Background Appearance Modelling with Applications to Visual Object Detection in an Open Pit Mine

Alex Bewley
School of Electrical Engineering and Computer Science
Queensland University of Technology,
Brisbane, Australia.
alex.bewley@hdr.qut.edu.au

Ben Upcroft *
ARC Centre of Excellence for Robotic Vision,
Queensland University of Technology,
Brisbane, Australia.
ben.upcroft@qut.edu.au

Abstract

This paper addresses the problem of detecting people and vehicles on a surface mine by presenting an architecture that combines the complementary strengths of deep convolutional networks (DCN) with cluster based analysis. We highlight that using a DCN in a naïve black box approach results in a significantly high rate of errors due to the lack of mining specific training data and the unique landscape in a mine site. In this work we propose a background model that exploits the abundance of background-only images to discover the natural clusters in visual appearance using features extracted from the DCN. Both a simple nearest cluster based background model and an extended model with cosine features are investigated for their ability to identify and suppress potential false positives made by the DCN. Furthermore, localisation of objects of interest is enabled through region proposals, which have been tuned to increase recall within the constraints of a computational budget. Finally, a soft fusion framework is presented to combine the estimates of both the DCN and background model to improve the accuracy of the detection. Our system is tested on over 11km of real mine site data in both day and night conditions where we were able to detect both light and heavy vehicles along with mining personnel. We show that the introduction of our background model improves the detection performance. In particular, soft fusion of the background model and the DCN output produces a relative improvement in the F1 score of 46% and 28% compared to a baseline pre-trained DCN and a DCN retrained with mining images respectively.

1 Introduction

While the mining industry pushes for greater autonomy, there still remains a need for human presence on many existing mine sites. This places significant importance on the safe interaction between human occupied and remotely operated or autonomous vehicles. In particular, for reliable collision avoidance it is necessary to maintain sufficient situational awareness. To improve the situational awareness in regards to collision avoidance, this work investigates computer vision based techniques for detecting other vehicles and personnel in the workspace of heavy vehicles such as haul trucks.

Traditionally, methods for detecting other vehicles and personnel from heavy mining equipment have relied on radio transponder based technologies. Despite transponder based sensors being mature and reliable for ideal conditions, in practise their reliability is circumvented by practical issues around their two way active

*http://www.roboticvision.org/

nature, portable power requirements, limited spatial resolution and human error. In contrast, computer vision techniques for object detection aim to recognise the presence of other objects by observing a video camera feed. This offers a unique alternative that operates passively and can utilise the available cameras on existing vehicles. Furthermore, it is expected that reliable collision avoidance systems would utilise multiple sensing technologies to provide a high integrity solution with redundancy.

Detecting and recognising objects has long been a topic of research within the computer vision community, which has advanced tremendously in recent years as measured by standard benchmarks (Deng et al., 2009; Lin et al., 2014). These major advancements can be largely attributed to both the availability of huge annotated datasets (Fei-Fei et al., 2007; Torralba et al., 2008; Deng et al., 2009; Lin et al., 2014) and developments in data driven models such as deep convolutional networks (DCN) (Krizhevsky et al., 2012; Sermanet et al., 2013). A DCN effectively models visual appearance through a huge set of parameters which are tuned by training on a large set of annotated images with relevant objects of interest. In this work we utilise the DCN of (Krizhevsky et al., 2012) which has shown astonishing performance on the ImageNet recognition benchmark (Deng et al., 2009) and repurpose it toward recognising personnel and vehicles from the background in an open pit mining environment. However, naïvely applying an off-the-shelf DCN to images collected in a mining environment results in a significant number of false positives due to the differences in appearance between the training set and the target domain. See Figure 1.

Adapting DCNs to different domains typically requires a large training set relevant to the target domain (Zeiler and Fergus, 2013; Yosinski et al., 2014). When the amount of training data is small, data driven approaches tend to over-fit the training samples and not generalise to unseen images. In this work we utilise a pre-trained DCN using millions of images from ImageNet and experiment with both a naïve remapping of ImageNet to mining classes and compare this to retraining the network with limited data.

With a vehicle mounted camera the focal length is fixed and the viewing angle is rigidly coupled to the vehicle's orientation. This distinguishes it from the ImageNet recognition problem where typical images collected were implicitly pointed at regions of interest and appropriately zoomed. Additionally, due to the wide field of view the majority of the images are background with zero to potentially multiple objects of interest visible in any given frame. To address these multi-scale and object localisation issues, we employ a similar strategy to (Girshick et al., 2014) by applying an initial step for finding likely object locations through a region proposal process before performing object recognition with the DCN.

Given that the majority of images collected on a mine site contain zero objects of interest, we can efficiently collect a huge amount of background data suitable for training a linear classifier. Using this newly trained classifier in conjunction with the DCN ensures robustness and drastically reduces spurious detections. This classifier is based on k-means clustering offering a convenient way to implicitly partition the background data into different categories. This approach accurately captures the characteristics of the background, enabling the discovery of novel non-background objects. In light of this, two approaches are presented for using this background model to either suppress false background detections or to correct the detection confidence.

Building off the preliminary work in (Bewley and Upcroft, 2015), this article presents an alternative fusion framework where hard decisions on suppressing background are replaced with a soft probabilistic model. This enables a soft fusion between the likelihood a sample was generated from the novel background model and the class prediction probability produced by the DCN classifier. Furthermore, various stages of the pipeline proposed in (Bewley and Upcroft, 2015) have been optimised to increase either the recall or precision in order to improve the overall detection performance. Additionally, a new method for discriminating between background and novel objects is proposed and incorporated to the fusion framework. Finally, a thorough examination of each component both individually and jointly is performed to highlight the major contributing factors and gain insight into how the vision based detector behaves in a mine site environment.

This paper is organised as follows: Section 2 provides a short review of related literature in the areas of mine-site sensing, pedestrian and general object detection, and information fusion. Section 3 describes the various steps in our approach including analysis for region proposals, adapting DCNs, formulation of the background

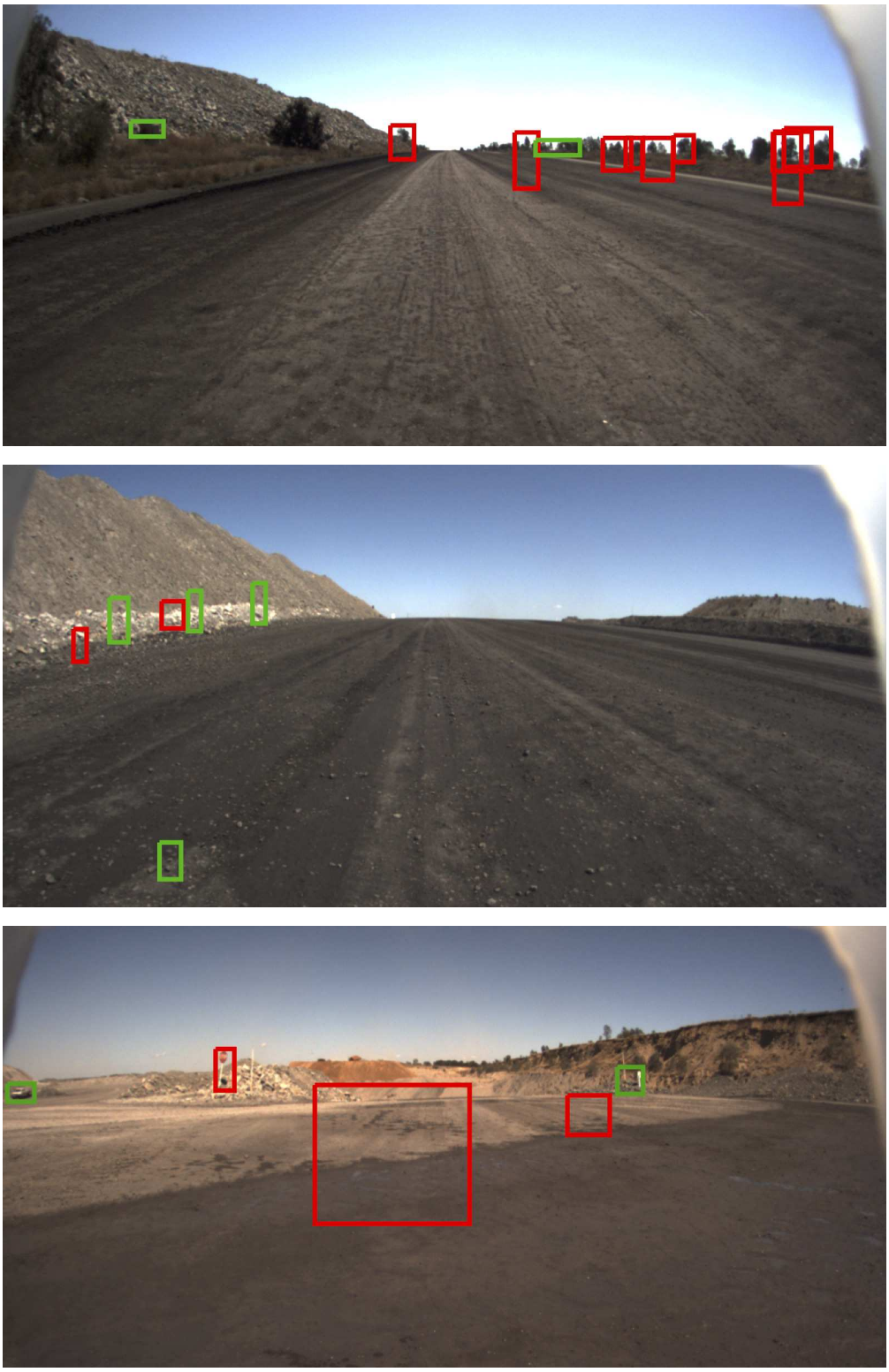

Figure 1: An example of the number of false positives when applying a DCN off-the-shelf in a mining environment. The colours represent different classes, e.g. person (*red*) and vehicle (*green*). The only true positive is the light vehicle on the left in the bottom image. Best viewed in digital form.

novelty detector, and information fusion. In Section 4 the detection performance of the proposed method is analysed on a challenging set of mining videos. Finally, Section 5 concludes the paper with a summary of the learnt outcomes and discussion for future improvement.

2 Related Work

Here we briefly review object detection methods that are not reliant on two way communication before covering some related work using DCN for generic object detection. Early work has focused on range based techniques such as LiDAR (Roberts and Corke, 2000; Marshall and Barfoot, 2008) commonly used for mapping fixed obstacles such as buildings or underground tunnel walls. Applying these sensors to detecting personnel and vehicles fitted with retro-reflectors is found to be sensitive to the dynamics of the sensor platform (Phillips et al., 2013). In this work we focus specifically on detecting potentially dynamic obstacles including vehicles and particularly people from vision based data. To this end, the more relevant prior work is that of (Mosberger and Andreasson, 2012) which exploits the standardised requirement for personnel on mine sites to wear high-visibility clothing equipped with retro-reflector patches. This enables a single IR camera with active flash to highlight personnel in view which can then be used for tracking (Mosberger et al., 2013). In this work, we do not restrict the detection system to specifically finding retro-reflectors, but rather formulate the problem as a recognition task focusing on the overall appearance of personnel and mining vehicles.

Pedestrian detection using computer vision techniques is an actively researched topic (Dalal and Triggs, 2005; Dollár et al., 2009; Benenson et al., 2013; Dollar et al., 2014; Zhang et al., 2015). These techniques take a sliding window approach, where for each image location, multiple rectangles with different scales and aspect ratios are examined for the presence of a pedestrian. This examination is generally characterised by extracting features specifically designed for pedestrians (Dalal and Triggs, 2005) before using a binary classifier (Dalal and Triggs, 2005; Dollár et al., 2009; Benenson et al., 2013) or cascade of classifiers (Dollar et al., 2014; Zhang et al., 2015) to determine if the rectangle represents a pedestrian.

Recent popularity of big data and deep learning have dominated the object recognition problem. Among these data driven approaches, deep convolutional networks (DCN) with recognition performance quickly approaching human levels (Krizhevsky et al., 2012; Donahue et al., 2013; Razavian et al., 2014; Sermanet et al., 2014) are selected for use in this work. DCNs themselves have been used for over 20 years (LeCun et al., 1989) for tasks such as character recognition. Over recent years DCNs have made an astonishing impact on the task of object recognition within the computer vision community (Krizhevsky et al., 2012; Farabet et al., 2013; Razavian et al., 2014; Girshick et al., 2014; Donahue et al., 2013) largely contributed to the availability of huge labelled image sets such as ImageNet (Deng et al., 2009). In this work, we use a network based on the work of (Krizhevsky et al., 2012) that was pre-trained on ImageNet which we re-purpose for the mine-site environment with limited label data and large sequences of unlabelled background data. While even deeper network architectures (Simonyan and Zisserman, 2014; Szegedy et al., 2015) have recently surpassed the (Krizhevsky et al., 2012) model for the ImageNet task, we continue to use the (Krizhevsky et al., 2012) model for efficiency and argue that a comparison of network architectures is beyond the scope of this work.

Recognising what objects are in an image is only half of the object detection problem. The other half is locating the objects within the image. Sermanet et al. (Sermanet et al., 2014) sample over multiple scales and exploit the inherently spatially dense nature of the convolutions within DCNs to identify regions with high responses. Similarly, (Farabet et al., 2013) also perform convolutions over multiple scales and combine the responses over superpixel segmentation (Felzenszwalb and Huttenlocher, 2004). Another popular approach and the one that we base this work off is the region convolutional neural network (RCNN) of (Girshick et al., 2014). The RCNN framework efficiently combines the DCN of (Krizhevsky et al., 2012) with an object proposal method: selective search (Uijlings et al., 2013). Generic object proposal methods aim to efficiently scan the entire image at different scales and aspect ratios to reduce potentially millions of search windows

down to hundreds (Hosang et al., 2014) of the most likely candidates. In this work we use `edge box` object proposals (Zitnick and Dollár, 2014) as the accuracy is higher and significantly faster according to a recent survey of object proposal methods (Hosang et al., 2014).

In this work we make use of information fusion for combining information from both the background model and the DCN output. In (Bewley and Upcroft, 2015) simple logic was used for combining the output of both systems, such that if the DCN responded with a `car` *or* `person` *and* the sample was *not* `background` then it would indicate a detection. This approach creates a hard decision where a failure in either system results in a miss detection. In contrast, probablistic approaches allow for a softer fusion of information (Dempster, 2008). In this work, we utilise probablistic variants of the DCN and background model in order to formulate the fusion of information in a classical Bayesian inference framework (Durrant-Whyte and Henderson, 2008).

3 Methodology

In this section, we describe our approach to vision based object detection where we highlight similarities and differences to the inspiring RCNN pipeline (Girshick et al., 2014). Our method consists of three key phases: 1) Region proposals with non-maximum suppression (NMS), 2) DCN recognition and finally, 3) Detections are validated by checking for novelty against the background model. See Figure 2 for a high-level overview of this pipeline. We bypass the problem of over-fitting on a small dataset by using a pre-training DCN and map its output to mining relevant classes.

The detection method is then extended with the incorporation of background modelling techniques to improve and adapt to the mining environment. From this background model we investigate two different novelty measures approximating the probability a detection was not generated by a background sample. This novelty measure produces a single value on the interval $[0, 1]$ where low values represent a high similarity with the background model and the high values represent that the sample is novel with respect to the background model. Finally, we describe how information from both the DCN prediction and the background model are combined using a probabilistic fusion framework.

3.1 Region Proposals

The aim of region proposals is to efficiently scan the image to eliminate millions of potential windows, keeping only the regions that are likely to contain an object of interest. We use the `EdgeBoxes` region proposal method (Zitnick and Dollár, 2014) over the `selective search` (Uijlings et al., 2013) used in the original RCNN work as this method is orders of magnitude faster with comparable accuracy. For a detailed comparison of region proposal methods we refer the reader to (Hosang et al., 2014).

Region proposal methods are regarded as a generic object proposal technique that is independent of object class (Hosang et al., 2014). However, in practice these methods are sensitive to parameter tuning, requiring data relevant to the target domain for optimal performance (McMahon et al., 2015). The default parameters for `EdgeBoxes` were adjusted to return a fixed 1000 proposals and an additional step of non-maximum suppression (NMS) is applied and described next.

3.2 Non-maximum Suppression

The region proposals provided by the `EdgeBoxes` method are further reduced through a process of NMS. The NMS process considers the score produced by the `EdgeBoxes` (EB) method and the overlap with other bounding boxes. As the name suggests it then greedily suppresses all but the maximum scoring proposal for all adjacent overlapping regions where the intersection-over-union (IOU) area is greater than a set threshold. In contrast to applying NMS after the DCN (Girshick et al., 2014), this way we can speed up the detection

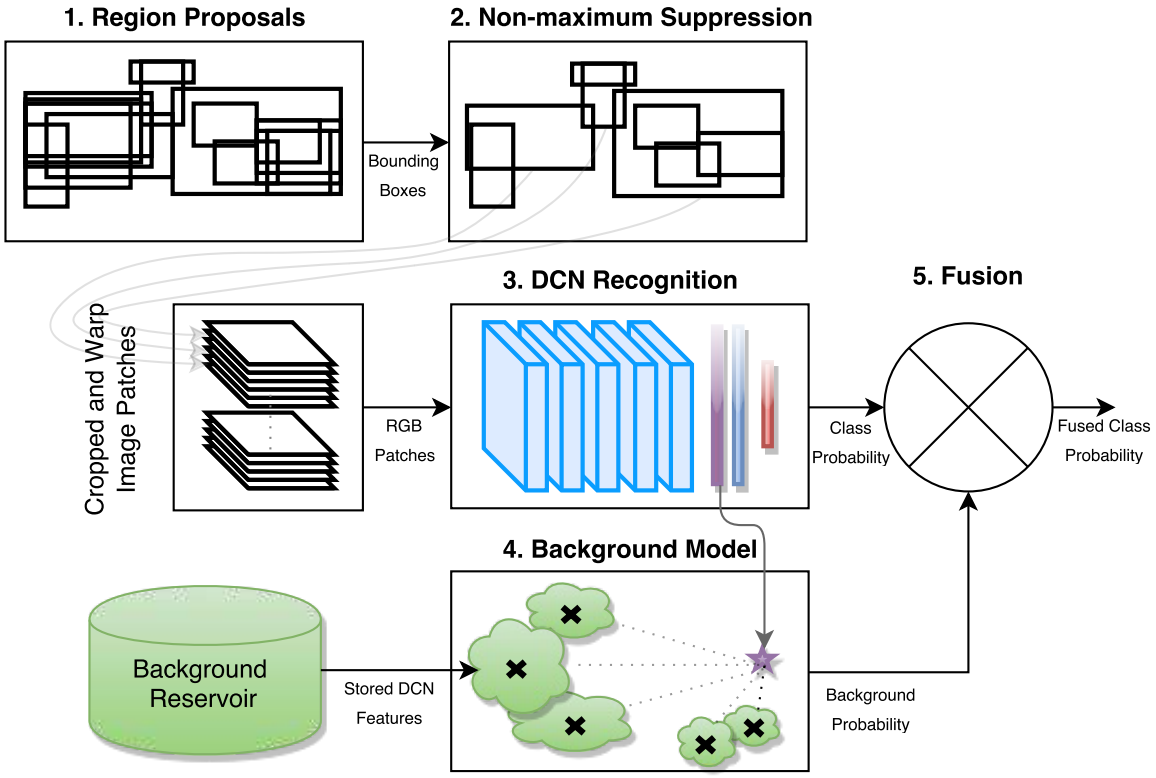

Figure 2: An illustration of the detection pipeline used in this work. When a new image is provided, the detection pipeline begins with the region proposal step. The system parameters are highlighted in blue and green which are learnt offline from an off-the-shelf network and background only images respectively. Purple represents the responses from an intermediate layer (also depicted as the purple star), that is compared with the cluster centers of the background model indicated with black crosses. The red output layer of the DCN contains scores for each class type. The output from the DCN and background model are fused to produce the systems final output.

pipeline by reducing the number of proposals going into the DCN while maintaining comparable coverage over the image.

To better understand the effect of applying NMS to the EB scores, the trade-off between the number of proposals and its effect on the recall is further investigated. Ideally, it is preferable to have as few proposals needed to cover all visible objects in order to minimise both the computational load of the later DCN and the overall rate of false positives. Since object proposals is used as the first step in the pipeline, it is paramount to have a good coverage of the true objects in the image since missed objects will never be recovered (Hosang et al., 2014). Figure 3 shows a comparison between applying NMS to the region proposals and reducing the number of proposals using the EB score. The EB+NMS curve shows that applying small amounts of NMS rapidly reduces the number of proposals while maintaining high recall. However, when suppressing proposals with overlaps lower than 0.3 IOU, results in lower performance compared to simply thresholding the EB score. The S shape of the EB+NMS curve shows that there exist a non-linear relationship between the localisation accuracy around true objects and the number of proposals. The overlap threshold in this work is set at 0.5 to produce approximately 260-300 proposals per frame where there is a considerable reduction in proposals (possibilities for false positives) for a moderate drop in recall.

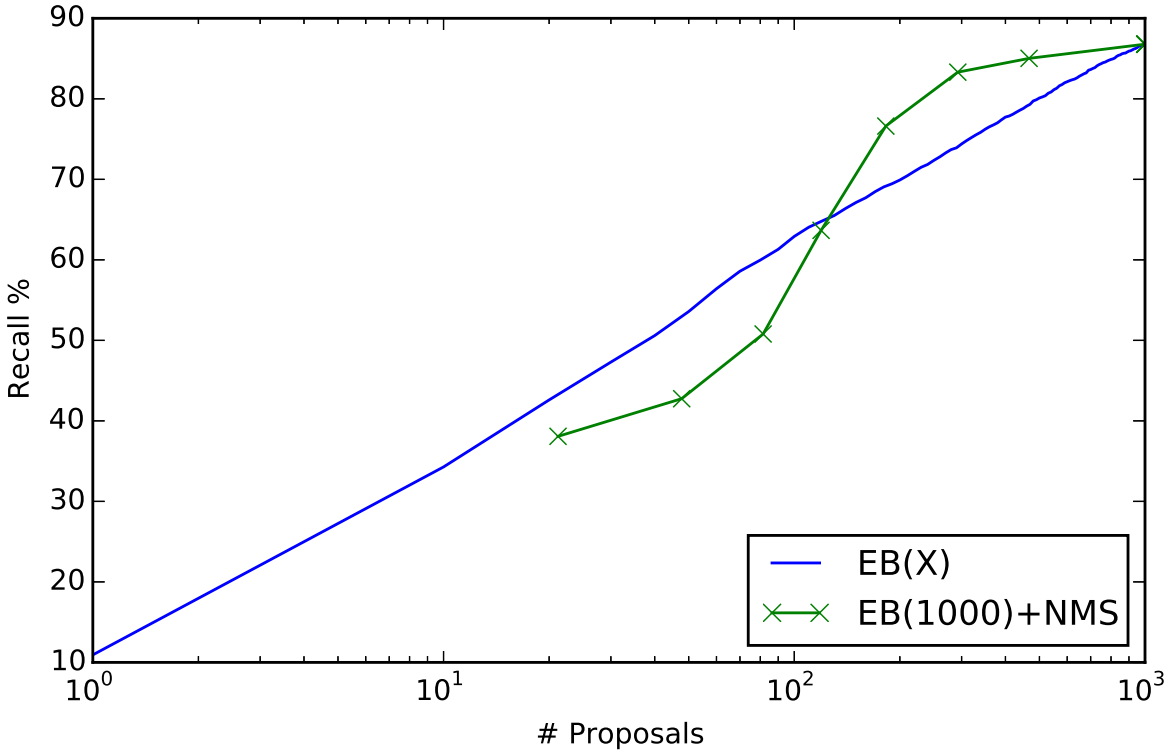

Figure 3: Performance of EdgeBoxes (EB) region proposal method along with non-maximum suppression applied with overlap thresholds set to $[0, 0.1, 0.2, ..., 0.7]$ shown with crosses. Note that the EB+NMS curve converges with the EB curve at NMS threshold of 0.7 since that is the limit used in the EB default parameters.

3.3 Region Classification

Having selected regions of the image that have the general characteristics of an object, we now perform object recognition to distinguish the object category. For this we apply the DCN from RCNN (Girshick et al., 2014) which is based on the winning architecture (Krizhevsky et al., 2012) for the ImageNet Large Scale Recognition challenge in 2012. For this work, we used the RCNN implementation provided with the Convolutional Architecture for Fast Feature Embedding (`caffe`) (Jia et al., 2014) framework out-of-the-box and apply it to classifiying the content of the object proposals generated as previously described.

Since the work in this paper largely gravitates around the recent success of DCNs, we briefly describe their workings. A DCN consists of multiple layers, each performing a transformation of their input data in the form of a linear projection followed by a differentiable, non-linear activation function to produce an output response. A set of weights and biases govern the linear projection step in each layer which essentially performs an inner product with the input and the weight parameters. The output responses for each layer forms the input for the following layer in a feed-forward fashion, where the first input is the data and the last represents a score for each class.

The architecture of (Krizhevsky et al., 2012) used in RCNN consists of eight layers in total with five being convolutional (conv1-conv5) and three fully-connected (fc6-fc8). In the convolutional layers, the weights only connect to a small receptive field in the input – acting like a 2D filter (or kernel) – that is shifted across the lateral dimensions of the input to perform a 2D convolution. The fully-connected layers on the other hand, remove the spatial structure of the data by flattening the transformed input before applying an inner product with their weights to project the entire input into a new high dimensional space. The intuition of

this DCN architecture is that the convolutional layers transform the input into low level visual descriptors representing local object parts, while the fully-connected layers are responsible for the high-level task putting the parts together to classify the entire image (Azizpour et al., 2015).

The RCNN architecture consists of around 60 million parameters across all eight layers which were optimised for classification via *deep learning*. Deep learning is the process of first applying a each layer transformation in turn to the input data to produce the network output. The classification error of the network (or loss) is then used to adjust the weights by an amount proportional to the error with respect to the layer input. As each layer is differentiable, the chain-rule enables the error to be propagated back to update the parameters in all layers. Given the high number of parameters, this network was trained using the large ImageNet dataset consisting of 1.3 million labelled images. Training a model on this scale is enabled through the use of Stochastic Gradient Descent (SGD).

The original detection task for the `caffe` RCNN model was to predict one of 200 classes that represent common objects found in images taken from the internet. For this application we are only interested in distinguishing between four high level categories, namely: `background`, `person`, `light vehicles(LV)` and `heavy vehicles (HV)`. Using a DCN model trained on ImageNet in a mining context raises several issues that need addressing:

1. The majority of the 200 ImageNet classes are indoor/domestic related objects including Food (apple,burrito, etc.), Musical Instruments (accordion, saxophone, etc.) or various animals (bird, camel, etc.)
2. How to associate mining classes with ImageNet classes?
3. Semantically, the `background` is significantly different from many of the existing object specific classes.

To gain some insight, we use a small validation set of 200 mining related images to investigate the output of the DCN out-of-the-box. This set is made up of cropped mine-site images containing the classes `person`, `LV` and `HV` along with 90 interesting region proposals extracted from background only images taken on mine sites. In Figure 4 we show the results of naïvely applying the pre-trained RCNN model to this image set. To better visualise the output we applied a soft-max transform to approximate the output class prediction as a probabilistic estimate. [1]

Not surprisingly, the `person` and `LV` classes are well represented and can be directly mapped from the `person` and `car` ImageNet classes used to train the original DCN. On the other hand, the `background` closely resembles uniform random sampling of classes as there are no relevant classes in the existing model such as trees, buildings, or road signs etc. Similarly, the `HV` class prediction also mostly resembles a uniformly random distribution with a slight bias towards the ImageNet classes `snowplow`, `cart` and `bus`.

3.3.1 Remapped Model

While the `person` or `car` class outputs can be simply mapped to `person` and `LV`, distinguishing between `background` and `HV` from the output is without avail. From a practical perspective, detecting `person` and `LV` is of higher importance for a passive computer vision system since `HV` are equipped with active protection devices. With this in mind, we simply assign all 198 non `person` or `car` outputs as background and accept that not detecting `HV` is a limitation of this remapping approach.

With this simple class mapping approach and assuming that falsely picking one of the positive classes is in fact uniformly random, we expect to eliminate 99% of all the proposed background regions. However, when

[1] It is important to note that this is for visualisation purposes only and that the y-axis does not represent the true probability since the finalsupport vector machine (SVM) layer of RCNN was not calibrated for probabilistic outputs.

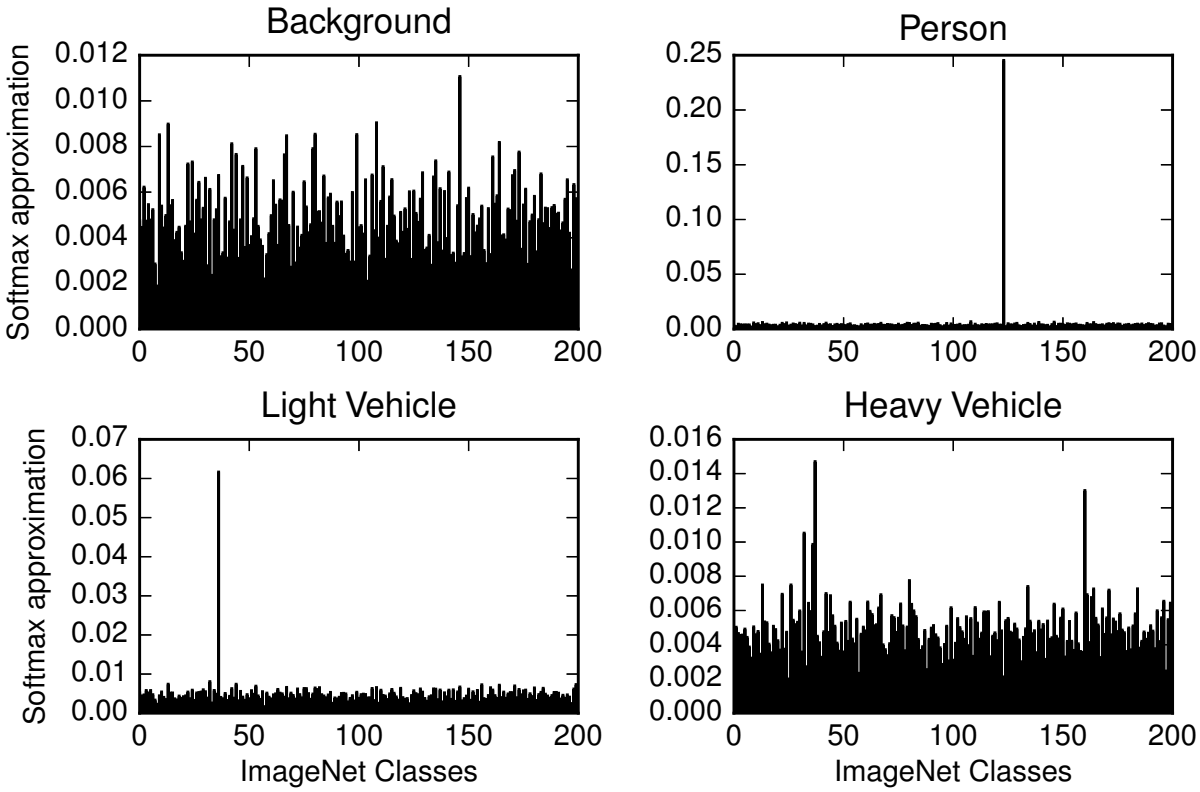

Figure 4: The average class estimate for a set of mining related images. Notice that person (class 123) and light vehicle/car (class 36) are existing classes for the pre-trained network and can be used directly. The background and the heavy vehicle classes are novel and show a wider spread as they are not modelled with the pre-trained DCN.

processing around 100 proposals per frame, the expected false positive rate is once per frame. Later, we propose a simple background model that reuses the DCN computation to provide a background likelihood estimate for reducing this false positive rate.

3.3.2 Retrained Model

An alternative is to train the DCN specifically on mining images for the target classes, including the `HV` class. This does raise the issue of learning the DCN parameters from the limited mining specific training data available. When using a small training set, the DCN can settle into a state that fails to recognise the wide variety of patterns that occur in the real world. This is largely caused by the massive number of connections inside a network that tend to converge on the few unique patterns in the training set. To overcome this effect of over-fitting the training data, many researchers (Ahmed et al., 2008; Aytar and Zisserman, 2011; Oquab et al., 2014; Yosinski et al., 2014; Azizpour et al., 2015) choose to take a network pre-trained on a massive dataset such as ImageNet and retrain only the last few layers for the task of predicting a set of domain specific classes. The intuition behind this is that the visual knowledge gained from the larger dataset is transferred to the target task to maintain diversity in low level feature extraction while the higher level classification task is allowed to adapt to the new domain.

3.3.3 Remap vs Retrain

In this work, we took the off-the-shelf DCN with 200 ImageNet outputs and remapped the mining classes referred to as the *remapped* DCN. Additionally, we train another model by replacing the last layer entirely and retrain the network keeping all but the last two layers fixed to prevent over-fitting. This network was trained until the loss reached zero on the set of 200 labelled mining specific examples. This model is referred to as *retrained*. Since the number of labelled examples used for training is small, Figure 5 shows the *retrained* model has negligible performance improvement over the *remapped* version which does not require any training. However, retraining the network with a softmax loss, effectively trains the model to accurately produce a probabilistic distribution over the target classes in contrast to using an SVM loss for the last layer as in (Girshick et al., 2014). As we will later show, this is important when fusing the DCN output with the proposed background model. Additionally, by retraining the DCN with outputs responsible for predicting mining related classes, it can be trained to also detect `HV` with negligible computational overhead.

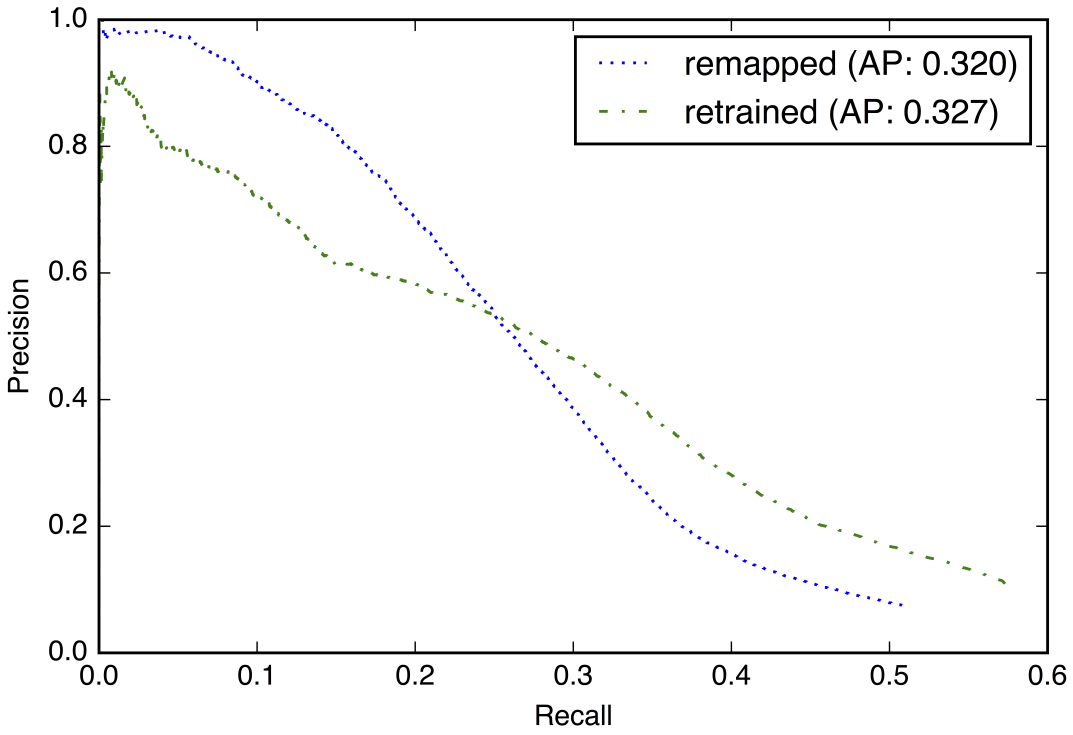

Figure 5: A comparison between simply remapping the ImageNet classes to Mining classes based on the activations shown in 4, versus retraining the network with mining images for classifying the target mining related categories directly. The average precision (AP) is provided in the legend.

3.4 Background Modelling

While the landscape on a mine-site constantly changes over time in a geometric perspective, the bleak visual appearance of the background remains predominantly consistent. The aim of this model is to capture the visual appearance of common background regions which are poorly characterised by the DCN used for classifying object proposals. At test time a proposal's similarity to this model provides an additional measure of confidence to be fused with the prediction output of the DCN model.

For a given background region, it generally belongs to one of an arbitrary set of categories, such as the

semantic categories of rock, sky, tree etc. Rather than using supervised techniques that require a set of manually annotated images, we instead partition the background data without explicit semantic labels. To do this, we exploit the assumption that intra-category samples generally appear visually similar to each other, yet may be distinctively different to other background categories. Put another way, the background regions form natural clusters enabling us to employ unsupervised techniques to model their visual appearance. See Figure 6 for an illustration of the natural background clusters found by applying a clustering approach to a mining dataset.

To describe the visual appearance of each region, the intermediate layers of the DCN provide a free and compact representation suitable for this task. Additionally, these features have been shown to be robust against lighting and viewpoint changes without any re-training (Sünderhauf et al., 2015). We refer the interested reader to (Krizhevsky et al., 2012) for an illustration of the DCN's inner workings. In general, the first layer of a DCN extracts simple colour and texture features in the first layer, and through subsequent layers, these features eventually transition to the learnt specific task (Yosinski et al., 2014) such as classifying the 200 ImageNet classes. Along the way irrelevant visual information for the original task (e.g. features describing sky) are lost once it reaches the final layer. With this intuition, we reuse the transformed data from one of the DCN's intermediate layers as an appearance descriptor for input to our background model.

To learn this cluster based model, a reservoir of negative samples is required. Gathering background data is a relatively simple task since only inspection for the presence of target objects is necessary. Specifically, any image sequence not containing any of the target objects can be used to build an extremely large reservoir by extracting proposals from each frame.

Ideally this background reservoir should be large enough to contain sufficient diversity to capture the variation in visual appearances, while also small enough for efficient nearest neighbour querying. To achieve this, a process called "hard"-negative-mining (Felzenszwalb et al., 2010) is utilised to focus only on the most informative image regions. Specifically, we run the detection pipeline with the pre-trained RCNN model of (Girshick et al., 2014) over these background only sequences and keep only regions that falsely classified as either a `person` or `car`. A sufficiently large reservoir can be built by also including near false positive regions that fall within the SVM margin of the pre-trained RCNN model. In doing this we are left with a collection of background region patches that are challenging in the sense that the DCN is unable to confidently classify them as background.

3.4.1 Nearest K-means Cluster Model

The first background model simply consists of a non-parametric model where the Euclidean distance to the nearest background example is used as a measure of novelty. Since the background reservoir is large, potentially redundant, and computing Euclidean distances for high dimensional data is slow, it is desirable to represent the background samples as a few exemplar points. With the intuition that the background forms natural clusters we choose to represent the background appearance as a set of clusters. Using the response output from an intermediate layer of the DCN as a visual feature $\mathbf{f}$ for a region proposal, the k-means algorithm is applied to group the reservoir points into clusters. This facilitates the selectable size of the background model M where the background sample points are the cluster centroids computed as follows:

$$\mathbf{m}_i = \frac{1}{|k_i|} \sum_j \mathbf{f}_j, j \in k_i, \tag{1}$$

where $\mathbf{f}_j$ is the visual feature of sample j which is a member of the cluster k_i.

At test time, each `person` or `LV` predicted patch is verified by measuring the Euclidean distance between its intermediate feature $\mathbf{f}_t$ and the nearest cluster mean

$$d_t = min_i(\|\mathbf{f}_t - \mathbf{m}_i\|). \tag{2}$$

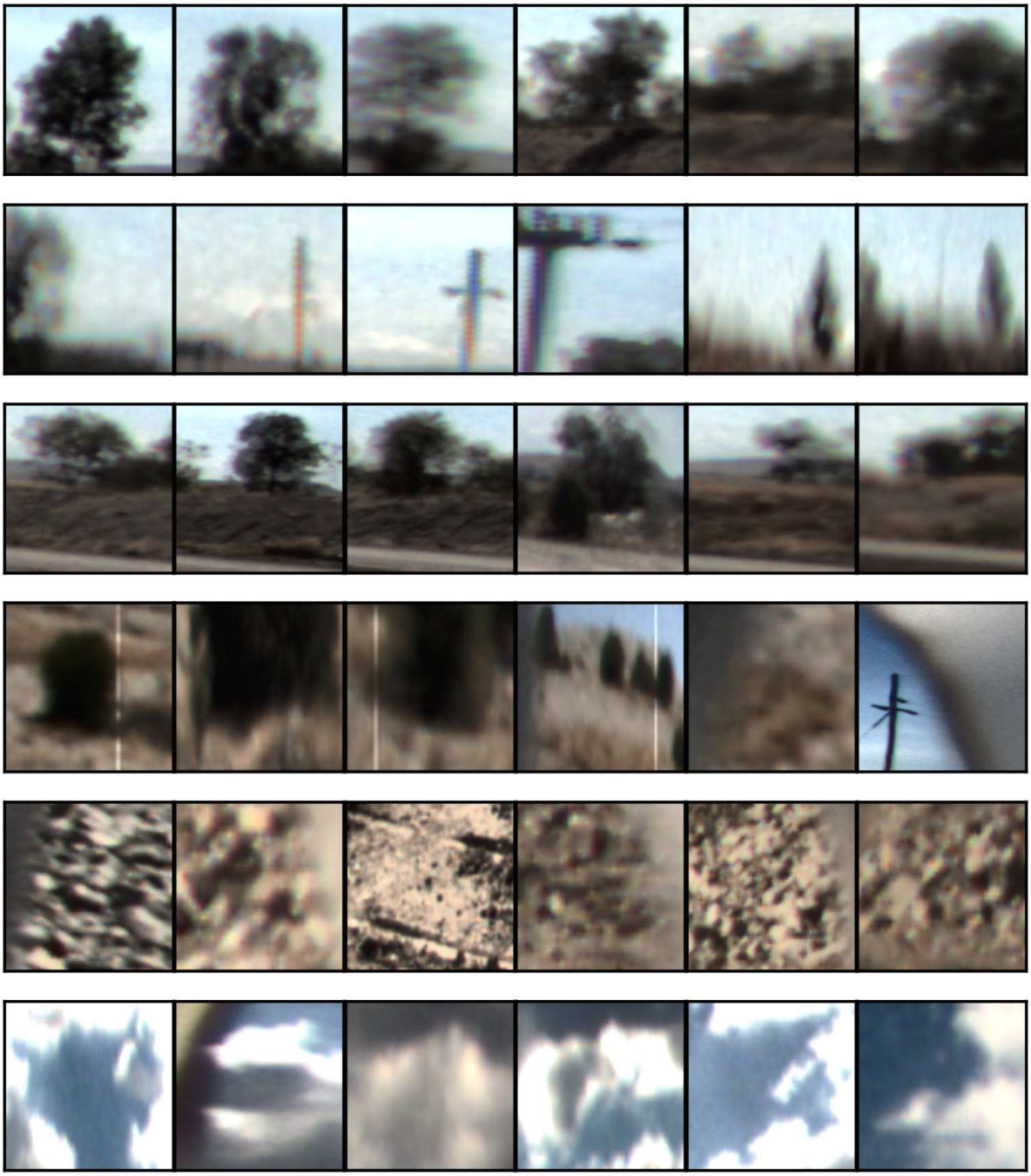

Figure 6: An illustration showing six of the most common types of background region proposals out of a total of 128 clusters. The rows represent different clusters while the columns show a random background region which is a member of the associated cluster. Each cluster gathers samples with similar visual appearance such as centred on a tree (top row) or centred on sky with an adjacent vertical structure (second row). Qualitatively, k-means clustering naturally formed mostly pure clusters with the exception of the forth row which also covers a number of patches occurring with relatively low frequency.

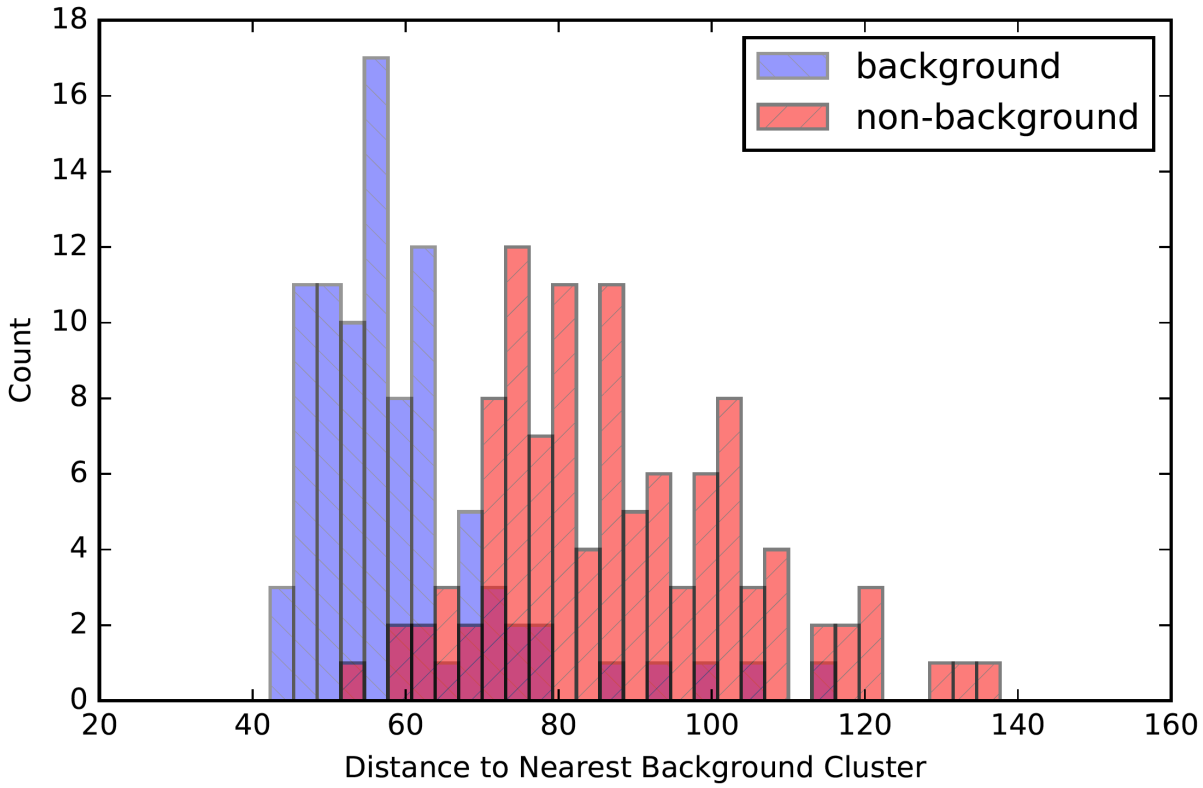

Figure 7: Detailed view of the distance to nearest cluster distribution for validation images composed of background and non-background images.

If the nearest background cluster is close in this feature space, i.e. is visually similar, then the likelihood that the test patch is background should increase. Conversely, it is expected that a patch corresponding to an object of interest should be dissimilar to the background model, which is signified by a greater distance to the nearest cluster centre. This expectation is corroborated with empirical evaluation on a small set of held out background patches and 110 images of non-background objects as shown in Figure 7.

While a hard threshold on the distance between a test patch and the nearest cluster centre can be used to suppress background detections (Bewley and Upcroft, 2015), a probabilistic version is favoured to enable soft fusion with other information such as the DCN output. To achieve this, the large collection of background samples is modelled using a parametric distribution. Figure 8 shows three different parametric distributions overlaid on the empirical histogram of distances between the background patches and their corresponding cluster centre. The *gamma* distribution best fits this data and is therefore used to approximate the likelihood of distances conditioned on a background sample being supplied. The cumulative density function is then used as a surrogate to the probability a test sample was not generated by the background as it has the property of monotonically increasing with the minimum distance to any background cluster. This soft probabilistic estimate of the test feature $\mathbf{f}_t$ being a non-background sample is computed using the Euclidean distance d_t to the nearest cluster as:

$$P(y_{BGM}|c \neq BG) = \frac{\gamma(\alpha, \beta d_t)}{\Gamma(\alpha)}, \tag{3}$$

where γ and Γ are the lower and upper incomplete gamma functions following the standard cumulative distribution function of a *gamma* distribution. The parameters α and β represent the shape and rate

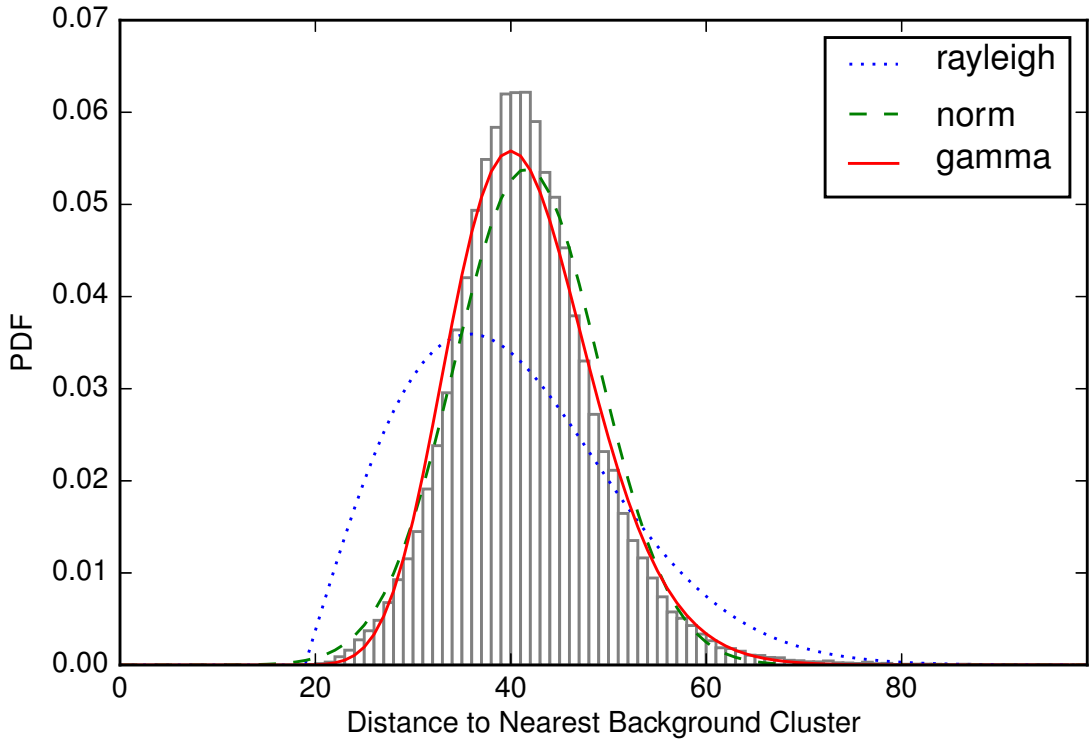

Figure 8: Distribution of background samples as a distance from nearest cluster centre.

parameters of the gamma distribution. These parameters can be estimated by fitting to the reservoir data as shown in Figure 8.

3.4.2 Joint Cosine Similarity Model

This alternate approach considers the similarity of all background clusters jointly rather than only the most similar cluster. Here, the similarity from a test sample to all cluster centres from the k-means model is used to discriminate between background and non-background samples. This enables a non-uniform weighting for each of the clusters effectively transforming the extracted DCN feature into a new feature space. This weighting is learnt by formulating the background model as a binary classification where the goal is to discriminate between background and non-background samples.

Here we reconsider the use of the Euclidean distance for measuring novelty from the background clusters. Particularly, given the high dimensionality of the DCN features, the Euclidean distance is easily compromised by the *curse of dimensionality* (Aggarwal et al., 2001). This would also explain the peculiar property visible in Figure 8 where the distance from the background samples to their nearest cluster centroid is distributed far from zero, with low variance. To investigate this, an alternative similarity measure is introduced based on the *cosine* distance to the cluster centres, which is considered to be more robust in high dimensions. In this new background model a weighted sum of all similarity measures between the test sample and the clusters is used to form a new feature representation.

Algorithm 1, details how the transformation of DCN features into a joint similarity representation along with learning the weight and bias parameter for a logistic regression. The $k \times n$ matrix S is used to store and represent a set of labelled features F in their joint similarity form. This transformation essentially

Algorithm 1 Logistic Cosine Background Model Learning

Input: M ▷ Set of all cluster means (Eqn. 1)
Input: F ▷ DCN features of example labelled regions.
Input: $\mathbf{l}$ ▷ Label vector corresponding to F.
Output: θ, b ▷ Weights and bias for logistic regression model.

1: **function** $(M, K, \mathbf{l})$
2: $S = \mathbf{0}_{k \times n}$ ▷ Initialise similarity matrix
3: **for** $\mathbf{f}_j \in F$ **do**
4: **for** $\mathbf{m}_i \in M$ **do**
5: $s_{i,j} = cosine\_similarity(\mathbf{f}_j, \mathbf{m}_i)$
6: $(\theta, b) = cross\_validation\big(train\_LR(), 10, (S, \mathbf{l})\big)$ ▷ Perform 10 fold cross validated learning
return θ, b

compares all n samples in F with each cluster mean vector using the standard cosine similarity measure (lines 3-5). Finally, a logistic regression (LR) classifier is learnt on this transformed data using the labels $\mathbf{l}$ provided (line 6). The weight and bias parameters; θ and b respectively, are optimised using SGD with $k$fold cross validation to select an appropriate amount of regularisation. At test time, the probability that a DCN feature $\mathbf{f}_t$ represents a non-background patch is computed by first transforming into the similarity form $\mathbf{s}_t$, and then fed into the LR classifier as follows:

$$P(y_{BGM}|c \neq BG) = \frac{1}{1 + exp^{-(\theta^T \mathbf{s}_t + b)}}, \tag{4}$$

where $\mathbf{s}_t$ is the joint similarity representation vector of the test sample computed as in lines 4 and 5 of Algorithm 1.

This LR classifier is trained using a held out validation set containing positive and negative patches, where the learnt separation is shown in Figure 9. In contrast to the k-means model where each cluster is treated with equal variance, this joint similarity based model captures the intuition that some clusters may have higher importance when discriminating background for non-background. Additionally, the LR classifier produces a probabilistic estimate enabling it to naturally fit within a Bayesian framework for fusing the background model estimate with the DCN class prediction.

3.5 Fusing Measurements

So far we have described how we can adapt a DCN for making predictions of the mining related class corresponding to each detected proposal. Additionally, a method for extracting features from the DCN, coupled with a clustering approach forms a good representation for identifying background samples. Here, we derive how these two measures can be combined with the goal of producing better estimates than either method individually.

Given the output predictions from both the DCN y_{DCN} and the background model y_{BGM}, we can treat these two predictions as two separate decision makers. While y_{DCN} and y_{BGM} are not independent i.e. $P(y_{DCN}, y_{BGM}) \neq P(y_{DCN})P(y_{BGM})$, they are conditionally independent on the object class c expressed as follows:

$$P(y_{DCN}, y_{BGM}|c) = P(y_{DCN}|c)P(y_{BGM}|c). \tag{5}$$

This independence enables us to fuse the DCN output, which has good target discrimination capabilities, while the background model is good at suppressing non-target proposals.

Using Bayes' Rule the class probability conditioned on the joint output of both the background model and

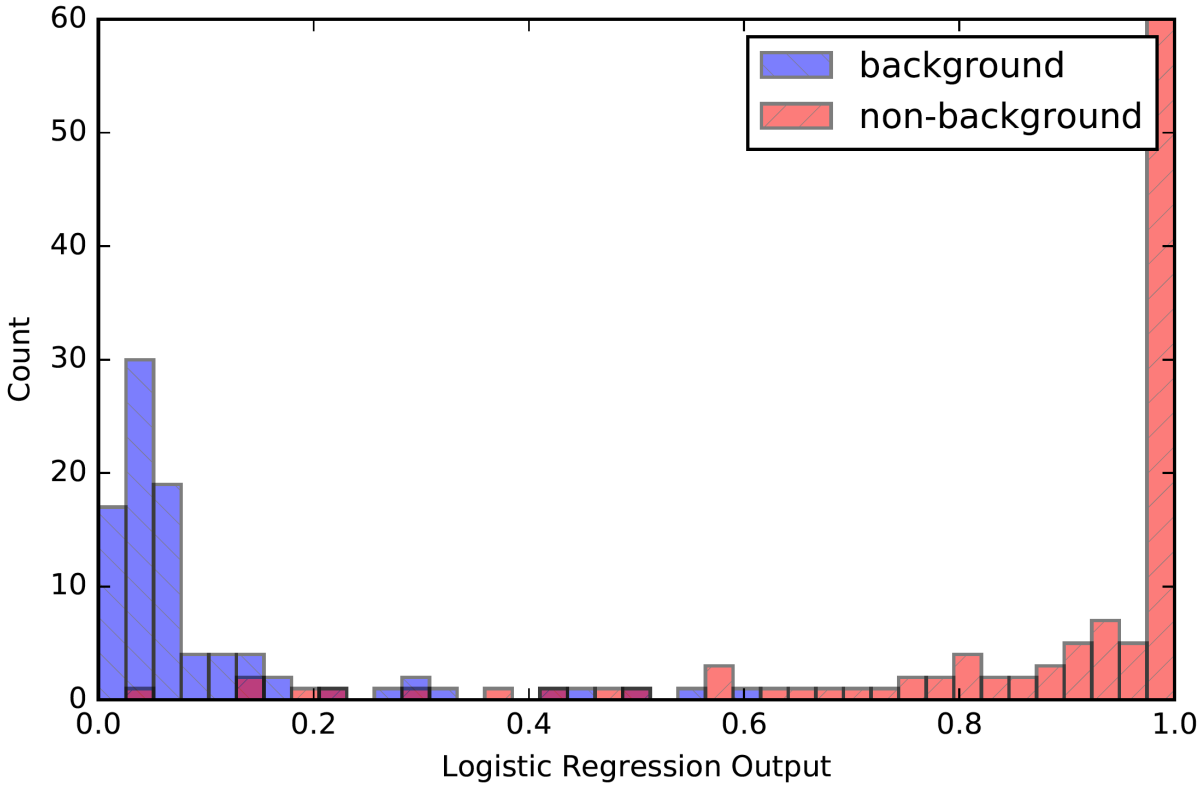

Figure 9: An illustration of the separation learnt using a logistic regression model on a small set of 90 background and 110 non-background samples.

the DCN is expressed as:

$$P(c|y_{DCN}, y_{BGM}) = \frac{P(y_{DCN}, y_{BGM}|c)P(c)}{P(y_{DCN}, y_{BGM})}, \tag{6}$$

where $P(c)$ is the class prior and $P(y_{DCN}, y_{BGM})$ is the joint probability of the observation. When selecting the most likely class, the joint probability of observations can be ignored since it is common for all classes c_i.

$$c^* = argmax_i[P(y_{DCN}, y_{BGM}|c_i)P(c_i)]. \tag{7}$$

4 Experiments

4.1 Mining Dataset

The dataset use for evaluating this work was collected from a camera mounted to a light vehicle operating in an active mine-site, see Figure 10. While the motivation is to put vision based sensing on a heavy vehicle, a light vehicle is more practical for gathering a diverse set of visual sequences. The cameras used in this study were 0.9 MP BlackFly cameras produced by PointGrey. The frame-rate was fixed to 10 Hz and maximum exposure time set to 20 ms to prevent blur. The lens focal length was set to 1.8 mm, translating a to 94 degree horizontal field of view. The dataset contains both static and dynamic instances of a `person`, `LV` or `HV`.

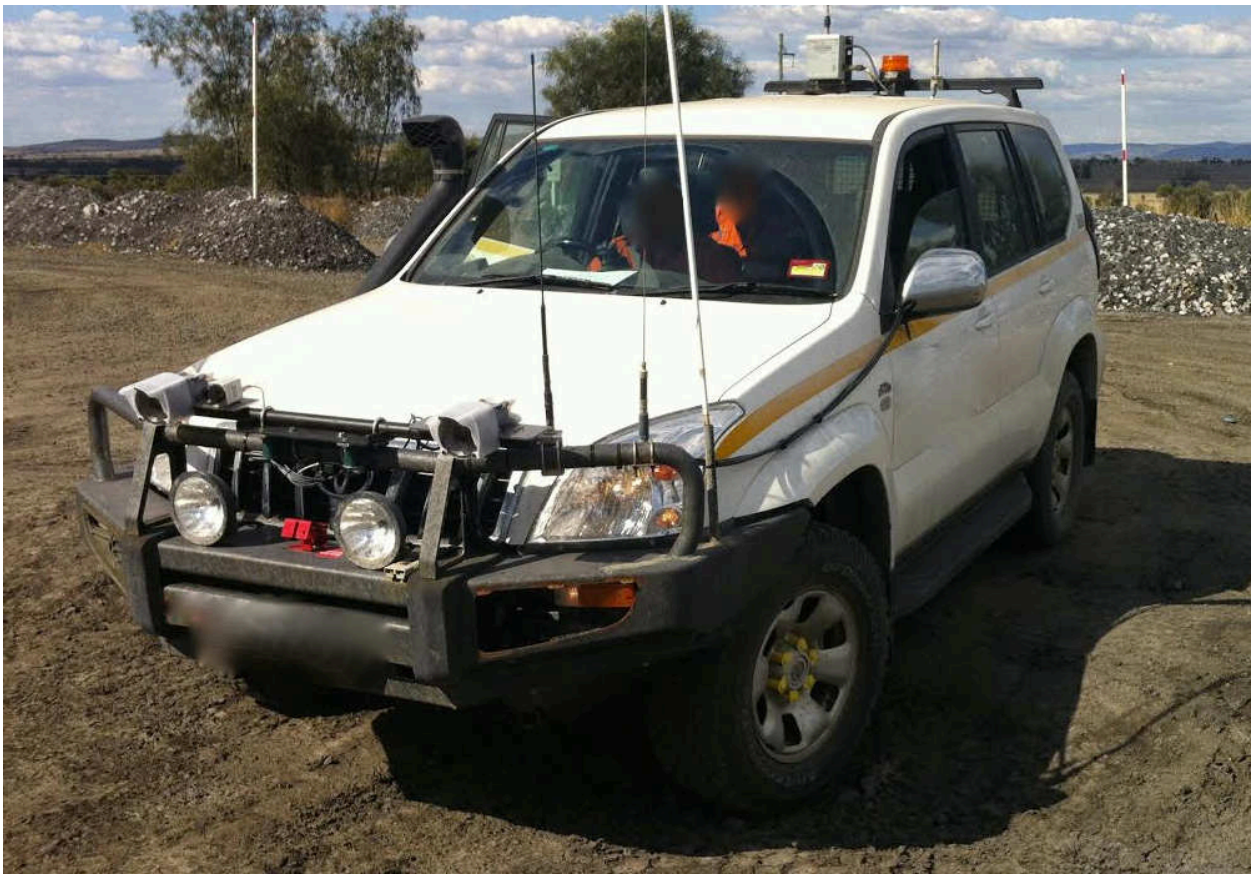

Figure 10: The experimental dataset gathering vehicle with cameras mounted to the bullbar. Note: all images used in this paper were captured from the camera on the left hand side of the vehicle.

Continuous video was gathered with and without the camera in motion and on various haul roads and a few light vehicle only zones to capture variation in the environment. This video data was captured at 10 Hz and partitioned into various sequences. In this work we use 5 sequences where no people or vehicles are visible to build our background model. Collectively these background sequences make up 8952 frames in total (approximately 14 km). These sequences are referred to as Train 1-5 in Table 1.

To evaluate the performance we use another 6 sequences with several instances of `person`, `LV` or `HV`, that were manually annotated using the tool developed by (Vondrick et al., 2012). These annotated sequences contain 11150 frames in total (approximately 11 km) and referred to as Test 1-6 in Table 1.

In addition to the train and test sequences captured in the field, we made a small validation set of 200 manually cropped images containing mining objects. These validation images were collected from various sources including the internet along with a few captured at night from the same mine site but in different locations to the test sequences. This validation set was used throughout the design process to generate Figures 4, 7, 9, 11, and Figure 12. Table 2 lists the parameters of the system with links to the empirical studies performed in this work and the dataset sequences used. Note that the study performed for selecting the NMS did use the test sequences since they are the only sequences available with substantial set of annotated mining object classes in real life scenarios. However, since the same NMS is applied to all the classifier models we consider the later comparisons in this experimental section to remain valid.

Table 1: Experimental Image Sequences

Name	Total Frames	Purpose	Environment	Conditions	Content
Train 1	2433	Learning Clusters	Haul road between pits	Morning, Sunny	BG Only
Train 2	1975	Learning Clusters	In-pit haul road	Midday, Sunny	BG Only
Train 3	1500	Learning Clusters	In-pit haul road	Afternoon, Semi-overcast	BG Only
Train 4	1669	Learning Clusters	Haul road between pits	Afternoon, Semi-overcast	BG Only
Train 5	1375	Learning Clusters	Haul road between pits	Evening, Overcast	BG Only
Validation	200	Parameter Tuning	Hand cropped mining images from various sources[a]	Day and night	BG, People, LV, HV
Test 1	1463	Evaluation	Store yard and processing plant	Morning, Sunny	People, LV
Test 2	2951	Evaluation	In-pit, haul road and LV area	Midday, Sunny	People, LV, HV
Test 3	600	Evaluation	Haul road between pits	Midday, Sunny	People, LV
Test 4	2827	Evaluation	In-pit and haul road	Midday, Sunny	LV, HV
Test 5	1569	Evaluation	LV only area	Evening, Overcast	People, HV
Test 6	1740	Evaluation	LV/HV parking area	Night, Overcast	People, LV, HV

[a] The 200 validation images are made up from both internet images and several challenging background and night images which were collected in the field but outside of the test sequences 1-6.

4.2 Background Model Validation

Here we describe the experiments performed to design our background modelling system explained in the previous section. From the 5 background sequences, we applied the region proposal and remapped DCN detection framework to find challenging region proposals from every tenth frame. While some of the false objects may be observed in multiple frames, the time difference is sufficient to capture a variety of view points for these distracting objects. We lowered the detection threshold to collect region proposals if the remapped DCN predicted either a `person` or `car` in the top 5 out of 200 ImageNet class responses. With this configuration we collect around 8000 hard negatives for our background reservoir. We held out 90 of the most interesting background regions and added them to the validation set.

To address the design decisions for the background cluster model, we perform an empirical study using the reservoir containing only negatives and the validation set with both negative and positives. We jointly test

Table 2: System Parameters

Parameter(s)	Values	Tuning Method	Data Used
EB Max Count	1000	Assumed	None
NMS Overlap	0.5	Empirical (see Fig. 3)	Test 1-5
DCN Remapping	*	Empirical (see Fig. 4)	Validation
DCN Retraining	*	SGD using classification loss	Validation
BGM Clusters (k)	128	Empirical k-means (see Fig. 11)	Train 1-5
BGM Euclidean (α, β)	(0.015,0.111)	Empirical (see Fig. 8)	Validation
BGM Cosine (θ, b)	(*,-38.9)	SGD with 10 fold cross validation	Validation
Fusion prior ($BG, Person, LV, HV$)	(0.7,0.1,0.1,0.1)	Assumed	None

* indicates a high dimensional parameter set.

different combinations of DCN layer features and number of clusters by evaluating their performance on the validation set. For the distance threshold we set this to the distance corresponding to a 95% recall on the positive set. With the recall fixed, the overall performance of the background model is measured by the precision at which it can identify a true negative.

In all experiments we use the DCN structure of (Krizhevsky et al., 2012), which is an eight layer network with five convolutional layers, followed by three fully connected layers. The last layer was adapted to the detection task (Girshick et al., 2014) with 200 outputs corresponding to the ImageNet detection dataset. For extracting a feature to describe the background appearance, we consider the response of layers in the network. Figure 11 shows the relative performance of the different DCN layers when sweeping over different background model sizes measured by the number of clusters used. This experiment shows that the `fc6` layer exhibits substantially higher precision across all model sizes. This could be put down to its position in the network, where it is the first layer to incorperate the global visual information, as opposed to the convolutional layers. These findings are in agreement with a recent related study for investigating the performance of transfering layers where the 6th layer regularly achieves the highest performance on the target task (Azizpour et al., 2015).

Figure 11 shows that each layer produces a dog-leg shaped curve with a common turning point at 128 clusters. This signifies a point where the diversity of the environment is sufficiently represented. While increasing the number of clusters beyond 128 clusters continues to improve precision, the rate of improvement is small. E.g. for 16x more clusters the improvement is only 1% (89% for 128 versus 90% for 2048 clusters) compared to a 3% drop when using 16x fewer clusters (86% for 8 clusters). In terms of processing time, both the nearest neighbour lookup for the k-means Euclidean model and the cosine similarity model is linear in the number of clusters [2]. Since the `fc6` layer with 128 provides a reasonable trade-off between precision and computation speed, these parameters are used in all remaining experiments. A detailed view of the distances between the validation samples and the nearest cluster centre using these parameters can be seen in Figure 7.

Figure 12 illustrates the samples in this validation set with the largest error. The false negatives are mostly night images which can be put down to the fact that similar images are rare if not non-existent in the ImageNet samples used to train the DCN. For the false positives, these are mostly signs which make up a minority of the scene. From these samples we can describe our background model as a form of novelty detection where interesting parts of the scene such as signs are distinguished from the general background. This finding along with the unsupervised clustering shown in Figure 6 are a testament to the DCN's expressive capabilities in representing visual similarity.

For the cosine based background model, `fc6` features were used and the number of clusters was set to 128, matching the Euclidean k-means model. The logistic regression was trained using stochastic gradient descent optimisation with regularisation penalty selected though 10-fold cross validation. The resulting learnt separation between background and non-background was previously shown in Figure 9.

4.3 Detection Evaluation

The system is evaluated on a set of five daytime sequences and one night sequence, where the task is to detect and locate both people and vehicles within each frame. In this evaluation we follow the same criteria as described in (Bewley and Upcroft, 2015), where we consider a true detection if at least 50% of the detection region is covered by a single ground truth object. This differs from the intersection-over-union (IOU) definition of overlap, as we accept detecting a `person`'s head and shoulders without their whole body while IOU would count this as both a miss detection and a false positive. Additionally, if multiple detections overlap a single ground truth instance, we count this as a single true positive and neither of the overlapping detections are false. A concrete example would be if a person's head is covered by a single detection and their body another. In such a scenario, both are considered valid detections but only count as a single true

[2] Efficient data structures like kd-trees where considered but due to the high dimensionality the practical improvement was negligible.

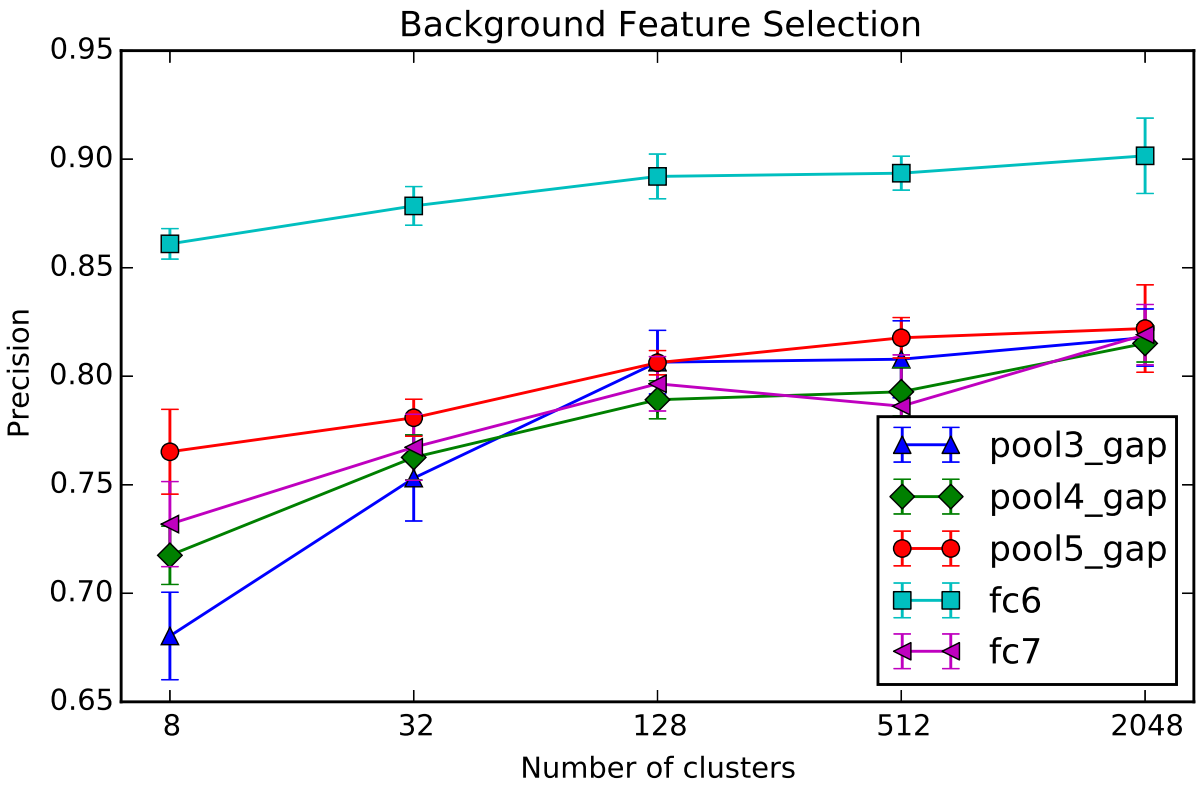

Figure 11: Cross-validation precision at a fixed recall of 95% for different DCN layers and the number of clusters used to represent the background. Each point shows average of 5 trials. Pool 3-5 are derived from convolutional layers 3-5 respectively while the fc6 and fc7 are direct copies of the fully connected layer responses. Note: the x-axis is in log_2 scale.

positive per object. It should be also noted that any detection or miss detection of an object labelled as partially occluded in the ground truth is ignored in this evaluation. Only the retrained model and the fused models are capable of detecting `HV` so were omitted from the valuation across the different models for fair comparison. To accommodate this, any detections which overlap with `HV` objects are considered as neither true or false and are excluded from the evaluation.

Figure 13 shows both a failure example on the left and valid detections on the right for each sequence. Sequence 1 (top row), is located near a store yard and a processing plant where there are a number of regions with varying appearance making this sequence particularly challenging for the background model. This sequence has a lower precision as there are a number of novel regions in this environment that mitigate the benefits provided from the background model trained on common background sequences. The *remapped* DCN model is less effected by this as some of the 200 ImageNet classes contain sufficient diversity to compete with the `person` and `LV` class responses. The second sequence was captured around an in-pit go-line, where mine workers transition between operating LV and HV. In this environment, the fused model performs exceptionally well in detecting `person` and `LV`, however, `HV` would occasionally be missed or misclassified as `LV`. The third and forth sequences were captured along main haul roads, between pits and in the pit (respectively), representing the typical operating environments for `HV`. In this environment the fusion of the background model nearly elimintates all the false positives caused by trees and other common road side objects. However, novel objects such as signs and cones are sometimes detected (e.g. in the fourth row of Figure 13 a distant "Reduce Speed" sign is mistaken for a `person`). The fifth sequence was taken in the late afternoon in semi-overcast weather conditions and consists of multiple mine workers in view of the

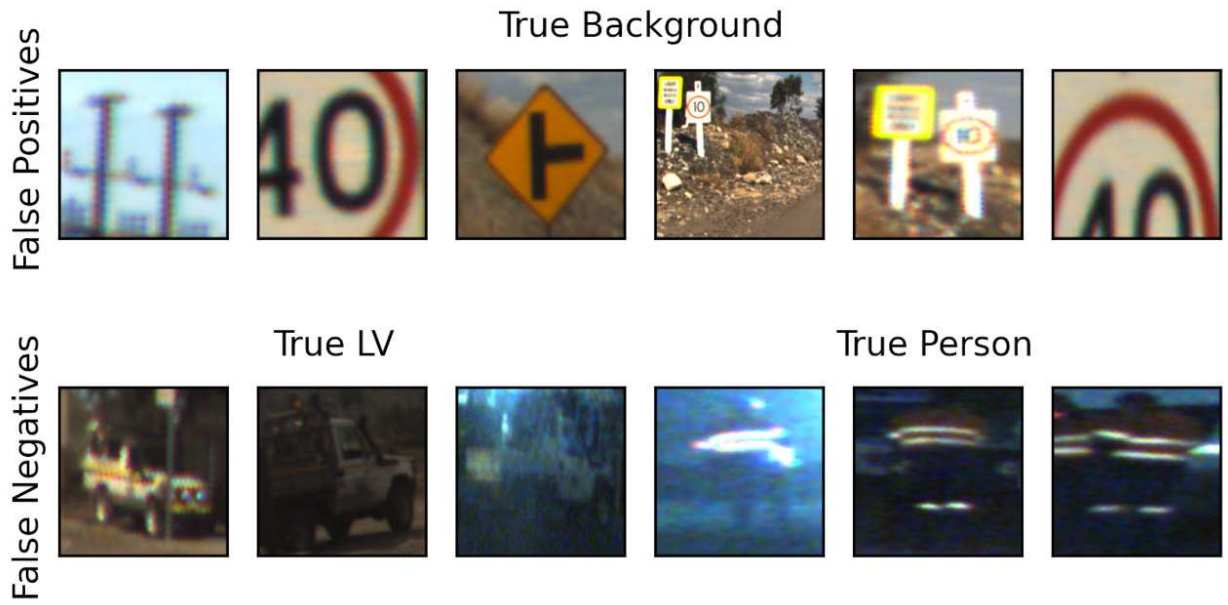

Figure 12: Validation samples where the nearest k-means background cluster model failed. Images are shown in their warped form, representing the DCN input. The four right false negatives were collected at night.

camera. The final sequence was shot at night around the mine's LV parking area where significant activity was observed – coinciding with a shift change. Under these conditions the fused model is able to detect both people and LV while it also appears to have learnt to distinguish between yellow lights mounted to vehicles and white lights such as the bright spot corresponding to a saturated retro-reflective sign to the left of the distance LV in the bottom right image of Figure 13.

Table 3 shows a detailed breakdown in the performance of all individual components and the fused method. Here we have separated the performance of each component to isolate the behaviour of the DCN and the background models (BGM). For all models the detection probability threshold was set to 0.5, except the (Bewley and Upcroft, 2015) model which was based on a Euclidean distance to achieve 95% recall on the validation set. The DCN output generally has a high recall but low precision as the DCN does not consider the class prior probability.

The two BGM models behave considerably differently from each other. Particularly the model based on the Euclidean distance to the nearest neighbour presents high recall as this was by design in the threshold selection. However it is interesting that the performance in both precision and recall significantly dropped compared to the held out validation set used in tuning the distance distribution parameters. On the other hand, the Cosine based BGM has significantly improved the recall, resulting in an overall F1 score that is compatible to the DCN output. Despite using a roughly assumed class prior, the soft fusion between the retrained DCN and cosine BGM, enables the fused model to take advantage of both the high DCN recall and the BGM's precision to produce an overall top F1 score.

5 Conclusions and Future Work

Visual object detection has the potential to make a significant impact to improving safety in the mining industry. The goal of this work is to address the gap in situational awareness of heavy mining vehicles to sense other vehicles and personnel without active transponders. Towards achieving that goal, this paper presented a passive, vision-only detection system that takes advantage of recent developments in computer

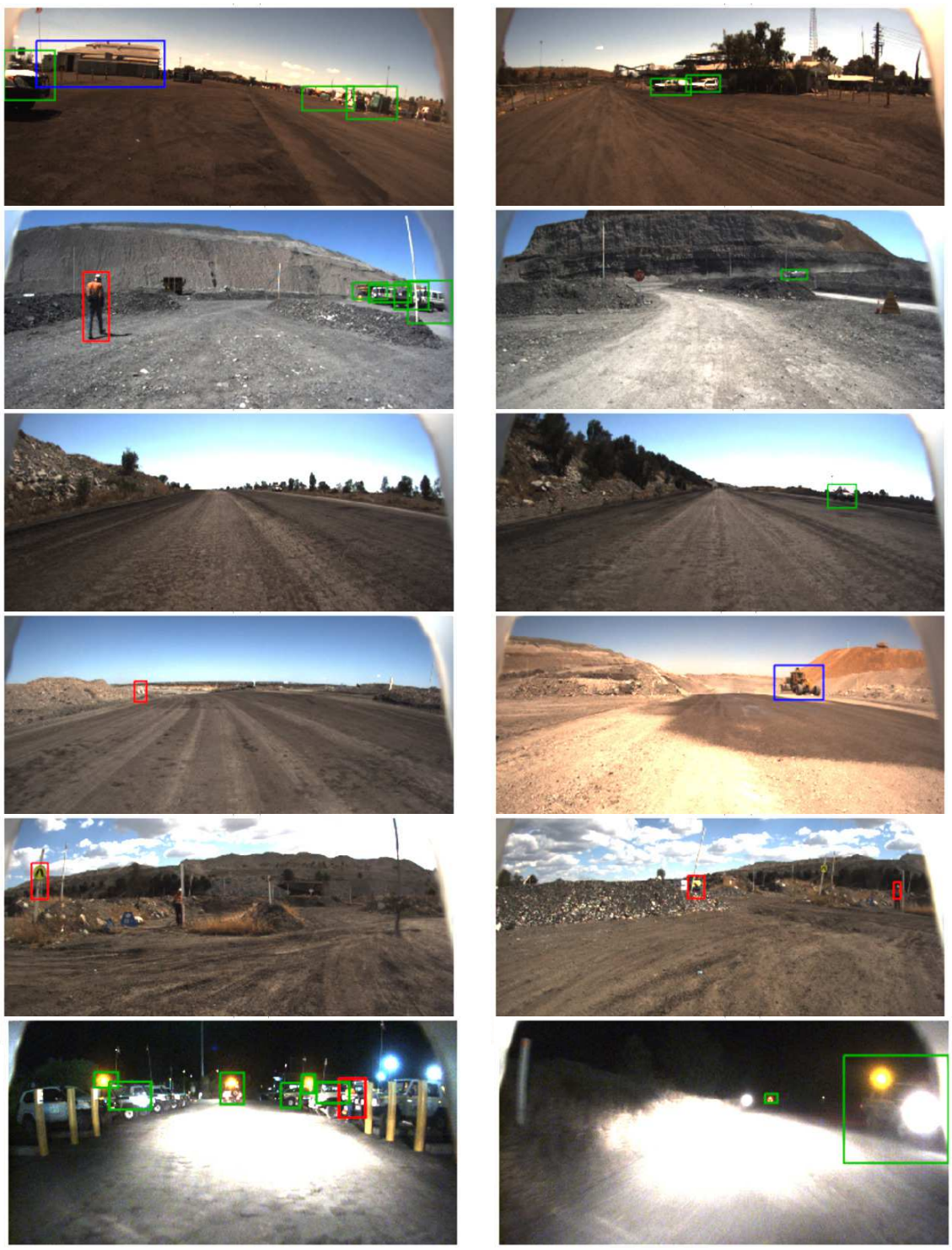

Figure 13: Exemplar images of *fused* model with images containing at least one false positive or miss detection in the left column and only true positives on the right. Each row represents the sequence order 1 to 6 corresponding to the results in Table 3. The colours denote the predicted class of the object with: *red* for Person, *green* for LV and *blue* for HV. Note: detecting and recognising HV is attributed to the *retrained* component of the *fused* model.

Table 3: Detection Performance on Mining Sequences
F1 (Precision, Recall)

Sequence (frames)	DCN Remapped[a]	DCN Retrained	BGM Euclidean	BGM Cosine	Fused
Test 1	**0.33** (**0.24**,0.54)	0.17 (0.10,**0.72**)	0.07 (0.04,0.67)	0.23 (0.14,0.57)	0.27 (0.17,0.67)
Test 2	0.81 (0.72,0.93)	0.80 (0.68,**0.98**)	0.22 (0.12,0.92)	0.73 (0.69,0.78)	**0.90** (**0.87**,0.94)
Test 3	0.04 (0.02,0.46)	0.09 (0.05,**0.68**)	0.01 (0.01,0.55)	0.12 (**1.00**,0.06)	**0.69** (0.90,0.56)
Test 4	0.26 (0.15,0.89)	0.42 (0.27,**0.96**)	0.10 (0.05,0.93)	0.56 (0.46,0.73)	**0.68** (**0.58**,0.82)
Test 5	0.43 (0.33,0.62)	0.72 (0.58,**0.93**)	0.25 (0.15,0.89)	0.60 (0.79,0.48)	**0.83** (**0.88**,0.78)
Test 6	0.57 (**0.54**,0.60)	**0.61** (0.49,0.83)	0.47 (0.32,**0.86**)	0.60 (0.48,0.80)	0.59 (0.49,0.75)
Overall	0.475 (0.39,0.75)	0.541 (0.42,**0.90**)	0.202 (0.12,0.86)	0.550 (0.56,0.66)	**0.692** (**0.65**,0.80)

[a] This is a soft probabilistic version of the method presented in (Bewley and Upcroft, 2015).
Bold indicates the best performance across models.

vision and machine learning to detect both personnel and mining vehicles. This sensing approach was evaluated in an active open-pit mine site environment across six different parts of the mine and at both day and night. Challenges around over-fitting on a small dataset were addressed by exploring a fusion framework incorporating a pre-trained DCN and exploring both remapping and retraining the final layers for adapting to the mining environment.

The experiments show that the in-pit environment is suitable for vision based detection utilising object proposals and DCNs, along with background modelling techniques. The experiments also show that the amount of NMS applied to the proposed regions can increase the proposal recall compared to the same number of proposals selected using the proposal score. However, it was also shown that this improvement has limited range in the amount of NMS applied. This characteristic, to a degree, could be contributed to the differences in the mining environment compared to the typical internet based images used in the development of the `EdgeBox` proposal method used in this work.

When applying an off-the-shelf DCN pre-trained for the ImageNet detection task to mining imagery we highlight that it performs poorly on both the background and the mining specific `HV` class. While this method has the advantage of retaining knowledge of the large amounts of data from ImageNet for the common classes, it is restricted in its ability to distinguish new and novel classes. On the other hand, retraining the final layers using mining images overcomes this limitation allowing the `HV` class to be distinguished from the `background`, `person` and `LV` classes. However, with the limited amount of labelled training data available, this retrained model is comprable to simply remapping the classes from the pre-trained ImageNet model.

The use of a DCN alone for visual object detection is shown to perform poorly when presented with typical mining imagery such as a haul road environment. To overcome this limitation, two variants of a background model are presented in this work for the purpose of suppressing the false positives produced by the DCN. The soft probabilistic variant of the background model when fused with the DCN output offers superior performance over the logic based combination previously used for background suppression in (Bewley and Upcroft, 2015). Quantitatively, the fused model presented achieves a relative improvement in F1 score of 46% over a pre-trained DCN and 28% over a DCN retrained with mining images.

This article also presents an improved variant of the background model shown to have several desirable attributes. Firstly, the visual feature descriptor is already computed as part of the DCN recognition. Secondly, it only required the bare minimum labelling effort to build a large reservoir of background only samples. It is made robust to the high dimensional nature of the DCN features by using the cosine similarity followed by logistic regression. All its parameters can be estimated from a small set of labelled images. Also the cosine variate provides superior background discrimination over the simpler k-means model, achieving a detection accuracy comparable to the best DCN baseline model on its own. Moreover, the higher precision of the cosine background model makes it complementary to the high recall DCN when used in a fusion framework. However, in sequences where the background differed significantly from the background only training sequences its performance dropped considerably. This combined with the use of a static offline trained background limits this approach to having either a comprehensive training set or restricting the detection to similar environments, e.g. only haul roads. Future work in addressing this issue could include learning the background cluster online by incorporating additional information from other sources.

The Bayesian fusion framework for combining the outputs of the DCN and the background model is basic and assume independence. Despite its simplicity, the improvement in F1 score of the fused model over its counterparts suggest there is value in combining these methods. Future work could include alternative fusion techniques such as cascade architectures to improve computational efficiency or introducing other sensor information.

Another characteristic of this work is that it does not retain any temporal information. In particular the feed-forward architecture of the DCN treats each frame in the video like it is seeing it for the first time. An advantage of this is that the framework is robust to motion experienced by either the object or camera by performing the detection of each frame independently. The disadvantage of this is that temporal memory is useful in video as objects move gradually in the image which can be exploited through tracking techniques to improve recall. Also, while this work is only concerned with single camera based sensor data we see many opportunities to combine techniques incorporating, motion segmentation (Bewley et al., 2014), odometry (Hawke et al., 2015), stereo (Bewley and Upcroft, 2013) or range-based sensors for improved robustness. Additionally, as more labelled mining image data becomes available we expect to be able to design and fine-tune a DCN that performs better in this domain than the existing network designed for the ImageNet benchmark.

Acknowledgments

This research was funded by the Australian Coal Association Research Program (ACARP). The authors would also like to acknowledge Anglo American for allowing data collection at the Dawson operation. Acknowledgement also goes to the high performance computing group at Queensland University of Technology for both support and use of their services when conducting the experiments in this paper.

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *International Conference on Database Theory*, pages 420–434. Springer Berlin Heidelberg.

Ahmed, A., Yu, K., Xu, W., Gong, Y., and Xing, E. (2008). Training Hierarchical Feed-Forward Visual Recognition Models Using Transfer Learning from Pseudo-Tasks. In *European Conference on Computer Vision (ECCV)*, pages 69–82. Springer Berlin Heidelberg, Berlin, Heidelberg.

Aytar, Y. and Zisserman, A. (2011). Tabula Rasa: Model Transfer for Object Category Detection. In *International Conference on Computer Vision (ICCV)*.

Azizpour, H., Razavian, A. S., Sullivan, J., Maki, A., and Carlsson, S. (2015). From Generic to Specific

Deep Representations for Visual Recognition. In *Computer Vision and Pattern Recognition (CVPR) Workshops*.

Benenson, R., Mathias, M., Tuytelaars, T., and Van Gool, L. (2013). Seeking the strongest rigid detector. In *Computer Vision and Pattern Recognition (CVPR)*.

Bewley, A., Guizilini, V., Ramos, F., and Upcroft, B. (2014). Online Self-Supervised Multi-Instance Segmentation of Dynamic Objects. In *International Conference on Robotics and Automation*, pages 1296–1303, Hong Kong, China. IEEE.

Bewley, A. and Upcroft, B. (2013). Advantages of Exploiting Projection Structure for Segmenting Dense 3D Point Clouds. In *Australian Conference on Robotics and Automation*.

Bewley, A. and Upcroft, B. (2015). From ImageNet to Mining : Adapting Visual Object Detection with Minimal Supervision. *Proceedings of the 10th International Conference on Field and Service Robotics (FSR)*.

Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

Dempster, A. P. (2008). A Generalization of Bayesian Inference. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 73–104. Springer Berlin Heidelberg, Berlin, Heidelberg.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Dollar, P., Appel, R., Belongie, S., and Perona, P. (2014). Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2009). Pedestrian detection: A benchmark. *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Donahue, J., Hoffman, J., Rodner, E., Saenko, K., and Darrell, T. (2013). Semi-supervised Domain Adaptation with Instance Constraints. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 668–675.

Durrant-Whyte, H. and Henderson, T. C. (2008). Multisensor Data Fusion. In *Springer Handbook of Robotics*, pages 585–610. Springer Berlin Heidelberg, Berlin, Heidelberg.

Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013). Learning Hierarchical Features for Scene Labeling. *Pattern Analysis and Machine Intelligence*, 35(8):1915–1929.

Fei-Fei, L., Fergus, R., and Perona, P. (2007). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–45.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*.

Hawke, J., Gurau, C., Tong, C. H., and Posner, I. (2015). Wrong Today , Right Tomorrow : Experience-Based Classification for Robot Perception. In *Field and Service Robotics*.

Hosang, J., Benenson, R., and Schiele, B. (2014). How good are detection proposals, really? In *British Machine Vision Conference (BMVC)*.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. In *International Conference on Multimedia*, pages 675–678.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages (Vol. 1, No. 2, p. 4).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1:541–551.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755. Springer International Publishing.

Marshall, J. A. and Barfoot, T. D. (2008). Design and field testing of an autonomous underground tramming system. *Springer Tracts in Advanced Robotics*, 42:521–530.

McMahon, S., Sünderhauf, N., Upcroft, B., and Milford, M. (2015). How Good Are EdgeBoxes, Really. In *In Workshop on Scene Understanding (SUNw), Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Mosberger, R. and Andreasson, H. (2012). Estimating the 3d position of humans wearing a reflective vest using a single camera system. In *International Conference on Field and Service Robotics (FSR)*.

Mosberger, R., Andreasson, H., and Lilienthal, A. J. (2013). Multi-human Tracking using High-visibility Clothing for Industrial Safety. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 638–644.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*.

Phillips, T., Hahn, M., and McAree, R. (2013). An evaluation of ranging sensor performance for mining automation applications. In *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics: Mechatronics for Human Wellbeing, AIM 2013*, pages 1284–1289.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN Features off-the-shelf: an Astounding Baseline for Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Roberts, J. M. and Corke, P. I. (2000). Obstacle detection for a mining vehicle using a 2D laser. In *Proceedings of the Australian Conference on Robotics and Automation*, pages 185–190.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2014). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *International Conference on Learning Representations (ICLR 2014)*.

Sermanet, P., Kavukcuoglu, K., Chintala, S., and LeCun, Y. (2013). Pedestrian Detection with Unsupervised Multi-stage Feature Learning. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv technical report*.

Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., and Milford, M. (2015). On the Performance of ConvNet Features for Place Recognition. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, Germany.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going Deeper with Convolutions. In *Computer Vision and Pattern Recognition (CVPR)*.

Torralba, A., Fergus, R., and Freeman, W. (2008). 80 Millions Tiny Images: a Large Dataset for Non-Parametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1958–1970.

Uijlings, J. R. R., Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. (2013). Selective Search for Object Recognition. *International Journal of Computer Vision*, 104(2):154–171.

Vondrick, C., Patterson, D., and Ramanan, D. (2012). Efficiently Scaling up Crowdsourced Video Annotation. *International Journal of Computer Vision*, 101(1):184–204.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks ? In *Advances in Neural Information Processing Systems (NIPS)*.

Zeiler, M. D. and Fergus, R. (2013). Stochastic Pooling for Regularization of Deep Convolutional Neural Networks. In *International Conference on Learning Representations*.

Zhang, S., Benenson, R., and Schiele, B. (2015). Filtered Channel Features for Pedestrian Detection. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Zitnick, C. and Dollár, P. (2014). Edge Boxes: Locating Object Proposals from Edges. In *European Conference on Computer Vision (ECCV)*.