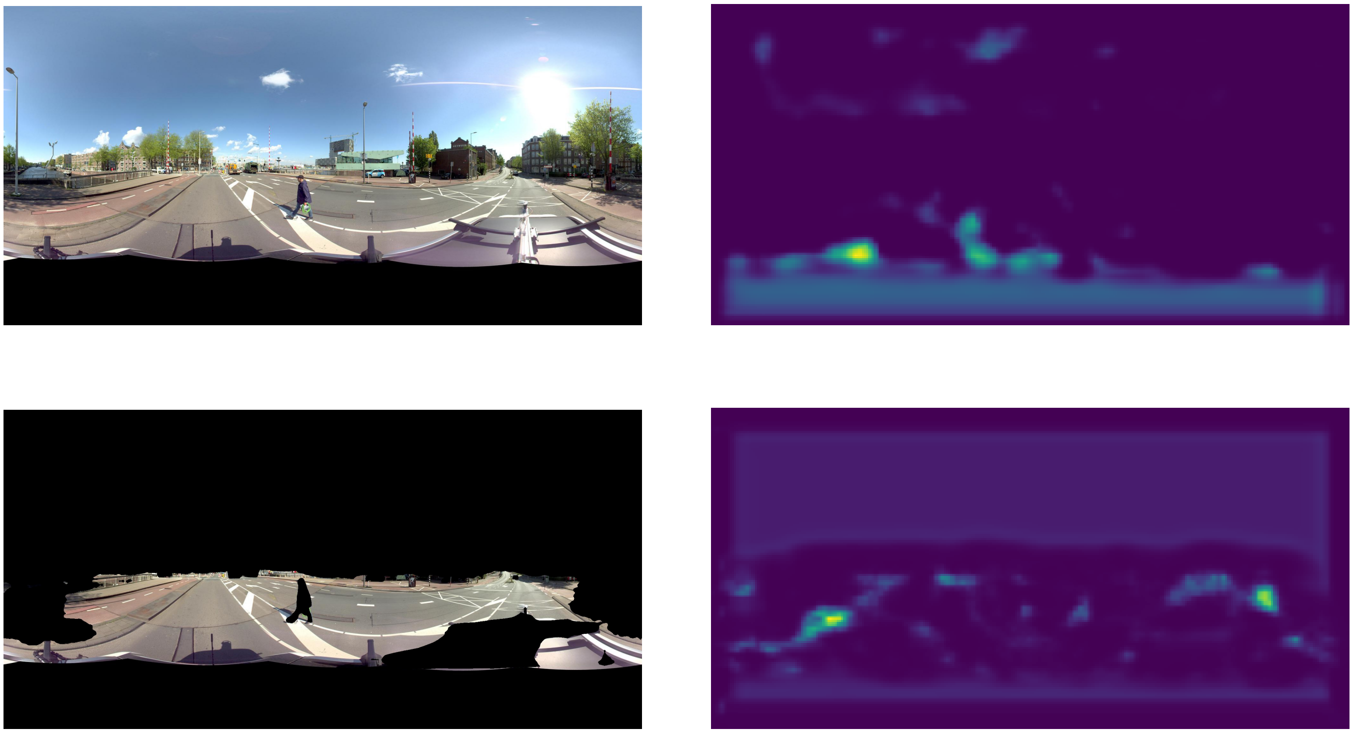
Context-aware stacked convolutional neural networks for road crack detection with weakly labelled data

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

Owen Winter
13376152

MASTER INFORMATION STUDIES
DATA SCIENCE
FACULTY OF SCIENCE
UNIVERSITY OF AMSTERDAM

2021-07-30



	Internal Supervisor	External Supervisor	External Supervisor
Title, Name	Dr. Stevan Rudinac	Maarten Sukel MSc	Iva Gornishka MSc
Affiliation	UvA, ABS	UvA, Gemeente Amsterdam	Gemeente Amsterdam
Email	s.rudinac@uva.nl	m.sukel@amsterdam.nl	i.gornishka@amsterdam.nl

Universiteit van Amsterdam

Gemeente Amsterdam

Context-aware stacked convolutional neural networks for road crack detection with weakly labelled data

Owen Winter
University of Amsterdam
owen.winter@student.uva.nl

ABSTRACT

Computer vision has developed rapidly in recent years. One potential application is detection of road damage. In this paper we consider the use of image recognition with transfer learning to detect road cracks in Amsterdam with weakly labelled data. While computer vision models have been shown to be effective at detecting damage when trained on annotated data, we combine manual inspection data, found at the road-level, with multiple panoramic street images per road. With weakly labelled data, there are a number of ways to introduce more contextual information for model training. We consider different forms of transfer learning including Region-of-Interest (ROI) masks, pre-training and stacked networks as ways to mitigate visual noise. We draw from research in medical imagery to consider a stacked architecture which uses a fully convolutional neural network (CNN) trained on annotated examples of cracks to produce dense prediction maps from unannotated panoramas, passed to a second CNN. Through an experimental process, we show some potential for the use of ROI-masks and stacked networks with weakly labelled data, as well as the limitations and difficulties of these approaches.

KEYWORDS: computer vision, convolutional neural networks, deep learning, image recognition, road condition monitoring, transfer learning, stacked networks, region-of-interest masks

1 INTRODUCTION

Rapid expansion in computer vision and availability of street-level imagery has opened the door to many implementations of street-level image classification, segmentation and object detection. In this paper, we test the efficacy of context-aware stacked networks for image classification using street-level panoramas labelled at the road-level. Using stacked networks, we hope to learn both global contextual and localised features of the panoramas to enable crack detection.

Most current approaches to computer vision rely on Convolutional Neural Networks (CNN) [1]. CNNs have significantly improved the effectiveness of image recognition in recent years and are now dominant in the field. In 2020, all of the highest scoring proposals for the Global Road Damage Detection Challenge utilised CNNs in some form [2]. These CNNs typically propose bounding boxes and classifications in either one or two stages [3]. However, using CNNs for localisation generally requires large volumes of training data with road damage annotated image-by-image.

In Amsterdam's case, training data is labelled at the road-level with varying numbers of panoramic images per road (from 1 to 172 - see Appendix A). Each image may or may not include the damage which corresponds to the road-level label, making this task a form of weak supervision.

There are many approaches to the problem of weak or limited labels, including transfer learning, feature extraction and unsupervised learning. In our case, the problem is compounded by the fact that road damage detection is a form of fine-grained classification - that is, we are seeking to classify whole images on the basis of very small regions of visual distinction, with high intra-class and low inter-class variation [4, 5].

We draw from existing learning using pre-trained models to make context-aware networks which can identify regions of interest for crack detection. This approach is inspired by the work of Bejnordi et al. [6] in medical imagery and Wang et al. [7] in fine-grained categorisation, as well research using weakly annotated videos [8, 9].

We experiment with different forms of transfer learning. This includes models trained on known images of road cracks, from the Global Road Damage Detection Challenge (GRDDC) dataset, either in one network with frozen parameters or as part of a stacked network [2, 10]. We also experiment with the use of region-of-interest (ROI) masks, generated by a pre-trained segmentation model, to remove non-road pixels.

Stacked networks are formed from two CNNs [6]. The first is trained on small, high-resolution image segments to detect local instances of a class. This network is applied to larger input images to produce dense prediction maps. Dense prediction maps are passed to a second CNN, which learns patterns of global distribution of the class, and outputs an image-level classification.

The research question we seek to address is:

Can context-aware stacked neural networks effectively classify weakly labelled images?

This research has implications for the use of image recognition across many domains. Current methods for inspecting road surfaces are slow, labour-intensive, expensive and subjective. The City of Amsterdam alone is responsible for almost 2,000km of road surfaces, which are manually inspected annually [11]. Image recognition with street-level imagery has the potential to expedite this process, especially in cities such as Amsterdam which regularly collect street-level panoramic images [12].

Beyond road damage, these techniques are relevant to fine-grained image classification problems involving street-level imagery, for example detecting facade damage or graffiti. Although for the purposes of this paper we have focused on binary classification, similar architectures could be explored for multi-class problems.

2 RELATED RESEARCH

In this section, we review some developments in the broader field of computer vision before focusing on methods relevant to this project in semantic segmentation, fine-grained image classification, and the use of transfer learning - specifically stacked CNNs - for road damage detection.

The application of machine learning techniques to visual data is one of the most rapidly growing areas of artificial intelligence. With greater computing power, increased volumes of labelled visual data, and more sophisticated machine learning techniques, computer vision has become both more effective and more accessible [1, 13, 14].

Alongside the increased use of computer vision is the growth of street-level imagery. This has facilitated research from house number recognition, to mapping architectural correspondences, to predicting election results by frequency of car models [15–17]. Increased street-level imagery has been critical for the development of autonomous driving, which has particularly relied on computer vision techniques such as object detection and semantic segmentation [18–20].

2.1 Convolutional Neural Networks

Recent image recognition projects overwhelmingly rely on CNNs [1]. CNNs are multi-layered artificial neural networks in which images are inputted as matrices of colour values. These matrices typically pass through matrix transformations (convolutions) and activation functions to simplify the image into recognisable patterns, known as feature representations, before linear layers are used to return the predicted class [15, 21, 22]. CNNs have proven to be very effective for a number of purposes, including image classification, object detection and pixel-level segmentation [1, 23].

Since the late 20th century, there have been rapid advances in the use of CNNs. Prominent innovations include the use of gradient descent backpropagation by LeCun et al. [24], with LeNet used to identify handwritten numbers; the move towards GPU processing suggested by Steinkraus et al. [13]; and the use of Rectified Linear Units as part of a deeper neural network suggested by Krizhevsky et al. [25]. Since 2012, deep CNNs have become the gold standard for image classification problems and progress has continued, as demonstrated in the annual ImageNet Challenge [14, 26]. In 2016, He et al. [27] introduced ResNet, utilising batch normalisation and crucially skip connections to allow for residual learning and even deeper networks.

2.2 Semantic Segmentation

Semantic segmentation is a key technique within computer vision, especially with regard to street-level images.

Similar to other computer vision problems, deep learning has become dominant in the semantic segmentation field in recent years [28]. Shelhamer et al. [23] proposed the Fully Convolutional Network (FCN) in 2015, utilising convolutions and skip connections to produce detailed segmentation outputs based on both deep semantic and local features. Other innovations have included trying to incorporate global context by various means, such as context vectors or conditional random fields, and using encoder-decoder methods to deconvolve convolutional layers [29–32].

One of the most important drivers of innovation for street-scene segmentation has been the development of autonomous driving [20, 32]. Many of these models are trained on increasingly expansive training data such as the CityScapes or COCO datasets [33, 34].

2.3 Fine-Grained Classification

While image classification and semantic segmentation have been shown to be highly effective, an additional longstanding challenge in computer vision is fine-grained classification [4, 5]. Fine-grained classification involves classifying images not only based on instances of high-level classes but also subordinate-level categories such as breed of dogs, model of car, or damage to surfaces. This poses unique difficulties because of high intra-class variation and low inter-class variation [4, 5].

Proposed solutions often include some form of localisation which is used to direct attention to visually discriminating elements of images. Wang et al. [7], for example, leverage the other labels in subordinate-level objects' ontology trees to train a series of classifiers at different levels of granularity. These networks combine information from each level of granularity to segment images by high-level class as well as subordinate-level discriminative features.

For road damage detection, many of the highest scoring entries in the GRDDC take a two-stage approach similar to this, using R-CNN or Faster R-CNN for region proposal, sometimes combined with road segmentation [2]. Others, including the winning entry, take a one-stage approach, using an Ultralytics-YOLO (You Only Look Once) architecture [35]. YOLO works by dividing images into a grid and predicting boundary boxes with confidence scores centred on each grid square, in a single network [3, 36].

Unlike the multiple granularity approach taken by Wang et al. [7], these road damage detection techniques require annotated training data. They also perform best when applied to images in which road surfaces form a large proportion of the total image - that is, when variation in the higher class is limited.

2.4 Stacked Networks

One novel approach to fine-grained classification is the idea of stacked CNNs proposed by Bejnordi et al. [6]. Their domain is medical imagery, particularly the detection of breast carcinomas. In this case, the higher level class is tumours, which are relatively easy to detect by their structure. The subordinate class is whether the tumour is malignant, which can be identified using cellular information. However, the size of image required to include both structural and cellular information is too large to take the parallel approach proposed by Wang et al. [7].

Stacked CNNs resolve this problem by training the model in two stages. The first stage is to train a fully convolutional classifier for the subordinate class with high resolution image segments. The second is to train a classifier using dense prediction maps produced by running the first convolutional model across full sized images. In this way, the global structure of different cells could be learned across the images with much lower computational cost.

This approach builds on Bejnordi et al. [37]'s earlier technique, segmenting medical imagery by pixel similarity before individually classifying the segments. A similar method by Li et al. [38] has also proven to be effective for detecting and classifying road markings. Using segmentation to propose ROIs is another way to provide further contextual information [39, 40].

3 METHODS

To address the problem of using image recognition to identify road cracks without annotated images, we develop a context-aware stacked architecture. Our approach draws from experience from other domains, especially medical computer vision [6]. We hope that by introducing contextual information through the stacked architecture and ROI masks, we can overcome some of the problems of weakly labelled data.

While GRDDC results suggest that road damage detection can be effective on images with relatively low variation in high-level classes (roads), we hope our model will be effective in images with high variation in both high-level (road) and subordinate (cracking) classes, as seen in Figure 1.

Figure 1: Street-level Panorama from Amsterdam

3.1 ResNet

For all of the models trained, we use ResNet-34 as the backbone [27]. The ResNet architecture has a number of innovations which make it well suited to our problem, including batch normalisation and skip connections.

Batch normalisation works by reducing internal covariate shift (the effect of changing distributions of inputs between each layer) to decrease training times [41]. Batch normalisation is applied to activation functions in the network so that mini-batches are normalised and therefore there is lower variation for backpropagation, allowing higher learning rates. This in turn allows for deeper networks.

As well as passing the output of each layer to the next, ResNet uses "skip connections" to pass the identity of the previous input to later layers. This solves the problem of degradation in very deep networks because once degradation occurs the model can rely on the identity of earlier inputs. This means ResNet can include a very high number of layers without degradation occurring. He et al. [27] define a building block as:

$$y = \mathcal{F}(x, \{W_i\}) + x.$$

Where x and y are the input and output vectors of the layer considered.

ResNet is also well suited to our architecture because it is fully convolutional until the final layers. This means the model can easily be adapted for our stacked network and convolutional layers can be used to produce a dense prediction map for input images of any size.

3.2 Stacked CNN

The primary contribution of this paper is the consideration of stacked CNNs. This architecture is inspired by the work of Bejnordi et al. [6] and Wang et al. [7] and is used to create a context-aware network which can identify damage using both the local features of image segments and the global features of the panorama.

This architecture is chosen because road damage in street-level images is challenging for a single CNN to classify. Firstly, there is high variation in the high-level classes included (cars, people, vehicles, buildings etc) in each image. Secondly, there is high intra-class variation in subordinate classes. Cracks take many forms and vary by road surface type, position, direction, etc. Finally, there is low inter-class variation. Cracks may be visually similar to other objects, such as antennae, natural debris or pavement joints.

For these reasons, it is necessary to train a first CNN with known examples of road cracks before considering larger unannotated images. The stacked CNN is intended to learn the local features of road cracks, from high resolution image segments, and apply this across the larger images to learn global contextual features. By pre-training the first model, the convolutional layers can be applied across the larger image to produce a dense prediction map without calculating gradients. This means the stacked network can take much larger images without becoming computationally overburdened.

The stacked CNN is trained in two distinct stages. The structure is summarised in Figure 2.

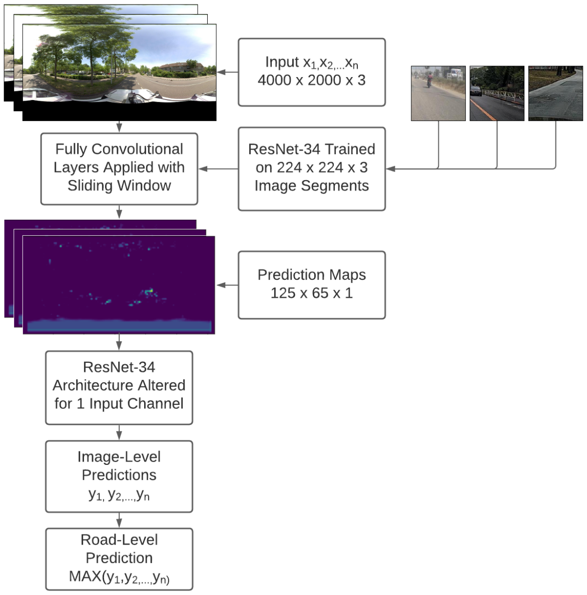
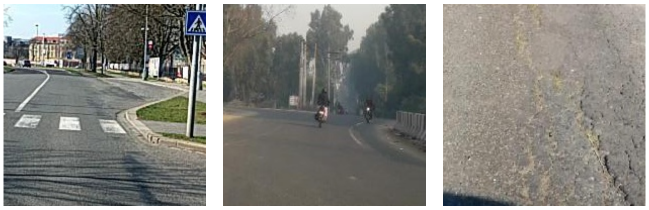

Input $x_1, x_2, ... x_n$
4000 x 2000 x 3
Fully Convolutional Layers Applied with Sliding Window
ResNet-34 Trained on 224 x 224 x 3 Image Segments
Prediction Maps 125 x 65 x 1
ResNet-34 Architecture Altered for 1 Input Channel
Image-Level Predictions $y_1, y_2, ..., y_n$
Road-Level Prediction MAX($y_1, y_2, ..., y_n$)

Figure 2: Stacked Network Structure

3.2.1 *Local Image Segments.* A ResNet-34 CNN is first trained on small image segments, taken from the GRDDC dataset [2, 10]. 16,832 200 × 200 image segments are cropped from the GRDDC set, either centred on the annotated road cracks, or from a random location in



(a) Czechia - 0

(b) India - 0

(c) Japan - 1

Figure 3: Examples of Image Segments with Target Classes

Layer	Segment Input 224 × 224 × 64	Panorama Input 4000 × 2000 × 64
conv1	112 × 112 × 3	2000 × 1000 × 3
conv2_x	56 × 56 × 64	1000 × 500 × 64
conv3_x	28 × 28 × 128	500 × 250 × 128
conv4_x	14 × 14 × 256	250 × 125 × 256
conv5_x	7 × 7 × 512	125 × 63 × 512
conv6	7 × 7 × 1	125 × 63 × 1
product	1 × 1	

Table 1: ResNet-34 Output by Layer [27]

images which do not include damage (Figure 3). Cropped segments are resized from 200 × 200 to 224 × 224 to fit the ResNet architecture and augmented using horizontal flip, random optical distortion, random brightness and contrast to create a training set of 50,496 images.

We replace the final, fully-connected layer of the ResNet CNN with a 2D convolution which takes 7 × 7 × 512 as an input and returns 7 × 7 × 1. Finally, a sigmoid activation is applied and the product is used to predict the image-level probability of road cracks. That is:

$$M = f(x)$$

$$y = 1 - \prod_{m \in M} (1 - m)$$

This method is applied, rather than taking the maximum, to reduce the confidence of the model when applied across larger images and to encourage the model to learn predictions of damage which are quasi-independent.

3.2.2 Global Image Features. The fully convolutional layers of the first CNN (trained on image segments) are then applied to panoramic images. Because the panoramas are much larger (2000 × 4000) than the image segments (224 × 224), this produces what is effectively a dense prediction map output (125 × 63, see Table 1 and Figure 4b).

A second CNN is then trained using dense prediction maps. The second CNN is intended to identify global features of the dense prediction maps, such as the location of predicted damage and the overall structure. For example, this might be recognising that particular patterns of detected cracks is more likely to be a pavement joint, or that cracks in certain areas are more likely to be the car's antennae.

The second CNN is another adjusted ResNet-34 architecture. The first 2D convolution of the model is replaced so that it takes 1 input channel rather than 3, and returns the same output (112 × 112 × 64). This allows the rest of the ResNet-34 structure to be retained. The dense prediction maps are resized to 224 × 225 before being passed to the second CNN.

3.3 Panoptic Segmentation

Another technique we experiment with to improve classification is panoptic segmentation for ROI filtering. This is an attempt to remove visual 'noise' which may impede classification, however it comes at the expense of potential loss of contextual information [39].

Panoptic segmentation combines the tasks of semantic segmentation and instance segmentation to give both pixel-wise semantic segmentation and instance detection. In 2019, Kirillov et al. [42] introduced Panoptic Feature Pyramid Networks using ResNet-101 as a backbone (R101-FPN). R101-FPN uses the same Feature Pyramid Network for both instance segmentation and semantic segmentation, scoring a mask average precision of 38.5% on the COCO test set, a very high score despite being a relatively simple and fast model [34].

We apply the pre-trained R101-FPN model on our training, validation and test data. Visual assessment of a sample of images suggests the model performs moderately well on panoramic images of Amsterdam and the GRDDC dataset, although in a small number of cases (around 2% of panoramas and 1% of GRDDC images) no road surface is identified. This is usually either because there is indeed no road surface, due to an error in data collection, or because the road is mislabelled as 'dirt' or 'pavement'.

Once road pixels have been identified, they are converted into a boolean mask. This is multiplied by the original image to filter all non-road surface pixels from the image (as can be seen in Figure 4c).

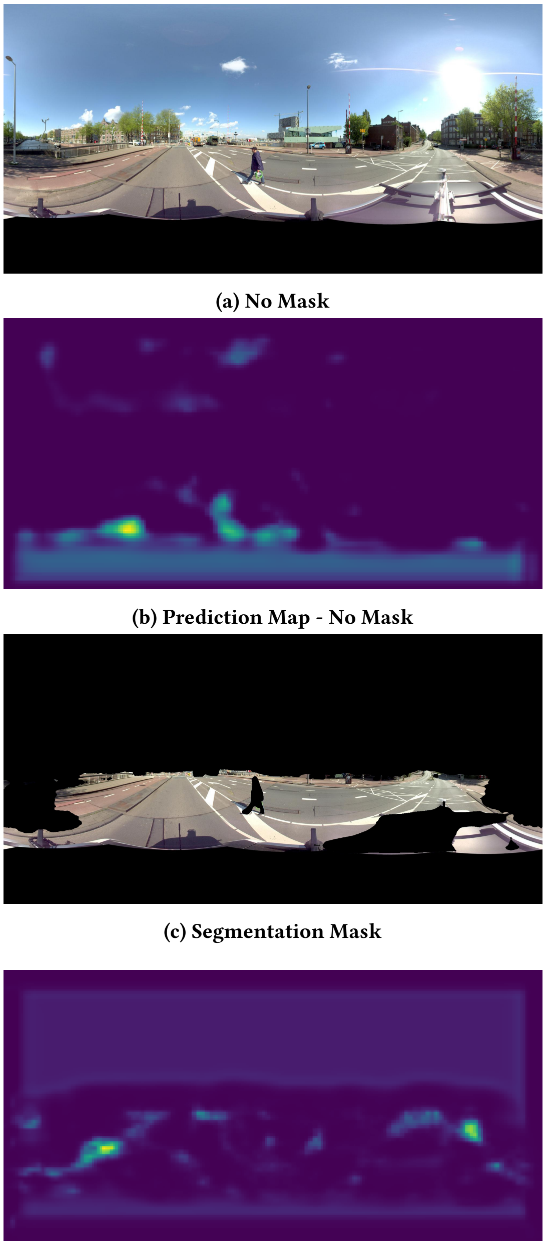
4 EXPERIMENTAL SETUP

To consider the use of these methods on real world data we address the problem of road crack detection in the City of Amsterdam. This means using weakly labelled data based on manual inspections by municipality employees. We also compare with annotated data from the GRDDC [10].

4.1 Data Collection

4.1.1 Manual Inspections. The source of road damage data in this study is manual inspections conducted by the Amsterdam Roads and Mobility Department. These manual inspections are undertaken by domain experts along guidelines set out by CROW [43]. Data is provided in the form of inspection-level classes linked to road-section shapefiles. The data includes 23,919 inspections undertaken on 11,112 separate stretches of asphalt road between January 2014 and February 2020.

Inspection-level data includes a unique road-section identifier, the date of inspection, inspection notes and classes for 8 different



(a) No Mask

(b) Prediction Map - No Mask

(c) Segmentation Mask

(d) Prediction Map - Segmentation Mask

Figure 4: Example of ROI Mask Used to Remove Non-Road Pixels and Corresponding Prediction Maps

forms of road damage: transverse unevenness, unevenness, fraying, edge damage, cracking, setting, joint filling and joint width. Of these, we focus on the 'cracking' class because it is the most visually distinctive and can be directly compared with the cracking class used in GRDDC data. In future we hope our methods will be generalisable to multi-class problems.

Road section data includes corresponding unique road-section identifiers, which link to inspections, as well as information on road surface, colour, construction year, surface area, and road usage. The coverage of inspection road sections is mapped in Appendix B.

4.1.2 Panoramic Images. Panoramic images of Amsterdam streets are available from the City of Amsterdam [12]. Images of each street are taken approximately annually at intervals of a few metres and are available with metadata including the latitude and longitude, timestamp, surface type, roll, pitch and heading. The images offer a full 360° view of the street scene, as shown in Figure 1.

4.1.3 Combined Dataset. For each manual inspection, images were requested from the Amsterdam panorama API taken before the inspection within the road-section.

To improve the class balance of the dataset, we oversampled cracked roads. For each inspection with no cracking, we requested images from within 30 days before the inspection, within 100m of the centre of the road-section, and from the first 100 API results. For inspections of cracked roads, we requested images within 60 days before the inspection, 500m from the centre of the road-section, and from the first 200 API results. For each result of this query, images with a latitude and longitude within the target road shapefile were added to the dataset. While cracked roads made up 16.4% of inspections, 28.7% of roads in the training set are cracked and 35.0% of images were taken from cracked roads.

This resulted in a combined dataset of 36,447 images corresponding to 2,117 inspections. Multiple images are available for each inspection (see Appendix A) meaning that coverage of the damage detected by the manual inspection is more likely. However, this means the data is noisy, with many images of damaged roads including no visual evidence of damage. The 36,447 image locations are visualised in Figure 5.

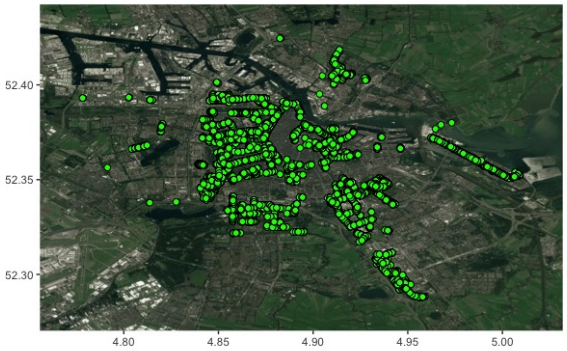


Figure 5: Locations of 36,447 panoramic images of Amsterdam roads, overlaid on satellite images from Google Maps [44]

4.1.4 International Road Damage Dataset. Alongside panoramas of Amsterdam, we utilise images of road surfaces collected from Czechia, India and Japan as part of the Global Road Damage Detection Challenge (GRDDC) [2, 10, 45]. Labels are provided in the form of bounding box coordinates and labels for instances of different types of damage in each image. The dataset includes 26,620 square images, most of which are 600 × 600 pixels.

4.2 Model Selection

In order to train and evaluate crack detection models, the datasets are split randomly into train, validation and blind test sets, of size 80%, 10%, 10% respectively. For the Amsterdam data we used grouped cross-validation, splitting the data by road so that no stretch of road appears in two different sets. This is because there may be group dependencies by road, misleading the classifier and resulting in unrepresentative evaluation scores. All three of the sets are imbalanced, with 28-33% of roads featuring cracks (Appendix C).

All of the models are trained using Binary Cross-Entropy (BCE) Loss as a pseudo evaluation metric. For image segment models, forming the first part of the stacked CNNs, models are selected by their validation BCE Loss. This is to ensure that the best calibrated outputs are produced and passed to the second CNN.

For all other models, selection is done by image-level F1 score on the validation set.

4.3 Road-Level Aggregation

Because the panoramic data is weakly labelled, with multiple panoramas for each road section considered, damage is unlikely to be evenly distributed across the panoramas. At the training stage, we simply propagate the road-level labels onto the images. However, in evaluation it is more relevant to consider road-level predictions, which can be compared to the ground truth.

In manual inspections, a road section is marked as ’cracked’ if cracking appears at any point within that section. To emulate this process, we use the ’MAX’ voting method - taking the maximum predicted probability of road cracks of all images per road. If a crack is present in one image, the section is predicted to include at least one crack, so the whole section is labelled as cracked. Although there is a risk of this voting method being too sensitive - leading to false positives - it is the choice which best reflects data collection.

Learned voting methods are beyond the scope of this paper. Of the fixed voting methods, mean, median and majority all contain an implicit assumption that instances of a class are evenly distributed across inputs, which is not always the case. The product rule is less effective in the presence of correlated errors, which are likely in our case [46]. In practice, this distinction makes little difference, with all of the fixed voting methods achieving similar scores when using a scale-invariant metric during validation.

4.4 Evaluation

4.4.1 F1-Score. We use F1-score for road-level and image-level evaluation because it incorporates both the precision and recall of predictions. This is useful because there is often a trade-off between the two. The F1-score is designed to include both, such that a model with high recall and high precision scores highest. The formula for F1-score is:

$$F_1 = 2 * \frac{(precision * recall)}{(precision + recall)}$$

Where:

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Precision = \frac{TruePositive}{TruePositive + FalseNegative}$$

4.4.2 ROC Curve and AUC. In addition to F1 scores, receiver operating characteristic (ROC) curves can allow us to consider the effectiveness of classifiers at different classification thresholds. This is useful in our case, particularly with road-level predictions, because while probabilities may be well calibrated at the image-level, they are unlikely to be well-calibrated after MAX voting has taken place at the road-level. When paired with AUC (area under the curve) scores, the ROC curve can be used to consider classifiers in a scale and threshold invariant way, meaning calibration of scores is not necessary.

4.5 Baselines

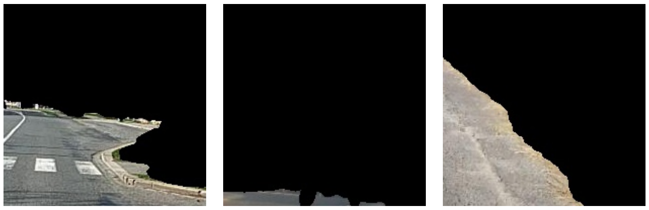
For comparability, we train a baseline model by which we can measure improvements. This baseline model uses the ResNet-34 architecture and is initialised with pre-trained weights from training on the ImageNet dataset [25, 27]. The final layer of this CNN is a fully connected layer with 512 inputs and 1 output, followed by a sigmoid transformation.

4.5.1 State of the Art Comparison. We also compare our results with completely pre-trained models from the Global Road Damage Detection Competition [2].

We applied IMSC’s winning model to the Amsterdam test set. This model takes an ensemble learning approach relying on Ultralytics-YOLOv5 as a backbone [35]. Panoramas were resized to 640 × 640 and we used the model with a threshold of 0.22 (the optimal threshold on GRDDC data) to create road-level predictions, taking the prediction of cracks in any image as a road-level prediction of cracking.

For the GRDDC test set, we used the IMSC architecture again, but this time trained on our GRDDC training set, to avoid testing on already-seen data. In this case, the model did not use ensemble predictions but was a single network initialised with YOLO-v5 weights and trained with batch size 16 on 640 × 640 images with default hyper-parameters. Images were classified at the image-level based on whether damage was detected with threshold 0.22.

4.5.2 Pre-trained Model. To evaluate the importance of weight initialisation and pre-training, we also train models with Amsterdam training data which have been pre-trained on GRDDC data. We take the same ResNet-34 architecture used for the baseline, this time trained on known instances of road damage from the GRDDC set, and freeze the first 20 parameters. All remaining parameters are adjusted for Amsterdam data. This approach is consistent with the use of transfer learning from across a broad range of computer vision problems, including road damage detection with international transfer learning [45, 47, 48].



(a) Czechia - 0

(b) India - 0

(c) Japan - 1

Figure 6: Examples of Image Segments (from Figure 3) with Target Classes After ROI Mask Applied

4.6 Experiments

To consider the performance of the methods discussed above, we experiment by training and testing 12 models. This includes baseline, pre-trained, state-of-the-art and stacked architectures, with and without ROI masks.

Five models are tested on GRDDC data, for which image-level predictions are used to calculate evaluation metrics (accuracy, recall, precision and F1) with a threshold of 0.5.

Seven models are tested on Amsterdam data, with road-level predictions calculated by the MAX image-level prediction per road (see 4.3). These predictions are used to calculate accuracy, precision, recall and F1 scores with a threshold of 0.5, as well as AUC which is scale and threshold invariant (see 4.4.2).

5 RESULTS

Results for all of the models tested on blind test sets are set out in Tables 3 and 4.

5.1 Image Segment Models

For the two stacked networks, two models are trained using image segments from the GRDDC annotated data. These models both use ResNet-34 architectures, which have been slightly adapted so that the fully-connected final layer is replaced with a 2D convolution, meaning the model is fully convolutional. The results are reported in Table 2.

The first is trained using segments of GRDDC images (Figure 3), either from a random crop or centred on an annotated example of a road crack. The second is trained using segments of GRDDC images which have already had an ROI mask applied to exclude non-road surface pixels (Figure 6).

5.2 Amsterdam Test Set

The Amsterdam test set proved challenging for all of the models (Table 3). No model had a higher F1 score than 0.510 or AUC than 0.665, indicating that they were not much of an improvement over a random prediction.

To a certain extent, low F1 scores are to be expected due to to the combination of weak labels and aggregation. The models are trained on data in which all images on a cracked road are labelled as cracked meaning the model is likely to be oversensitive. Then, in aggregation, the maximum prediction is taken per road, so the poorly calibrated image-level predictions result in a high number of false positives. This explains why all of the experiments result in much higher recall scores than precision - the models are too sensitive. For this reason, a scale invariant measure such as AUC (see 5.2.1) is more appropriate.

Similarly, accuracy scores for all of the models are low. This is partly because of the same sensitivity issue outlined above, and partly because of the imbalance of the test set (Appendix C). With an imbalanced set, a high accuracy score could be achieved simply by assigning the majority class. In our case, this is a poor measure of model efficacy.

The best performing model on the Amsterdam test set by AUC was the Stacked Network applied with an ROI mask. The next best performing model was the ResNet-34 pre-trained on GRDDC data, which outperformed the stacked network without a mask.

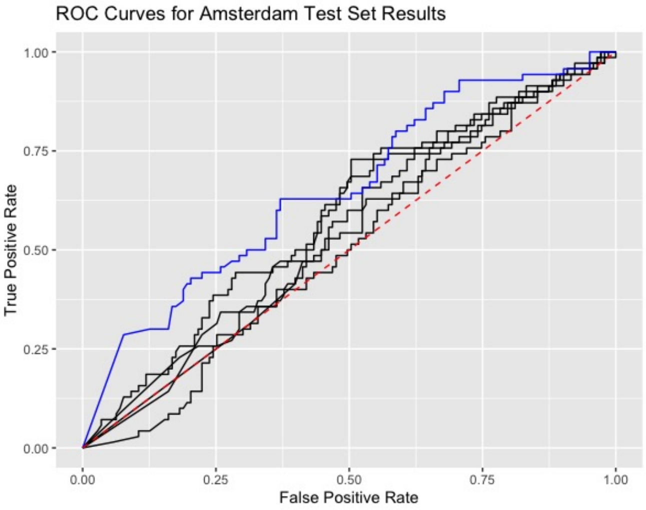


Figure 7: ROC Curves for Amsterdam Test Set

5.2.1 *ROC Curves.*

On the ROC curve plot we can see the best performing model (Stacked Network with ROI Mask) plotted in blue, compared to the other four models in black as well as a 'no skill' curve plotted as a dashed red line. Here we can see how all of the models have limited ability to discriminate between classes but that the Stacked Model with ROI mask performs best. The other four models show very limited improvement over the 'no skill' line.

There are many reasons why this might be the case, discussed in Section 6.

5.3 GRDDC Test Set

On the GRDDC test set, the best model was the Baseline ResNet-34 with ROI mask (Table 4). This model achieved an F1-Score of 0.59 and an accuracy of 74.3%. Both of the Baseline ResNet-34 models performed significantly better than the Stacked Network, which had very limited scores with or without ROI masks. All of the models had higher recall than precision, suggesting the models might be slightly over-confident of positive cases.

As expected, on the GRDDC test set the IMSC Ultralytics-YOLOv5 model performed very well, with an accuracy of 81.5% and F1-score of 0.718. This reflects the model's strengths on annotated data with

Model	ROI Mask	Accuracy	F1-Score	BCE Loss
Fully Convolutional ResNet-34	No	91.7	0.859	0.236
Fully Convolutional ResNet-34	Yes	90.7	0.834	0.294

Table 2: Stacked CNN First Networks - Trained on Image Segments

Model	ROI Mask	Accuracy	Precision	Recall	F1-Score	AUC
Baseline ResNet-34	No	38.0	0.339	0.929	0.496	0.553
Baseline ResNet-34	Yes	36.6	0.354	0.914	0.510	0.595
Pre-Trained ResNet-34	No	39.4	0.324	0.914	0.498	0.599
Pre-Trained ResNet-34	Yes	38.0	0.337	0.914	0.492	0.508
Stacked Network	No	39.4	0.342	0.900	0.498	0.586
Stacked Network	Yes	38.5	0.342	0.943	0.502	0.665
IMSC Ultralytics-YOLOv5	No	42.6	0.329	0.727	0.452	

Table 3: Amsterdam Blind Test Set Results

Model	ROI Mask	Accuracy	Precision	Recall	F1-Score
Baseline ResNet-34	No	73.8	0.534	0.601	0.565
Baseline ResNet-34	Yes	74.3	0.568	0.622	0.594
Stacked Network	No	48.6	0.328	0.772	0.460
Stacked Network	Yes	61.5	0.382	0.581	0.461
IMSC Ultralytics-YOLOv5	No	81.5	0.632	0.831	0.718

Table 4: GRDDC Test Set Results

less variation in high-level content. The scores listed here are not directly comparable with GRDDC scores because this is a binary classification problem rather than multi-class object detection [2]. To evaluate the IMSC model as a binary classifier, the inferred results are translated so that an image is classified as cracked if a single crack is detected within the image at a threshold of 0.22 (which the IMSC team give as the optimal threshold).

6 DISCUSSION

On the Amsterdam test set, we saw limited results from all of the models tested. This partially reflects the difficulty of using weakly labelled data and potentially the quality of data collected, but experiments from the GRDDC set suggest limits to the methods independent from the Amsterdam problem.

6.1 IMSC Ultralytics-YOLOv5

The IMSC Ultralytics-YOLOv5 model [35], winner of the GRDDC 2020 [2], performed poorly on Amsterdam panoramas. This is to be expected for a number of reasons so is not necessarily a reflection on the IMSC model, but demonstrates the need for fine-tuning when transferring road damage detection models to other locations.

One reason that this is expected is that the panorama data is significantly different from GRDDC training data. To use the IMSC model, we resized panoramas from 2000×4000 to 640×640, resulting in some distortion. The panoramas also include a much larger visual field than GRDDC training data, so road damage might be too small to detect. In general, road surfaces take up a much higher proportion of GRDDC images than the Amsterdam panoramas, making directly transfer difficult.

Another reason is that the 0.22 threshold for detection is optimised for GRDDC data. With road-level detection, this is unlikely to be the optimal threshold. It was not possible to obtain image-level probabilities from the IMSC model, so it was not possible to use a scale invariant metric such as AUC.

These issues aside, however, the results of the IMSC model show the importance of fine-tuning when transferring computer vision learning. For example, inspection of outputted bounding boxes suggests that the IMSC model routinely incorrectly identified car antennae on the camera-mounted vehicle as road cracks. In Figure 8 we plot heatmaps of outputted bounding boxes for false positive versus true positive classifications. Here we can clearly see the cluster of false positive bounding boxes in the narrow band where antennae frequently appear.

However, without annotated images of Amsterdam, it is not possible to simply fine-tune the same architecture. Using the same architecture trained with only road-level labels is also unlikely to provide satisfactory results, given the fine granularity of road cracks and the small proportion that road surfaces form in each panorama. It is for this reason that we pursued solutions which involved reducing visual "noise" and introducing contextual awareness.

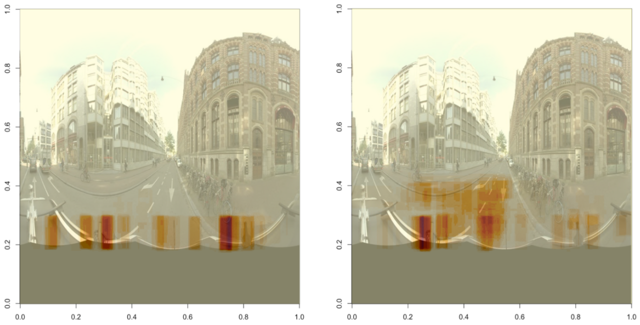


Figure 8: Heatmap of Bounding Boxes for False (left) vs True (right) Positive Road Crack Predicted Classifications, Overlaid on an Example Panorama for Reference

6.2 Stacked Network

In our results, the performance of stacked networks is limited, both in the Amsterdam and GRDDC test sets (Tables 3 and 4 respectively). In the Amsterdam test set, stacked models yielded a moderate improvement in AUC compared with the baseline models, while in the GRDDC set it resulted in a dramatic fall in F1.

This is likely because the first network (trained on image segments) is less effective than expected and provides too little information to the second network. To some extent, this is intentional. With weakly labelled data such as the Amsterdam test set, removing extraneous visual information may work as a means to focusing the second network only on relevant sections of the image, using the annotated image segments to bypass over-interpreting poorly labelled data. However, in the GRDDC set it seems that this removed too much information for the classifier. There are a number of reasons why this might be the case.

Firstly, the image segment network is based on randomly cropped GRDDC images (see Figure 3). It may be that the image segment model scores (F1 = 0.86 for whole image segments, F1 = 0.83 for ROI image segments) are artificially inflated by the fact that the training set includes non-road surface segments alongside undamaged and damaged road surface segments. If the image segment model can discriminate between road surface and non-road surface, it would be able to achieve high scores without effectively assessing road surface for damage.

Indeed, in the architecture proposed by Bejnordi et al. [37], the image segment network is trained with segments labelled both according to high-level classes (i.e. whether the cells are from a tumour or, in our case, whether the segment features a road) as well as subordinate classes (i.e. whether the tumour is malignant, or whether the road is cracked). If the image segment model passed a dense prediction map consisting of three classes (cracked road, non-cracked road, non-road) it might perform better than simple probability of road cracks.

Another potential drawback of our approach is that the image segments model is trained on segments for which the size is arbitrarily selected (200 × 200 within the 600 × 600 whole images). This may not be the most effective segment size, and as Bejnordi et al. [37] themselves point out, it may be preferable to have a learned size. This is particularly difficult to discern when applying the convolutional patch learned from one dataset to another with different dimensions.

6.3 Region-of-Interest Masks

In both the Amsterdam and GRDDC data, ROI masks tended to yield a small improvement in model performance. This is as we expected, with evidence of ROI mask effectiveness from many domains [39, 40, 49, 50]. The ROI mask can remove extraneous visual features which might affect classification. In our case, it also provides new information explicitly on the structure of the road, but also implicitly on the quality of the image.

One potential drawback of our approach is the total removal of non-road surface pixels. This means we are relying entirely on the quality of the segmentation model, which is not always assured, as well as removing potential contextual information which may help explain the presence or absence of road cracks.

As Eppel [39] argues, there may be merit in highlighting rather isolating the relevant pixels so that contextual information is not lost. They show how in some cases adding rather than multiplying the binary ROI mask, so that relevant pixels appear more visually striking, may allow the model to treat the ROI mask as a recommendation without losing contextual features. In future, we could also experiment with including the ROI mask either as an additional channel or applying it in a higher layer.

6.4 Road-Level Aggregation

Another challenge, which is reflected in the divergent scores between Amsterdam and GRDDC data, is road-level aggregation due to weak labels and sequence inputs in the Amsterdam data. Our approach was to propagate road-level classes onto all images for that road, but this poses some obvious problems. It may be that the model is effective at learning features of roads which are likely to be damaged, without learning anything about the cracks which are present in a minority of positively labelled images.

One approach to tackle this would be some form of candidate selection for training or evaluation. In some ways, the Amsterdam data is comparable to video data, in that it includes individual frames taken at intervals travelling along a road. Given the sparsity of image data on some roads, it has not been possible to utilise a video-like approach, but given information about heading and geo-location is available, it is theoretically possible to combine panoramas into a road-level video.

If this were the case, we could draw from approaches such as that taken by Prest et al. [8], which focus on candidate selection within weakly labelled videos. Prest et al. [8] identify spatio-temporal tubes which are then used as the input for a classifier. Similarly, Karpathy et al. [9] use a multi-resolution, spatio-temporal approach by having foveated streams of information from input frames. With these methods, road damage could be located and assessed across multiple frames. However, it is likely that these approaches would require a much more extensive dataset than ours from Amsterdam, which has only 2,117 inspections as ground truth.

Candidate selection would also sidestep the problem of voting rules. Given our results, we still consider the MAX voting method the most appropriate as it mirrors manual inspections. However, its main drawback is that it can be arbitrary because it is not calibrated to the ground truth. A learned voting method might be able to improve the calibration of image-level predictions as well as comparability with the ground truth [46].

6.5 Semi-Supervised Techniques

Given the limits of learning using weak supervision, another approach could be to introduce some forms of supervision without the need for a fully annotated training set. For example, Chun and Ryu [51] train a segmentation model to segment road damage from 6,756 manually annotated images before applying this model to a much larger dataset of unlabelled images to produce a 68,567 images with pseudo-labels.

In our case, a similar approach could be to use the classifiers we have trained to re-label a certain number of images which may be mislabelled by weak supervision. For cracked roads with many images (see Appendix A), it is likely the cracks do not appear in every image. A new training set could be created by re-labelling images under a certain threshold as having no cracks.

Another approach would be introducing a human-in-the-loop [52, 53]. In our case, a system could be developed to allow a domain expert to contribute to labelling as the network is trained, for example being fed the images that the model is most confident about for manual inspection and re-labelling. This would greatly reduce amount of manual annotations required while introducing expert interactivity.

7 CONCLUSIONS

Returning to our research question, **"Can context-aware stacked neural networks effectively classify weakly labelled images?"**, we have provided mixed evidence for the use of stacked networks. We employed an experimental approach, applying a variety of techniques to both annotated GRDDC data and weakly labelled panoramas from Amsterdam. The results varied widely across the two sets, showing the importance of altering hyper-parameters for different domains.

On the Amsterdam test set, the most effective model was the stacked network with ROI masks. This shows the potential for the technique for problems involving weakly labelled data. However, the test set results were weak and were not reflected in the GRDDC set, suggesting limited effectiveness in general. As set out in our discussion, this does not discount the use of stacked networks in future. The use of techniques such as learned segment sizes, improved image segment models and multi-class prediction maps may improve the effectiveness of stacked networks.

We also found more conclusive evidence to support the use of ROI masks. Four of the five models which used ROI masks outperformed their respective counterparts. With continually improving panoptic and semantic segmentation models, ROI masks can provide a way of reducing visual noise and improving classification models.

Our results also show the importance of annotated data. In our case, the models struggled to learn from images labelled at the road-level. With high levels of variation within the panoramas, image-level or more granular annotations would be very helpful. This could potentially be achieved with reduced workload by using human-in-the-loop techniques, combining expert manual recommendations with machine learning.

Altogether, this research shows the possibilities for refining and improving stacked networks for fine-grained image classification. Whether in road damage detection or related fields, these methods have the potential to make a significant contribution to image recognition and wider society.

REFERENCES

[1] Navdeep Kumar, Nirmal Kaur, and Deepti Gupta. Major Convolutional Neural Networks in Image Classification: A Survey. *Lecture Notes in Networks and Systems*, 116:243–258, 2020. doi: 10.1007/978-981-15-3020-3{_}23. URL https://link.springer.com/chapter/10.1007/978-981-15-3020-3_23.

[2] Deeksha Arya, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, Hiroshi Omata, Takehiro Kashiyama, and Yoshihide Sekimoto. Global Road Damage Detection: State-of-the-art Solutions. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 12 2020. ISBN 978-1-7281-6251-5. doi: 10.1109/BigData50022.2020.9377790.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:779–788, 12 2016. doi: 10.1109/CVPR.2016.91.

[4] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing 2017 14:2*, 14(2):119–135, 1 2017. ISSN 1751-8520. doi: 10.1007/S11633-017-1053-3. URL https://link.springer.com/article/10.1007/s11633-017-1053-3.

[5] Xiu-Shen Wei, Jianxin Wu, and Quan Cui. Deep Learning for Fine-Grained Image Analysis: A Survey, 7 2019. URL https://arxiv.org/abs/1907.03069v1.

[6] Babak Ehteshami Bejnordi, Guido Zuidhof, Maschenka Balkenhol, Meyke Hermsen, Peter Bult, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, and Jeroen van der Laak. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *Journal of Medical Imaging*, 4(04):1, 12 2017. ISSN 2329-4302. doi: 10.1117/1.jmi.4.4.044504. URL https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-4/issue-4/044504/Context-aware-stacked-convolutional-neural-networks-for-classification-of-breast/10.1117/1.JMI.4.4.044504.fullhttps://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-4/issue-4/044504/Context-aware-stacked-convolutional-neural-networks-for-classification-of-breast/10.1117/1.JMI.4.4.044504.short.

[7] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple Granularity Descriptors for Fine-Grained Categorization. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 12 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.276.

[8] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3282–3289. Institute of Electrical and Electronics Engineers (IEEE), 6 2012. ISBN 9781467312264. doi: 10.1109/CVPR.2012.6248065. URL https://www.research.ed.ac.uk/en/publications/learning-object-class-detectors-from-weakly-annotated-video.

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei Fei Li. Large-scale video classification with convolutional neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 9 2014. doi: 10.1109/CVPR.2014.223.

[10] Deeksha Arya, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, and Yoshihide Sekimoto. RDD2020: An annotated image dataset for automatic road damage detection using deep learning. *Data in Brief*, 36:107133, 6 2021. ISSN 2352-3409. doi: 10.1016/J.DIB.2021.107133.

[11] Lengte van wegen; wegkenmerken, regio, 12 2020. URL https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70806ned/table?ts=1627223935877.

[12] Gemeente Amsterdam. Panoramabeelden, 2021. URL https://data.amsterdam.nl/data/panorama/TMX7316010203-001187_pano_0000_001517/.

[13] Dave Steinkraus, Ian Buck, and Patrice Y. Simard. Using GPUs for machine learning algorithms. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, volume 2005, pages 1115–1120, 2005. ISBN 0769524206. doi: 10.1109/ICDAR.2005.251.

[14] Ahmed Ali Mohammed Al-Saffar, Hai Tao, and Mohammed Ahmed Talab. Review of deep convolution neural network in image classification. In *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pages 26–31. IEEE, 2017. ISBN 9781538638491.

[15] Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks. In *ICLR2014*, 2013.

[16] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes Paris look like Paris? *Communications of the ACM*, 58(12):103–110, 2015. ISSN 00010782.

[17] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences - PNAS*, 114(50):13108–13113, 2017. ISSN 0027-8424.

[18] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing (Amsterdam)*, 338:321–348, 2019. ISSN 0925-2312.

[19] Weichao Xu, Baojun Li, Sun Liu, and Wei Qiu. Real-time object detection and semantic segmentation for autonomous driving. In Jayaram K. Udupa, Hanyu Hong, and Jianguo Liu, editors, *MIPPR 2017: Automatic Target Recognition and Navigation*. SPIE, 2 2018. ISBN 9781510617193. doi: 10.1117/12.2288713.

[20] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2 2019. doi: 10.1109/TITS.2020.2972974. URL http://arxiv.org/abs/1902.07830http://dx.doi.org/10.1109/TITS.2020.2972974.

[21] K Fukushima. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980. ISSN 0340-1200.

[22] Yann Lecun and Yoshua Bengio. Convolutional networks for images, speech, and time-series. In M A Arbib, editor, *The handbook of brain theory and neural networks*. MIT Press, 1995.

[23] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017. ISSN 0162-8828.

[24] Y LeCun, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, and L D Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural computation*, 1(4):541–551, 1989. ISSN 1530-888X.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 5 2017. ISSN 0001-0782. doi: 10.1145/3065386.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 12 2015. ISSN 15731405. doi: 10.1007/s11263-015-0816-y. URL https://link.springer.com/article/10.1007/s11263-015-0816-y.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 770–778. IEEE Computer Society, 12 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90.

[28] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1 2020. URL https://arxiv.org/abs/2001.05566v5.

[29] Wei Liu, Andrew Rabinovich, and Alexander C Berg. ParseNet: Looking Wider to See Better. *CoRR*, 2015.

[30] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 4 2018. doi: 10.1109/TPAMI.2017.2699184.

[31] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning Deconvolution Network for Semantic Segmentation. *CoRR*, abs/1505.04366, 2015. URL http://arxiv.org/abs/1505.04366.

[32] Michael Treml, José Arjona-Medina, Thomas Unterthiner, Rupesh Durgesh, Felix Friedmann, Peter Schuberth, Andreas Mayr, Martin Heusel, Markus Hofmarcher, Michael Widrich, Bernhard Nessler, and Sepp Hochreiter. Speeding up Semantic Segmentation for Autonomous Driving. In *29th Conference on Neural Information Processing Systems*, Barcelona, 2016.

[33] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. doi: 10.1109/CVPR.2016.350.

[34] Tsung-Yi Lin, Michael Maire, Serge J Belongie, Lubomir D Bourdev, Ross B Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

[35] Vinuta Hegde, Dweep Trivedi, Abdullah Alfarrajeh, Aditi Deepak, Seon Ho Kim, and Cyrus Shahabi. Yet Another Deep Learning Approach for Road Damage Detection using Ensemble Learning. In *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, pages 5553–5558. Institute of Electrical and Electronics Engineers Inc., 12 2020. ISBN 9781728162515. doi: 10.1109/BigData50022.2020.9377833.

[36] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv*, 4 2020. URL http://arxiv.org/abs/2004.10934.

[37] Babak E. Bejnordi, Geert Litjens, Meyke Hermsen, Nico Karssemeijer, and Jeroen A. W. M. van der Laak. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *Medical Imaging 2015: Digital Pathology*, volume 9420, page 94200H. SPIE, 3 2015. ISBN 9781628415100. doi: 10.1117/12.2081768. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9420/94200H/A-multi-scale-superpixel-classification-approach-to-the-detection-of/10.1117/12.2081768.fullhttps://www.spiedigitallibrary.org/conference-proceedings-of-spie/9420/94200H/A-multi-scale-superpixel-classification-approach-to-the-detection-of/10.1117/12.2081768.short.

[38] Cheng Li, Ivo M. Creusen, Lykele Hazelhoff, and Peter H. N. de With. Detection and recognition of road markings in panoramic images. *Video Surveillance and Transportation Imaging Applications 2015*, 9407:940708, 3 2015. doi: 10.1117/12.2081395. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9407/940708/Detection-and-recognition-of-road-markings-in-panoramic-images/10.1117/12.2081395.fullhttps://www.spiedigitallibrary.org/conference-proceedings-of-spie/9407/940708/Detection-and-recognition-of-road-markings-in-panoramic-images/10.1117/12.2081395.short.

[39] Sagi Eppel. Classifying a specific image region using convolutional nets with an ROI mask as input, 2018. URL https://arxiv.org/pdf/1812.00291.pdf.

[40] Marcus Wallenberg and Per Erik Forssen. Attentional masking for pre-trained deep networks. *IEEE International Conference on Intelligent Robots and Systems*, 2017-September:6149–6154, 12 2017. doi: 10.1109/IROS.2017.8206516.

[41] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

[42] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:6392–6401, 1 2019. URL https://arxiv.org/abs/1901.02446v2.

[43] Handleiding globale visuele inspectie, 12 2011. URL https://www.crow.nl/publicaties/handleiding-globale-visuele-inspectie-2011.

[44] David Kahle Wickham and Hadley. ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013.

[45] Deeksha Arya, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, Alexander Mraz, Takehiro Kashiyama, and Yoshihide Sekimoto. Transfer learning-based road damage detection for multiple countries. *arXiv preprint arXiv:2008.13101*, 2020.

[46] Robert P.W. Duin and David M.J. Tax. Experiments with classifier combining rules. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1857:16–29, 2000. ISSN 16113349. doi: 10.1007/3-540-45014-9{_}2. URL https://link.springer.com/chapter/10.1007/3-540-45014-9_2.

[47] Pengcheng Wu and Thomas G. Dietterich. Improving SVM accuracy by training on auxiliary data sources. In *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pages 871–878, 2004. ISBN 1581138385. doi: 10.1145/1015330.1015436.

[48] Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. Heterogeneous Transfer Learning for Image Classification. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI'11, page 1304–1309. AAAI Press, 2011.

[49] Sagi Eppel. Setting an attention region for convolutional neural networks using region selective features, for recognition of materials within glass vessels, 2017.

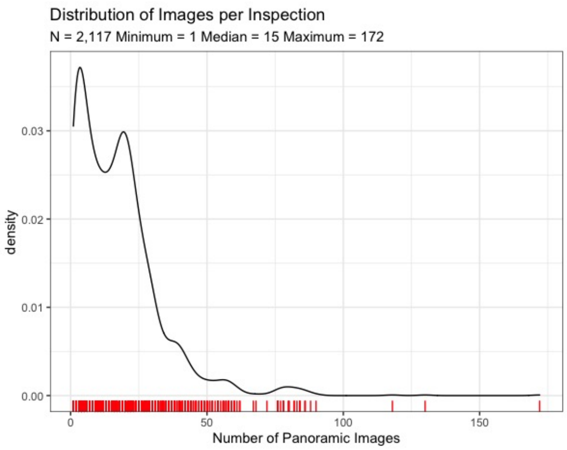
[50] Kan Chen, Jiang Wang, Haoyuan Gao, and Wei Xu. ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering, 2016.

[51] Chanjun Chun and Seung Ki Ryu. Road surface damage detection using fully convolutional neural networks and semi-supervised learning. *Sensors (Switzerland)*, 19(24), 12 2019. ISSN 14248220. doi: 10.3390/s19245501. URL /pmc/articles/PMC6961057//pmc/articles/PMC6961057/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6961057/.

[52] O S Khan, B Þ Jónsson, S Rudinac, J Zahálka, H Ragnarsdóttir, Þ Þorleiksdóttir, G Þ Guðmundsson, L Amsaleg, M Worring, J M Jose, E Yilmaz, J Magalhães, P Castells, N Ferro, M J Silva, and F Martins. Interactive Learning for Multimedia at Large. *Lecture notes in computer science*, 1:495–510, 2020. ISSN 0302-9743.

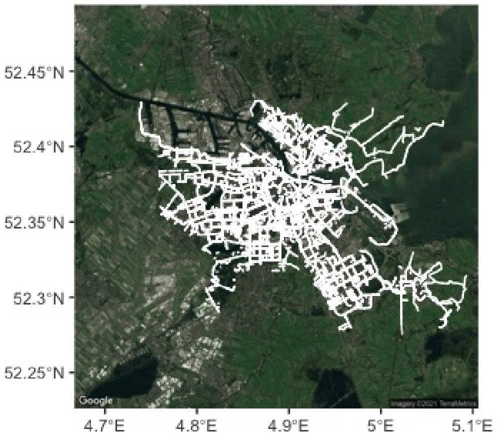
[53] Jan Zahalka, Stevan Rudinac, Bjorn Tor Jonsson, Dennis C Koelma, and Marcel Worring. Blackthorn: Large-Scale Interactive Multimodal Learning. *IEEE transactions on multimedia*, 20(3):687–698, 2018. ISSN 1520-9210.

A IMAGES PER INSPECTION



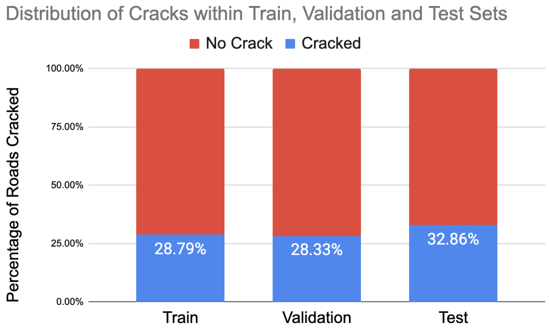
Distribution of Images per Inspection

B AMSTERDAM ROAD SECTIONS



Amsterdam road sections covered by inspection data, overlaid on satellite images from Google Maps [44]

C CLASS DISTRIBUTION



Distribution of Cracks by Train, Validation and Test Sets