INFRASTRUCTURE DEGRADATION AND POST-DISASTER DAMAGE DETECTION USING ANOMALY DETECTING GENERATIVE ADVERSARIAL NETWORKS

S. M. Tilon [1,*], F. Nex[1], D. Duarte[2,3], N. Kerle[1], G. Vosselman[1]

[1] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands - (s.m.tilon, f.nex, n.kerle, george.vosselman)@utwente.nl
[2] INESC-Coimbra, Institute for Systems Engineering and Computers at Coimbra, University of Coimbra, Coimbra, Portugal
[3]Department of Mathematics, University of Coimbra, Coimbra, Portugal - diogoavaduarte@mat.uc.pt

Commission II, WG II/4

KEY WORDS: Generative Adversarial Networks, anomaly detection, degradation, damage, infrastructure monitoring, post-disaster.

ABSTRACT:

Degradation and damage detection provides essential information to maintenance workers in routine monitoring and to first responders in post-disaster scenarios. Despite advance in Earth Observation (EO), image analysis and deep learning techniques, the quality and quantity of training data for deep learning is still limited. As a result, no robust method has been found yet that can transfer and generalize well over a variety of geographic locations and typologies of damages. Since damages can be seen as anomalies, occurring sparingly over time and space, we propose to use an anomaly detecting Generative Adversarial Network (GAN) to detect damages. The main advantages of using GANs are that only healthy unannotated images are needed, and that a variety of damages, including the never before seen damage, can be detected. In this study we aimed to investigate 1) the ability of anomaly detecting GANs to detect degradation (potholes and cracks) in asphalt road infrastructures using Mobile Mapper imagery and building damage (collapsed buildings, rubble piles) using post-disaster aerial imagery, and 2) the sensitivity of this method against various types of pre-processing. Our results show that we can detect damages in urban scenes at satisfying levels but not on asphalt roads. Future work will investigate how to further classify the found damages and how to improve damage detection for asphalt roads.

1. INTRODUCTION

Infrastructure and urban services are essential for societies and economies. However, they are increasingly prone to disruptions caused by climate change-induced extreme weather events, rising population numbers and general ageing of structures (Hallegatte et al., 2019). The key to reducing the impact of these disruptions is to increase the resilience of the structures and services. In this context, mapping degradation and damage plays an important role. Degradation mapping allows for efficient and timely maintenance resource allocation, thus prolonging infrastructure service life and raising its level of service (Frangopol, 2011). Post-disaster damage mapping aids the assessment of damages, which in turn aids faster post-disaster relief and recovery (Eguchi et al., 2009). For these reasons, damage mapping has been an active field of research for decades.

The field of damage and degradation detection has been strongly advanced by deep learning. Instead of handcrafting damage features, they can now be learned from the data themselves. So far, research on damage mapping using deep learning has regarded damage mapping as a supervised classification problem. However, a primary issue is that training datasets are hard to obtain. This is because most datasets are tailored to fit specific research areas, which severely limits research to compare methods or to transfer methods to other geographical locations or different typologies of damages (Nex et al., 2019b). Several benchmark databases have been created in an attempt to alleviate these issues (Gupta et al., 2019). Nonetheless, quality and quantity of training data remain important issues. Quality suffers from manual annotations, introducing bias and error. Varied quantities of samples in classes lead to class imbalance. Urban areas and infrastructures are visually varied. Damages therefore also appear in various shapes, sizes or contexts. Obtaining a sufficient quantity of the full extent of possible damages is impractical (Nex et al., 2019b). Additionally, assuming that most of the time structures are healthy, damages are occurring sparingly in time and space, which makes them intrinsically the minority class. On top of that, sampling the never before seen damage is impossible, which means that a trained deep learning model is unprepared for new scenarios (Akçay et al., 2018).

To solve these issues with training data, we propose to consider damages as novelties or anomalies from the undamaged state and to use anomaly detecting Generative Adversarial Networks (GANs) for damage detection (Akçay et al., 2018). GANs consists of two Convolutional Neural Nets (CNNs). The first tries to generate false samples by learning to approach the distribution of real input data, while the second aims to classify input images as either fake or real. They compete against each other in a two-player zero-sum game and gradually improve in conjunction. By learning only the distribution of healthy scenes, unhealthy scenes are difficult to approximate, which then serves as a measure of degree of anomaly. The main advantages of anomaly detecting GANs are that 1) they do not require any labelled training data but apart from data of healthy scenes, which are present in abundance, 2) they are adept in recognizing the never before seen damage, and 3) when trained on a variety of healthy scenes, a trained model can be used for inference at different geospatial locations or typologies of damages, thereby assisting preparedness in post-disaster scenarios (Akçay et al., 2018). These advantages alleviate the issues related to training data mentioned before. One limitation of anomaly detecting GANs is that they are not able to differentiate between types of anomalies. However, we believe that the benefit of not needing labelled training data outweighs the lack of obtaining descriptive labels.

[*] Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2020-573-2020 | © Authors 2020. CC BY 4.0 License.

Moreover, this method will improve resilience by increasing preparedness in both the post-disaster and routine scenario.

To our knowledge, GANs are mainly used for training data generation or augmentation (Antoniou et al., 2018; Zhu et al., 2019). Anomaly detecting GANs have only been applied to curated datasets which represent simpler tasks than datasets coming from real-world scenarios (Beggel et al., 2019; Zenati et al., 2018; Zhu et al., 2019). To the best of our knowledge, our study is the first to apply anomaly detecting GANs for damage detection in real world remote sensing scenarios.

In this study we aimed to investigate 1) the ability of anomaly detecting GANs to detect road infrastructure degradation using high resolution Mobile Mapper imagery and detect building damage using post-disaster aerial imagery, and 2) the sensitivity of this method against various types of pre-processing. Specifically, we aimed to detect (but not classify) asphalt degradation such as potholes and cracks in a routine infrastructure monitoring scenario, and to detect (but not classify) damages such as collapsed buildings and rubble piles in a post-earthquake scenario. To assess the sensitivity of the anomaly detecting GAN method, we investigated how performance varied with different levels of pre-processing. Specifically, we pre-processed training data to remove samples that could contain objects that in other remote sensing deep learning tasks have shown to cause difficulties: vegetation and shadows. Additionally, we tested the transferability of models that were trained on these different datasets, to assess the practical limitations of anomaly detecting GANs. To this end, we evaluated the trade-off between the practical constraints of training an anomaly detecting GAN and its performance.

The remainder of the paper is organized as followed. Section 2 discusses related work, section 3 describes the methodology and section 4 discusses our experimental setup, datasets and results. Finally, a conclusion and practical considerations are presented in section 5. Throughout the paper we will use the term damage to refer to both routine degradation and post-disaster damages.

2. RELATED WORK

2.1 Deep learning based damage mapping

Traditional damage mapping has been greatly improved by advances in earth observation (EO) systems. Traditionally, damage mapping was achieved through ground surveys. Accessibility to structures, but also the extent of the damages or amount of structures to inspect, made these surveys difficult and slow (Eguchi et al., 2009). Satellites and airborne remote sensing systems allowed for fast and large-scale damage mapping. Developments in Unmanned Aerial Vehicles (UAVs) increased the possibilities of faster and more detailed monitoring (Kerle et al., 2020). EO systems allowed for the collection of a variety of data such as optical, Synthetic-aperture Radar (SAR) or Light Detection and Ranging (LiDAR) data. Nowadays, optical data collection using UAVs is the low-cost alternative for damage mapping, aided by the rise in popularity of cheap UAVs (Nex et al., 2019a). However, for small-scale degradation monitoring on road infrastructures, monitoring systems are required that can observe damages in higher resolution. Therefore, Mobile Mapping systems or low-cost cameras are used to acquire optical imagery of asphalt road surfaces in high resolution (Eisenbach et al., 2017; Maeda et al., 2018).

Despite that in many cases image analysis for damage mapping still relies on visual interpretations, the process from data acquisition to information retrieval has been vastly accelerated by automated image analysis. Automated post-disaster image analysis makes use of edge, texture or colour to detect damaged building areas. A detailed overview of remote sensing based building damage mapping techniques can be found in Dong and Shan (2013). For infrastructure crack detection, various approaches have been proposed including a variety of image filtering techniques (Yeum and Dyke, 2015; Zalama et al., 2014). A comprehensive review on infrastructure damage detection can be found in Zakeri et al. (2017). The downside to most methodologies is the sensitivity of the methods to imaging conditions such as lighting, noise, angles and distortions. Shadows and varying lighting conditions make it difficult to detect damaged buildings or cracks on infrastructures (Dong and Shan, 2013).

There is a considerable amount of literature on damage detection, localization and identification using deep learning in the post-disaster and infrastructure monitoring domain. Infrastructure degradation mapping has focused on the detection of cracks, corrosion or lamination using CNNs. Cha et al. (2018) applied Faster R-CNN to detect and locate multiple types of damages: steel lamination, steel and bolt corrosion and concrete cracks. Zhang et al. (2017) developed CrackNet, where depth information from laser scans was used to produce 2.5D images to detect cracks on roads at pixel level. In most studies, both the inability to compare methods as well as extensive manual data collection and labelling was reported. In the post-disaster domain, debris or building façade damage detection is the main feature of interest. Multi-resolution airborne imagery and CNNs have been used to detect building damages after an earthquake event (Duarte et al., 2018). Adding 3D information to optical data improved subtle damage detection using CNNs (Vetrivel et al., 2018).

Most studies towards damage detection considered it a supervised problem, and thus struggled with issues related to training data as discussed in the introduction. Additionally, it was observed that few studies addressed anomaly detection using deep learning. The few that did, were focussed on traffic or crowd incident management (Lopez-Fuentes et al., 2018). The proposed anomaly detecting GAN approach shows similarities to one-class classification method such as one-class Support Vector Machines (OCSVM) (Scholkopf et al., 2001). In OCSVM, there is only interest in detecting a specific class, while all other classes are ignored and labelled as anomalies. Few remote sensing studies adopted this method such as the study of Li et al. (2011) towards one-class road extraction. The downside of OCSVM is that, albeit fewer, labelled training data are still required.

2.2 Generative Adversarial Networks

Generative Adversarial Networks were introduced by Goodfellow et al. (2014) as a way of generating new samples. As mentioned in the introduction, two CNNs are trained in conjunction: a Generator and a Discriminator. The Generator consists of an encoder and a decoder. It aims to learn the Generators distribution p_g, such that it matches the input data distribution p_{data}. The encoder maps the high-dimensional image data (x) to a lower-dimensional latent space (z) to the distribution p_z. The decoder maps the latent vector back to the image space x_g with probability p_g. The function of the Generator is $G(x;\ \theta_g)$, where θ_g denotes the architecture of the Generator. The Discriminator consists of an encoder that outputs a single scalar (z). It aims to distinguish whether the input (x) is either coming from p_{data} or p_g. It is described by $D(x;\ \theta_d)$, where θ_d denotes the encoder CNN. The discriminator aims at reducing $D(G(z))$, whereas the Generator tries to minimize it according to: $\log(1 - D(G(z)))$ (Goodfellow et al., 2014). Both models are locked in a so called two-player zero-sum game. In order to improve, the features used to correctly detect false samples are passed from the Discriminator to the Generator. Vice versa, the properties used to fool the Discriminator are passed to the Discriminator. Ideally both models simultaneously improve over time. Once p_g is equal to p_{data}, the models cannot improve any further (Goodfellow et al., 2014).

2.2.1 Anomaly detection using GAN. As mentioned earlier, anomaly detecting GANs are only trained on non-anomalous images (x_{normal}) with probability $p_{data\text{-}normal}$. During inference, whenever an anomalous image coming from probability $p_{data\text{-}abnormal}$ is fed to the Generator, it is expected that the Generator will fail to construct a realistic sample x_g and thus that p_g is far removed from $p_{data\text{-}normal}$. In contrast, if the input image came from $p_{data\text{-}normal}$, the distances are expected to be low. This distance can be used to score the degree of anomalousness. Using a threshold criteria (φ), these distances can be used to classify and quantify anomalies. There are several ways to calculate the distance that leads to an anomaly score. For example, we can calculate the distance between the generated image (x_g) and input image (x) or the distance between distribution p_g and $p_{data\text{-}normal}$, which is the Wasserstein distance (Arjovsky et al., 2017; Goodfellow et al., 2014).

AnoGAN was one of the first unsupervised deep convolutional generative adversarial nets (DCGANs), created to detect anomalies in retina images (Schlegl et al., 2017). It consists of a single encoder as the Generator and a single decoder for the Discriminator. Anomalies were identified using the residual distance and the discriminator loss, reaching precision and recall scores of 0.88 and 0.73, respectively. F-AnoGAN builds on the latter framework (Schlegl et al., 2019). It uses the same CNN architecture for the Discriminator and Generator, however now the Discriminator makes use of the Wasserstein distance. In addition, an encoder was trained seperately to learn the mapping from image to latent space, with the sole purpose of speeding up inference time. Schlegl et al. (2019) experimented with different loss functions to train the encoder and found that the best loss function yielded a precision of 0.79, performing slightly better than AnoGAN.

Efficient Gan Based Anomaly Detection (EGBAD) makes use of the AnoGAN structure; however, it now employs bidirectional learning to train the Encoder simultanuously with the Generator and Discriminator (Zenati et al., 2018). This way, the mapping of image to latent-space and vice versa is learned in a single step, improving on runtime. Higher precision, recall and F1 scores were obtained using this method compared to AnoGAN.

GANomaly was developed by Akcay et al. (2018). The Generator consists of a standard auto-encoder and the Discriminator of a standard encoder network, similar to what has been described in 2.2. Unique to this framework is the third encoder network, whose aim is to map the generated image to feature space $\hat{z}$ in order to implement a loss function which minimizes the distance between $\hat{z}$ and z. GANomaly was tested on x-rays of lugage where firearms or weapons were the anomaly class. The False Positive Rate (FPR) and True Positive Rate (TPR) were used to calculate the Area under the Receiver Operator Curve (AUC). GANomaly yielded higher AUC scores compared to AnoGAN. It achieved the lowest runtime compared to AnoGAN and EGBAD.

Skip-GANomaly builds on GANomaly. In Skip-GANomaly the Generator is replaced by a skip-connected encoder-decoder framework (U-net) (Akçay et al., 2019; Ronneberger et al., 2015). Features from the encoder layers are copied and concatenated to features in the sibling layers in the decoder. Therefore, information from varying resolutions is retained in subsequent convolutional layers in de decoder, yielding a high quality output image. Skip-GANomaly yielded impressive results and outperformed GANomaly, AnoGAN and EGBAD on the CIFAR10 and the firearms and weapons dataset. Considering these results, the state-of-the-art GANomaly and Skip-GANomaly are adopted in this research and discussed in the remainder of this paper.

3. METHODOLOGY

3.1 GANomaly and Skip-GANomaly architectures

The architectures of GANomaly and Skip-GANomaly can be found in Akcay et al. (2018) and Akçay et al. (2019). The loss function of both GANomaly and Skip-GANomaly are defined by three different loss functions: the adversarial loss, the contextual loss and the latent loss. Latent loss is also being called the encoder loss in Akcay et al. (2018). The adversarial loss steers the Generator to create realistic images that will fool the discriminator (Eq. 4). The contextual loss steers the Generator to not only create images that will fool the Discriminator, but to create images that are contextually sound. To this end, the input and generated images are compared at pixel level (Eq. 5). The latent loss steers the encoders inside the Generator and Discriminator to construct robust latent representations of the input and generated image (Eq. 6).

$$L_{adv} \| f(x) - f(\hat{x}) \|_2 \tag{4}$$

where, $f(.) = \mathbb{E}_{x \sim p_x} [\log D(.)]$

$$L_{con} = \| x - \hat{x} \|_1 \tag{5}$$

$$L_{lat} L_{lat} = \| z - \hat{z} \|_2 \tag{6}$$

The overall objective function is defined by:

$$L = w_{adv} L_{adv} + w_{con} L_{con} + w_{lat} L_{lat}$$

Where w_{adv}, w_{con} and w_{lat} are weights that control the influence of the individual losses to the objective function.

3.2 Anomaly classification

The intersection between the distribution of anomaly scores of normal and abnormal samples was found to determine the threshold criterion. This criterion is used to classify the test-samples into either normal or abnormal (Akçay et al., 2018). By plotting the histograms of both distributions, this threshold could be visualized. Ideally, the distributions are non-overlapping, meaning that both anomalies and normal samples are well distinguishable. The distributions can therefore serve as a measure of descriptive success.

3.3 Performance metrics

Performance metrics that were considered include Recall, Precision, Accuracy, and F1-score (Eq. 9-12).

$$Recall = \frac{tp}{tp + fn} \tag{9}$$

$$Precision = \frac{tp}{tp + fp} \tag{10}$$

$$Accuracy = \frac{tp + tn}{tp + fn + fp + fn} \tag{11}$$

$$F1 = \frac{2tp}{2tp + fp + fn} \tag{12}$$

Where, *tp* is the number of true positives, *tn* is the number of true negatives, *fp* is the number of false positives and *fn* is the number of false negatives.

Recall was deemed most important since it is an indicator of how many of the total damages were actually retrieved. However, precision should also be considered, since it is an indicator of how many of the positive samples were actually damages. In addition, AUC was calculated to allow comparison with other anomaly detecting GANs. As mentioned in 2.2.1, AUC is the area under the curve when the TPR and FPR are plotted against each other.

3.4 Sensitivity analysis

As explained in the introduction, we aimed at showing how anomaly detecting GANs performed with different levels and types of pre-processing. When there are no time or capacity constraints, such as in infrastructure monitoring scenarios, users could choose to reach for maximum accuracy or other performance metrics. To this end, the original datasets were modified and the model performances were compared.

As was mentioned in the introduction, modifications to the post-disaster dataset include the removal of images from the dataset that contain vegetation and shadows. These classes are generally difficult to classify in deep learning classification tasks. Shadows can make crack detection or building damage detection difficult. Moreover, in temperate regions, vegetation changes during the year. The variety of contrast could make it difficult to learn the data distribution of normal images. To this end, the Canopy Shadow Index (SI) (Eq. 7) and the Green-Red Vegetation Index (GRVI) (Eq. 8) was calculated for each image (Azizi et al., 2008; Motohka et al., 2010). To determine the maximum amount of positive pixels that could contain these classes, a threshold criterion was determined empirically. The images that exceeded this threshold were removed from the dataset. In addition to vegetation and shadow removal, one post-disaster dataset was converted from RGB to greyscale to investigate how sensitive anomaly detecting GANs are to colour-information.

$$SI = \sqrt{(256 - B2) * (256 - B3)} \tag{7}$$

$$GRVI = \frac{\rho_{green} - \rho_{red}}{\rho_{green} + \rho_{red}} \tag{8}$$

In the road infrastructure dataset, we observed that damages were not located on road markings. Therefore, we argued that they do not convey information on degradation. For that reason, any image containing a road marking was removed from the dataset. In this dataset road markings were displayed as large areas of white pixels in geometric shapes. Images containing road markings were identified using Gaussian blurring, dilation-and-erosion and thresholding techniques. The threshold criteria were again determined empirically.

3.5 Transferability of trained models

We further build on the aim explained in section 3.3. When there are time and capacity constraints, a user might choose to balance quality and speed. For this purpose a trained model is tested on the test-sets of other datasets, varying in degree of complexity. Comparing the resulting metrics should give an impression to what extent we can improve preparedness in time-sensitive scenarios by training models in advance.

3.6 Damage segmentation

The qualitative performance of anomaly detecting GANs was evaluated by creating damage segmentation maps. As explained earlier, the fake and real images are likely to be more different when the real image contains an anomaly. The pixel-wise difference between the real and generated image were calculated to obtain pixel-level anomaly scores. It was expected that the pixels located on damages have higher anomaly scores than non-damaged pixels. To select which pixels to visualize, a threshold criterion was determined empirically. The actual spatial distribution of damages was expected to surface by visualizing them. These visualizations were evaluated manually for correctness since no pixel level annotation were available to evaluate the segmentation maps against.

3.7 Comparison against other methods

We compared anomaly detecting GANs against one supervised and one unsupervised classification method. The supervised method makes use of transfer learning and fine-tuning. Transfer learning is a method where a model trained on a different task, is applied to another task. Fine-tuning refers to the process where only the final layer of the pre-trained model are re-trained for the new task using the new dataset. The advantage of transfer learning is that general concepts do not have to be learned anymore, if the pre-trained model was trained on a generalizable dataset. The advantage of fine-tuning is that time and resources can be spend to learn specific concepts for the task at hand. In this research we fine-tuned Densenet161 which was pre-trained on the ImageNet dataset (Huang et al., 2017).

The unsupervised method makes use of OCSVM which was mentioned in earlier. SVM aims to separate distinctive classes by finding the decision boundary which separates the classes. In SVM the aim is usually to distinguish two classes from each other. In OCSVM, the aim is to distinguish a specific class from others. OCSVM can be used in anomaly detection to detect the target class (for example: "apples") and label all the other samples as outliers ("non-apples"). In this specific case, we use OCSVM aim to distinguish normal cases ("undamaged") from abnormal cases ("damaged"). Again, the advantage is that no damaged training data is needed. The disadvantage is that this method is sensitive to the tuning of the OCSVM parameters, or the number of features (Li et al., 2011). Because the images in our datasets contain many features (image height x image width x nr. of channels), instead of applying OCSVM on all the image features, we first extracted image features from the pre-trained DenseNet161. To reduce the amount of features, the most important features were selected using Principal Component Analysis. The remaining features were used in OCSVM.

4. EXPERIMENTS

4.1 Datasets

4.1.1 German Asphalt Pavement Distress dataset. The infrastructure dataset consists of patches from the German Asphalt Pavement Distress (GAP) dataset v1.0 (Eisenbach et al., 2017). GAP v1.0 consists of 1969 high resolution greyscale images of road surfaces. They were acquired using a mobile mapping system, a vehicle mounted with stereo-cameras facing the surface. Each image covers approximately 2.84 x 1.0 m in a 1920 x 1080 resolution with a Ground Sampling Distance of 1.2 mm x 1.2 mm. The images were acquired at driving speed and shadows were kept to a minimum by using a synchronized lighting unit. The images were pixel annotated by experts for road surface damages. These include: cracks, potholes or patching. Even though, in GAP v1 only two labels exist: damaged or undamaged. From these images, patches were extracted of 64x64 pixels (Figure 1). Each patch is labelled as either damaged or undamaged.

The GAP dataset has a high signal-to-noise ratio. The damage feature is mainly expressed in brightness values in a geometrical shape (linear, round, etc.). The surrounding pixels, however, are depicted as random variation of brightness values, making it hard to extract the damage signal. In this research noise was not treated explicitly and no feature enhancement techniques were added.

4.1.2 Buildings dataset. For the Buildings dataset we selected patches of 80x80, derived from a collection of aerial imagery (Figure 1). Several images represent healthy urban scenes in Europe, while others represent post-earthquake urban scenes in New Zealand, Italy and Haiti (Nex et al., 2019b). Damages consist of collapsed buildings, or damaged roofs and facades. Patches were manually drawn, extracted and labelled.

4.1.3 Pre-processing. As explained earlier, the datasets were treated using various levels of pre-processing to investigate the performance of the proposed method. Figure 2 shows an example of pre-processing on a pavement patch where the amount of pixels containing road markings exceeds the threshold of 10%. Figure 3 shows an example where the amount of vegetation and shadow pixels exceed the threshold of 10%. Table 1 describes all the datasets that were used in this research. In GAP2 the road markings were removed from the original GAP. In Buildings 2 (B2), shadows and vegetation were removed. The threshold criterion was set low, in order to remove even small patches of vegetation or shadows in a strict manner. In Buildings 3 (B3), only vegetation was removed. In Buildings 4 (B4), only shadows were removed. In Buildings 5 (B5) both vegetation and shadows were removed similar to B2, only this time higher threshold values were used.

In all datasets, the undamaged patches were split in 80/20 train and test set. As mentioned before, only undamaged patches were used for model development. The undamaged patches in the test dataset in combination with all the damaged patches were used for evaluation.

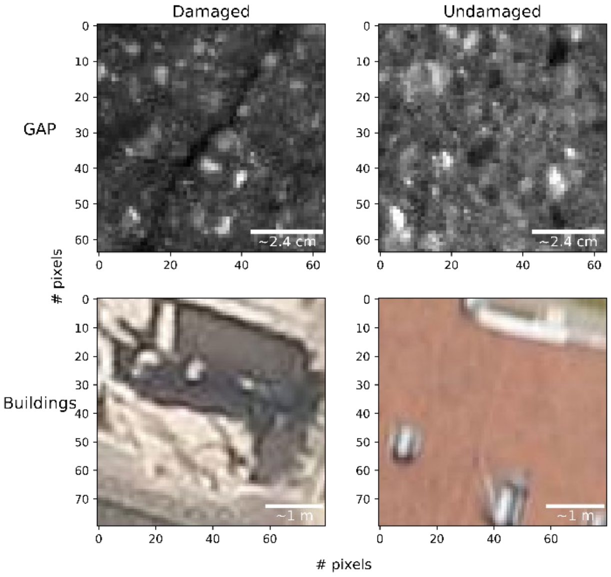

Figure 1. Example of damaged and undamaged samples within the GAP and Buildings dataset.

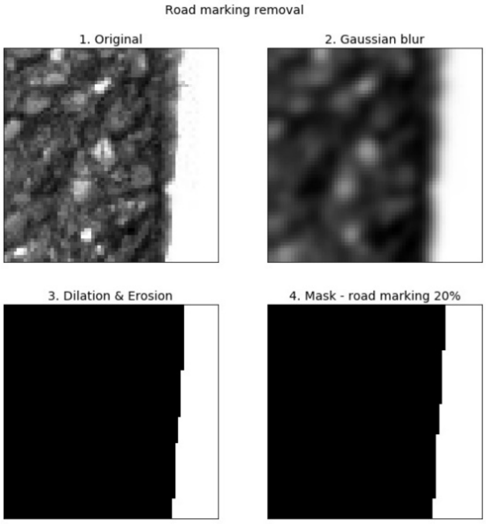

Figure 2. Example of road marking removal.

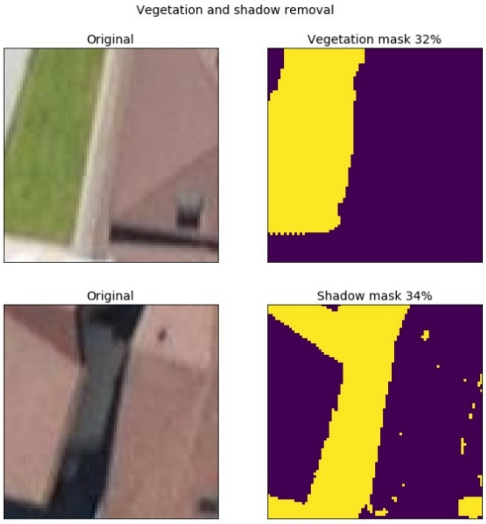

Figure 3. Example of vegetation and shadow removal.

Dataset	Description	# undamaged patches	# damaged patches
GAP	Original GAP dataset	5.221.249	679.154
GAP2	No road markings	5.079.787	567.250
B	Original buildings dataset	430.475	3.113
B2	No vegetation or shadows (strict)	12.830	575
B2_grey	No vegetation or shadows (strict) in greyscale	12.830	575
B3	No vegetation	55.509	1.703
B4	No shadows	62.622	636
B5	No vegetation or shadows (lenient)	91.126	3.113

Table 1. Description of the different datasets used in this research plus the number of undamaged and damaged patches.

4.2 Hyper-parameter tuning

In order to ensure that a model performs optimally for each dataset and each architecture, hyper-parameter tuning was required. Hyper-parameters influence how fast or how efficient the objective function can be reached. The main parameters tuned for Skip-GANomaly were the loss weights (w_{adv}, w_{con} and w_{lat}) and the size of the latent vector z. The size of z influences the amount of information retained in z and subsequently the encoder loss. In addition, GANomaly was tuned for the number of extra layers present in the Generators encoder and decoder. The size of the encoder and decoder influences the amount of convolutional layers. The models were trained on a single GPU (TITAN XP) and on 16 CPU cores. On average, a model was trained on 12.000 samples for 10 epochs within 18 hours. During inference, deriving labels of a batch containing 64 samples, averaged 4.2 milliseconds.

4.3 Results

4.3.1 Performance metrics. To evaluate the results, we inspected the performance metrics, the generated images and the anomaly scores distribution. The performance metrics are reported in Table 2. The best performing models were selected based on the recall-precision trade-off. Practically speaking, we wanted to retrieve all damaged samples, while at the same time not burden first responders or maintenance workers with manually eliminating false positives. We therefore valued recall over precision. Comparing GANomaly and Skip-GANomaly, Skip-GANomaly performed better on all metrics with the occasional exception for precision. While Skip-GANomaly reached recall values of up to 0.95, GANomaly did not reach values higher than 0.86. This trend is in line with findings in Akçay et al. (2019) where AUC values between 0.68 and 0.94 were found.

	Recall	Precision	Accuracy	F1-score	AUC
G_{GAP}	0.180	0.514	0.537	0.266	0.57
S_{GAP}	**0.478**	**0.642**	**0.631**	**0.547**	**0.68**
G_{GAP2}	0.032	0.421	0.560	0.059	0.58
S_{GAP2}	**0.401**	**0.545**	**0.593**	**0.462**	**0.62**
G_{B}	0.608	0.26	**0.857**	0.364	0.82
S_{B}	**0.686**	**0.298**	0.870	**0.416**	**0.90**
G_{B2}	0.863	**0.92**	0.937	0.89	0.98
S_{B2}	**0.955**	0.873	**0.945**	**0.912**	0.98
S_{B2_grey}	0.883	0.849	0.919	0.866	0.98
G_{B3}	0.501	**0.732**	**0.838**	0.595	0.85
S_{B3}	**0.775**	0.591	0.819	**0.671**	**0.89**
G_{B4}	0.723	0.526	0.913	0.609	0.94
S_{B4}	**0.951**	**0.623**	**0.941**	**0.753**	**0.98**
G_{B5}	0.568	**0.838**	0.860	0.677	0.93
S_{B5}	**0.814**	0.738	**0.877**	**0.774**	**0.94**

Table 2. Recall, Precision, Accuracy, F1-score and AUC-score for all datasets (denoted by subscript) using GANomaly (G) and Skip-GANomaly (S).

The generated images are shown in Figure 4. GANomaly consistently failed to generate realistic samples. Skip-GANomaly, on the other hand, showed that it is able to create fake samples that are indistinguishable from reality. We believe that the skip-connections in U-net successfully allowed multi-scale information to be passed forward to the decoder network.

Interestingly, despite the inability of GANomaly to generate realistic images, some models obtained high performance metrics such as G_{B2}. This can be explained by inspecting the distribution of anomaly scores (Figure 5). As explained in section 3.2, well distinguishable anomaly distributions can be considered as a measure of descriptiveness. Figure 5 shows that the anomaly scores distributions of G_{B2} are largely overlapping. The threshold value of 0.078, which divides abnormal from normal samples, can be considered non-descriptive. A majority of the abnormal test samples were correctly classified as normal. Consequently, a high *tp* and low *fn* value is obtained, therefore yielding high recall and precision values. Nonetheless, we conclude that the descriptive value of the model is low. In contrast we found that S_{B2} has a high descriptive value (Figure 6). The anomaly distributions are well distinguishable and most samples are correctly classified, resulting in high recall and precision values. The results showed that building damages were distinctively different from normal. However, besides damages, the GAP and Buildings datasets did not contain anomalies other than damages. Non-damaged anomalies could include: tire-marks, debris or on purpose demolished buildings. We argue that would they be present in the dataset, they would also be classified as being anomalous. Therefore, the *fp* score for damages would increase since no distinction is made between damaged or non-damaged anomalies. Nonetheless, as discussed in the introduction, we argue that the benefit of not needing labelled training data, outweighs the lack of obtaining descriptive labels. Moreover, we argue that retrieving non-damaged anomalies still provides information on deviations from normal. Future research will focus on anomaly classification, in order to provide end-users with more qualitative information about the found anomalies.

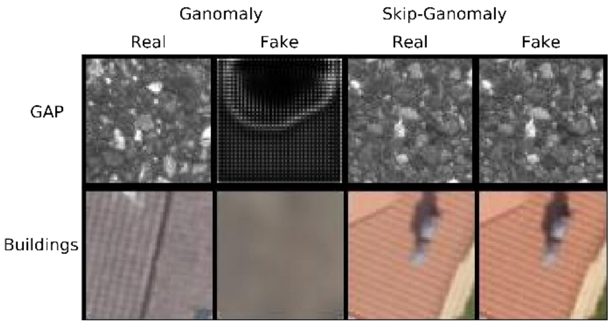

Figure 4. Real and generated fake images for GANomaly and Skip-GANomaly.

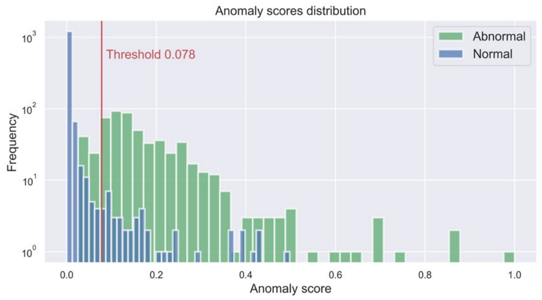

Figure 5. Anomaly scores distribution for G_{B2}

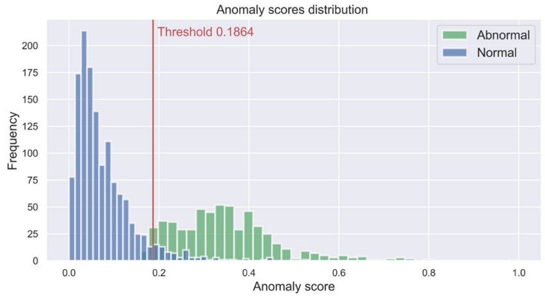

Figure 6. Anomaly scores distribution for S_{B2}.

4.3.2 Sensitivity analysis. As explained in section 3.4 we investigated the sensitivity of this method to varying levels of pre-processing, since it would shed light on practical considerations in operational settings. Only the performance on the Buildings datasets improved once the complexity was reduced (Table 2). Specifically, Skip-GANomaly models trained on the least complex datasets (B2, B3, B4 and B5) performed better than when trained on the most complex dataset (B). Skip-GANomaly, trained on the least complex and heaviest pre-processed dataset B2, outperformed the rest, yielding 0.955 recall and 0.873 precision. Despite these results, in practice, one should consider the effort spent on curating B2. Less effort was spent on creating B5, while yielding the respectable results of 0.814, 0.738 for recall and precision, respectively. Despite the performance on B2, B5 poses a viable alternative, balancing effort, time and performance, without sacrificing a lot on performance, which is preferable in post-disaster scenarios. Other findings showed that the Skip-GANomaly model trained without shadows (B4) performed better than when trained without vegetation (B3), suggesting that the anomaly detecting GANs were more sensitive to shadows. We speculate that this has to do with the lack of contrast and colour inside the shadowed areas. As mentioned in the introduction, this has been reported to inhibit the detection of edges inside shadowed areas. We argue that less information was retained in the CNN's feature maps and latent vector (z) which led to a rise in contextual loss. In line with this reasoning, vegetation did exhibit contrast and colour. Therefore, it added information to intermediate feature maps and value to the contextual loss. Evidence of the importance of colour was also found when comparing S_{B2} and S_{B2_grey} (Table 2). This showed that performance drops once colour information is removed. More evidence was found in a converging contextual loss of approximately 0.5 and 0.7 for S_{B2} and S_{B2_grey}, respectively, showing that greyscale images yield higher loss. A final finding was that, surprisingly, removing road markings did not improve performance. They might have unexpectedly provided contextual information throughout the patch which leads to a better approximation of $P_{data\text{-}normal}$.

4.3.3 False Negatives, False Positives. We investigated the False Positives (FPs) and False Negatives (FNs) to understand where both GANomaly and Skip-GANomaly failed in giving the correct classification (Figure 7). Most FNs in the Buildings dataset, depicted major geometrical features or areas that still depict a (blurred) pattern. Uninterrupted blurred features and areas may resemble normal images too closely. FPs mostly consisted of large homogeneous areas and colours. Patches with little contrast did not fit the learned distribution of normal patches. Improvements could be achieved by increasing the patch size, such that more contextual information is added to the model. Distinguishing damages on asphalt patches was even more problematic. Damages were sometimes too subtle to detect. Notice for example the red outlined patch in Figure 7. Thorough selection of patches that clearly represent damages is expected to improve performance although in practice this would be impractical.

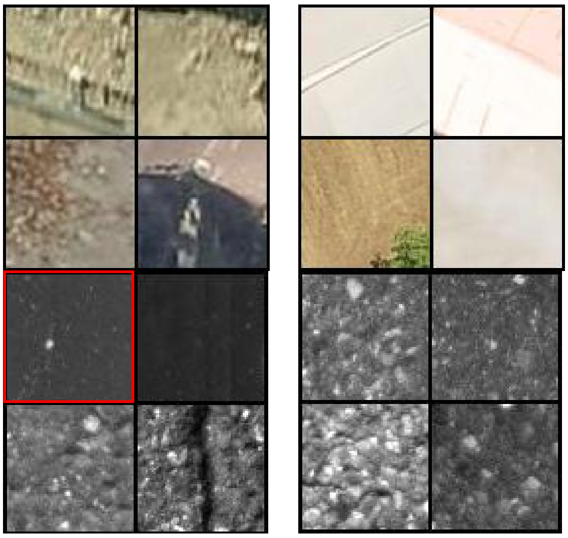

Figure 7. False Negatives (left) and False Positives (right) of S_{B2} (top) and S_{GAP} (bottom).

4.3.4 Transferability of trained models. As explained in section 3.5, we tested trained models on the test sets of other datasets to investigate to what degree preparedness can be achieved. Figure 8 shows the performance of Skip-GANomaly models that are tested on the test-sets of Building datasets on which they were not trained. A primary observation is that a model trained on a pre-processed dataset yielded satisfying results when tested on other pre-processed datasets on which it was not trained. S_{B1}, the model trained on the original dataset, yielded worse results on pre-processed datasets (B2, B3, B4 or B5). On the other side of the spectrum, the model trained on the heavily pre-processed dataset (B2) did not perform well on either the pre-processed (B3, B4 or B5) or the original datasets (B1). We observed that a compromise yielded the best outcome. Specifically, S_{B3} is a moderately pre-processed dataset which did not perform well on the neither the heaviest pre-processed (B2) nor the original dataset (B1), but performed satisfyingly on equally pre-processed datasets (B4 and B5). Similarly, S_{B4} seemed to perform on par on other equally pre-processed datasets, except on B3 which seemed to suggest that shadows were exhibiting a significant effect on model performance. Lastly, S_{B5} performed well on all pre-processed datasets, but not on the original dataset (B1).

4.3.5 Damage segmentation. In order to evaluate whether anomaly scores were indeed higher on damaged pixels, we visualized the pixel level anomaly scores based on a threshold criterion. Two criteria were found empirically and set at the second and third quantile.

Figure 9 shows anomaly segmentations of a damaged Buildings and GAP patch. Although no clear outline of the damaged sections are found, a higher density of anomaly pixels was observed around damaged areas. In the buildings patch, the anomaly density is higher in the upper-left corner where a collapsed roof and debris is visible and less dense in the lower-right corner where the roof is intact. In the pavement patch, a higher density is observed around the crack. This suggests that two of our assumptions are correct. The anomaly detecting GAN is less adept in generating damaged images and damaged pixels yield higher pixel level anomaly scores than others. The network can not only identify, but also locate damages inside images. A disadvantage of this visualization method is that the threshold is determined empirically. Future research should focus on extracting the threshold automatically instead of setting it manually.

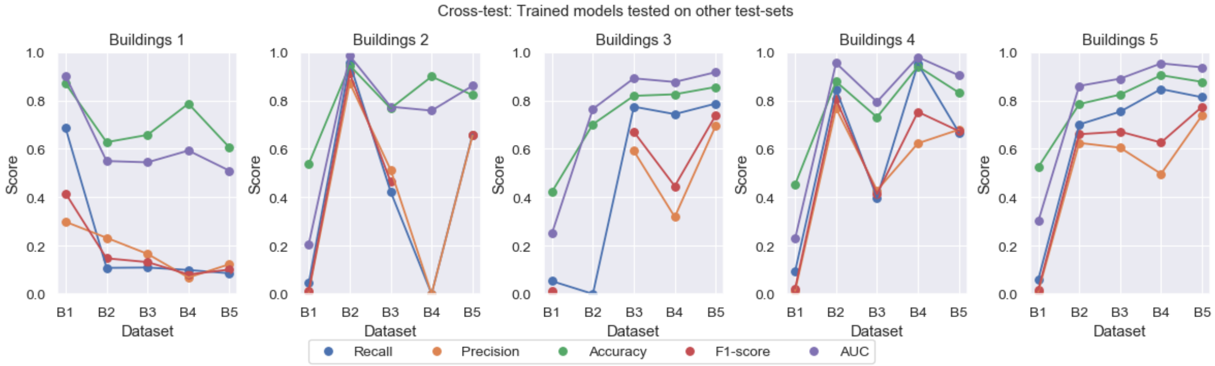

Figure 8. Cross-test results. Trained Skip-GANomaly Building models are tested on other Building test-sets which were not used for training.

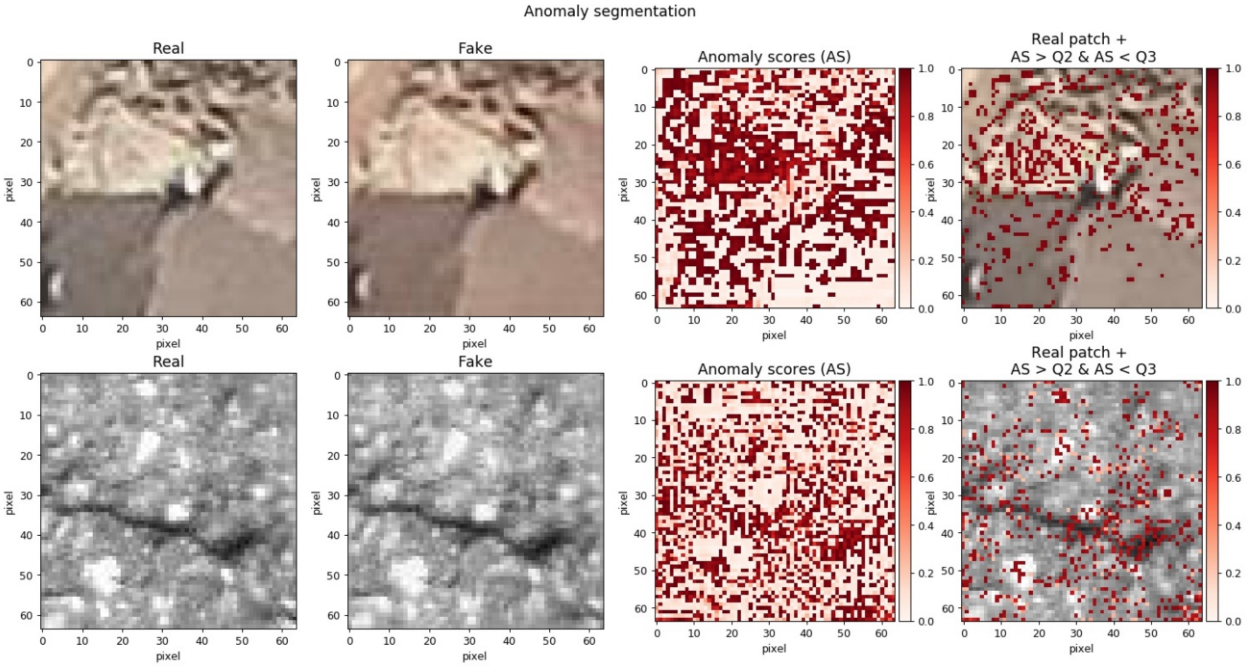

Figure 9. Anomaly segmentation for a damaged building (top row) and a damaged pavement (bottom row).

4.3.6 Comparison against state of the art. Table 3 shows how our method compared against other classification methods. OCSVM performed worse than our proposed method. This has likely to do with the nature of the target class. Normal buildings are diverse in visual appearance. This diversity makes it difficult to maximize the distance between the target class and anomalies. While our method scored high on all performance metrics, the fine-tuned DenseNet outperformed our method. The advantage of a pre-trained network is that less effort is required to train a model. However, more importantly, the main disadvantage is that damaged examples are needed. As discussed earlier, this is a problem for most damage mapping tasks. Our method has therefore a clear advantage while minimally sacrificing on performance.

	Recall	**Precision**	**Accuracy**	**F1-score**
OCSVM	0.074	0.777	0.107	0.088
DenseNet	**1.000**	**0.957**	**0.957**	**0.978**
S_{B2}	0.955	0.873	0.945	0.912

Table 3. Comparison of our method against other unsupervised and supervised methods.

5. CONCLUSION

In the context of increasing resilience in the public domain, this study investigated 1) the ability of anomaly detecting GANs to detect degradation in road infrastructures using Mobile Mapping imagery and building damages in urban settings using post-disaster aerial imagery, and 2) the sensitivity of this method against various types of pre-processing. Two distinctive datasets were used: a high resolution road surface mobile mapper dataset (GAP) and a post-earthquake urban aerial imagery dataset (Buildings). Two state of the art anomaly detecting GANs were applied: GANomaly and Skip-GANomaly.

Only Skip-GANomaly performed satisfyingly when detecting damages in the buildings dataset. GANomaly was consistently not able to find the descriptive features of damages and therefore to detect damages. We conclude that the U-net architecture of Skip-GANomaly's Generator plays the largest role in providing good results.

In order to test the sensitivity of the models against different levels of pre-processing, the datasets were pre-processed to reduce the complexity of the datasets. Specifically, shadows and vegetation were removed from the Buildings dataset and road markings were removed from the GAP dataset. No improvements were visible for the GAP dataset. For the Buildings dataset however, reducing its complexity improved recall and precision measures. Shadow removal improved performance the most.

We investigated, to what extent trained models can be transferred to other datasets. This would show to what extent preparedness in operational settings can be facilitated using anomaly detecting GANs. Results showed that models trained on pre-processed datasets were able to infer satisfying results on less pre-processed datasets.

We visualized the spatial location of high pixel level anomaly scores. We found that the density of anomalous pixels is higher at damaged locations. This shows that anomaly detecting GANs are not only able to identify damages, but also to locate damages. Finally, we compared anomaly detecting GANs with other commonly used classification methods. Our method performed better than the unsupervised OCSVM, although it was outperformed by supervised transfer learning and fine-tuning. The main disadvantage of the latter method, is that training examples of damages are required. As explained in the introduction, this is difficult to obtain in most real world classification tasks. Our method yields results close to the supervised method, while not needing any training examples but from healthy scenes. We argue that our method is therefore the most suitable for damage mapping tasks.

These results allow us to make practical suggestions for efficient damage detection in post-disaster scenarios using anomaly detecting GANs: a model should be trained on datasets in which at least shadows are removed. Too heavy pruning might give adverse effects. Inference should be done on datasets in which at least some pre-processing has been done. Removing shadows is advised as it produces the best results, however this comes at the price of not being able to detect damages in shadowed areas. Moreover, depending on the size of your training data, training may take a significant amount of time, which might not be available in post-disaster scenarios. Inference time however is fast. Therefore, we suggest to train models in advance, so that inference can be done immediately when needed. Finally, we advise end-users to remember that retrieved samples can be considered anomalous but not all can be considered damages. No practical suggestions can be made for asphalt damage detection. This task remains a challenge. We argue this is mainly caused by the difficult nature of asphalt damages which are difficult to distinguish. Adding an additional source of information, such as depth from Light Detection and Ranging (LiDAR) sensors, might aid in distinguishing physical damages from normal.

This study is the first to show that unsupervised damage detection using anomaly detecting GANs is possible without the need of any prior damage information. This eases model development and facilitates preparedness in post-disaster damage mapping scenarios. We argue that this framework can be used to detect damages in a large range of damage or post-disaster typologies, no matter the nature of the damage. Therefore, this work is valuable, as it signals the start of a shift from task-specific supervised damage mapping to a uniform unsupervised damage mapping approach. Future work will investigate the transferability of this method to other geographical locations or typologies of damages. Furthermore, we will investigate how to reduce the signal-to-noise ratio in asphalt imagery, how to distinguish non-damaged anomalies from damaged anomalies and how to further segment and classify damages.

ACKNOWLEDGEMENTS

Financial support has been provided by the Innovation and Networks Executive Agency (INEA) under the powers delegated by the European Commission through the Horizon 2020 program "PANOPTIS–Development of a decision support system for increasing the resilience of transportation infrastructure based on combined use of terrestrial and airborne sensors and advanced modelling tools", Grant Agreement number 769129. We thank NVIDIA for their donation of one TITAN XP to support our research.

REFERENCES

Akçay, S., Atapour-Abarghouei, A., Breckon, T.P., 2019. Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection, *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

Akçay, S., Atapour-Abarghouei, A., Breckon, T.P., 2018. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training, *Asian Conference on Computer Vision*. 622–637. doi.org/10.1007/978-3-030-20893-6_39

Antoniou, A., Storkey, A., Edwards, H., 2018. Data Augmentation Using Generative Adversarial Network. arXiv Prepr. arXiv1711.04340v3.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN. arXiv Prepr. arXiv1701.07875v3.

Azizi, Z., Najafi, A., Sohrabi, H., 2008. Forest Canopy Density Estimating, Using Satellite Images, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.* 1127–1130. doi.org/10.13140/2.1.2953.6967

Beggel, L., Pfeiffer, M., Bischl, B., 2019. Robust Anomaly Detection in Images using Adversarial Autoencoders. arXiv Prepr. arXiv1901.06355.

Cha, Y.-J., Choi, W., Suh, G., Mahmoudkhani, S., Büyüköztürk, O., 2018. Autonomous Structural Visual Inspection Using Region-Based Deep Learning for Detecting Multiple Damage Types. *Comput. Civ. Infrastruct. Eng*. 33, 731–747. doi.org/10.1111/mice.12334

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS J. Photogramm. Remote Sens.* 84, 85–99. doi.org/10.1016/j.isprsjprs.2013.06.011

Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2018. Satellite Image Classification of Building Damages using Airborne and Satellite Image Samples in a Deep Learning Approach, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences.* 89–96. doi.org/10.5194/isprs-annals-IV-2-89-2018

Eguchi, R.T., Huyck, C.K., Ghosh, S., Adams, B.J., Mcmillan, A., 2009. Chapter 15: Utilizing New Technologies in Managing Hazards and Disasters, in: Showalter, P.S., and Lu, Y. (Eds.), *Geospatial Techniques in Urban Hazard and Disaster Analysis.* Springer, Netherlands, 295–323. doi.org/10.1007/978-90-481-2238-7_15

Eisenbach, M., Stricker, R., Seichter, D., Amende, K., Debes, K., Sesselmann, M., Ebersbach, D., Stoeckert, U., Gross, H.M., 2017. How to get pavement distress detection ready for deep learning? A systematic approach, *Proceedings of the International Joint Conference on Neural Networks.* IEEE, 2039–2047. doi.org/10.1109/IJCNN.2017.7966101

Frangopol, D.M., 2011. Life-Cycle performance, management, and optimisation of structural systems under uncertainty: Accomplishments and challenges. *Struct. Infrastruct. Eng.* 7, 389–413. doi.org/10.1080/15732471003594427

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets, *Advances in Neural Information Processing Systems.* 2672–2680.

Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M., 2019. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.* 10–17.

Hallegatte, S., Rentschler, J., Rozenberg, J., 2019. Lifelines. The Resilient Infrastructure Opportunity, Sustainable Infrastructure Series. The World Bank, Washington, DC. doi.org/10.1596/978-1-4648-1430-3

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4700–4708. doi.org/10.1109/CVPR.2017.243

Kerle, N., Nex, F., Gerke, M., Duarte, D., Vetrivel, A., 2020. UAV-based structural damage mapping: A review. *ISPRS Int. J. Geo-Information.* 9, 1–23. doi.org/10.3390/ijgi9010014

Li, W., Guo, Q., Elkan, C., 2011. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Trans. Geosci. Remote Sens.* 49, 717–725. doi.org/10.1109/TGRS.2010.2058578

Lopez-Fuentes, L., van de Weijer, J., González-Hidalgo, M., Skinnemoen, H., Bagdanov, A.D., 2018. Review on computer vision techniques in emergency situations. *Multimed. Tools Appl.* 77, 17069–17107. doi.org/10.1007/s11042-017-5276-7

Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., Omata, H., 2018. Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images. *Comput. Civ. Infrastruct. Eng.* 0, 1–15. doi.org/10.1111/mice.12387

Motohka, T., Nasahara, K.N., Oguma, H., Tsuchida, S., 2010. Applicability of Green-Red Vegetation Index for remote sensing of vegetation phenology. *Remote Sens.* 2, 2369–2387. doi.org/10.3390/rs2102369

Nex, F., Duarte, D., Steenbeek, A., Kerle, N., 2019a. Towards real-time building damage mapping with low-cost UAV solutions. *Remote Sens.* 11, 1–14. doi.org/10.3390/rs11030287

Nex, F., Duarte, D., Tonolo, F.G., Kerle, N., 2019b. Structural Building Damage Detection with Deep Learning : Assessment of a State-of-the-Art CNN in Operational Conditions. *Remote Sens.* 11, 1–17. doi.org/10.3390/rs11232765

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 234–241. doi.org/10.1007/978-3-319-24574-4_28

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Med. Image Anal.* 54, 30–44. doi.org/10.1016/j.media.2019.01.010

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, *International Conference on Information Processing in Medical Imaging.* 146–147. doi.org/10.1007/978-3-319-59050-9_12

Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution, *Neural computation.*

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., Vosselman, G., 2018. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS J. Photogramm. Remote Sens.* 140, 45–59. doi.org/10.1016/j.isprsjprs.2017.03.001

Yeum, C.M., Dyke, S.J., 2015. Vision-Based Automated Crack Detection for Bridge Inspection. *Comput. Civ. Infrastruct. Eng.* 30, 759–770. doi.org/10.1111/mice.12141

Zakeri, H., Nejad, F.M., Fahimifar, A., 2017. Image Based Techniques for Crack Detection, Classification and Quantification in Asphalt Pavement: A Review. *Arch. Comput. Methods Eng.* 24, 935–977. doi.org/10.1007/s11831-016-9194-z

Zalama, E., Gómez-García-Bermejo, J., Medina, R., Llamas, J., 2014. Road Crack Detection Using Visual Features Extracted by Gabor Filters. *Comput. Civ. Infrastruct. Eng.* 29, 342–358. doi.org/10.1111/mice.12042

Zenati, H., Foo, C.S., Lecouat, B., Manek, G., Chandrasekhar, V.R., 2018. Efficient GAN-Based Anomaly Detection. arXiv Prepr. arXiv1802.06222.

Zhang, A., Wang, K.C.P., Li, B., Yang, E., Dai, X., Peng, Y., Fei, Y., Liu, Y., Li, J.Q., Chen, C., 2017. Automated Pixel-Level Pavement Crack Detection on 3D Asphalt Surfaces Using a Deep-Learning Network. *Comput. Civ. Infrastruct. Eng.* 32, 805–819. doi.org/10.1111/mice.12297

Zhu, D., Xia, S., Zhao, J., Zhou, Y., Jian, M., Niu, Q., Yao, R., 2019. Diverse sample generation with multi-branch conditional generative adversarial network for remote sensing objects detection. *Neurocomputing.* doi.org/10.1016/j.neucom.2019.10.065