than CenterNet. The performance achieved by Rocky-CenterNet on both the Hammer-Rocks dataset and the Scaled Front View dataset show that this algorithm is useful for performing rock detection from images.

We conclude that Rocky-CenterNet is a suitable choice for detecting rocks in mining applications when real-time operation, and a good approximation of a rock's shape are needed, as in the case of autonomous loading with LHDs or autonomous rock breaking using hammers.

CRediT authorship contribution statement

Patricio Loncomilla: Conceptualization, Methodology, Resources, Data Curation, Writing - original draft, Writing - review & editing, Supervision. **Pavan Samtani:** Methodology, Software, Investigation, Data Curation, Writing - original draft, Writing - review & editing, Visualization. **Javier Ruiz-del-Solar:** Conceptualization, Methodology, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially funded by ANID (Agencia Nacional de Investigación y Desarrollo), through grants CONICYT PIA grant AFB18004, FONDECYT 1201170, and FONDEF IDEA ID19I10142. We also thank the journal's reviewers for the valuable feedback offered in the review stage.

References

Bickel, V. T., Conway, S. J., Tesson, P.-A., Manconi, A., Loew, S., & Mall, U. (2020). Deep Learning-Driven Detection and Mapping of Rockfalls on Mars. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 13*, 2831–2841. https://doi.org/10.1109/JSTARS.2020.2991588

Bickel, V. T., Lanaras, C., Manconi, A., Loew, S., & Mall, U. (2018). Automated Detection of Lunar Rockfalls Using a Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens., 57*(6), 3501–3511. https://doi.org/10.1109/TGRS.2018.2885280

Fanara, L., Gwinner, K., Hauber, E., & Oberst, J. (2020). Automated detection of block falls in the north polar region of Mars. *Planet. Space Sci., 180*. https://doi.org/10.1016/j.pss.2019.104733

Furlán, F., Rubio, E., Sossa, H., & Ponce, V. (2020). CNN Based Detectors on Planetary Environments: A Performance Evaluation. *Front. Neurorob., 14*, Article 590371. https://doi.org/10.3389/fnbot.2020.590371

Glorot, X., & Bengio, Y. (2010). *Understanding the difficulty of training deep feedforwardneural networks., 9*, 249–256.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* The MIT Press. doi: 0.5555/3086952.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.322

Khachiyan, L. G. (1996). Rounding of polytopes in the real number model of computation. *Math. Operat. Res., 21*, 307–320.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. Lake Tahoe, Nevada: Curran Associates Inc.

Li, Y., Dong, H., Li, H., Zhang, X., Zhang, B., & Xiao, Z. (2020). Multi-block SSD based on small object detection for UAV railway scene surveillance. *Chin. J. Aeronaut., 33*(1), 1747–1755. https://doi.org/10.1016/j.cja.2020.02.024

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*, (pp. 740-755).

Liu, X., Wang, H., Jing, H., Shao, A., & Wang, L. (2020). Research on Intelligent Identification of Rock Types Based on Faster R-CNN Method. *IEEE Access, 8*, 21804–21812. https://doi.org/10.1109/ACCESS.2020.2968515

Lobos, K., Loncomilla, P., & Ruiz-del-solar, J. (2019). Rocky-YOLO: Detección de rocas usando deep learning para aplicaciones en minería. *SIMIN 2019*.

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis. Comput., 22*(10), 761–767. https://doi.org/10.1016/j.imavis.2004.02.006

Niu, L. (2020). *Improving the Visual Perception of Heavy Duty Manipulators in Challenging Scenarios*. Tampere University. Doctoral thesis.

Niu, L. C., Jia, K., & Mattila, J. (2019). Efficient 3D Visual Perception for Robotic Rock Breaking. In *2019 IEEE 15th International Conference on Automation Science and Engineering* (pp. 1124–1130). https://doi.org/10.1109/COASE.2019.8842859

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 6517-6525). doi: 10.1109/CVPR.2017.690.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., & A.. (2016). You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *ArXiv, abs/1804.02767*.

Salvador, C., Mascaró, M., & Ruiz-del-Solar, J. (2020). Automation of Unit and Auxiliary Operations in Block/Panel Caving: Challenges and Opportunities. *8th Int. Conference on Mass Mining - MassMin 2020*. Santiago, Chile.

Somua-Gyimah, G., Frimpong, S., Nyaaba, W., & Gbadam, E. (2019). A Computer Vision System for Terrain Recognition and Object Detection Tasks in Mining and Construction Environments. *2019 SME Annual Conference and Expo and CMA 121st National Western Mining Conference*.

Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as Points. *ArXiv, abs/1904.07850*.