Don’t Blindly Trust Your CNN: Towards Competency-Aware Object Detection by Evaluating Novelty in Open-Ended Environments

Rhys Howard[⋆], Sam Barrett[⋆] and Lars Kunze[⋆]

***Abstract*— Real-world missions require robots to detect objects in complex and changing environments. While deep learning methods for object detection are able to achieve a high level of performance, they can be unreliable when operating in environments that deviate from training conditions. However, by applying novelty detection techniques, we aim to build an architecture aware of when it cannot make reliable classifications, as well as identifying novel features/data. In this work, we have proposed and evaluated a system that assesses the competence of trained *Convolutional Neural Networks (CNNs)*. This is achieved using three complementary introspection methods: (1) a *Convolutional Variational Auto-Encoder (VAE)*, (2) a latent space *Density-adjusted Distance Measure (DDM)*, and (3) a *Spearman’s Rank Correlation (SRC)* based approach. Finally these approaches are combined through a weighted sum, with weightings derived by maximising the correct attribution of novelty in an adversarial ‘meta-game’. Our experiments were conducted on real-world data from three datasets spread across two different domains: a planetary and an industrial setting. Results show that the proposed introspection methods are able to detect misclassifications and unknown classes indicative of *novel* features/data in both domains with up to 67% precision. Meanwhile classification results were either maintained or improved as a result.**

I. Introduction

Object detection is essential for many robotic applications including autonomous driving, industrial inspection, and planetary exploration. Over the last decade, deep learning has revolutionised the field and has set new standards in classification performance. However, deep learning methods require an enormous amount of training data and while they are generally applicable, they struggle with dataset shifts, i.e., when the distribution of inputs and outputs differs between training and deployment. This problem raises the question to what extent autonomous robot systems should rely on trained neural networks when deployed in open-ended environments where dataset shifts are bound to occur and how these systems might cope with such dataset shifts.

This work primarily relates to the space robotics project ADE[1] [1], which concerns Mars exploration and nuclear decommissioning tasks. However, we additionally verify our methods on the publicly available Mars novelty detection Mastcam labeled dataset[2] [2]. In ADE, one of the robot’s task is to detect objects of known (or typical) classes in their environment and at the same time, evaluate if an object is novel. To this end, we calculate the likelihood that a reliable classification can be made for a given candidate. This information allows us to assess the competency of trained models as well as detecting novel objects/features that are unrepresented by any classifiers.

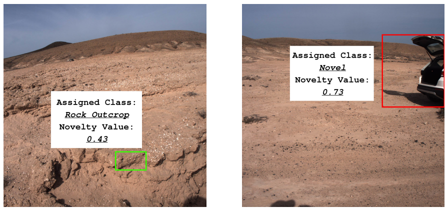


Fig. 1: Competency-aware object detection for robot missions in open-ended environments. The images show detected objects and their classifications made by *Convolutional Neural Networks (CNNs)* which were trained on *typical* objects in a desert environment. During deployment/testing, our approach also evaluates the novelty of object candidates based on the training distribution, thereby detecting cases where reliable classification may not be possible. *Left:* Low novelty score for an in-distribution object. *Right:* High novelty score for an out-of-distribution object.

In order to accomplish the tasks set out by ADE, we propose an original architecture that aims to augment a set of baseline classifiers with additional introspective components that consider the competency of the model in the current situation. The baseline is composed of a stack of *Convolutional Neural Network (CNN)* classifiers, each of which provides a single confidence value for a given object/feature class. The presented approach aims to boost the performance given by this baseline with three complementary methods: A convolutional *Variational Auto-Encoder (VAE)*, a latent space *Density-adjusted Distance Measure (DDM)* and a *Spearman’s Rank Correlation (SRC)* based approach. By combining these methods with the baseline *CNN* stack we aim to identify *novel* features which the *CNN* stack may struggle to classify reliably. In this paper, *novel* features concern both the features of unknown classes and features of known classes that are not sufficiently represented in their corresponding *CNN* classifiers (e.g. a new shape/colour of rock outcrop is a novel feature, even if the class is known and represented). This novel image data could additionally prove *interesting* to a human observer in scientific exploration or inspection tasks.

In the following, a literature review is provided (Sec. II), followed by a discussion of the proposed architecture (Sec. III), an explanation of the novelty detection approach (Sec. IV), and an experimental evaluation (Sec. V), before we finally conclude (Sec. VI).

II. Related Work

Object and feature detection is a well-studied research area with roots in pattern recognition and scene congruence [3].

[⋆] Rhys Howard, Sam Barrett and Lars Kunze are with the Oxford Robotics Institute, Department of Engineering Science, University of Oxford, 17 Parks Road, Oxford OX1 3PJ, UK. Please use rhyshoward@live.com for correspondence.

[1] https://www.h2020-ade.eu

[2] https://zenodo.org/record/3732485

Given the field's extensive history and rapid recent developments through the widespread adoption of deep neural networks this work will not discuss the full scope of this field. A comprehensive review is given in [4].

Our approaches are rooted in *CNN*-based object/feature classification as in [5], [6]. Furthermore, we utilise *Selective Search* [7] as a region proposal method which has several benefits for our application including low processing times, full-image coverage, and no requirements for training data. However, Edge Boxes [8] and Region Proposal Networks (RPN) [9] provide alternative approaches which could have been also considered in this context.

The performance of object detection during robot deployments was recently addressed by [10]. While they use a threshold-based approach by monitoring the per-frame mean average precision, our work uses a combination of three complementary introspection methods for such analysis.

The first form of novelty analysis in our proposed competency-aware approach is based upon VAEs [11] for detecting *novel* and anomalous data [12], [13], [14]. Given the emphasis on planetary exploration, the presented approach is derived from our previous work [15] on applying VAEs for novelty anaylsis of multispectral images from the Mars *Curiosity* rover [2], [16]. This paper builds upon these contributions in two regards. Firstly, the architecture proposed describes a full system which can be deployed in real-world conditions [1] and therefore experiments are carried out on datasets of rover trial locations. Secondly, this work proposes not only the combined application of traditional novelty analysis and classification techniques, but additionally several means of augmenting the overall performance either by combining this data or by factoring in new metrics of novelty.

A further novelty analysis technique we consider is based upon density-based clustering. Density-based clustering methods [17], [18], [19] are popular and robust methods for analysing data, without making assumptions about the densities of the individual clusters yet also accounting for possibility of cluster hierarchies. These methods have also garnered attention for novelty and outlier detection [20], [21]. Being inspired by the *OPTICS* algorithm [17], we suggest a modification which is suitable to our novelty detection analysis by making a connection with metric space theory.

The third novelty analysis technique adopted in this paper is based upon the Spearman's Rank Correlation Coefficient. The combination of *CNN* confidence values at prediction time is compared with the correlation of various classes from training time in order to provide a metric of how *novel/typical* a given set of confidence values is.

The simplest way to combine the outputs of these three methods into a single scalar output would be to take their mean output. However this assumes the models are all equally effective, which may not reflect reality. Instead we use the evaluation methods in [22] where a game-theoretic mode of evaluation is proposed: a two player zero-sum game where one player aims to select the best model for a task, and the other player aims to fool the models with the most difficult tasks. [22] propose that the *maximum entropy Nash equilibrium* of this game is a robust measure of the efficacy of each model (and correspondingly, the difficulty of each task). They show that this type of equilibrium has nice theoretical properties which are directly relevant to our problem.

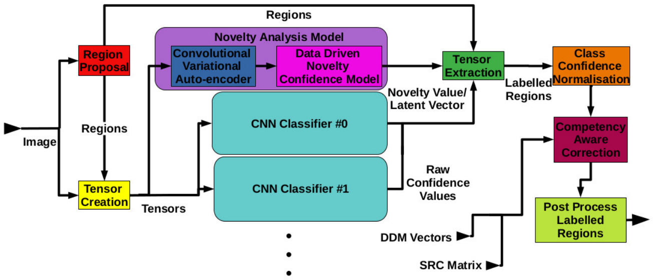

Regions
Novelty Analysis Model
Region Proposal
Convolutional Variational Auto-encoder
Data Driven Novelty Confidence Model
Tensor Extraction
Class Confidence Normalisation
Labelled Regions
Novelty Value/ Latent Vector
Image
Regions
CNN Classifier #0
Competency Aware Correction
Raw Confidence Values
CNN Classifier #1
Tensor Creation
Tensors
Post Process Labelled Regions
DDM Vectors
SRC Matrix

Fig. 2: A high level diagram of the system architecture. The input image is first processed to propose regions of interest. These regions of the image are then packed into tensors and fed through a novelty analysis networks as well as a *CNN* classifier for each class. A novelty value as well as raw confidence values for each class are then extracted from the resulting tensors. The raw class confidences are then normalised before competency aware correction is carried out through the *VAE*, *DDM* and *SRC* methods detailed in this work. Finally post processing is carried out to filter the labelled regions before they are output from the system.

III. Object Detection System Architecture

A. System Overview

We first give an overview of the object detection system before we describe its components in subsequent sections.

Given an input image, the system first generates a set of region proposals (bounding boxes) (Sec. III-B). Each region proposal is then processed by a set of *CNNs* (each trained for a specific class of interest) (Sec. III-C). Additionally, each proposal is evaluated by three complementary methods: a novelty assessment (Sec. IV-A), a density-adjusted distance measurement (Sec. IV-B), and a statistical analysis (Sec. IV-C). Finally, the probabilistic outcomes of these three methods are combined using learnt weights (Sec. IV-D). For each region proposal, the distribution of object classes is reported (including a probabilistic measure of their novelty).

B. Region Proposal Generation

The region proposal generation is based on *Selective Search* [7]. The approach first applies graph-based segmentation [23] before performing hierarchical clustering based upon a variety of distance metrics, any combination of which form a strategy. Due to the lack of necessity for high precision bounding boxes and a requirement for fast processing speed we employed the *Single Strategy* [7]. Finally, regions are filtered based upon their area. Regions that are either too small to contain sufficient data or so big as to make a spatially localised classification meaningless are removed.

C. CNN-based Classification

To classify the individual region proposals each of them is passed through a stack of trained Convolutional Neural Networks (*CNNs*). While one could also train a single *CNN* to classify regions, we have opted for a set of models as in [24]. Thereby our approach is parallelisable, requires little hyperparameter tuning, is easily scalable to *novel* classes, and allows us to swap out models at runtime (e.g. to save energy). Work by [24] has also shown that an ensemble of models provides high quality predictive uncertainty estimates and is more robust against dataset shifts. However, overall, the taken approach is not critically important (and can be replaced) as our main focus is on the introspection methods which are described in the next section.

Each *CNN* was trained for a single class, while all other classes are treated as negative labels. An image region is passed through successive convolutional layers of shape $(32 \times 5 \times 5)$ and $(64 \times 5 \times 5)$. Each convolutional layer is followed by a (2×2) max pool layer, a batch normalisation layer, a leaky ReLU layer and a 25% drop out layer. After the convolutional layers the output is flattened and passed through two dense layers of size 500 and 1 respectively. The networks are trained in a supervised fashion using a weighted cross-entropy function:

$$L_{CNN} = -\frac{1}{N}\sum_{i=0}^{N} wy_i \ln(\sigma(x_i)) + (1 - y_i)\ln(1 - \sigma(x_i)) \quad (1)$$

where N is the batch size, y and x are the set of corresponding labels and produced logits, and w is the weighting with emphasis on recall proportional and precision inversely proportional to this value respectively. The confidence values from each expert are intended to reflect a likelihood of the region corresponding to a given class. To consolidate the outputs from the different experts, we normalise their respective confidence values by dividing them by their overall sum. When introducing the *novel* class confidence value for a region, the value can be introduced simultaneously with the normalisation step, or afterwards before renormalising. We experimented with both approaches but for brevity will only evaluate results from the simultaneous introduction, due to it typically performing better in our experience.

While we classify all region proposals using this CNN-based approach, we explain in the next section how these results can be assessed, analysed, and improved.

IV. Competency-aware Object Detection

In this section we first present three complementary introspection methods to assess and analyse the performance of CNN-based classifiers. Finally, in Sec. IV-D, we explain how these methods can be optimally combined.

A. VAE-based Novelty Detection (VAE)

Our aim is to detect a dataset shift, i.e. *novel* objects that were not part of the training distribution. In this work, we have adopted the approach by [15] for novelty detection based on Variational Autoencoders (VAE). To this end, the *VAE* model was only trained on *typical* data from the target environment. The underlying idea being that the *VAE* 'knows' how to process *typical* data, but not *novel* data. At test time, the *VAE* model aims to reconstruct any given region proposal. Each proposal **X** is first encoded into the VAE's latent space representation **z**. Then, the VAE's decoder creates a reconstructed image $\hat{\mathbf{X}}$ from the latent space representation. The encoder consists of convolutional layers with shapes $(32 \times 7 \times 7)$, $(128 \times 5 \times 5)$, $(64 \times 3 \times 3)$, $(6 \times 3 \times 3)$ and identical layers inbetween each convolutional layer as the CNN classifiers. This is again followed by flattening and passing through two dense layers except this time both have size double the number of latent dimensions in order to produce the mean and variance for each dimension. Once the mean and variance is calculated for a given image region, a latent vector representation is sampled from a multivariate Gaussian distribution, this is then passed into the decoder, which is defined symmetrical to the encoder. We employ a latent representation consisting of 768 dimensions, based upon the findings of [15]. In order to train the VAE and to assess the novelty of a region proposal, we use a negative ELBO loss that combines reconstruction errors in the image space as well as losses obtained from the VAE's latent space:

$$L_{VAE} = -ELBO = -E_{q(\mathbf{z}|\mathbf{X})}(\ln p(\mathbf{X}|\mathbf{z})) + KL(q(\mathbf{z}|\mathbf{X})||p(\mathbf{z})) \quad (2)$$

$$p(\mathbf{X}|\mathbf{z}) \sim \sum_{x_{i,j,k}\in\mathbf{X}} \mathcal{N}(x_{i,j,k}, 0.04) \quad (3)$$

$$p(\mathbf{z}) \sim \mathcal{N}(0, 1) \quad (4)$$

The loss provides a confidence value how *novel* an object is. For further discussion of the VAE approach please refer to [15]. At prediction time, once we have derived a loss value for a region, we compare it against a histogram of loss values from training time and use the placement relative to these values to determine the novelty, i.e. a value of 1.0 for loss greater than any training time case, a value of 0.0 for loss less than any training time case.

B. Density-Adjusted Distance Measure (DDM)

While the *VAE* does take into account an image region's relationship with its latent space, the means by which it does so is rudimentary. The approach described here is *Density-Adjusted Distance Measure (DDM)* for the *VAE's* latent space.

Before describing our method, let us briefly state some preliminaries. A *metric space* is a set X together with a distance function $d : X \times X \longrightarrow \mathbb{R}$ such that $\forall x, y, z \in X$, $d(x, y) \geq 0$, $d(x, y) = d(y, x)$, $d(x, y) = 0$ if and only if $x = y$, and finally $d(x, z) \leq d(x, y) + d(y, z)$. The *OPTICS* algorithm [17], [18] is a density-based clustering algorithm which takes a collection of points from a metric space as input, and produces an ordering on the dataset from which a cluster assignment is recoverable.

An observation is made in [18] that *OPTICS* is closely related to Prim's algorithm for finding minimum spanning trees. *OPTICS* takes a point cloud from a metric space as input; then if we interpret this as a fully-connected weighted graph where the weights are pairwise distances, then *OPTICS* starts by *redefining* the weights on the edges by taking local density into account via its use of core distances. Then a greedy algorithm, reminiscent of Prim's algorithm, is executed on this re-weighted graph.

More specifically, given a collection of points $x_1, ..., x_n \in X$ where (X, d) is a metric space, define the *core distance* of a point $x \in X$ to be:

$$c(x) := \inf\{\epsilon > 0 : |B_\epsilon(x)| \geq N\} \tag{5}$$

$$B_\epsilon(x) := \{x_i : d(x, x_i) < \epsilon\} \tag{6}$$

where $N \in \mathbb{N}, 1 < N < n$ is a hyperparameter. Following this, *OPTICS* defines the re-weighting or *reachability distance* to be:

$$d_o(x_i, x_j) := \max\{d(x_i, x_j), c(x_i)\} \tag{7}$$

However the reader may observe that the two above definitions *are not distance functions* in general, hence (X, d_o) is not a metric space. As [18] notes: d_o is not symmetric, so is not a metric. This closes off many avenues of analysis, as metric spaces have been heavily studied. So instead, we pick an alternative re-weighting:

$$d_m(x_i, x_j) := \max\{d(x_i, x_j), \lambda\,|c(x_i) - c(x_j)|\} \tag{8}$$

for a hyperparameter $\lambda > 0$.

Lemma 1. *d_m is a distance function.*

Proof: Clearly d_m is symmetric in its arguments. Furthermore, $d \geq 0$ means $d_m \geq d \geq 0$. Positivity follows from the fact that $d_m(x_i, x_i) = \max\{d(x_i, x_i), \lambda\,|c(x_i) - c(x_i)|\} = \max\{0, 0\} = 0$. The triangle inequality follows from:

$$\begin{aligned} d_m(x_i, x_j) &\leq \max\{d(x_i, x_k) + d(x_k, x_j), \\ &\quad \lambda\,|c(x_i) - c(x_k)| + \lambda\,|c(x_k) - c(x_j)|\} \\ &\leq \max\{d(x_i, x_k), \lambda\,|c(x_i) - c(x_k)|\} + \\ &\quad \max\{d(x_k, x_j), \lambda\,|c(x_k) - c(x_j)|\} \\ &= d_m(x_i, x_k) + d_m(x_k, x_j) \end{aligned}$$

In addition to being a metric, d_m is an attractive choice for outlier analysis for qualitative reasons. In addition to being a metric, the density dependent nature of d_m means that clusters of uniform density are not changed much, whereas the boundaries of clusters are affected due to an abrupt change in density. For sufficiently small λ, the majority of the pairwise distances will be exactly equal to their distance with respect to d.

We now end the analogy with *OPTICS* and use this density-adjusted distance measure d_m to test for outliers. We define a point to be an *outlier* if its distance from its nearest neighbour distance is *significantly* higher than the rest of the dataset's nearest neighbour distances[3]. If we calculate this with respect to d_m rather than d, differences in density are captured naturally, as opposed to when directly applied to d.

For each point x_i, define the nearest-neighbour distance

$$m_j := \min_{k \in \{1..n\}, k \neq j} d_m(x_j, x_k) \tag{9}$$

and the average nearest-neighbour distance differences

$$T_i := \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} m_j - m_i \tag{10}$$

Then to query if x_i is an outlier, define

$$p_i = \frac{|\{j \neq i : T_j > T_i\}|}{n-1} \tag{11}$$

Then p_i is the novelty value for the point x_i, so $1 - p_i$ is how confident we are in the point *not* being an outlier [4].

C. Spearman's Rank Correlation (SRC)

Each class represented in the *CNN* stack will bear some likeness or unlikeness to the other classes in the *CNN* stack. This approach attempts to leverage this by deriving a novelty value based upon the deviation from class confidence correlations seen at training time.

To analyse the correlation we calculate the Spearman's rank correlation matrix M for the ensemble of *CNN* classifiers across the entire training dataset of regions. Then for a runtime prediction vector $v \in [0, 1]^n$ of the classifiers, we calculate the value $v^T M v$. Since M is positive semidefinite, the mapping $v \mapsto \sqrt{v^T M v}$ is a seminorm, therefore giving us a notion of length. We then translate this notion of length to give us an indication of novelty, according to the formula

$$v \mapsto \exp\left(-\lambda v^T M v\right) \tag{12}$$

for some hyperparameter $\lambda > 0$.

D. Combining Introspection Methods

Rather than taking a single approach to detecting novelty, we propose that utilising an ensemble of approaches could lead to more robust indicator of novelty. One could simply take the average of novelty scores, however a weighted sum might be more appropriate to avoid the assumption that the methods are equally proficient.

In line with the evaluation methods in [22], we set up a 'meta-game' between two players: one player chooses a novelty detection method, another chooses an image region. As each region is labelled *novel* or *typical*, this allows us to use the model's confidence value against the ground truth label as a measure of correctness. Hence the payoff matrix of this game is defined as the log-odds[5] that a given model will correctly classify a given image region. Then we compute the maximum entropy Nash equilibrium of this game, and used this distribution as the weighting of the models.

An attractive feature of the redundancy invariance means that if the given novelty models have comparable performances their weight will be spread evenly. This invariance is desirable for the image regions too: the drastic class imbalances often found in novelty detection problems cause no skew in the output, i.e. the variety of classes have equal consideration.

V. Experiments

A. Datasets

We conducted experiments on three datasets taken from two different domains: planetary and nuclear. The planetary domain relates to geological exploration tasks on remote planets (e.g. Mars), while the nuclear domain covers automated surveying of nuclear decommissioning projects. The ADE project contributes a dataset for each of these domains,

[3] While this is a rather rudimentary definition, its power comes not from the definition of an outlier, but from the measure of distance we choose.

[4] The calculations in the above method are similar to that of the *permutation test*; however that test assumes independence of samples, and the quantities we are averaging over in the definition of T_i are not independent.

[5] To prevent infinities appearing in the payoff matrix, we smoothed the model output probabilities before applying the log-odds function, by a factor of 0.01.

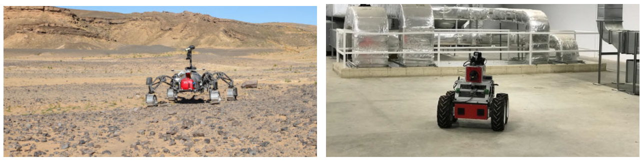


Fig. 3: Rovers used within ADE. Both rovers are equipped with a high-resolution camera as used for data collection in this work. *Left:* SherpaTT rover in a desert environment (Photo: Thomas Frank, DFKI GmbH). *Right:* Foxizirc rover in a test campaign at GMV facilities (Photo: GMV).

TABLE I: Dataset Class Distributions

Train+Validation		Test	
Class	Count	Class	Count
ADE Planetary (Image Regions / Images)			
Background	6667	Feature Dense	30
Lighting	311	Feature Sparse	20
Vegetation	1967	Novel	10
Float Rock	6781		
Rock Outcrop	1737		
ADE Nuclear (Image Regions / Images)			
Background	305813	Water Leaks	20
Water Leaks	226	Dry	10
		Novel	10
Mastcam (Image Regions / Images Regions)			
Background	9358	Background	1756
Broken Rock	60	Broken Rock	16
Drill Hole	50	Drill Hole	12
Dirt	89	Dirt	22
Dump Pile	75	Dump Pile	18
Bedrock (*Novel*)	5	Bedrock (*Novel*)	6
Float Rock (*Novel*)	6	Float Rock (*Novel*)	12
Meteorite (*Novel*)	11	Meteorite (*Novel*)	23
Veins (*Novel*)	10	Veins (*Novel*)	20

meanwhile we use the previously mentioned publicly available Mars novelty detection Mastcam labeled dataset. Fig. 3 shows pictures of the rovers used with the ADE project within these domains.

For all datasets, training sets were constructed from regions cropped out of images collected in their respective domains. For the ADE datasets the cropped regions were the result of labelling carried out by 10 individuals, several of which were experts in geology who offered their insight while constructing the planetary dataset. Dataset augmentation was also carried out by flipping the cropped image regions horizontally to effectively double the training sets. The mastcam dataset consists entirely of image regions cropped by experts as part of an earlier work [2], [16]. Test sets were also constructed, although the contents of these varied between datasets, namely in the case of the ADE datasets these consisted of whole images upon which region proposal would take place, meanwhile for the mastcam dataset, the images were already cropped and therefore the images were evaluated directly without a region proposal step. For a detailed breakdown of the distribution of image/image regions across classes and train/test sets, see Tab. I. In the case of the mastcam dataset because it consists purely of labelled cropped images, we opted to select the classes with the fewest available images to act as part of the novel class, these classes are indicated in the table.

B. Experiment Parameters & Platform

All neural networks were trained with the Adam optimiser with learning rates of 10^{-3} and 10^{-5} for the *CNN* and *VAE* respectively. The CNN and VAE were trained for 60 and 10 epochs respectively. The cross-entropy weight was configured to correspond to the difference in exponent for the positive and negative training counts.

The experiments were carried out on a desktop PC with a Ryzen 9 3950x CPU and 32 GB of 3600 MHz DDR4 RAM. OpenCV was configured to operate with a maximum of 8 threads, as was the TensorFlow library (2 inter op threads & 8 intra op threads). The KNN-CPP library was set to auto-detect the number of threads to use through the use of OpenMP. A time limit of 15 s was allocated to each image, however most images were processed in approximately 5 s.

C. Quantitative Evaluation

After our architecture has been run on the test sets of images/image regions we end up with a collection of image regions labelled by the architecture to either be one of the known classes or the novel class. In the case of the mastcam dataset we already have a ground truth label for each image region, however for the ADE datasets because a comprehensive labelling of all images would be infeasible in many cases (e.g. an image with thousands of rocks of varying scales in it) we instead label the proposed regions. Due to the fact the *uninteresting* labels (e.g. background, lighting, etc.) can often number in the hundreds per image, we cap the number of *uninteresting* regions to 50, as determined by the class label awarded to the region by the architecture.

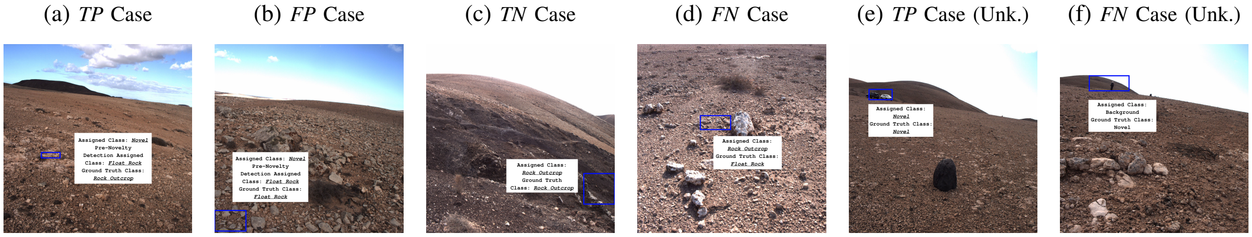
Once the architecture assigned and ground truth labelling is established for each image region, the results can be quantitatively evaluated. Each image region is considered to be a *True Positive (TP)*, *False Positive (FP)*, *True Negative (TN)* or *False Negative (FN)* for each class based upon the relationship between the labels for a given image region. The assignment of these four categories to each class for each image region follows standard conventions except for the novel class. If an image region is assigned to the novel class that would have been improperly classified without novelty analysis then this is considered a *TP* for the novel class, although it will still be registered as a *FN* for the ground truth class, this is to reflect the novelty analysis successfully preventing a false class label assignment, even though the ground truth class label could not be determined. Similarly, if an image region is improperly classified with novelty analysis in place, it would be registered as a *FN* for the novel class.

We evaluate a baseline *CNN* stack, the *CNN* stack combined which each introspection approach individually, and the *CNN* stack combined with the introspection approach ensemble. The results for each of these approaches are shown in Tab. II. The trend shown throughout the results is that our approach leads to an increase in precision at the expense of recall, this results in a marginal overall gain in performance, as measured by WACC. As our work surrounding the ADE project is concerned with opportunistic exploration, precision is desirable over recall, since pursuing misclassified features

TABLE II: Weighted accuracy (WACC) / Precision / Recall per model & class for each dataset

Class	*CNN*			*CNN+VAE*			*CNN+DDM*			*CNN+SRC*			*CNN+VAE+ DDM+SRC*		
ADE Planetary															
Mean (ex. *Novel*)	0.53	0.32	**0.33**	**0.54**	**0.37**	0.22	**0.54**	**0.37**	0.23	**0.54**	0.36	0.27	**0.54**	**0.37**	0.23
Background	0.76	0.78	**0.64**	**0.77**	**0.84**	0.51	**0.77**	0.83	0.55	0.76	0.81	0.57	**0.77**	0.83	0.53
Vegetation	0.61	0.25	**0.28**	**0.66**	**0.35**	0.24	**0.66**	**0.35**	0.24	0.65	0.33	0.24	**0.66**	**0.35**	0.24
Lighting	0.53	0.06	**0.05**	0.54	0.09	0.03	**0.55**	**0.11**	0.03	0.54	0.09	0.03	0.54	0.09	0.03
Float Rock	0.7	0.55	**0.69**	**0.71**	**0.65**	0.38	0.7	0.64	0.37	**0.71**	0.62	0.52	**0.71**	0.63	0.42
Rock Outcrop	0.56	0.28	**0.31**	0.57	0.3	0.17	0.57	0.3	0.18	**0.58**	**0.31**	0.23	0.57	0.3	0.18
Novel		–		0.62	0.57	**0.55**	0.62	0.59	0.52	**0.63**	**0.62**	0.38	0.62	0.58	0.5
ADE Nuclear															
Mean (ex. *Novel*)	0.53	0.52	**0.59**	0.53	0.52	0.55	0.53	0.52	0.56	0.53	0.52	0.56	0.53	0.52	0.56
Background	0.52	0.94	**0.22**	0.52	0.94	0.2	0.52	0.94	0.2	0.52	0.94	0.21	0.52	0.94	0.2
Water Leaks	**0.54**	0.09	**0.95**	0.53	0.09	0.91	0.53	0.09	0.91	0.53	0.09	0.91	0.53	0.09	0.91
Novel		–		**0.47**	**0.67**	**0.06**	**0.47**	0.66	0.05	0.46	0.65	0.04	**0.47**	0.66	0.05
Mastcam															
Mean (ex. *Novel*)	0.55	0.26	**0.59**	0.55	0.27	0.4	0.55	0.28	0.47	**0.56**	**0.29**	0.5	**0.56**	0.28	0.51
Background	**0.59**	0.99	**0.7**	0.57	**1.0**	0.58	0.57	**1.0**	0.55	0.57	**1.0**	0.58	0.58	**1.0**	0.62
Broken Rock	0.54	0.09	**0.81**	0.5	0.0	0.0	**0.55**	**0.11**	0.38	0.54	0.09	0.5	**0.55**	0.1	0.5
Drill Hole	0.56	0.13	0.25	0.61	0.23	0.25	0.59	0.19	0.25	**0.62**	**0.25**	0.25	0.6	0.21	0.25
Dirt	0.54	0.08	0.91	0.54	0.08	0.91	0.54	0.08	0.91	0.54	**0.09**	0.91	0.54	0.08	0.91
Dump Pile	0.51	0.02	0.28	0.51	**0.03**	0.28	0.51	**0.03**	0.28	0.51	**0.03**	0.28	0.51	**0.03**	0.28
Novel		–		0.58	0.45	**0.31**	0.53	0.37	0.26	0.55	0.4	0.25	**0.59**	**0.47**	0.22

Fig. 4: Qualitative Novelty Detection Results


(a) *TP* Case
(b) *FP* Case
(c) *TN* Case
(d) *FN* Case
(e) *TP* Case (Unk.)
(f) *FN* Case (Unk.)
Assigned Class: Novel
Pre-Novelty Detection Assigned Class: Float Rock
Ground Truth Class: Rock Outcrop
Assigned Class: Novel
Pre-Novelty Detection Assigned Class: Float Rock
Ground Truth Class: Float Rock
Assigned Class: Rock Outcrop
Ground Truth Class: Rock Outcrop
Assigned Class: Rock Outcrop
Ground Truth Class: Float Rock
Assigned Class: Novel
Ground Truth Class: Novel
Assigned Class: Background
Ground Truth Class: Novel

might be a waste of time/energy. However, the novelty detection results of the architecture are the most significant findings. If we consider the *CNN+VAE+DDM+SRC* model, the lowest precision seen is 0.47 for the Mastcam dataset, while the planetary and nuclear ADE datasets have precision 0.58 and 0.66 respectively. This makes sense given the Mastcam dataset was the smallest of the three used to train the architectures. However, what this means is that typically at least $\sim$ 50% of the image regions classified as novel indeed contain novel content, either as an unknown class or as new information for an existing class (i.e. a misclassification). This opens up the opportunity for the introspective methods described here to facilitate a self improving system. This could either be achieved by an additional manual labelling step for the novel cases, or involve running a more precise classifier with greater processing times in the background.

D. Qualitative Evaluation

In Fig. 4 qualitative results for several novelty detection cases. Cases (a) through (d) show instances of where novelty detection succeeded and failed in avoiding a misclassification, meanwhile cases (e) and (f) show a success and failure case for dealing with an unknown class i.e. A pair of cars and a human respectively.

The most significant factor demonstrated by these images is the inherent difficulty of determining what should be considered novel information. Consider cases (b) and (d), despite the visual resemblance of the labelled image regions being very similar one is successfully classified by the *CNN* stack, while the other is not. Despite this, the novelty detection only assigns the successfully classified case to the novel class. The difficulty is in the ability to anticipate the behaviour of *CNNs* combined with the preexisting ambiguity between classes (e.g. Float Rock vs Rock Outcrop). In order to address this difficulty further work will need to be conducted that takes into account no only the training dataset, but also the behaviour of *CNNs* trained from this data.

VI. Conclusions

We have presented an approach to competency-aware object detection for robot missions in open-ended environments. Through experiments on real-world data in two domains, we have demonstrated how introspection methods can assess the outcome of CNN-based classifiers without a loss in performance. By exploring introspection for object detection, our work makes steps towards autonomous robot systems which operate in changing, real-world environments.

Acknowledgment: This work has been supported by the EU Framework Programme H2020 (grant agreement No 821988, ADE), the AAI Programme (Lloyd's Register Foundation/York University), and the NVIDIA Corporation (donation of Titan Xp/V GPUs).

References

[1] J. Ocón, I. Dragomir, A. Coles, A. Green, L. Kunze, R. Marc, C. Perez, T. Germa, V. Bissonnette, G. Scalise, *et al.*, "Ade: Autonomous decision making in very long traverses," in *Proceedings of the 2020 International Symposium on Artificial Intelligence, Robotics and Automation in Space*. i-SAIRAS, 2020. [Online]. Available: https://www.hou.usra.edu/meetings/isairas2020fullpapers/pdf/5033.pdf

[2] H. R. Kerner, K. L. Wagstaff, B. D. Bue, D. F. Wellington, S. Jacob, P. Horton, J. F. Bell, C. Kwan, and H. B. Amor, "Comparison of novelty detection methods for multispectral images in rover-based planetary exploration missions," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1642–1675, 2020.

[3] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on computers*, vol. 100, no. 1, pp. 67–92, 1973.

[4] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[6] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[7] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[8] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*. Springer, 2014, pp. 391–405.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[10] Q. M. Rahman, N. Sünderhauf, and F. Dayoub, "Performance monitoring of object detection during deployment," 2020.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[12] A. Vasilev, V. Golkov, M. Meissner, I. Lipp, E. Sgarlata, V. Tomassini, D. K. Jones, and D. Cremers, "q-space novelty detection with variational autoencoders," *arXiv preprint arXiv:1806.02997*, 2018.

[13] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, vol. 2, no. 1, pp. 1–18, 2015.

[14] R. Yao, C. Liu, L. Zhang, and P. Peng, "Unsupervised anomaly detection using variational auto-encoder based feature extraction," in *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*. IEEE, 2019, pp. 1–7.

[15] L. Sintini and L. Kunze, "Unsupervised and semi-supervised novelty detection using variational autoencoders in opportunistic science missions," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[16] H. R. Kerner, D. F. Wellington, K. L. Wagstaff, J. F. Bell, C. Kwan, and H. B. Amor, "Novelty detection for multispectral images with application to planetary exploration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9484–9491.

[17] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 49–60. [Online]. Available: https://doi.org/10.1145/304182.304187

[18] E. Schubert and M. Gertz, "Improving the cluster structure extracted from optics plots," in *LWDA*, 2018.

[19] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172.

[20] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers." vol. 29, 06 2000, pp. 93–104.

[21] ——, "Optics-of: Identifying local outliers," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 1999, pp. 262–270.

[22] D. Balduzzi, K. Tuyls, J. Perolat, and T. Graepel, "Re-evaluating evaluation," 2018.

[23] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6402–6413.